Introduction to medical biometry and statistics / by Raymond Pearl.

Contributors

Pearl, Raymond, 1879-1940.

Publication/Creation

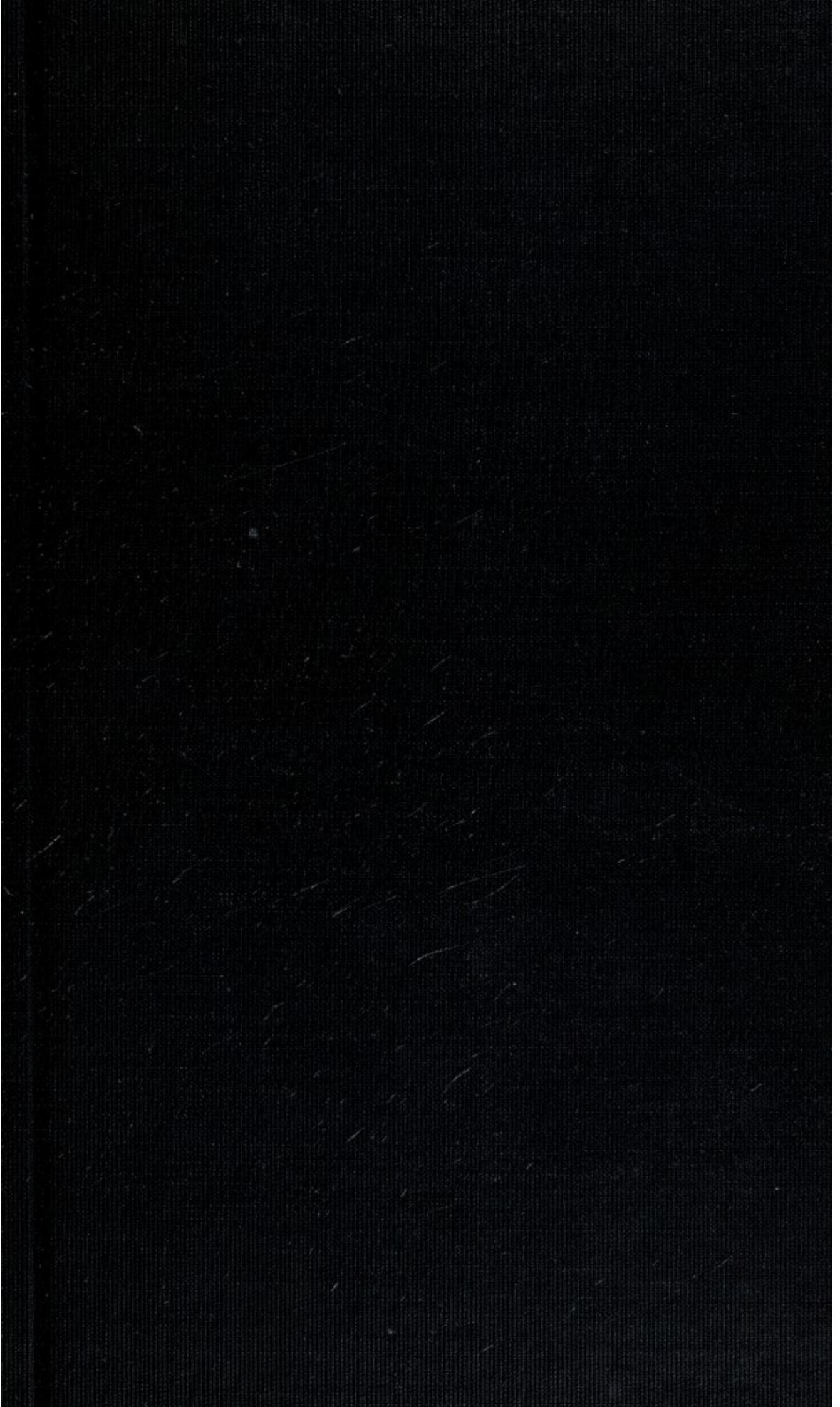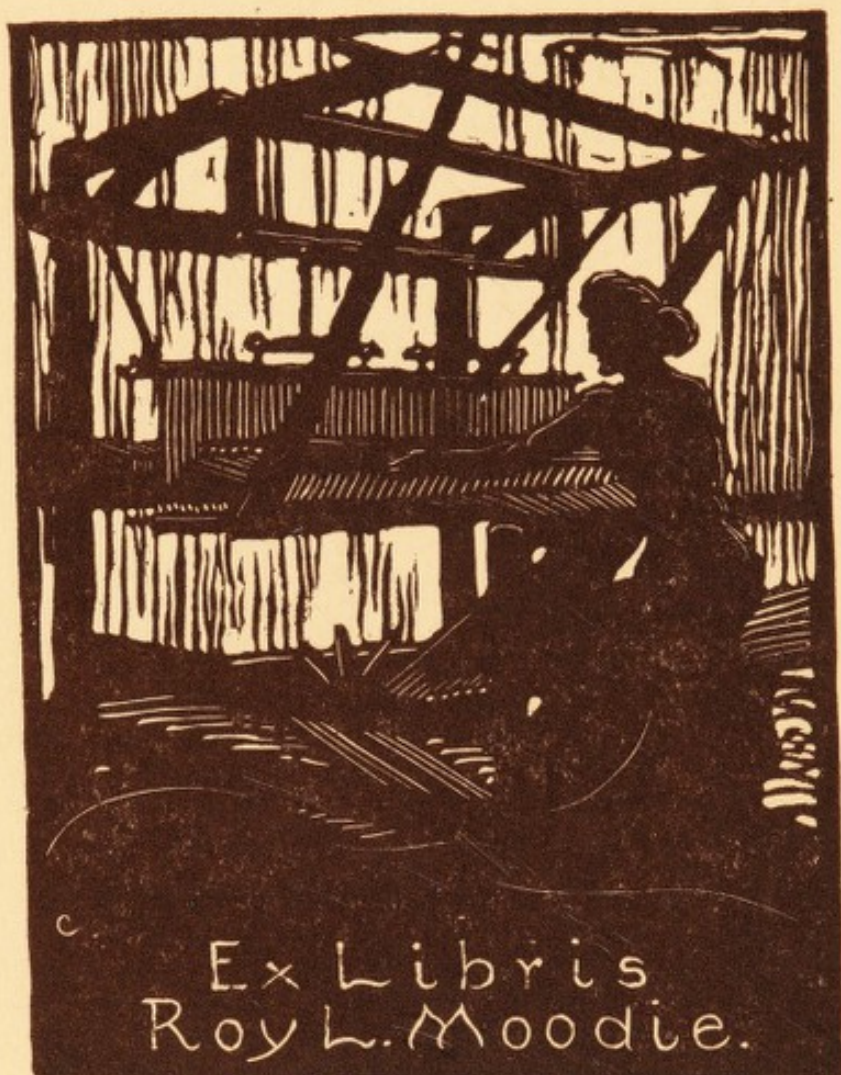Philadelphia ; London : W.B. Saunders company, 1930.

Persistent URL

https://wellcomecollection.org/works/d7xmprr8

License and attribution

Ex Libris
Roy L. Moodie.

W. H. M M.

C. xviii

20/p

Introduction to
Medical Biometry and Statistics

By

Raymond Pearl

*Professor of Biology in the School of Hygiene
and Public Health, and in the Medical
School,*

The Johns Hopkins University

SECOND EDITION, REVISED AND ENLARGED

Copyright, 1923, by W. B. Saunders Company.   Reprinted September, 1927.
Revised, reprinted, and recopyrighted October, 1930

Copyright, 1930, by W. B. Saunders Company

MADE IN U. S. A.

TO

WILLIAM HENRY WELCH

ARDENT ADVOCATE AND STRONG SUPPORTER
OF QUANTITATIVE IDEALS AND METHODS
IN THE MEDICAL SCIENCES

THIS BOOK IS DEDICATED BY ITS AUTHOR
AS A SLIGHT TOKEN OF HIS
DEEP AFFECTION AND ADMIRATION

PREFACE TO SECOND EDITION

THE changes which have been made in this edition have as their purpose the better adaptation of the book for class-room teaching of the elements of statistical methods useful in the biostatistical and medical fields. These changes have taken the form of additions, omissions, and rearrangements of the material. To a considerable extent the book has been rewritten. On account of the widespread interest in the matter at the present time a chapter has been added dealing with the logistic curve.

As before, I am deeply indebted and grateful to my colleagues for help in the preparation of this volume: especially to Prof. Lowell J. Reed and to Dr. John Rice Miner. Their suggestions, advice, and criticism have been invaluable, and they have also given much aid in matters of computation, etc. To the artist of the staff of the Department of Biology, Mr. Arthur Johannsen, I am indebted for the new illustrations, and to Miss Hermine Grimm for help in the details of manuscript preparation and proof-reading. Two former students, Dr. R. B. Tewksbury and Dr. T. J. LeBlanc, have been helpful with critical suggestions for the revision. I am very grateful to Prof. Haven Emerson of Columbia University and Dr. T. F. Murphy, Chief Statistician for Vital Statistics of the Census Bureau, for help in getting the new material incorporated in Chapter III. I am obliged to the Macmillan Company for permission to reprint in modified form the material in Chapter I under the heading "The Nature of Statistical Knowledge" from an earlier book, "Modes of Research in Genetics," of which that company owns the copyright. Finally to my old and dear friends G. Udny Yule, F. R. S., and Major Greenwood, F. R. S., I am deeply grateful for their permission to reproduce their portraits in this volume.

RAYMOND PEARL.

October, 1930.

PREFACE TO FIRST EDITION

THIS book is the result of many years' experience in attempting to teach biometric methods to biologists and medical men. Its faults and its merits, if any, both derive mainly from that experience. Perhaps nearly, if not quite, every traditional canon of supposedly sound pedagogy in the teaching of mathematics is done more or less violence to in the pages that follow. For this, as an admirer in some degree of tradition in general, I am sorry. My only plea in extenuation is a merely pragmatic one. The mode of exposition of the subject followed in this book *works*. I know because I have tried it, many times and on many people. Our students seem to like the subject, and to feel that they get something of value out of our presentation of it. Perhaps a teacher ought not to ask any more than this. Certainly I am not disposed to of men and women whose primary interest is, and will continue to be, in biology and medicine, and most certainly not in mathematics.

And there is this further to be said on the point: whether the mathematician likes it or not, there are now and there will continue to be, many biologists and medical men who are going to use biometric methods in their work whether they have had any special mathematical training or not. If we, who are charged with the elementary teaching of these persons, insist on a rigorous mathematical approach to the subject at every point, with complete analytical proofs of every step, the net result with the vast majority of students will simply be to disgust them, and drive them away from such sound elementary training as they might otherwise be willing to accept, and from which they, my colleagues, and I, at least, agree that they do profit. In writing this book, therefore, I have tried to present the mathematical matters necessarily involved in a language and with a logical method of ap-

proach which is not only capable of being understood by the primarily biologic or medical reader, but to which persons of this type of mind and training are sympathetic.

This book, as its title indicates, is and is intended to be, only an *introduction* to the subject. Many matters are omitted which might properly find a place in it. It is my belief, however, that in the present state of development of biometry itself, and in the use which is actually being made of its principles in biology and medicine by those who are not, and never will be, primarily specialists in this field, there is more need for a simple exposition of the basic elements of the subject than for an exhaustive treatise. The latter will, of course, come in time, but for the present it seems to me better to ground the student in elementary principles, and give him an introduction to the original sources, which he may follow up then for himself, to any degree he likes. In this connection there may be some inclined to criticize because of the brevity, and sometimes derivative character, of the reading lists at the ends of the chapters. The proper policy to pursue in this matter has greatly puzzled me. I have in manuscript a tolerably extensive and penetrating bibliography of vital statistics and biometry. I might easily have printed the whole of it herein. But again, the policy I have actually chosen to follow, after much deliberation, is based upon my teaching experience, which is to the effect that one can cajole a busy student into only a definitely limited amount of collateral reading. It is my conviction that it is, in a practical sense, better to recognize this fact frankly, and choose carefully a limited list of references, than to incorporate into a book which is not in any sense an original source an extensive bibliography. I am, in this particular case, the more happily led to this conclusion because of the splendidly thorough bibliography of the important original sources which already exists in Yule's "Introduction to the Theory of Statistics," which is, of course, the classic, model text-book of modern statistical methods, and is available to everyone.

This book is written for the medical reader primarily. The illustrations of method are mainly chosen from that field. Biometric methods already have a secure place in general biology.

Their use is developing in the medical field with extraordinary rapidity just now. It has seemed to me on this account that an elementary introduction to the subject designed primarily and directly for medical readers might be found particularly useful at this time.

I am indebted to various persons in many ways for help in the making of this book, though for its defects I am alone responsible. First of all, to my colleagues in this laboratory, who have loyally helped in the organization and development of our teaching work to its present stage, I owe a debt which I cannot adequately describe. We have worked out *together* our present method of teaching the subject. More specifically, I am deeply grateful to Professor Lowell J. Reed for reading critically the manuscript and catching up a number of errors which otherwise might have slipped by, and for discussing with me the most appropriate methods of presentation of many points, both in this book and in our courses of instruction. To Dr. John Rice Miner, Miss Agnes Latimer Bacon, and Dr. Flora D. Sutton I am indebted for the arithmetic work on many of the numerical illustrations of method. The wisdom and sagacity of Dr. William Travis Howard, Jr. in the broad fields of pathology, public health administration, and vital statistics have been freely at my disposal, and of inestimable aid in the whole development of the Department of Biometry and Vital Statistics of the School of Hygiene and Public Health, of which development this book is an integral part.

Finally, I wish most sincerely and gratefully to acknowledge something of what I owe to the great master and creator of biometry, Professor Karl Pearson. When, nearly twenty years ago now, I spent a winter in his Biometric Laboratory at University College, London, I got a fund of inspiration from first-hand contact with the working of his mind, which the passing years have never lessened or dimmed, and which I have tried to pass on to my students. If we have sometimes differed on biologic matters in these years, it has meant no slightest diminution of my deep and sincere admiration for one whose sheer intellectual power has rarely been equaled in the whole history of science. Feeling this way it is a great gratification and pleasure to me that Professor

Pearson has allowed me to present to the readers of this book the splendid portrait which appears on page 58.

In the little verse on page 16 the "file" which Robert Recorde was writing about was "geometrie." Such a "fresshe fine witte" as that old worthy's, however, would perceive and enjoy, I am sure, the peculiar aptness of the application of his lines to biometry today.

RAYMOND PEARL.

CONTENTS

ILLUSTRATIONS

All fresshe fine wittes by me are filed;
  All grosse, dull wittes wishe me exiled.
Though no mann's witte reject will I,
  Yet as they be, I wyll them trye.

—Robert Recorde

An Introduction to
Medical Biometry and Statistics

CHAPTER I

PRELIMINARY DEFINITIONS AND ORIENTATION

To an ever-increasing degree modern science is becoming quantitative in its methods of thought and activity. The history of science from the beginning shows that the earliest development of any discipline is purely qualitative, and that only as it emerges from this state and passes over into the quantitative phase, in greater or less degree, does it begin to take an assured place in the hierarchy of the established sciences. Recent examples of this change from a qualitative to a quantitative point of view are found in psychology and sociology. With the development of knowledge and of an appropriate technic eventually any natural phenomenon which can be observed can also be quantitatively measured. The entire history of medicine shows that there has been almost from the first an earnest desire and effort, on the part of some of its leaders, to develop quantitative modes of thought and methods of work. The large measure of progress which has been made in this direction is sufficiently evidenced by the number of items of diagnostic and clinical significance which are measured and recorded in quantitative terms.

In the ever-increasing specialization which occurs in science, and the multiplication of technical journals which such differentiation of interest necessarily entails, it is difficult, not to say impossible, for one to keep abreast of all the newer developments even in his own science, to say nothing of cognate subjects. This is particularly true for the practitioner and investigator in the field of medicine. The consequences are unfortunate. One often fails to get the benefit of applying, in his own subject, what might be

very useful methods or ideas from another science. This lack of familiarity with even the simplest technical terminology of one of the newer special fields may be so complete as to be embarrassing in a general scientific gathering or discussion of any sort. It is only fair that any one proposing to set out the bearings of one of the newer and somewhat highly specialized branches of science upon an older and established field and to discuss its methods, should begin by clearly defining at least the more general technical terms he intends to use.

DEFINITIONS

Biometry is a term which came into general use in the late nineties, to designate that branch of science which studies by methods of exact measurement on the one hand, and precise and refined mathematical analysis on the other hand, *the quantitative aspects of vital phenomena*. It is a term co-ordinate with biology in its comprehensiveness. Indeed, it may perhaps happen that with the passage of time the term "biology" will be used to cover only qualitative phases of vital phenomena, while biometry will be the identifying term for all discussions of measurements or counts of living things in the widest sense of the words. The general tendency of all science is to proceed always toward greater and greater precision of results and reasoning. It has elsewhere been pointed out that "the real purpose of biometry is the general quantification of biology. Its fundamental point of view is that, without a study of the quantitative relations of biologic phenomena in the widest sense, it will never be possible to arrive at a full and adequate knowledge of those phenomena. This point of view insists that a description which says nothing about the magnitude of the thing described is not complete, but, on the contrary, lacks an element of primary importance. It insists, also, that an experiment which takes no account of the probable error of the results reached is inadequate and as likely as not to lead to incorrect conclusions."

Biometry, as a definitely recognized branch of biologic science, owes its origin and establishment primarily to the efforts of two men—the late Sir Francis Galton, and Karl Pearson, Galton Professor of Eugenics in University College, London. In a later

chapter the part played by each of these men will be set forth with greater particularity.

The definitions of *statistics* given by Yule, in his well-known *Introduction to the Theory of Statistics*, which is by all odds the best general elementary introduction to the subject, are extremely clarifying and helpful. He says: "By *statistics* we mean quantitative data affected to a marked extent by a multiplicity of causes.

"By *statistical methods* we mean methods specially adapted to the elucidation of quantitative data affected by a multiplicity of causes.

"By *theory of statistics* we mean the exposition of statistical methods.

"The insertion in the first definition of some such words as 'to a marked extent' is necessary, since the term 'statistics' is not usually applied to data, like those of the physicist, which are affected only by a relatively small residuum of disturbing causes. At the same time 'statistical methods' are applicable to all such cases, whether the influence of many causes be large or not."

There is another way in which we may define statistics, which has important bearing upon the logical development of the subject. It may be said that:

Statistics is that branch of science which deals with the *frequency* of occurrence of different *kinds of things*, or with the *frequency* of occurrence of different *attributes* of things.

If we discuss the case incidence of typhoid fever we are dealing with the frequency of occurrence of things, for what we say is that of N people constituting a population or group, a certain number, A, have typhoid fever within a given interval of time, while during the same interval another number, $B = N - A$, do not have typhoid fever. Here, then, are two *kinds of things*, namely, people who have typhoid fever and people who do not. And so similarly for all other cases where the figures with which we are presented are simple *counts* of the number or frequency of occurrence of physically discrete entities.

Let us now look at the other side of the case. Stature is one attribute of a man, in the sense that the word "attribute" is here used. Suppose we measure carefully the stature of each of 1000 men.

We can then sort these measures (the attributes) into a series of groups such that each group shall contain only statures which are nearly alike, say differing by not more than 0.5 cm. Then, if we count the number of cases in each group, we shall have the *frequency* of occurrence of each particular kind of attribute (*i. e.,* particular stature) within the original group of 1000. From these frequencies we may then calculate, by simple processes to be fully explained farther on, certain derivative constants like the *average* stature, etc. But these derived functions are all implicit in the frequencies, and have no validity beyond that which inheres in the original counts.

All statistics are comprised within one or the other of these two categories, frequencies of things themselves, or of the attributes of things.

The separateness of things which makes them countable for statistical purposes may be relative either to space, or to time, or to both space and time. If, upon the same day, as in a census, we count the number of cases of typhoid fever existing in a city, we shall have gathered statistics of the frequency of persons with typhoid fever, *upon a space base*. The underlying differentiant factor which makes these cases countable is that each is, at the same instant of time, located at a particular and unique region in space. Suppose, on the other hand, we consider as a universe of discourse 1000 particular persons and observe these same persons every day for a year to see whether typhoid occurs among them, it being premised that they do not move about at all. We shall then have at the end of the year the frequency of occurrence, within the group, of persons with typhoid fever, *upon a time base*. Another example may perhaps help to clarify the point. We may study, as the writer once did, the variation of milk production by dairy cows in two ways. If we examine the differences in amount or quality of milk produced by each individual cow in a large herd on the same day, we shall be studying the variation in milk production *on a space base*, since each cow is a spatially separate entity. But suppose, with this same herd, we pour each cow's milk each day into one big vat, mix it thoroughly with the milk of all the other cows in the herd, and then weigh or measure the whole amount of milk in the vat each

day, and by drawing a sample from it determine the butter-fat percentage, etc. The amount and quality of this *herd's* milk, the *herd* now being one single spatial entity, will vary from day to day throughout the year. If now we examine this *daily* variation, we shall be studying the variation of milk production *upon a time base*.

The statistical method is essentially a *technic*, which finds its justification in its usefulness in helping to solve the problems of the basic sciences, physics, chemistry, biology, etc. Statistics, in any proper sense, has no, or at best few, problems of its own. Its technical problems are really problems of mathematics. The statistical method is, or should be, a working tool of science, just as is the microscope or the kymograph. But it is probably of wider utility than any other single technical method which science has discovered or devised. For it has an applicability and a usefulness, direct or indirect, in virtually every problem. It is, in short, a fundamental element of scientific methodology.

Biometry deals with statistics derived from living things, or things which have at some time been living, and applies statistical methods, in the broadest sense, to such data.

"Vital statistics," for which a better term is *biostatistics*, is the special branch of biometry which concerns itself with the data and laws of human mortality, morbidity, natality, and demography.

In this book the attempt will be made to show, by concrete examples, how the point of view of biometry, and the application of modern statistical methods, may be of use to the medical man in helping him to draw correct conclusions from his facts, and to solve problems constantly arising in his work, which he cannot possibly hope to solve correctly without such methods. It is not to be expected, or perhaps even desired, that every medical practitioner or investigator shall be an accomplished mathematician. But it is evident enough to every thoughtful observer that clinical medicine is proceeding by great strides along the quantitative, scientific pathway. Every step in this direction adds to the necessity of the medical man having at his command the necessary elementary principles for dealing easily, confidently, and accurately with quantitative data.

IMPORTANCE OF BIOMETRIC IDEAS AND METHODS IN MEDICINE

The growing recognition by medical men themselves of the importance of modern biometric methods and viewpoint for work in medicine was forcibly expressed a few years ago by the distinguished clinician, Dr. Lawrason Brown, in the following words[*]:

"None of you will contradict me when I say that statistics are very dry, but some of you may dispute me when I say that only by statistics does the world, lay or medical, advance. Consider what knowledge is and you will see how inseparable it is from statistics. Medicine is no exact science, and diagnosis rests largely upon the law of probability which, in turn, is statistical. All scientific experiments are statistical arguments in favor of or in opposition to certain inductions or deductions. Further, statistics lend the authority that is necessary for their acceptance.

"The trouble in medicine does not lie with the statistical method, but with the medical men who do not know how to use it. I regret to state that I belong to this class and have felt keenly that in medical school I did not have an opportunity to attend a course on medical statistics. The day will come, gentlemen, when such courses will be given, when the law of probability will help in diagnosis, when the coefficient of correlation, now explained by most authorities in such terms that in a few minutes my idea of my relation to my surroundings has become totally insufficient—when, I say, all these things will be understood by the medical graduate. At that time medical men will cease to do such foolish things with statistics as to try to add cabbages and cows, or, what is nearly as bad, to try to solve problems in heredity by finding how many parents had the disease from which the offspring suffers without due respect to many other very important and possibly contradictory details. What would you think of a bookkeeper who after years of personal experience would gather up the bills in the cash drawer and go to the bank with the statement that his personal experience led him to believe that the roll of bills amounts to $1000. The receiving teller would quickly apply the statistical method and few would venture to side with the bookkeeper, no matter how large his experience had been.

"Do not misunderstand me. This is not an argument in favor of dry statistical articles which we all prefer to avoid reading. But if I can make you see how important it is for us to cease using the pet phrase 'my personal experience' except when we have sufficient data to support it, I shall have accomplished what I had hoped for."

The point of view from which medical problems should be attacked by quantitative, biometric methods has been well set forth by Greenwood[11] in the course of a discussion of some animadversions of Sir Almroth Wright upon quantitative methods, when he describes the method by which a therapeutic problem ought to be investigated. Greenwood remarks:

[*] Brown, Lawrason: American Review of Tuberculosis, September, 1920, vol. iv.

"Let us suppose that the question is whether a certain treatment is of advantage in acute lobar pneumonia. We must first inquire whether the morbid state connoted by the phrase 'acute lobar pneumonia' is clinically recognizable. The question is answered in the words of Sir William Osler: 'No disease is more readily recognized in a large majority of cases. The external characters, the sputum, and the physical signs combine to make one of the clearest of clinical pictures. The ordinary lobar pneumonia of adults is rarely overlooked.'

"The next point to be investigated is the variation of fatality in cases not treated by the method under investigation.

"(a) *Influence of Age.*—That the fatality increases with the age of the patient is well known and evidence need not be quoted here. Naturally, in comparing fatalities it will be necessary to correct for age.

"(b) *Sex.*—The influence of sex is not so marked, but allowance can similarly be made for it.

"(c) *Secular Variations.*—It would appear that these are of minor importance. It also appears that the fatality of hospital cases from different institutions in the same country during the same period varies but little.

"(d) *The Influence of Social Class.*—Evidence capable of being analyzed has been sparingly published. The 873 cases recorded by the British Medical Association's Collective Investigation Committee in 1886 show a corrected fatality rate of 17 per cent., which is below the London Hospital rate for the same period. The results of Huss at Stockholm, more than forty years ago, suggest that the fatality in the Military Hospital was about seven-elevenths of the rate obtaining in the General Hospital.

"(e) *Influence of Race or Climate.*—We find striking differences in the hospital fatality rates of different countries, the rate at the Stockholm Hospital in the 'fifties' of last century being far below that recorded for the same period at Vienna or Basel. There is a less striking difference between the recent London figures and those of Chatard from Baltimore.

"In view of what has been said, it will be plain that in comparing a series of treated cases with 'general experience' attention will have to be paid to the differences noted, all of which can be tested by the statistical method. When a true control series is available, it will still be necessary to allow for race and environment. An inquiry into these points would seem a necessary prelude to an evaluation of the effects of any specific treatment.

"These are all questions of great moment, and cannot be answered by appeal either to authority or to the introspective notions yielded by the 'experiential method.'

"Having made due allowance for these difficulties, we shall proceed to compare the rate of mortality in the treated and untreated cases. This will involve a careful sifting of the material, since we must reject such cases as died in consequence of some accident in no way connected with the evolution of the disease. The criteria of exclusion must be defined, and no case excluded without the grounds of such exclusion being clearly stated and the particulars published in full to give others an opportunity of judging the sufficiency of the criterion.

"Next, we shall in some cases be able to compare the percentages and determine the probability that such difference as results might be an 'error of random sampling.' This will by no means complete the task, however, since it might

happen that the treatment, although not associated with a significant reduction of fatality, did influence the course of the disease. The features which it is desired to measure having been determined on, we can by the method of multiple correlation endeavor to connect the variations of such features with each other and with those of the therapeutic factor we are studying. Since in general it will be difficult to secure controls and treated samples absolutely alike in other respects, the method of correlation is likely to be required in most cases. We shall, indeed, be fortunate if we are able to 'express the final result in the form of a percentage.'

"I have outlined the process by which, as I think, such a problem may be investigated. The essence of the whole matter is to ask ourselves at every turn, Is the control a real control? What is the probability that such and such an event is due to such and such a cause? There is no intrinsic merit in numbers and percentages or in coefficients of correlation, their value is in aiding us to think clearly and compelling us to express conclusions in a language which all may master if they choose."

Dr. Alfred E. Cohn, of the Rockefeller Institute for Medical Research, in a recent letter regarding the work of the Heart Committee of the New York Tuberculosis and Health Association, discusses the significance of statistics in medicine from a still different angle. He has kindly permitted quotation from this letter here.

"The value of these investigations, statistical in nature, has often been made the subject of solicitous, not to say sceptical enquiry. The Research Committee is nevertheless convinced of the value of its enterprises. It sees in them a continuation of that effort at classification of diseases, related to the heart in this case, which has always been recognized to be a sound and valuable tradition in the history of medicine, and indeed an indispensable method in the history of science in general. That these studies should be reliable in the sense in which statistical studies are believed not to be so, has been a matter of great solicitude on our part. The method of work has been described; it depends on securing reliability by attending to uniformity in nomenclature, by following specific criteria for naming diseases, by recording phenomena in a uniform fashion on history and physical examination forms which have been carefully constructed, by supervising and by collating the material once it is recorded through a staff of statistical clerks which now has had a detailed experience of several years duration. So far as we can judge, foresight and precaution have done their share in assuring satisfactory results.

"If there is still doubt of the value of results such as these, the doubt must, it seems, rest on popular conceptions—perhaps popular misconceptions—of the nature of statistics. What physicians require and what after all they must have in the practice of medicine is information twofold in nature. They must know about the general movements of diseases, or their natural history; but they will also desire a knowledge of methods which make applicable to individual cases the general considerations to which reference has been made. The difficulty is here. Except in so far as the former aids an understanding of the latter, it can scarcely

be pretended that the statistical or general method can be a useful practical instrument. Why this is so demands perhaps some discussion.

"The search for law in biology and of course in medicine, rests on the conception that the discovery of laws has served the physical sciences in extraordinarily useful ways. There can be no doubt of the soundness of this belief. But precisely what the analogy is between law and the individual in the physical world, and law and the individual in the biological one, requires precise definition —more precise indeed than is usually accorded to this matter. If there is a difficulty it lies, we believe, in misunderstanding this relation of law to the individual in the physical world. The view we take may be illustrated. In the case of the gas laws for instance, beginning with Boyle, a number of statements have been made which permit accurate predictions of the behavior of *volumes* of gas; that these laws do not describe the behavior of individuals within the volume is amply demonstrated by reflecting on the fact that the kinetic theory assumes violently diverse and unpredictable behavior on the part of the individual molecules in these volumes. The laws apply to the mass, the volume, the average; they make no statements concerning individual performance. And yet the laws are invaluable; they and their kind are the basis of calculation in the practical as well as in the theoretical world. In biology and in medicine, just as in physics, it is not to the individual that the laws, whatever they are or may be discovered to be, apply. Individuals represent deviations from any law both in biology and in physics; if a law is sound, the deviations from the average must not however exceed a certain maximum, the probable error. If it does there is no law of value. Deviation is the fate of the individual; uniformity in the sense of identity either of being or of behavior scarcely exists in any world. The general behavior of patients afflicted by typhoid fever, the general behavior of mobs may be known, but to know these phenomena has, relatively speaking, little meaning in understanding or in predicting the conduct of any individual in a mob or in making an accurate prognosis, based on general experience in the case of any typhoid fever patient.

"And yet no one denies that general statements can be made and are useful in physics and in psychology nor that general statements on prognosis and on the natural history of diseases have value. General statements and inference in individual instances each have their domain of eminent usefulness. Harm results only when the nature and objects of the two are confused. This is the direction in which the Research Committee believes it comprehends the function of its labors. Whether in the natural history of any one of the heart diseases, or in an estimation of prognostic values, or in the measure of success of a therapeutic agent, or of the degree of relevance of social or economic advice, its aim is the attempt to understand *general* movement. No other indeed is possible. It believes, because it is plain teaching of the history of science, that the ability to attain orientation of this sort is indispensable in envisaging the probable course of any individual life or of any individual act. Thought and action would otherwise be chaos."

THE NATURE OF STATISTICAL KNOWLEDGE

There is a very general tendency, including in its operation not only the layman but also the professional man of science, toward

the notion that there is a special virtue, a sort of transcendent heuristic worth, in such knowledge as is reached by the examination of large numbers of cases. There seems to be a feeling, sometimes apparently almost mystic in its origin and in its strength, to the effect that statistical knowledge is a higher and better kind of knowledge than any other. Numberless quotations might be cited to show the prevalence of this view. Every one has seen passing, as it were in review, the line of problems, which, if we may trust the assertions of the interested individuals, "can only be solved" by the application of the statistical method.

Now this attitude toward statistical knowledge and statistical ideas (which, of course, include besides the compilation of large numbers of individual instances, the concepts of averages, approximation, and probability) may be entirely right and justifiable and certainly is so in considerable part. Indeed, a cautious person is bound to be very chary about even suggesting any criticism of it when he considers the eminence of some who have espoused it. But the statistical method, as an organized and formulated scientific technic, came only relatively lately into the field. A realistic examination of its powers, sympathetic if critical, cannot do any harm.

It is the object of the following remarks to discuss statistical concepts and methods with the purpose of trying to see what these methods are, in fact, capable of doing. In this discussion let us endeavor to avoid dogmatic assertion, since, in the first place, assertion does not really get us far in the search for truth, and, in the second place, the writer himself feels in regard to these questions very far from that serene consciousness of being quite unassailably right which is essential to proper dogmatism. Indeed, it is for the purpose of definitely formulating some doubts, which have grown in his mind during many years, that this discussion is written. Very likely some will not agree with its reasoning or its tentative conclusions, but even in such event, it may help the disagreeing reader to the more complete ordering of his own ideas about statistical concepts.

Let us first consider this question: What caused the development of the statistical viewpoint and method, which in science had such an important growth in the nineteenth century? For what purposes did men turn to the statistical method? This question

has been very ably discussed by Theodore Merz in the second volume of his *History of European Thought in the Nineteenth Century*, and we cannot do better than follow his development of the matter. Speaking of the origin of statistics, Merz says (*loc. cit.*, pp. 554, 555):

"That which everywhere oppresses the practical man is the greater number of things and events which pass ceaselessly before him, and the flow of which he cannot arrest. What he requires is the grasp of large numbers. The successful scientific explorer has always been the man who could single out some special thing for minute and detailed investigation, who could retire with one definite object, with one fixed problem into his study or laboratory and there fathom and unravel its intricacies, rising by induction or divination to some rapid generalization which allowed him to establish what is termed a law of general aspect from which he could view the whole or a large part of nature. The scientific genius can 'stay the moment fleeting'; he can say to the object of his choice, 'Ah, linger still, thou art so fair'; he can fix and keep the star in the focus of his telescope, or protect the delicate fiber and nerve of a decaying organism from succumbing to the rapid disintegration of organic change. The practical man cannot do this; he is always and everywhere met by the crowd of facts, by the relentlessly hurrying stream of events. What he requires is grasp of numbers, leaving to the professional man the knowledge of detail. Thus has arisen the science of large numbers or statistics, and the many methods of which it is possessed."

Further on the same author says of the origin of the science of probability (*loc. cit.*, pp. 567, 568):

"The necessity of having recourse to elaborate countings, to registrations of births, deaths, and marriages, to lists of exports and imports, to records of consumption and production of foodstuffs and many other items, forced upon those who were intrusted with the gathering and using of these data the observation that all such knowledge is incomplete and inaccurate. Owing to the variability, within certain limits, of recurring events and the errors of counting and registration, we have to content ourselves always with approximation instead of certainty. Error bulks very largely in all statistics, and vitiates them; and as regards coming events, our minds are in a state of expectation rather than of assurance. But events can be more or less probable, errors can be greater or smaller, cumulative or compensatory, and our expectations may be well- or ill-founded. And so there has arisen the science of Probabilities and of Chances, and the Theory of Error, two subjects intimately interwoven. The former arose in the seventeenth century out of the frivolous or vicious practice of betting and gambling, whilst the latter was founded when astronomical observations accumulated, and the question presented itself how to combine them so as to arrive at the most reliable result."

Now from these two quotations, which may certainly be considered as fairly stating the case, it is apparent that those circum-

stances which led men to turn to statistical methods of reasoning and investigation were not such as grow out of an increasing precision and certainty of knowledge about the events or things under consideration, but rather were quite the opposite. In other words, the statistical point of view, in the first instance, was adopted as an admittedly imperfect means of getting some sort of knowledge about a class of events concerning which it was difficult or impossible to get by other methods the precise or particular kind of knowledge which was wanted. To take a concrete example. A life table tells, with considerable and commendable accuracy, when men aged fifty-six, say, will die, on the average. But what the family, his business associates, his physician, and a good number of other people would *like* to know is when John Particular Smith, aged fifty-six, will die, and this statistics are unable precisely to tell. No honest and intelligent person can be deluded into the belief that "in general" or "on the average" knowledge is as satisfactory or useful as "individual" knowledge would be if he could get it, when it is individuals he is concerned about, as is mostly the case.

A careful consideration of the history of statistical science, as well as of the present-day application of these methods, leads to the conclusion that statistical methods are used for two sorts of purposes, or to gain two sorts of knowledge about events or things.

(*A*) On the one hand the statistical method finds one of its chief uses in furnishing a method (and the only one known in science) of describing a *group* in terms of the group's attributes, rather than in terms of the attributes of the individuals which compose the group.

What sort of positive, definite, and exact knowledge do statistics give us?

1. Precise knowledge of the *composition* of groups or masses. This is the knowledge gained by *counting*. Suppose we find a basket containing a number of balls of several different colors, and proceed to count them with the following results:

> 7 Reds
> 9 Whites
> 2 Blacks
> 1 Green

Such a count furnishes us at once with a great deal of perfectly definite and precise information about this group or population of balls. For example, the count tells us that it will never be possible to take away from the basket more than one pair of balls of which one member is green. This is a definite attribute of this population which may be used to differentiate it from other populations. In this particular population only one green ball occurs.

This sort of knowledge derived by counting is perfectly definite and precise so far as relates to the particular group which it concerns in any particular case. It does not involve any approximation, or probability, and is as precise as knowledge of the individual. It, however, pertains to the group. It forms a part of a proper scientific description of a group to count the numbers of each of the different kinds of elements which compose it.

2. Knowledge of certain *abstract qualities* of groups. This knowledge is obtained by calculation from the data got by counting. The more important of the abstract qualities* of groups are:

(*a*) The *central* or *typical condition* of the group; or the condition about which the individuals composing the group cluster. This is variously measured: by the arithmetic mean or average, which gives the center of gravity of the group; by the median, which tells the point on either side of which exactly half the individuals fall; by the mode, which tells the point of greatest frequency of occurrence in the group, etc.

(*b*) The degree of *individual diversity* comprised in the group. This attribute, called the variability of the group, is again variously measured: by standard deviations, coefficients of variation, etc.

(*c*) The *degree of asymmetry* of the distribution of the individuals composing the group. This is measured by the skewness or other related constants.

(*d*) Various other attributes of distributions might be here included, such as, for example, the kurtosis, but for purposes of the present general analysis this is not necessary. Though some of these attributes involve very complex mathematical expressions for their determination, the general fact remains clear that they are

---

* A more detailed discussion of the following constants will be found in Chapter XIII.

all attributes of groups or masses which are described by the statistical constants.

One point should be quite clear. It is that the kind of knowledge discussed under this heading 2 is just as definite and precise, and involves as little approximation and indeterminism, as does any piece of individualistic knowledge, *so long as we confine attention solely to the particular group discussed in a particular single case.* It is the custom to state means, for example, with probable errors. But this is only because it is proposed, overtly or tacitly, to extend the conclusions beyond or outside of the particular group and the particular instance for which the mean was calculated. *For that group and that instance* the mean is perfectly exact and precise to that degree of precision denoted by the unit of measure used, assuming that no arithmetical mistakes have been made in its computation. Thus suppose one measures the stature of three men to the nearest inch, and then calculates the average. The result is, without any probable error, the average height, at the particular moment when they were measured, of *those three men exact to the unit of measurement used.* It describes and measures precisely an attribute of those men considered together as a group or trio. But if we were to consider this result from the viewpoint of whether it gave a reasonable measure of the average height of men in general, or from the viewpoint of whether it gave a proper value for the mean height of these men when repeatedly measured under varying conditions, it would clearly be subject to a large probable error. It would, in point of fact, have lost its character of precise and definite knowledge, and have become a more or less poor prediction, approximation, or guess, for the reason that three men are too meager a number to give any reliable indication of the attributes in general.

3. Knowledge of the *degree of association* or contingency between different events or characters within a group. This is furnished by the method of correlation in one or another of its various forms. By this general method it is possible to get a numerical index of the degree of likeness, in the direction and amount of the variation in two or more characters in the individuals composing a group. So long as attention is confined to the particular group on

which the measurement is made, and to that group alone, and to a single instance (in time) the knowledge gained is precise. It is a part of the description of the attributes of that group. But when we endeavor to predict from that particular group to other groups or individuals or to conditions in general, our results are no longer precise, but inferential and what are called probable errors tell us something about the degree to which the inference may be regarded as trustworthy.

Summarizing the results of the above analysis, we see that the statistical method can

1. Furnish precise descriptive knowledge about groups. This knowledge is of various sorts. It is definite and precise so long as attention is confined solely to the particular group and the particular instance on which it is based.

2. The knowledge gained by the statistical method, as we have analyzed it above, precise though it may be, *pertains to the group and not to the individual.* It is exact knowledge about the composition, or attributes, or contingencies of masses or groups.

3. This ability to describe groups in terms of the groups' own attributes, which is an unique property of the statistical method, is extremely useful in the practical conduct of scientific investigation. It makes the statistical method a valuable adjunct to every other scientific method, and particularly to the experimental.

(B) We may now turn to a wholly different aspect of the statistical method, wherein it is used for the purpose of predicting or estimating the probable or the approximate condition in the *individual* from a statistical examination of the condition in the mass or the group. Resort is had to the statistical method for this purpose primarily in those cases where the outcome of the event, or the condition of the thing in a particular individual case cannot be directly determined by direct examination of that particular individual, because of spatial or temporal limitations imposed by the nature of the problem; and also where the outcome of the event or the condition of the thing is determined by the combined action of a large number of small causes, each about equally influential upon the final result.

Originally the statistical method was only employed for this

second purpose in cases where, because of the multiplicity of the cause groups involved in the determination of the event, and the consequently small effect of each, it was impossible to make any reasonable prediction regarding an individual from an examination of that individual alone. Such employment might be considered legitimate, though not very fruitful, on the ground that any prediction so made, uncertain and doubtful as it may be, is after all perhaps better than no prediction at all. As time has gone on, however, there has been an increasing tendency to assume that this use of the statistical method has general a *priori* validity and can be profitably employed in all sorts of cases.

This leads us to consider carefully the general question of the validity, on the one hand, and the usefulness, on the other hand, of this whole second mode of employment of the statistical method. It is the one which has attracted the greatest attention because of its essentially spectacular nature coupled with a sort of mysteriousness bordering upon the miraculous. It seems a wonderful, indeed almost a superhuman, accomplishment to be able to say in the manner of the oracles of old, "So many men will commit suicide next year."

Since Clerk-Maxwell introduced statistical modes of reasoning into physical science there has been an ever-increasing tendency to regard the universe as organized on a statistical plan. This has come, by gradual evolution, to carry with it two implications, one of which seems quite fallacious and the other partly so.

The first of these is that the individual events, of which all the causes are not precisely known to us, are indeterminate. Such an assumption is, of course, unwarranted. Because we do not know all the causes leading to a particular event does not mean that that event is any less precisely determined by the course of antecedent events. Consider a box containing 100 consecutively numbered cards. Suppose one card were to be drawn and that it bore the number 36. It would be quite impossible to formulate precisely all the causes which led to the drawing of the number 36 on the particular occasion considered, but it is equally impossible to conceive that this particular drawing was not definitely "caused." In other words, there clearly was a whole train of antecedent circumstances, which taken all together definitely resulted, *and could*

only have resulted, in the drawing of the number 36. The too prevalent conclusion that the application of the statistical method or statistical modes of thought implies phenomenal indeterminism in the individual case seems to be totally fallacious.*

The second currently accepted implication of a statistical view of the universe is that in general a particular event or phenomenon is the outcome of the combined action of a great number of causes, each of which alone produced but a small part of the final total effect. There is clearly so much truth in this point of view as is included in the fact that individual events or phenomena do, in some degree or other, vary, and further these variations in general distribute themselves more or less in accord with well-known laws of error. But the assertion that events are individually the outcome of the action of great numbers of causes, each of which had a small part and a part significantly equal to that played by every other one of the causes concerned in the final result, appears upon examination to be true only if the "universe of discourse" is indefinitely extended in time. But *practically* science works in a definitely and rather narrowly limited universe of discourse so far as concerns time. It undoubtedly is true that a vast number of small causes do play a part in the determination of any particular event. But, in many of the events, at least, in which science is interested, these multitudinous minor causes do not play any *significant* part in the differential determination at a particular instant of time. There is in connection with the causation of most events some one or two, or at most a very few, outstanding cause groups, which, for all practical purposes, at a given moment completely determine their occurrence. The total effect of all the vast number of other minor causes concerned in the remote past is so minute, as compared with the part played by the really determinative ones at the moment, as to be negligible. In other words, all natural cause groups are not small, nor of equal (balanced) value in the final determination of the event to which they relate, provided we confine ourselves to the time limits of finite practical operations.

* It should, however, be said that there are a few men of science and philosophers who take the view of phenomenal indeterminism. Their view has not won general acceptance.

The fact that all natural causes or cause groups are not equally significant quantitatively is, of course, what makes the experimental method fruitful—one might perhaps even say possible—in science. The very essence of the experimental method is that the conditions for the happening of an event are so arranged that the influence of one putative causal factor or a very limited number of such factors may be tested at a time. If with a radical change in this one factor, whilst all others remain, so far as may be, constant, no change in the happening of the event is observed, the experiment has shown that this particular factor has no *significant* causal relation to the happening of the event. If a marked change in the happening of the event is observed always to follow the change of conditions of operation of the factor under investigation, then clearly this factor plays a determinative part.* In other words, it is a fundamental logical prerequisite of the experimental method if it is to be successful (that is, contribute to knowledge) that it operate in a universe in which all causal factors are not of equal quantitative significance at any given instant of time.

Clearly experimental analysis of this sort would have quickly discovered, if the common sense of men had not long previously shown, that the course which a particular event is going to take is not immediately the result of the action of an indefinitely large number of individually insignificant causal factors, but that it is the outcome of the action of a few immediately *determinative* factors, and the effect of the indefinitely large number of historically antecedent small causes is insignificant in the sense of being differential. Generalized, the point may be put in this way: an event A is about to happen. It may happen in any one of n different ways, each one of which ways may be designated by a letter, l, p, r, t, etc. Now an indefinitely large number of causes are concerned in bringing it about that the event A is going to happen, and that it can equally well happen either as l, p, r, t, etc. In other words, the setting of the stage for the event has involved a vast number of small and balanced causes. But the causes which are differential in the particular case, that is, which determine that A

* Cf. Jennings' valuable paper on radical experimental analysis, *American Naturalist*, vol. 47, pp. 349–360, 1913.

shall happen in the p way this particular time, and not in the l, the t, or any other way, are, in general:

1. Few in number.
2. Immediate in time.
3. Large in relative quantitative effect.

The point under discussion may perhaps be made plainer by a homely illustration. Suppose a man steps up behind a mule and prods the creature with his walking stick. The human intellect is unequal to the task of predicting exactly, in the particular case, what precise portion of the man's body the mule's hoof will land upon. A multitude of minor causes will affect this: the relative height of the man and the mule, the age of each, the place poked with the walking stick, the degree of fatigue of the mule, the temperature, the season of the year, and countless other things have an influence in determining just the precise spot where the mule's foot and the man's body come together. These could be investigated statistically and tables drawn up from which one could predict the part of the man which would probably receive the hoof. But what a silly, futile piece of business this all would be, since clearly the influence of all these small causes on what happens to the man is stupendously overshadowed by the results of two factors, namely, putting himself behind the mule and prodding the animal with a stick. Of course, a vast number of antecedent causes are involved in the setting of the stage, but these are not differential in the determination of the end-event of the series.

The preceding illustration has nothing directly to do with science, but the essential point involved operates, in the use of the statistical method as a weapon of scientific research. This method, being only a descriptive method, tells us nothing *directly* about the causes involved in the determination of any events or phenomena under consideration. It may be of great aid, in combination with the experimental method, in helping to arrive at such knowledge, but alone and of itself it cannot directly furnish knowledge of causes of individual events. Yet the statistical method, particularly in that phase of it which we have here under discussion, which essays to predict the probable condition of the individual from the knowledge of the mass, *seems* to furnish information about causes. It

wears a specious air of bringing a kind of knowledge which in reality it not only never does, but from the very nature of the case never can furnish.

Let us consider now a little more in detail the nature of the prediction of the probable condition of the *individual* from a knowledge of the *mass* or group. It has been shown that statistics give perfectly definite and precise, and often very useful, knowledge about masses or groups. We are now, however, not concerned with this as group knowledge, but rather with one use to which such knowledge has been put. This use is that which is comprised in the subject of statistical probabilities, and which involves the drawing of conclusions as to the *probable* condition of the individual, based on an *exact* knowledge of a particular mass or group.

In order to approach the subject in the simplest way let us consider a concrete case. Suppose a problem of the following sort were to be set for answer: What is the probability that, at some chosen moment of time, the next birth to occur in, let us say, the city of Baltimore, will be of a white child. Now if we look at this as a question of statistical probability the appropriate way, of course, to go about solving it is to turn up the registration reports for the city of Baltimore covering a period of years, and find out what is the proportion of white to colored births in that city. Then by the simplest theorem in the calculus of chance, the probability that any single birth in Baltimore, taken by itself, will be of a white child is conventionally regarded as given, in principle, by a fraction of which the numerator is the number of white children born in Baltimore and the denominator is the total number of children born in Baltimore, both figures including the same period of time. When we have obtained such a fraction we have a definite piece of statistical knowledge, but of just what use is it so far as concerns a *particular* individual case, the "next" birth? It implies no biological knowledge of any kind; no knowledge of the laws of heredity. It really adds essentially, it seems to me, to the sum total of the world's knowledge only one thing. That thing is the proper betting odds on what the color of the next child born in the city will be. This knowledge would be really useful, in a pragmatic sense, only provided some one wishes to gamble upon that event.

Of course the statistical count, on which the probability is based, in itself furnishes definite and precise information about the population of Baltimore, *as a population*. This may be useful. What we are now considering, though, is knowledge about individual cases.

Let us see what a totally different kind of ability to predict the future event in an individual case is gained when we take into account one single biological fact of an individualistic instead of a statistical character. Suppose, that is to say, that we are informed that the mother of the next baby to be born in Baltimore is a black woman. It needs no argument to show how much more precise will be the prediction as to the color of the next baby under these conditions.

This illustration brings out clearly the difference between the two possible bases for the prediction of a future event. On the one hand, such prediction may be based merely on statistical ratios. This means only a count of an indefinitely large past experience regarding the occurrence or failure of the event, but in no way takes into account the causes which underlie the happening of the event in any particular case. On the other hand, we have the prediction which is based on a definite knowledge of the determinative causes which bring about the happening of a particular individual event of the sort in which we are interested and about which we are to predict. There can be, it would seem, no comparison between the usefulness, in the pragmatic sense, of these two kinds of knowledge. The statistical knowledge on which a statistical prediction is made is essentially the most sterile kind of knowledge that one can possibly have *so far as concerns the individual event*. It merely gives one the betting odds for or against the occurrence of that event, and absolutely nothing more. Now a wager, however large, in the scientific sense neither discovers, expounds, nor is a criterion of the truth. Bets, in other words, are not evidence, though the statistician sometimes seems to forget this, and to deal with statistical ratios as though they had probative worth in regard to individual phenomena.

On the other hand, a prediction based on experimentally acquired knowledge of the determinative cause of the individual event brings with it a more realistic knowledge of a natural phe-

nomenon. The predictions so made may not always turn out correct, but when they do not, it incites us to investigate the particular disturbing factor which, under such circumstances, may overwhelm the normally determinative cause of a particular event.

Man soll das Kind nicht mit dem Bade verschütten. The critical reader may be inclined to think that this is exactly what the discussion in this section has done. If, as has there been suggested, that part of the statistical method which uses the calculus of probability as a basis for the prediction of future events, gives only a knowledge of betting odds, one may ask: What about the whole concept of probable error? The value of this concept in scientific research is unquestioned. Yet plainly the whole concept has its basis in the calculus of probability. Has not our discussion led us unwittingly into a serious contradiction?

I think not. Let us examine the probable error concept a little more carefully. Suppose we read that the mean length of the thorax of a thousand fiddler crabs is 30.14 ± 0.02 mm. Just what does this actually mean? Accepting the figures at their face value, or, put another way, assuming for the argument that the mathematical theory on which the probable error was calculated was the correct one, the figures mean something like this: If one were to take, quite at random, successive samples of 1000 each of fiddler crabs and determine the mean thoracic length from each sample, these means would all be different from each other by varying amounts. In other words, no single sample would give us the absolutely true value of the mean thoracic length of all fiddler crabs in the world. This true value is in an absolute sense unknowable, because, for one reason, always we must come at the finding of it by the way of random sampling, and sampling means variation. Now it is an observed fact of experience that the variations due to random sampling distribute themselves according to definite laws of mathematical probability. Knowing such laws, it is clearly possible to state the mathematical probability for (or against) any particular deviation or variation occurring as the result of random sampling. Exactly this is what the probable error does. It says, in the particular case here considered, that it is an even chance, that a deviation or variation in the value of the mean as great as or greater than

0.02 mm. above or below will occur as a result of random sampling. Or, put in another way, it is an even bet that the value of the mean thoracic length of fiddler crabs in general will fall between 30.14 + 0.02 = 30.16, and 30.14 − 0.02 = 30.12.

Now all the knowledge that this probable error furnishes is this: that if a man were to say, "I'll bet a thousand dollars that the true mean thoracic length of the population from which this sample of fiddler crabs was drawn is either over 30.16 mm. or under 30.12 mm." one would not be justified in offering odds. He could wager on even terms. Either party involved in the transaction would be as likely to lose (or to win) as the other.

Putting the case in this way, it is clear that this kind of knowledge which comes from an examination of probable errors is the same as that discussed above. It is a knowledge of betting odds. It has no necessary relation *per se* to any physical, chemical, or biological laws. It merely informs one how he may safely gamble on an event if he is so minded and can find some one else ready to do the same thing.

Wherein lies the value of the probable error concept for science, then? Simply in that it serves as a test or check on every mode of research in science. So far as I can see, the calculus of probability, in and of itself alone, is not and never can be an effective weapon of research for the discovery of truth in phenomenal science, be it physical or biological. Yet it operates as an ever-present test of the trustworthiness of the results obtained by modes of research which are in themselves adapted to making discoveries about phenomena. The student of probability says something like this to the experimentalist: "Yours is the way to find out the significant underlying causes of phenomena. Let it be practiced with all zeal, but let it be remembered that you operate in a finite universe, and consequently all your results are subject to such fluctuations and variations as experience has shown arise from random sampling. I regret that I cannot directly and alone discover significant causes, but at any rate I can furnish you a test whereby you may reasonably judge whether your result is significantly influenced by these fluctuations of random sampling."

To sum the whole matter up: It seems that the statistical method in science has been used to do two things.

The first of these is a unique function of the method—to furnish a description of a group of objects or events in terms of the group's attributes rather than those of the individuals composing the group. Herein lies the great value of the statistical method. It is, however, a descriptive method only and has the limitations as a weapon of research which that fact implies.

The second purpose that the statistical method has been called upon to accomplish is the prediction of the individual case from a precise knowledge of the group or mass. This involves something really additional to the statistical method *per se*, namely, the mathematical theory of probability. We have seen that this side of the statistical method gives only a somewhat sterile kind of knowledge so far as concerns individuals, namely, a knowledge of betting odds. The theory of probability grew up about the gaming table, not in the laboratory. Its place in the methodology of science is not an independent one. By it alone one cannot discover new truths about phenomena. But it is a highly important adjunct to other modes of research.

Plainly, however, one cannot regard statistical knowledge in general as a higher kind of knowledge than that derived in other ways. Nor is the statistical method to become the dominant or exclusive method of science, though it will always be useful, and in many fields an essential method. It will find its chief usefulness, first in its sphere of furnishing shorthand description of groups, and second in furnishing a test of the probable reliability of conclusions.

SUGGESTED READING

In lieu of any formal bibliography, there will be given at the end of each chapter some suggestions as to further reading along lines touched upon in the text.

1. Yule, G. U.: An Introduction to the Theory of Statistics, sixth edition, London (C. Griffin & Co.), 1922.
2. Jones, D. Caradog: A First Course in Statistics, London (G. Bell & Sons, Ltd.), 1921.
3. Mortara, G.: Lezioni di Statistica Metodologica, Città di Castello (Soc. Tipografica, "Leonardo da Vinci"), 1922.
4. Czuber, E.: Die statistischen Forschungsmethoden, Wien (L. W. Seidel & Sohn), 1921.
5. Mills, F. C.: Statistical Methods Applied to Economics and Business, New York (Holt), 1924.

6. Fisher, R. A.: Statistical Methods for Research Workers, Edinburgh and London (Oliver and Boyd), 1925.

7. Rietz, H. L.: Mathematical Statistics, Chicago (Open Court Publ. Co.), 1927.

8. Burgess, R. W.: Introduction to the Mathematics of Statistics, Boston (Houghton Mifflin Co.), 1927.

9. Niceforo, A.: La Méthode Statistique et ses applications aux sciences naturelles aux sciences sociales et à l'art, Paris (Giard), 1925.

(References 1 to 9 are excellent *general text-books* of statistics, not intended in any way for the medical or vital statistician *particularly*, but, on the whole, much better for his use than some of the books which have been especially prepared for him.)

10. Pearl, R., and Miner, J. R.: Variation of Ayrshire Cows in the Quantity and Fat Content of Their Milk, Jour. of Agr. Research, vol. 17, pp. 285–322, 1919. (Contains some discussion and application of the concept of variation on space and time bases.)

11. Greenwood, M.: On Methods of Research Available in the Study of Medical Problems, Lancet, 1, 158, 1913.

12. Hooper, W.: Article "Statistics," Ency. Brit., 11th edit., vol. 25, pp. 806–811, 1911.

13. Kilgore, E. S.: Relation of Quantitative Methods to the Advance of Medical Science, Jour. Amer. Med. Assoc., vol. 75, pp. 86–89, 1920.

14. Wolff, G.: Die statistische Methode in der Epidemiologie und medizinische Ursachenforschung, Klin. Wochenschr., Bd. 6, pp. 2025–2031, 1927.

CHAPTER II

SOME LANDMARKS IN THE HISTORY OF VITAL STATISTICS

In the earlier volumes of the Journal of the Royal Statistical Society—those mines of curious information—a favorite form of contribution was the "tabular résumé," which presented a series of more or less statistical facts on a chronologic base. So distinguished a precedent seems to justify the use of the same method to furnish a bird's-eye view of the development of biostatistics itself. Consequently the table which follows has been prepared.

TABULAR REVIEW OF THE HISTORY OF VITAL STATISTICS

This "tabular résumé" attempts to set forth in chronologic array what the passage of time has shown to be some of the most important landmarks in the history of biostatistics. To disarm in some measure criticisms, which from the standpoint of the professional historian would otherwise be undoubtedly merited, it may be said, first, that there has been no slightest thought of encompassing within this short table a complete history of the subject. Historic completeness and the tabular form of presentation do not go well together. The object of the present table is much simpler. It is to get before the student the briefest conspectus of the time relations of the development of the subject, on the one hand, and of the personalities concerned in a large pathbreaking way in this development, on the other hand. The precise manner in which such a purpose will be carried out will obviously be different for each person who attempts it. One person's estimate as to the relative historic significance of a particular event or personality will differ from another's. In any event, it seems clear that any historic review of vital statistics would be bound to contain at

TABULAR REVIEW OF SOME OF THE IMPORTANT EVENTS IN THE HISTORY OF VITAL STATISTICS

Year.	Event.	Personality concerned.	Authority for record.
1532	First definitely known compilation of weekly bills of mortality in London.	———	Hull, C. H., Econ. Writ. of Sir Wm. Petty, p. lxxxi.
1539	Beginning of official registration of baptisms, marriages and deaths in France.	———	Faure, F., Hist. Stat., p. 242.
1608	Beginning of oldest parish register in Sweden.	———	Arosonius, E., Hist. Stat., p. 537.
1657	Publication of *De Ratiociniis in Ludo Aleae*, the first *printed* work on games of chance.	Christiaan Huygens (1629–1695).	Walker, H. M., Hist. Stat. Meth., p. 7.
1662	Publication of first edition of "Natural and Political Observations Mentioned in a following Index, and made upon the Bills of Mortality."	Capt. John Graunt, Citizen of London (1620–1674).	Hull, C. H., Econ. Writ. of Sir Wm. Petty, p. 315.
1666	First Census of Canada (the earliest modern census of population).	———	Godfrey, E. H., Hist. Stat. p. 179.
1669	Application of mathematical theory of probability to expectation of human life.	Christiaan Huygens (1629–1695).	Stuart, C. A. V., Hist. Stat., p. 430.
1693	Publication of "Estimate of the Degrees of Mortality of Mankind," in the Philosophical Transactions of the Royal Society.	Halley, the astronomer (1656–1742).	Hull, *Loc. cit.*, p. lxxvii.
1713	Publication of "Physico-theology; or a demonstration of the Being and Attributes of God from his Works of Creation."	Rev. William Derham (1657–1735).	Hull, *Loc. cit.*, pp. lxxvi and lxxviii.
1718	Publication of the "Doctrine of Chances."	A. DeMoivre (1667–1754).	Art. DeMoivre, Encyc. Brit.
1733	Publication of *Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi*, the discovery of the normal curve.	A. DeMoivre (1667–1754).	Pearson, K., Biometrika, xvi, p. 402.
1735	Registration of vital statistics begun in Norway.	———	Kiaer, A. N., Hist. Stat. p. 447.
1741	Publication of "Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen, etc."	Johann Peter Süssmilch (1707–1767).	Hull, *Loc. cit.*, p. lxxviii.
1746	Publication of the first French tables of mortality under the title "Essai, sur les probabilités de la durée de la vie humaine."	Deparcieux.	Faure, F., *Loc. cit.*, p. 265.
1748	Beginning of Swedish official vital statistics.	———	Arosonius, E., Hist. Stat. p. 540.
1749	First complete Census of Sweden.	———	Rossiter, W. S., Cent. Pop. Growth, p. 2.
1753	First Census of population in Austria ordered.	———	Meyer, R., Hist. Stat., p. 85.
1769	First population Census of Denmark and Norway.	———	Jensen, A., Hist. Stat., p. 201.
1790	First federal Census of the United States.	———	
1795	First Census of the Netherlands.	———	Stuart, C. A. V., Hist. Stat., p. 43.
1797	Establishment of Danish-Norwegian Tabulating Office.	———	Jensen, A., *Loc. cit.*, p. 201.
1798	First complete Census of Spain.	———	Rossiter, W. S., Cent. Pop. Growth, p. 2.
1801	First complete Census of Great Britain.	———	Rossiter, W. S., *Loc. cit.*
1801	First complete Census of France.	———	Rossiter, W. S., *Loc. cit.*
1805	Formation of first statistical state office within boundaries of German Empire.	———	Würzburger, E., Hist. Stat., p. 3.
1810	First complete Census of Prussia.	———	Rossiter, W. S., *Loc. cit.*
1812	Publication of "Théorie analytique des probabilités."	Pierre Simon Laplace (1749–1827).	Encyc. Brit. Art., Laplace.
1812	Inauguration of civil registration of births, marriages and deaths in the Netherlands.	———	Stuart, C. A. V., Hist. Stat., p. 432.
1812	Publication of "Theoria combinationis observationum erroribus minimis obnoxia" (Least squares).	Karl Friedrich Gauss (1777–1855).	Encyc. Brit. Art., Gauss.
1815	First complete Census of Norway.	———	Rossiter, W. S., *Loc. cit.*
1815	First complete Census of Saxony.	———	Rossiter, W. S., *Loc. cit.*
1816	First complete Census of Baden.	———	Rossiter, W. S., *Loc. cit.*
1818	First complete Census of Austria.	———	Rossiter, W. S., *Loc. cit.*
1818	First complete Census of Bavaria.	———	Rossiter, W. S., *Loc. cit.*

TABULAR REVIEW OF SOME OF THE IMPORTANT EVENTS IN THE HISTORY OF VITAL STATISTICS—*Concluded*

Year.	Event.	Personality concerned.	Authority for record.
1825	Publication of "Mémoire sur les lois des naissances et de la mortalité à Bruxelles," Quetelet's first statistical paper.	Lambert Adolph Jacques Quetelet (1796–1874).	Lottin, Quetelet, p. xx.
1826	Establishment of statistical commission in Belgium.	Ed. Smits.	Julin, A., Hist. Stat., p. 126.
1829	First official Census of Belgium.	Ed. Smits.	Julin, A., Hist. Stat., p. 128.
1832	Publication of "Recherches sur la reproduction et sur la mortalité de l'homme aux differents ages et sur la population de la Belgique d'apres la recensement de 1829 (premier recueil officiel des documents statistiques)."	Quetelet and Smits.	Lottin, *Loc. cit.*, p. xxi.
1834	Royal Statistical Society (London) founded.	———	Title page of Journal
1835	Publication of "Sur l'homme et le développement de ses facultés, ou Essai de physique sociale."	Lambert Adolph Jacques Quetelet (1796–1874).	Lottin, *Loc. cit.*, p xxi.
1836	First complete Census of Greece.	———	Rossiter, W. S., *Loc. cit.*
1837	Civil registration of vital statistics in England. Establishment of office of Registrar-General.	———	Baines, A., Hist. Stat., p. 370.
1838	Publication of "Essay on Probabilities" in Lardner's Encyclopedia.	Augustus DeMorgan (1806–1871).	Encyc. Brit. Art., DeMorgan.
1838	Publication of first paper on the logistic curve of population growth.	P. F. Verhulst.	Yule, G. U., Jour. Roy. Stat. Soc., vol. 88, pp. 1–58. 1925.
1839	Appointment of William Farr as compiler of abstracts in the Registrar-General's Office.	William Farr (1807–1883).	Farr's Vit. Stat., Edit. Humphrey.
1839	Organization of American Statistical Association.	———	Hist. of Stat., p. 3.
1846	Publication of "Analyse mathématique sur les probabilités des erreurs de situation d'un point." Acad. des Sci. Mém. par div. sav. IIe. Sér. t. ix (Correlation).	Auguste Bravais (1811–1863).	Yule, Introd., p. 188.
1848	Foundation of the Institute of Actuaries of Great Britain and Ireland.	———	Encyc. Brit. Art., "Actuary."
1860	First complete Census of Switzerland.	———	Rossiter, W. S., *Loc. cit.*
1861	First complete Census of Italy.	———	Rossiter, W. S., *Loc. cit.*
1863	Austria establishes Central Statistical Commission.	Count Mercandin.	Meyer, R., *Loc. cit.*, p. 89.
1865	Publication of "History of Mathematical Theory of Probability from the Time of Pascal to that of Lagrange."	Isaac Todhunter (1820–1884).	Encyc. Brit. Art., Todhunter.
1867	First creation of independent official statistical organization in Hungary.	———	Buday, L. von, Hist. Stat., p. 395.
1869	Publication of "Hereditary Genius."	Sir Francis Galton (1822–1907).	Art. Galton, Encyc. Brit.
1869	Foundation of Société de statistique de Paris.	———	Title page of Journal.
1872	Opening of German Imperial Statistical Office.	———	Würzburger, E., Hist. Stat , p. 337.
1881	First general Census of India.	———	Baines, A., Hist. Stat., p. 421.
1887	Royal Statistical Society incorporated by Royal Charter.	———	Title page of Journal.
1890	First Census in which mechanical methods of tabulation were used.	John S. Billings and Herman Hollerith.	Rept. Supt. Census 1889, p. 8.
1894	Publication of first of "Contributions to the Mathematical Theory of Evolution" in Phil. Trans. Roy. Soc.	Karl Pearson.	Title page.
1897	Publication of paper "On the Theory of Correlation" in the Jour. Roy. Stat. Soc.	G. Udny Yule.	Jour. Roy. Stat. Soc., vol. lx, p. 812.
1897	First Census of Russia.	———	Kaufman, A., Hist. Stat., p. 481.
1900	First year of separately published official mortality statistics for Registration Area of United States.	———	Title page of "Mortality Statistics."
1901	Publication of first number of Biometrika.	Francis Galton, Karl Pearson, W. F. R. Weldon, C. B. Davenport.	Title page.
1902	Creation of permanent Census Bureau in the United States.	———	Cummings, J., Hist. Stat., p. 682.
1915	First year of separately published official birth statistics for Registration Area of United States.	———	Title page of "Birth Statistics."

least a good many of the items of the present table. More than this in the way of agreement among scholars on a historic matter it is doubtless idle to hope for.

In the second place it should be said that if the sources chosen for statement of reference as to the facts are obviously in some cases second-hand, and perhaps somewhat casual, this is so of deliberate purpose. It is hoped that by so choosing them it may perchance be possible to entice an unwary student or so to do a little reading about the men who have helped to develop modern statistics. I am quite sure that this will not happen if he is referred straight off to a ponderous and deadly "Geschichte der Statistik." Nor is there much chance that the embryo health-officer or medical man would make anything but heavy weather if he essayed a voyage into the "Théorie analytique." But if he will read the article in the Encyclopedia Britannica on Laplace he will tend to have a measure of wholesome respect for a great man, and will know a little at least of what that man meant in the history of science.

CAPTAIN JOHN GRAUNT

Vital statistics, in the modern sense of the term, may be said to take its origin from the publication, in 1662, of a remarkable book for any age, but particularly so for that time, entitled, *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*, by John Graunt, Citizen of London (1620–1674). Bills of mortality, consisting of lists of burials, marriages, and baptisms, had been compiled by the parish clerks for upward of a century before Graunt's time, but no one before him had conceived the idea of making an analytical study of these observations to the end of determining the basic laws of human mortality, natality, and movement of population. From his inadequate and meager material, as measured by present standards, Graunt successfully demonstrated four of the most important facts which the study of vital statistics to this day has disclosed. First, he made clear the *regularity* of certain vital phenomena which appear to be merely the play of chance in their individual occurrence. Second, he first pointed out the *excess of male over female births*, and the approximately equal numbers of

Natural and *Political*

OBSERVATIONS

Mentioned in a following I N D E X,

and made upon the

Bills of Mortality.

B Y

Capt. *JOHN GRAUNT*,

Fellow of the *Royal Society*.

With reference to the *Government*, *Reli-gion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said C I T Y.

———— *Non, me ut miretur Turba, laboro, Contentus paucis Lectoribus.*————

The Fourth Impression.

O X F O R D,

Printed by *William Hall*, for *John Martyn*, and *James Allestry*, Printers to the Royal Society, MDCLXV.

Fig. 1.—Facsimile (actual size) of the title-page of the first treatise on vital statistics

the sexes in the population.   Third, he demonstrated the relatively *high rate of mortality in the earliest years of life*, and finally he discovered that the *urban is higher than the rural death-rate* normally.

Besides the intrinsic value of its results, Graunt's book served for many years as the stimulator of other work in the same general field.   In particular it is probably safe to conclude that Graunt's

EDMUNDUS HALLEIUS R.S.S.
Astronomus Regius et Geometriæ Professor Savilianus.

Fig. 2.—Portrait of the eminent astronomer and mathematician, Edmund Halley (1656–1742), who was the first person to construct a life table on sound principles.

book was the inciting agency which led the astronomers and mathematicians, Huygens in Holland and Halley in England, to take up the problems of determining by appropriate mathematical methods the probable expectation of human life at any given age.   Halley constructed the first really significant mortality table.   Some of his results are shown in Fig. 3.

Age. Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons
1	1000	8	680	15	628	22	586	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427

Age. Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons	Age Curt	Perfons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Perfons
7	5547
14	4584
21	4270
28	3964
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
Sum Total.	

Fig. 3.—Survivorship distribution of the first life table (Halley's). Reproduced in facsimile from Baddam's "Memoirs of the Royal Society," vol. iii, p. 36. This table "shews the number of persons living in the age current annexed thereto."

THE MOST ANCIENT BILL OF MORTALITY

The earliest known bill of mortality is an interesting document. It was in manuscript form, and is preserved among the Egerton MSS. at the British Museum. It is shown in facsimile in Fig. 4.

Creighton[9] believes its date to be 1532 (week of November 16th to 23d), and gives evidence for his belief as to the year (Vol. I, p. 295): "The extant bill for the week 16th to 23d November is clearly one of a series; there are no good grounds for assigning it to an earlier date than the year 1532, while there are reasons for not placing it later. There are two other plague-bills extant, for August, 1535, written out in a more clerkly fashion, and bearing the marks of greater experience. The bill for the week in November is more primitive in appearance; and we may fairly take it as one of the series first ordered by the Council in 1532: for that was the most considerable year of the plague immediately preceding the outburst of 1535, to which the finished bills certainly belong." This earliest of official reports of vital statistics to be preserved is transcribed by Creighton (retaining the original spelling) as follows:
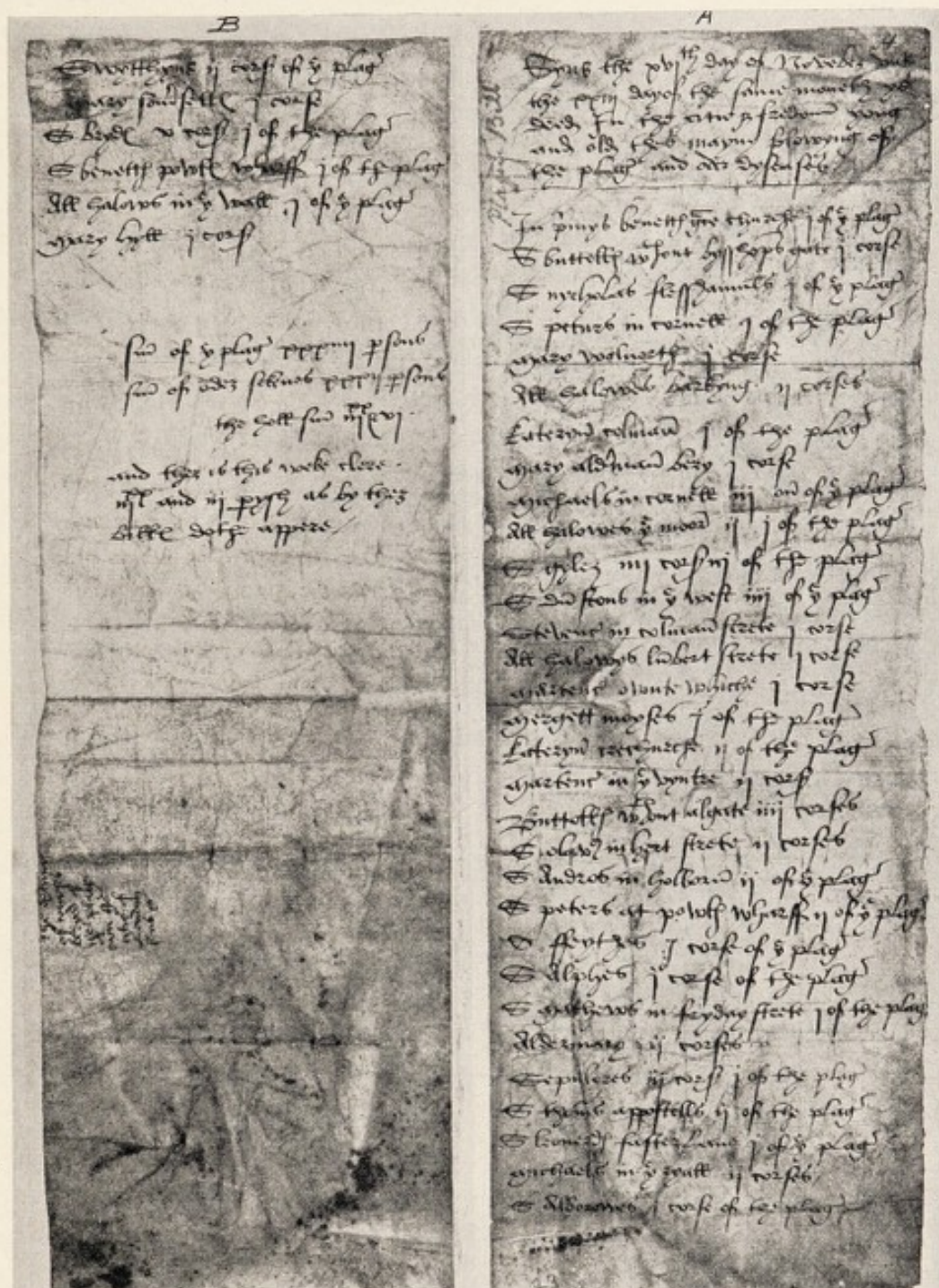
Fig. 4.—Photographic reproduction of the earliest known bill of mortality: *A*, obverse; *B*, reverse. Reduced to about one-half actual size. (For permission to publish the photographic reproduction of this interesting document I am obliged to Sir Frederick Kenyon, Director of the British Museum. The photographs were procured for me by Mrs. Onera A. Merritt Hawkes, to whom I am greatly indebted for this service.—R. P.)

Syns the xvith day of November unto the xxiii day of the same moneth ys dead within the cite and freedom yong and old these many folowyng of the plage and other dyseases.

Inprimys benetts gracechurch i of the plage
S Buttolls in front of Bysshops gate icorse
S Nycholas flesshammls i of the plage
S Peturs in Cornhill i of the plage
Mary Woolnerth i corse
All Halowes Barkyng ii corses
Kateryn Colman i of the plage
Mary Aldermanbury i corse
Michaels in Cornhill iii one of the plage
All halows the Moor ii i of the plage
S Gyliz iiii corses iii of the plage
S Dunstons in the West iiii of the plage
Stevens in Colman Strete i corse
All halowys Lumbert Strete i corse
Martins Owut Whiche i corse
Margett Moyses i of the plage
Kateryn Creechurch ii of the plage
Martyns in the Vintre ii corses
Buttolls in front Algate iiii corses
S Olavs in Hart Strete ii corses
S Andros in Holburn ii of the plage
S Peters at Powls Wharff ii of the plage
S Fayths i corse of the plage
S Alphes i corse of the plage
S Mathows in Fryday Strete i of the plage
Aldermary ii corses
S Pulcres iii corses i of the plage
S Thomas Appostells ii of the plage
S Leonerds Foster Lane i of the plage
Michaels in the Ryall ii corses
S Albornes i corse of the plage
Sywtthyns ii corses of the plage
Mary Somersette i corse
S Bryde v corses i of the plage
S Benetts Powls Wharff i of the plage
All halows in the Wall i of the plage
Mary Hyll i corse.
Sum of the plage xxxiiii persons
Sum of other seknes xxxii persons

The holl sum iii^{xx} & vi.

And there is this weke clere iii^{xx} and iii paryshes as by this bille doth appere

The execⁿ of corses buryed of the plage within the cite of London syns &c.

SÜSSMILCH, QUETELET, AND FARR

The next considerable contribution to vital statistics, as such, was the publication of *Die göttliche Ordnung in den Veränderungen des Menschlichen Geschlechts aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen, etc.*, by the Reverend Johann Peter Süssmilch (1707–1767). Süssmilch was stimulated by Graunt's *Observations* to apply the same general sort of method to the development of natural theology. This book exerted a great influence in fields other than theological, and was the logical fore-runner of the great work of the famous Belgian vital statistician, Lambert Adolph Jacques Quetelet (1796–1874), entitled *Sur l'homme et le développement de ses facultés, ou Essai de physique sociale*, published in 1835. Quetelet is the first great outstanding figure in the development of modern vital statistics. Trained as a mathe-matician, he brought to bear upon the data of human vital phenom-ena a more adequate methodology than had before been applied.

The present-day procedure in official vital statistics undoubtedly owes more to William Farr (1807–1883) than to any other person. Besides this he may fairly be regarded as the greatest *medical* statistician who has ever lived. Greenwood[14] says: "But if ultimately Graunt had a worthy disciple in the medical profession, it was not until he had been in his grave more than a century. He died in 1674 and William Farr was born in 1807."

In this paper just quoted Greenwood gives the best existing brief estimate of the significance of Farr in the history of medicine, and it may properly be reproduced here in full. He says:

"The real revolutionary was a licentiate of the Society of Apothecaries, a 'Mr. Farr, a gentleman of the medical profession,' who was appointed Compiler of Abstracts in the General Register Office on July 10, 1839. Although Mr. Noel Humphreys earned the gratitude of all medical men by his collection of Farr's writings, published in 1885, a really adequate edition of Farr has yet to be produced. We sometimes dream of such an edition; we picture it with an introduction by Farr's worthy successor, Dr. Thomas Stevenson, and with footnotes and appendices by Dr. John Brown-lee. But it is an idle dream; governments in England, so the newspapers tell us, often spend money in odd ways, but at least
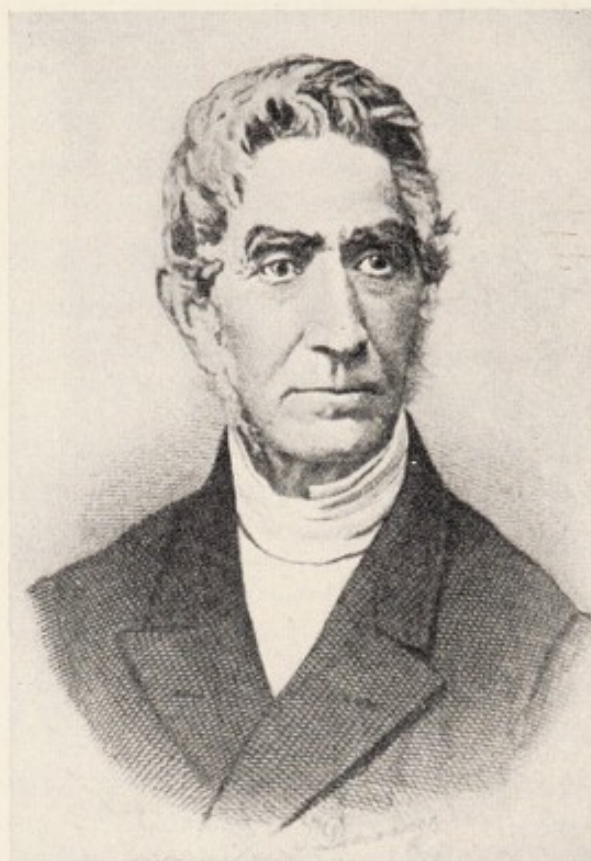
Fig. 5.—Portrait of Lambert Adolph Jacques Quetelet (1796–1874)

Fig. 6.—Portrait of Dr. William Farr (1807–1883).

they have never been so eccentric as to waste it on the publication of the collected works of great Englishmen. Farr was a very great Englishman, and the characteristics of his genius were precisely those which, in moments of self-esteem, we like to fancy are typically English. We can make our point clear by contrasting him with two great men who were at their prime when he was young, and both made important contributions to statistical knowledge, Siméon Poisson and George Boole. Poisson wrote a large treatise upon ostensibly the most practical of subjects, the best way to secure just verdicts in courts of law; Boole dealt with the very matter-of-fact problem of numerical approximation. But the most superficial reader of Poisson or of Boole—not that their works are very attractive to a hasty reader—will at once realize that the authors are far more interested in algebra than in the concrete applications of their algebra. Farr has left many pages which, to the aforementioned hasty reader, will offer almost as many algebraical difficulties as even Boole; but in the densest forest of symbols Farr never loses sight of, and never allows his companion to lose sight of, some perfectly definite and concrete end which he proposes to reach.

"No branch of medical or vital statistics needs for its cultivation a greater variety of algebraical tools than that concerned with the production of complete life tables; the natural faculty which characterizes the born mathematician is not, indeed, essential, but skill in the manipulation of symbols is. To Farr a life table was—

'An instrument of investigation; it may be called a biometer, for it gives the exact measure of the duration of life under given circumstances. Such a table has to be constructed for each district and for each profession, to determine their degree of salubrity. To multiply these constructions, then, it is necessary to lay down rules, which, while they involve a minimum amount of arithmetical labour, will yield results as correct as can be obtained in the present state of our observations.'*

"This was the spirit of all his work. He faced mathematical

* From a paper contributed to the Proceedings of the Royal Society in 1859. (See Farr's "Vital Statistics," ed. Humphreys, London, 1885, p. 492.)

difficulties with a courage which nothing could daunt—it takes some courage for a self-taught man to venture upon original research within the province of the oldest of the sciences—when they obstructed his progress toward a practical end. He never attempted to compete with the masters of pure analysis on their own ground. We have been the gainers. The greatest mathematical statisticians of the first half of the nineteenth century were not Englishmen; we have not to our credit any theoretical work of that date which will compare with the researches of Laplace and of Poisson in France or of Gauss in Germany; but of no civilized country can a record of fatal disease be constructed with the precision which appertains to the medico-statistical history of England and Wales since 1840.

"The practical advantages to the physician and the sanitarian are enormous. Matters which our great grandparents fiercely debated, topics respecting which only a very shrewd and experienced physician of 1820 could form an opinion, are now within the compass of a junior medical student. If Farr had been born a generation earlier and the General Register Office had been founded in 1807 instead of in 1837, the sanitary history of our manufacturing towns might have been different. If even the lessons he taught year by year had sunk into the minds of all members of our profession, many disappointments would have been spared and perhaps some false apprehensions quieted. The curious reader of old blue-books will find much of interest in the census reports of Lamb's friend Rickman, but Rickman was not a Farr. Rickman, for instance (in 1831), commented upon the apparent unhealthiness of the northern manufacturing districts, but he could not speak with much authority, for his basis of facts was no more than an abstract of burial and baptismal registers. These are the words of Farr (from the supplement to the thirty-fifth Annual Report):

'Take for example the group of 51 districts called healthy for the sake of distinction, and here it is found that the annual mortality per cent. of boys under five years of age was 4.246; of girls, 3.501. Turn to the district of Liverpool, the mortality of boys was 14.475; of girls, 13.429. Here it is evident that some pregnant exceptional causes of death are in operation in this second city of

England. What are these causes? Do they admit of removal? If they do admit of removal, is this destruction of life to be allowed to go on indefinitely? It is found that of 10,000 children born alive in Liverpool 5396 live five years, a number that in the healthy districts could be provided by 6544 annual births.'

"The 'dear old doctor'—as Mr. Humphreys called him—could round a period in the early Victorian style with the best; the classical quotations in his reports might have tempted William Pitt or Charles Fox to become statisticians; but he could also use very plain English indeed. Statistics with plain English as a propellent are formidable missiles.

"We could fill many columns with examples, but we must take leave of the greatest of medical statisticians with one observation. Farr's work has on it the seal of all supreme achievements; it is indestructible. It was, of course, a piece of good luck that his three successors, the late Dr. William Ogle, Dr. John Tatham, and Dr. Thomas Stevenson, were men having the same ideals and zealous to build higher upon his foundations. The nation, we hope, will always be fortunate enough to secure equally worthy spiritual descendants of the founder. But no weakness of human instruments or credible deteriorations of the system could ever take from the General Register Office the power of 'rendering immense service to sanitary science by enabling it to use exact numerical standards in place of the former vague adjectives.'* So far as records of mortality are concerned, the real reformer is one who treads accurately in the footprints of William Farr."

THE HISTORY OF BIOMETRY

The application of statistical methods to the study of biologic problems other than those of anthropology, and of vital statistics in the narrower sense, may be said to have begun with the work of the late Sir Francis Galton. Galton was a born statistician. He tells in his *Memories*[13] of the instinct, which he inherited from his father, to arrange, classify, and collect statistics about all sorts of things. At the same time he was deeply interested in problems of biology, particularly those having to do with inher-

* Simon: English Sanitary Institution, p. 212.

itance. His interest in this direction crystallized into definite activity at about the time that his cousin, Charles Darwin, was elaborating his theory of heredity, which was called pangenesis. Galton instantly realized that this conception of the physiology of the hereditary process was essentially statistical in character, and that statistical methods were demanded to test and broaden it. Upon this work he therefore embarked with the vigor and
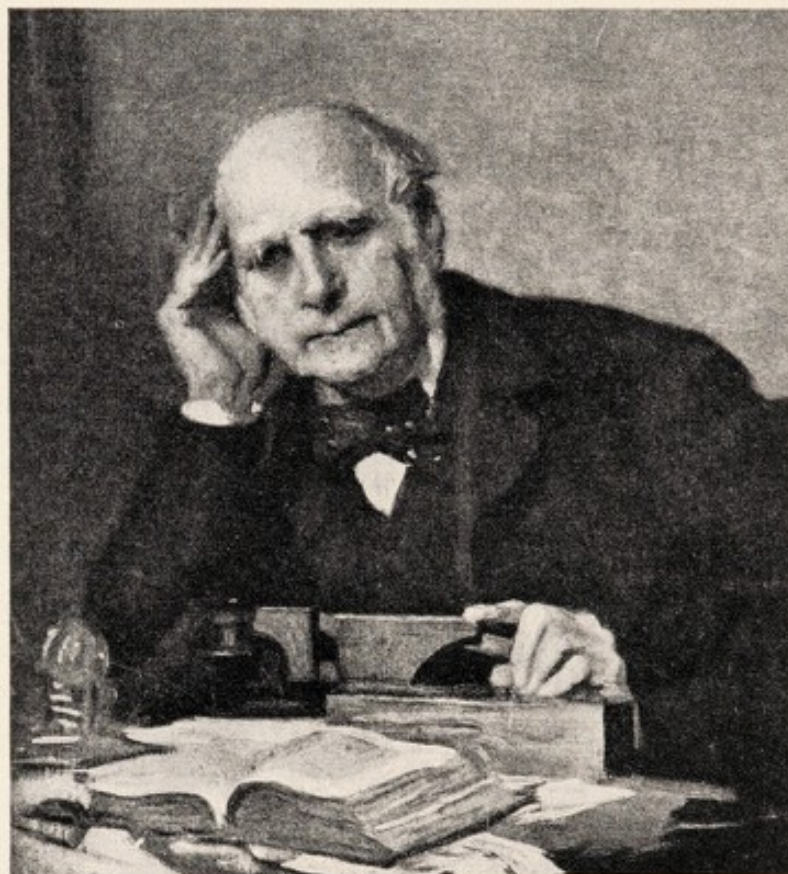
Fig. 7.—Portrait of Francis Galton (1822–1907). (For permission to publish this portrait here I am indebted to Dr. G. H. Shull, Editor of Genetics.)

ardent enthusiasm which characterized all of his scientific work. His results found expression in a series of memoirs and books which have become classics in biologic science. Of these the most important is perhaps *Natural Inheritance*, since in it are brought to a focus a number of different lines of work which engaged Galton's thought and energy for many years. In this book the attempt is made for the first time to determine, on a statistical basis, the degree of resemblance, in respect of bodily, mental, and tem-

peramental traits, which obtains between relatives of different degrees. Previously no attempt had been made to measure precisely these resemblances, which were, of course, a matter of common observation, though not of precise definition, to everyone.

In order to make the desired analysis of this problem it was necessary for Galton to devise new methods of dealing with statistics. The general mathematical foundations of statistical science had, to be sure, been laid by the mathematicians Laplace and

Fig. 8.—Portrait of Pierre Simon Laplace (1749–1827).

Gauss, and some progress in the application of these methods had been made by Quetelet. But none of these men had dealt specifically with the measurement of what are now known as correlated variations. From Galton's point of viewing the problem of heredity such a measure was an absolute necessity. He, therefore, devised one. It was not altogether a perfect one, but was practically usable, and led very shortly to developments which furnished the entirely adequate measure which he had sought.

To the end of his life Sir Francis Galton retained his interest in the science of biometry, of which he may truly be said to have been the founder. His keenness of interest served in great part as the primal inspiration and stimulus which led two other distinguished English workers to enter this field and begin to rear the super-

Fig. 9.—Portrait of Karl Pearson, F. R. S.

structure on the foundation already laid. These were Professor Karl Pearson of University College and the late Professor W. F. R. Weldon. To Professor Pearson belongs the very great credit of developing adequate and general mathematical methods for the analysis of biologic statistics. Statistical mathematics in the main fall within the realm of the calculus of probability. The founda-

tions of that calculus were laid by Laplace and Gauss, as has already been pointed out. Since their day the most notable fundamental advance in the mathematical theory of probability has, in the writer's judgment, been due to the genius of Karl Pearson. Starting from the sound position that the facts of nature are of more importance than any theory, Pearson in three classic memoirs, in his series of *Mathematical Contributions to the Theory of Evolu-*

Fig. 10.—Portrait of G. Udny Yule, F. R. S. (Photo: Russell.)

tion, developed a theory of skew frequency curves, and skew correlation, which took due account of the asymmetry so frequently seen in chance-determined phenomena. This system of skew frequency curves has now had the test of more than twenty-five years' usage. Every attempt at destructive criticism which has been made against it has failed. None of the substitutes, some of which have been proposed by eminent mathematicians, has shown any approach to the generality and elegance of these curves.

Few biologists have an adequate conception of the extent to which biometry is indebted to Professor Karl Pearson. If, as has been maintained, every real advance in science depends upon the discovery and perfection of a new technic, then, for whatever advance in biology may come through biometry, the debt to that distinguished investigator will be large for many years to come.

The English may perhaps fairly be said to have led the world in the development of modern statistical theory and practice. In
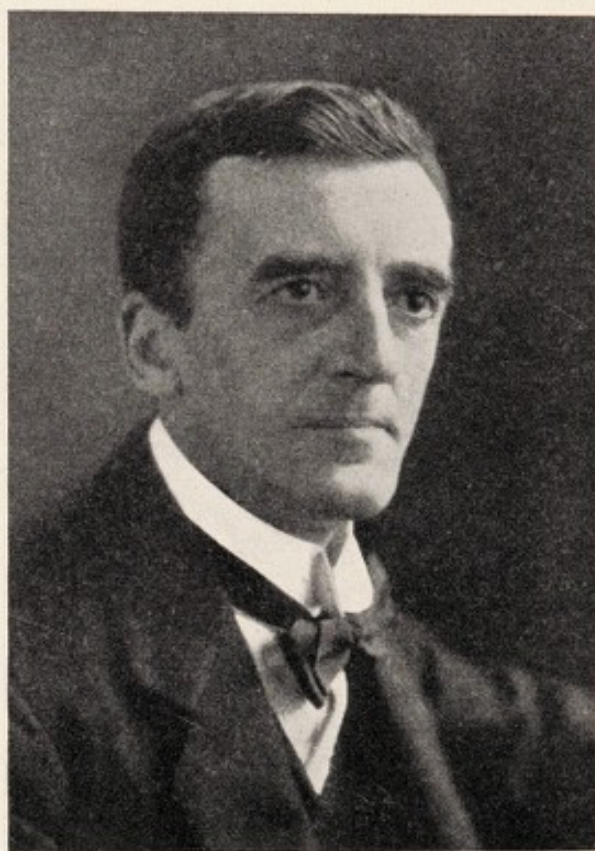
Fig. 11.—Portrait of Major Greenwood, F. R. S.

addition to the achievements of Graunt, Halley, Farr, Galton, Weldon and Pearson, who have been discussed in this chapter, some mention, at least, must be made of a number of other English workers, who have made fundamental contributions, notably De Moivre, F. Y. Edgeworth,[18] W. F. Sheppard, G. Udny Yule, L. Isserlis, and H. E. Soper. In the application of biometric methods to specifically medical problems, English workers, notably Prof. Major Greenwood of the London School of Hygiene and

Tropical Medicine, from whose work we have already quoted, and the late Dr. John Brownlee have taken a leading part. These workers and their associates have made notable contributions to the understanding of some of the most difficult problems of etiology and epidemiology.

Aside from the English perhaps the most outstanding school in the development of statistical theory and practice has been the Scandinavian. Here the important names are those of J. P. Gram, T. N. Thiele, C. V. L. Charlier, S. D. Wicksell, and Arne Fisher, who has for many years made his home in America. The Scandinavian school is chiefly noted for a system of skew curves based upon the semi-invariants of Thiele.

SUGGESTED READING

(This list includes, among other items, the expanded references to citations in the Tabular Review in the text.)

1. Encyclopedia Britannica, Eleventh edition (as cited).
2. Hull, C. H.: The Economic Writings of Sir William Petty, etc., Cambridge University Press, 1899, 2 vols.
3. Lottin, J.: Quetelet, Statisticien et Sociologue, Louvain and Paris, 1912.
4. Hankins, F. H.: Adolph Quetelet as Statistician, Columbia University Studies in History, Economics and Public Law, vol. 31, No. 4, pp. 1–134, 1908.
5. Rossiter, W. S.: A Century of Population Growth from the First Census of the United States to the Twelfth, 1790–1900, Washington, Government Printing Office, 1909.
6. The History of Statistics, Their Development and Progress in Many Countries, collected and edited by John Koren, New York (Macmillan), 1918.
7. Walker, Helen M.: Studies in the History of Statistical Method with Special Reference to Certain Educational Problems, Baltimore (Williams and Wilkins Co.), 1929. (This recent addition to the literature of the history of statistics is an extremely valuable one. Every student should read it.)
8. Yule, G. U.: Introduction to the Theory of Statistics, sixth edition, London (Griffith and Company), 1922.
9. Creighton, C.: A History of Epidemics in Britain, 2 vols., Cambridge, 1891. (A monumental work of the greatest importance to the student of the natural history of disease.)
10. Verhulst, P. F.: Notice sur la loi que le population suit dans son accroissement, Corr. math. et phys. publ. par A. Quetelet, T. X., pp. 113–121, 1838. (This first announcement of the logistic curve was a preliminary note only. Verhulst's definitive publications were: Recherches mathématiques sur la loi d'accroissement de la population. Nouv. mém. de l'Acad. Roy. des Sci. et Belles Lett. de Bruxelles. T. 18, pp. 1–38, 1845; Deuxième mémoire sur la loi d'accroissement de la population. *Ibid.*, T. 20, pp. 1–32, 1847.)

11. Yule, G. U.: The Growth of Population and the Factors Which Control It, Jour. Roy. Stat. Soc., vol. 88, pp. 1–58, 1925.

12. Pearson, K.: The Life, Letters, and Labors of Francis Galton, vol. i, Birth 1822 to Marriage 1853, Cambridge (University Press), 1914; vol. ii, Researches of Middle Life, Cambridge (University Press), 1924. (This biography, just being completed as this book is passing through the press, has already demonstrated that it is one of the greatest ever written in any language. Every student of science, whatever his branch, should read it.)

13. Galton, F.: Memories of My Life, New York (E. P. Dutton and Co.), 1909. (Every student must read this along with Pearson cited above.)

14. Greenwood, M.: Medical Statistics, Lancet, May 7, 1921.

15. Vital Statistics: A Memorial Volume of Selections from the Reports and Writings of William Farr, M. D., D. C. L., C. B., F. R. S. Edited by Noel A. Humphreys, London (Edward Stanford), 1885. (Now a rare book, but one which every student should read.)

16. Ogle, W.: An Inquiry into the Trustworthiness of the Old Bills of Mortality, Jour. Roy. Stat. Soc., vol. 55, pp. 437–460, 1892.

17. Pearl, R.: To Begin With. Being Prophylaxis Against Pedantry. Second edition, New York (Alfred A. Knopf), 1930. (This book may be read in connection with the present chapter. It was written for the specific purpose of guiding the student toward an interest in historical and related cultural matters.)

18. Bowley, A. L.: F. Y. Edgeworth's Contributions to Mathematical Statistics, London (Royal Stat. Soc.), 1928, pp. 1–139.

CHAPTER III

THE RAW DATA OF BIOSTATISTICS

BROADLY there are three ways in which statistical data are accumulated in the realm of human biology. These are:

1. The census method.
2. The registration method.
3. The *ad hoc* or case record method.

Of these the first two are the methods of official *vital statistics*, while the third is *par excellence* the method of medicine and biometry.

In the present chapter we shall discuss some aspects of the first two methods, while in Chapter V a more detailed discussion of the third method will be undertaken.

THE CENSUS METHOD ·

Theoretically a census is a count, made at a single specified instant of time, of a population in respect of certain attributes of the persons composing the population, or of things. Practically, of course, the "instant of time" is rather stretched out, but the endeavor is always made, and with a fair degree of success, to have the information gleaned referable to a single day.

All living things and all their affairs and concerns and attributes are continually *changing* with greater or less degrees of rapidity. The living world, in short, is in a state of continuous flux. It may be thought of as a vast stream, constantly added to by births, and subtracted from by deaths, diverted (but only slowly) from its previous pathway by divers impinging forces, but always and above all, moving, flowing.

Now a census attempts to acquire knowledge of the composition and characteristics of this great stream by examining carefully, at regular intervals of time (usually ten years apart), *an instantaneous cross-section of it*. What happened before the cross-section was

taken, or what will happen after it is taken, can only be inferred, when the census method of acquiring statistical information is employed, from the characteristics of the cross-section itself.

Censuses are taken either (*a*) by enumerators, (*b*) by questionnaires filled up by the victims themselves, or (*c*) by the two means in combination. The first method is the one chiefly employed in the United States. A person visits every household in a limited area on or near census day, and by personal inquiry elicits the desired information. The second method is the one chiefly employed in England, where there is placed in the hands of each householder a little time before census day a questionary form which he must truthfully and promptly fill in, under rather heavy penalty of the law for failure.

The data of value in biostatistics for which dependence is chiefly put on the census method at the present time are those relating to the living population, its numbers, age, sex, occupation, race, etc.

The path of census taking, while theoretically straightforward, is actually beset with difficulties and hazards, both general and specially technical. In the first place, there is an ancient and persistent opposition on the part of the people at large to being counted by the government. Ignorance and superstition combine to create antipathy to censuses. Who knows whether the government is not using this opportunity, by some subtle and diabolical machinations, to snoop about and pry out some information regarding the individual, or his business, or his wealth, or his love affairs, which may later be used to bring about his discomfiture? That this feeling, of which there is evidence in the most ancient historical records, persists to the present day is clearly manifest in the proclamation issued by President Herbert Hoover, on November 22, 1929, calling for the regular census of the United States in 1930. This proclamation reads as follows:

"*By the President of the United States of America.*
"A PROCLAMATION.
"*Whereas*, by the Act of Congress approved June 18, 1929, the fifteenth decennial census of the United States is to be taken beginning on the first day of April, nineteen hundred and thirty; and

"*Whereas,* A correct enumeration of the population every ten years is required by the Constitution of the United States for the purpose of determining the representation of the several States in the House of Representatives; and

"*Whereas,* It is of the utmost importance to the interest of all the people of the United States that this census should be a complete and accurate report of the population and resources of the nation;

"Now, therefore, I, Herbert Hoover, President of the United States of America, do hereby declare and make known that, under the law aforesaid, it is the duty of every person to answer all questions on the census schedules applying to him and the family to which he belongs and to the farm occupied by him or his family, and all other census schedules as required by law, and that every person refusing to do so is subject to penalty.

"The sole purpose of the census is to secure general statistical information regarding the population and resources of the country, and replies are required from individuals only to permit the compilation of such general statistics. No person can be harmed in any way by furnishing the information required. The census has nothing to do with taxation, with military or jury service, with the compulsion of school attendance, with the regulation of immigration or with the enforcement of any national, state, or local law or ordinance.

"There need be no fear that any disclosure will be made regarding any individual or his affairs. For the due protection of the rights and interests of the persons furnishing information every employee of the Census Bureau is prohibited, under heavy penalty, from disclosing any information which may thus come to his knowledge.

"I, therefore, earnestly urge upon all persons to answer promptly, completely and accurately all inquiries addressed to them by the enumerators or other employees of the Census Bureau, and thereby contribute their share toward making this great and necessary public undertaking a success.

"In witness whereof, I have hereunto set my hand and caused to be affixed the Great Seal of the United States.

"Done at the city of Washington, this 22d day of November, in the year of our Lord one thousand nine hundred and twenty-nine, and of the independence of the United States, the one hundred and fifty-fourth.

<div style="text-align:right">

"Herbert Hoover,

"President of the United States,

"By the President,

"Henry L. Stimson,

"Secretary of State."

</div>

The questions asked by the enumerator in the 1930 census, regarding which the proclamation is at such pains to reassure the people, covered the following points:

1. Relationship to head of family, including a statement as to the home-maker in each fam'ly.
2. Whether the home is owned or rented.
3. Value of home, if owned, or monthly rental, if rented.
4. Radio set? ("Yes" or "No.")

5

5. Does this family live on a farm?   ("Yes" or "No.")
6. Sex.
7. Color or race.
8. Age at last birthday.
9. Marital condition.
10. Age at first marriage.   (For married persons only.)
11. Attended school or college any time since September 1, 1929?   ("Yes" or "No.")
12. Whether able to read or write?   ("Yes" or "No.")
13. Place of birth of person.   (State or country.)
14. Place of birth of person's father.   (State or country.)
15. Place of birth of person's mother.   (State or country.)
16. Mother tongue of each foreign-born person.
17. Year of immigration to the United States.   (For foreign born only.)
18. Whether naturalized.   (For foreign born only.)
19. Whether able to speak English.   (For foreign born only.)
20. Occupation of each gainful worker.
21. Industry in which employed.
22. Whether employer, employee, or working on own account.
23. Whether actually at work.   (For each person usually employed but returned as not at work, additional information will be secured on a special unemployment schedule.)
24. Whether a veteran of the United States military or naval forces; and for each veteran, in what war or expedition he served.

Several of these questions were new in United States Census practice. Among the most important of these new questions is that calling for the value of the home if owned, or the monthly rental if rented. This makes possible a classification of families according to economic status, or, it is perhaps hoped, according to buying power. Such a classification was urgently desired by individuals and firms using the census figures as a basis for organizing their selling and advertising campaigns and will doubtless serve many other purposes. It was promised that the replies to these questions would be used only as a basis for classification of the families into broad groups.

Another new question is that which asks for the age at first marriage. This serves two purposes. In the first place it gives definite information as to the relative age at marriage of persons in different racial and economic groups. In the second place it makes possible a tabulation of important data on the size of families, such tabulation to be based on the number of children reported in the families of women who have been married a number of years.

The data so obtained should be of great value for the study of differential fertility and other population problems.

In the classification of gainful workers according to occupation and industry much greater stress than heretofore was put on the returns for industry. The enumerators were instructed to pay special attention to this section of the schedule.

Women doing housework in their own homes (or supervising such work done by servants) and carrying the other responsibilities of the home were designated as home-makers. This designation was entered in the family relationship column of the schedule, rather than in the occupation column, in order that those women who follow a profession or other gainful occupation, in addition to being home-makers, might be properly classified in respect to both lines of activity.

A special schedule for unemployment contained a number of questions designed to separate those not working into several classes, including, besides those absolutely unemployed, those who had a job but were for the time being on lay-off without pay, etc.

In the classification by color or race a special group was provided for Mexicans, in which were placed all persons of Mexican origin except those of strictly white ancestry, who were counted as heretofore with the whites, and possibly a small number who were classified as Indians.

Provision was again made for classifying the foreign born, which still form a very important element in the population, in five different ways: namely, by country of birth; by mother tongue (which is sometimes a better index of nationality than is country of birth); by year of immigration to the United States; by citizenship (that is, whether naturalized, having first papers, or alien); and by ability to speak English.

There are many technical difficulties in getting completely accurate results in census taking. Indeed only an approximation is ever obtained. Such approximation is probably closest in respect of the total number of the population, and is less good, in varying degrees, relative to such matters as age, occupation, national origin, etc. It is impossible to go here into detail regarding all the difficulties of the census method. One only may be discussed

somewhat fully, because it involves, next to the bare count, the most important datum for which the vital statistician is dependent upon the census. This is the age distribution of the living population.

It has been the universal experience in census work that there are two outstanding errors in census results respecting age. One is that the number of infants under one year of age, and between one and two years of age, is always understated in the census returns. It is certain that there are always more living infants of these ages than the census count shows. The second error is that the return always shows an excess of persons at certain particular years of age. This concentration is most marked on ages which are multiples of 5 and 10, but there is also observable a tendency in lesser degree to concentrate on even ages, as contrasted with odd, among persons less than twenty years of age. These concentrations are regarded as due to a well-known trait of human nature to state ages in round numbers, especially in cases where the enumerator is obliged to get his information second hand because the person concerned is away from the domicile when the call is made, and the person who does the answering literally has no exact knowledge of the absent one's age.

An excellent discussion of this matter may be quoted from *Population*, Part II of *Census Reports*, Volume II, of the Twelfth Census of the United States (1900), p. xxxv.

"Evidences of concentration were noticeable in the census returns of 1890, in spite of the fact that in the printed instructions at that census the attention of the enumerators was directed specially to these inaccuracies in the return of ages, and that they were cautioned not to accept such indefinite statements without first endeavoring to secure the exact year of age. This specific instruction had some effect, apparently, in lessening the extent to which ages were given in round numbers; but it is evident, as stated in the report for 1890, that 'no matter how specific the instructions to the enumerators on this point may be, the natural tendency is, and probably always will be, to give the nearest five or ten year period, especially where definite information is not at hand.' It was further suggested in the same report that probably this tendency could be obviated in part 'by requiring the return of ages, so far as possible, by the exact day, month, and year of birth and allowing a return of the age by the approximate year in only those cases where it is manifestly impossible to ascertain the date of birth, on the assumption that a fairly good approximation is better than no return at all.' "

As a means of determining by a practical test whether or not it was possible, under existing methods of census enumeration, to obtain a better age distribution of the population, an effort was made at the 1900 census to secure, wherever possible, a return of ages with a statement of the month and year of birth. As was to be expected, this did not prevent the return of an abnormally large proportion of persons as of the ages thirty, thirty-five, forty, forty-five, fifty, etc., but a comparison of the figures for 1900 with those for the two preceding enumerations shows that the more exact inquiry with respect to age in 1900 reduced materially the concentration at these ages.

Table 1 (page 70), a portion of one of the tables of the 1900 census (Table XX of the volume cited), shows this phenomenon of concentration in percentage form for native whites, foreign born whites, and colored, and demonstrates the improvement over the three censuses, 1880, 1890, and 1900. This table gives the percentage which the excess at each designated age (over the immediately preceding year of age) is of the number in 100,000 of the population at the designated age.

The conclusions to be drawn from Table 1 are first, that at the census of 1890 the most striking concentration in the earlier period of life was on two years of age, the excess on that age representing more than one-third of the relative number of males and females respectively, for each element of the population considered. The great concentration on this particular age was undoubtedly due to the form of inquiry in 1890, the schedule used at that census calling for a return of "age at nearest birthday" instead of "age at last birthday," as at earlier censuses and at the 1900 census. In 1900 the concentration on ages is very slight, and the improvement in the return of ages at that census is, generally speaking, apparent throughout the table, especially for the earlier periods of life.

For both native white and foreign white persons the concentration in 1900 on the five-year periods after twenty-five years, although considerable, is very much less than at the preceding census. For colored persons, however, the improvement in this respect is not so marked, the percentages of excess in 1900 being large, although somewhat less, in each case, than those shown for 1880 and 1890.

TABLE 1

EXCESS OF NATIVE WHITE, FOREIGN WHITE, AND COLORED MALES AND FE-
MALES, RESPECTIVELY, REPORTED FOR CERTAIN YEARS OF AGE IN THE
UNITED STATES CENSUSES OF 1880, 1890, AND 1900 [1]

Percentages

| Sex and ages. | Per cent. of excess of number in 100,000, for each specified age. | | | | | | | | |
| | Native white. | | | Foreign white. | | | Colored.[2] | | |
	1900	1890	1880	1900	1890	1880	1900	1890	1880
MALES.									
2 years	2.6	37.8	11.3	....	....	....	7.3	34.3	15.7
4 years	0.1	....	1.6	....	....	....	1.9	5.7	4.1
6 years	0.8	2.5	0.1	....	....	....	2.2	8.1	4.0
8 years	....	....	0.2	....	....	....	1.9	7.6	4.3
10 years	2.6	6.7	7.1	....	....	....	10.4	20.2	20.6
12 years	0.2	12.5	11.5	....	....	....	18.6	31.6	33.2
14 years	0.7	7.3	....	....	....	....	1.7	5.3	4.7
16 years	0.4	5.2	3.8	....	....	....	1.1	3.0	....
18 years	....	4.6	10.7	19.5	18.6	14.8	10.2	16.0	26.1
20 years	0.1	....	....	17.1	22.8	17.1	13.5	....	19.7
21 years	0.1	13.8	4.1	1.1	1.3	....	....	4.6	....
25 years	....	....	....	10.2	11.0	18.3	13.2	16.6	29.7
30 years	15.1	26.7	28.9	32.4	41.0	50.8	45.1	54.0	69.4
35 years	3.8	16.2	23.1	20.5	31.5	44.9	39.8	55.9	69.8
40 years	10.2	28.0	31.5	29.6	50.3	60.4	49.1	58.2	76.2
45 years	9.1	27.6	25.7	23.3	44.1	53.3	52.2	66.7	76.2
50 years	14.8	31.9	32.9	33.0	51.4	61.3	57.3	65.0	79.5
55 years	4.3	....	8.5	14.5	19.9	27.2	37.1	43.3	58.6
60 years	12.3	36.3	33.8	33.7	57.4	61.9	68.5	78.0	86.8
65 years	....	13.0	13.4	17.9	24.6	35.2	58.0	65.6	74.2
70 years	4.2	29.0	16.2	25.1	45.6	47.2	65.1	73.1	82.9
75 years	....	....	....	6.6	20.0	24.5	58.6	65.1	74.5
80 years	....	17.4	4.7	8.9	36.2	43.6	66.3	75.7	83.5
85 years	....	....	....	....	....	....	47.8	50.0	60.0
90 years	....	16.7	....	....	35.7	30.0	62.5	70.6	82.4
FEMALES.									
2 years	2.8	38.1	11.0	....	....	....	6.0	34.2	14.4
4 years	0.1	....	0.3	....	....	....	....	1.3	2.2
6 years	1.2	2.8	0.5	....	....	....	2.9	8.8	5.8
8 years	....	....	....	....	....	....	1.9	8.6	5.5
10 years	2.8	5.1	5.3	....	....	....	9.2	15.7	15.7
12 years	....	11.4	9.1	....	....	....	18.0	30.6	30.7
14 years	....	5.3	....	....	....	....	2.5	5.1	2.4
16 years	1.3	8.0	7.1	....	....	....	5.6	12.2	6.8
18 years	0.8	10.2	16.9	20.6	23.4	18.9	13.7	18.8	29.9
20 years	3.2	8.4	8.5	17.6	26.7	22.7	23.3	23.4	41.2
21 years	....	....	....	....	....	....	....	....	....
25 years	....	....	....	5.9	9.0	18.8	12.6	23.5	38.7
30 years	14.0	29.3	32.4	25.6	38.2	51.8	45.8	60.7	76.2
35 years	2.0	13.6	22.4	14.8	29.3	43.5	40.5	59.8	73.9
40 years	8.1	31.5	34.4	23.6	49.8	60.3	54.0	65.6	81.1
45 years	4.6	22.9	20.5	18.6	38.1	47.4	53.5	69.6	78.6
50 years	14.1	35.8	36.4	32.9	53.8	61.8	62.7	70.6	84.4
55 years	4.2	....	6.2	16.0	22.4	27.4	44.4	52.5	66.8
60 years	14.1	40.3	39.1	38.5	60.6	65.3	73.6	82.3	90.2
65 years	1.2	15.3	16.8	21.0	27.1	36.9	65.8	74.1	82.0
70 years	8.1	33.6	27.0	31.2	50.4	56.6	74.0	80.2	87.9
75 years	1.3	6.3	5.3	13.8	24.6	31.6	67.2	73.7	82.8
80 years	....	22.4	20.0	20.8	47.7	56.6	77.1	84.2	90.3
85 years	....	....	....	....	....	3.6	58.1	63.3	71.0
90 years	....	12.5	....	14.3	47.1	50.0	74.1	78.6	90.9

[1] For the mainland of the United States.
[2] Persons of negro descent, Chinese, Japanese, and Indians.

On account of the constantly increasing number of foreign white persons reported for each year of age, no attempt was made in Table 1 to measure the relative excess for this class of persons under eighteen years of age. There is considerable concentration on eighteen and twenty years and thereafter on the five- and ten-year periods, and the percentages given in Table 1, although not so conclusive as those for the native elements, are at least indicative of the extent to which concentration had lessened in 1900 as compared with the two censuses preceding.

The general conclusion drawn by the Census authorities, from a study of the figures of Table 1 was "that, as far as the concentration on certain ages is concerned, the attempt to secure in 1900 a return of age according to date of birth was partially successful, but it is apparent that concentration cannot be wholly avoided in any case. This is particularly true with respect to colored persons, and to a somewhat less extent with respect to foreign white persons; and for these two elements of the population it is probable that no great improvement can be expected under the present conditions of census enumeration."

That no great progress has been made since in this matter is indicated by the following statement in the reports of the Fourteenth Census, 1920 (vol. ii, pp. 145, 146):

"Irregularities of this character are due in large part to errors in the census returns. These errors result from three causes: (1) Some persons do not know their exact age. (2) The enumerators are obliged in many cases to obtain information relating to the persons enumerated from a third person, either some member of the family found at home or a person in charge of a hotel or boarding house, who can give the age only approximately. (3) In certain instances, apparently more frequent among women than among men, the age is intentionally misstated. Where the age is not accurately known there is a tendency to report it as a multiple of 2 or of 5, and especially, in the case of ages above 20, as a multiple of 10. There is also a tendency to concentrate on age 21 for men. In general, the degree of inaccuracy is greater for adults than for children and youths, and is greater for those classes of the population in which the proportion of illiterates is greatest. The returns also undoubtedly exaggerate the number of centenarians, particularly among the Negroes and Indians."

THE REGISTRATION METHOD

The theory of this method is to record or register each event in the ceaseless flow of the stream of life *as, and when, it happens.*

A mechanism is created in the body politic which makes certain individuals responsible for the prompt recording of each event when it happens. In the field of our present interest it is the physician who is thus held primarily responsible for the recording or registering with some central authority of the facts about births and deaths. If a person dies and no physician has been in attendance, the record is caught up through the necessity of a burial permit. The *corpus* of every deceased human being must be somehow disposed of. The central registration authority in each locality is the only person qualified to permit legal disposal. Therefore substantially all deaths must get registered. In the case of birth, the attending physician or midwife again is required by law to report the fact. Unfortunately, if the birth has not been attended by anybody but the mother and infant, it is not so easy as in the case of death to catch the record. There are growing up, however, various legal necessities for the possession of a birth certificate, so that ultimately the registration of births should become something like as accurate as the registration of deaths.

The heuristic advantages of the registration over the census method are apparent. The *course* of events can be followed. Registration gives us such knowledge as we have of births, deaths, sickness, marriages, divorces, etc., so far as concerns population aggregates.

THE *AD HOC* OR CASE RECORD METHOD

This is the ordinary method of science in general for getting a collection of pertinent quantitative data. In a defined universe of interest cases are recorded in respect of the points or attributes of interest. Thus some may record in all cases of typhoid fever the age, stature, body weight, daily temperature, etc., of the individual. Logically considered, it is a combination of the essential features of the census and the registration method confined to a particular universe of interest. In a later chapter more will be said of the making of medical records.

OFFICIAL REGISTRATION RECORDS

There are reproduced below in reduced facsimile the standard *birth and death registration certificates* as used in the United States

DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

STANDARD CERTIFICATE OF BIRTH

State File No. ....
Registered No. ....

MARGIN RESERVED FOR BINDING
WRITE PLAINLY WITH UNFADING INK—THIS IS A PERMANENT RECORD
N. B.—In case of more than one child at a birth, a SEPARATE RETURN must be made for each, and the number of each, in order of birth, stated.

1. PLACE OF BIRTH—

County.................... State ....
Township .... or Village ....
City.................... No. .... St. .... Ward
(If birth occurred in a hospital or institution, give the NAME instead of street and number)

2. Full name of child....................
(If child is not yet named, make supplemental report, as directed)

3. Sex | If plural births | 4. Twin, triplet, or other .... | 6. Premature........ | 7. Legitimate? .... | 8. Date of birth.................... 19....
5. Number in order of birth .... | Full term .... | (Month, day, year)

9. Full name — FATHER | 18. Full maiden name — MOTHER

10. Residence (usual place of abode) (If nonresident, give place and State) .... | 19. Residence (usual place of abode) (If nonresident, give place and State) ....

11. Color or race .... | 12. Age at last birthday .... (Years) | 20. Color or race .... | 21. Age at last birthday .... (Years)

13. Birthplace (city or place) .... (State or country) | 22. Birthplace (city or place) .... (State or country)

OCCUPATION
14. Trade, profession, or particular kind of work done, as spinner, sawyer, bookkeeper, etc.
15. Industry or business in which work was done, as silk mill, sawmill, bank, etc.
16. Date (month and year) last engaged in this work .... 19....
17. Total time (years) spent in this work ....

OCCUPATION
23. Trade, profession, or particular kind of work done, as housekeeper, typist, nurse, clerk, etc.
24. Industry or business in which work was done, as own home, lawyer's office, silk mill, etc.
25. Date (month and year) last engaged in this work .... 19....
26. Total time (years) spent in this work ....

27. Number of children of this mother (At time of this birth and including this child) (a) Born alive and now living .... (b) Born alive but now dead .... (c) Stillborn ....

28. If stillborn, period of gestation .... {months or weeks} | 29. Cause of stillbirth .... {Before labor .... During labor ....}

CERTIFICATE OF ATTENDING PHYSICIAN OR MIDWIFE

I hereby certify that I attended the birth of this child, who was .... at .... m. on the date above stated
(Born alive or stillborn)

{When there was no attending physician or midwife, then the father, householder, etc., should make this return.}
Given name added from a supplemental report ....
(Date of)

(Signed) ...., M. D.
or ...., Midwife
Address ....
Filed .... 19....
Registrar.

UNITED STATES STANDARD CERTIFICATE OF BIRTH

Why births should be registered.—There is hardly a relation of life, social, legal, or economic, in which the evidence furnished by an accurate registration of births may not prove to be of the greatest value, not only to the individual but also to the public at large. It is not only an act of civilization to register birth certificates but good business, for they are frequently used in many practical ways, some of which are listed below:

(1) As evidence to prove the age and legitimacy of heirs;

(2) As proof of age to determine the validity of a contract entered into by an alleged minor;

(3) As evidence to establish age and proof of citizenship and descent in order to vote;

(4) As evidence to establish the right of admission to the professions and to many public offices;

(5) As evidence of legal age to marry;

(6) As evidence to prove the claims of widows and orphans under the widows' and orphans' pension law;

(7) As evidence to determine the liability of parents for the debts of a minor;

(8) As evidence in the administration of estates, the settlement of insurance and pensions;

(9) As evidence to prove the irresponsibility of children under legal age for crime and misdemeanor, and various other matters in the criminal code;

(10) As evidence in the enforcement of law relating to education and to child labor;

(11) As evidence to determine the relations of guardians and wards;

(12) As proof of citizenship in order to obtain a passport;

(13) As evidence in the claim for exemption from or the right to jury and military service.

Statement of occupation.—Make some entry in this section for each parent. For a woman whose only occupation is that of home housework, write *housework* in answer to Question 23 and *own home* in answer to Question 24. For a person engaged in domestic service for wages, however, designate the occupation by the appropriate terms, as *housekeeper—private family*, *cook—hotel*, etc. For a person who has no occupation whatever write *none*.

To be complete, an occupation return must state:

14 and 23.—The trade, profession, or particular kind of work done.
15 and 24.—The industry or business in which the work is done.
16 and 25.—The month and year the person last worked at the occupation.
17 and 26.—The number of years the person followed the occupation.

In stating the occupation, avoid the use of such indefinite terms as "employee," "worker," "operative," etc. Find out the particular kind of work done and return that, as *spinner*, *weaver*, etc.

In stating the industry or business, avoid the use of such general terms as "store," "factory," "mill," etc. State the particular kind of store, factory, mill, etc., as *grocery store*, *soap factory*, *cotton mill*, etc.

Distinguish carefully the different kinds of engineers by stating the full descriptive titles, as *civil engineer*, *mechanical engineer*, *mining engineer*, *stationary engineer*, etc. Avoid the term "laborer" when a more precise statement of occupation can be secured. Do not use the word "mechanic," but give the exact occupation, as *carpenter*, *painter*, *machinist*, etc. Distinguish carefully between *retail merchants* and *wholesale merchants*. A person who sells goods should be called a *salesman* and not a *clerk*.

Registration Areas. They are introduced here in order that the reader may understand clearly what information is basically available in official vital statistics in the United States. In actual practice the extent to which the different items on the certificates are filled out depends upon the force and vigilance of the registration officials. In some communities there is a good deal of laxity in regard to such items as occupation, birthplace of parents, etc. But if the registra-

tion officials are sufficiently active and painstaking in their duties, all of the information called for on the certificates can be had.

The student of vital statistics should study the birth and death certificates with the most painstaking care. Indeed, he will find it advantageous to learn by heart every word and punctuation mark on them. From the certificates comes the raw material with which he is compelled to work. Whenever he deals with birth or death rates, in whatever connection, there should be a clear picture in his mind as to exactly how the basic data were got, and what they mean in the individual case.

The latest (1930) improved standard forms of birth certificates and death certificates, as officially approved by the Census Bureau of the United States, are shown on pages 73–75. In both cases the printed matter on the reverse of the certificate is reproduced, as well as the material on the face.

The new death certificate embodies a number of improvements over the one formerly used. These chiefly concern greater detail

UNITED STATES STANDARD CERTIFICATE OF DEATH

Statement of occupation.—Precise statement of occupation is very important, so that the relative healthfulness of various pursuits can be known. Make some entry in this section for every person aged 10 years or over. If the occupation had been given up or changed on account of the disease causing death, report the occupation prior to illness. If the deceased had retired from business, report the occupation prior to retirement. Children not gainfully employed may be returned as *at school* or *at home*. For a woman whose only occupation was that of home housework, write *housework* in answer to Question 8 and *own home* in answer to Question 9. For a person engaged in domestic service for wages, however, designate the occupation by the appropriate terms, as *housekeeper—private family, cook—hotel*, etc. For a person who had no occupation whatever write *none*.

To be complete, an occupation return must state:

8.—The trade, profession, or particular kind of work done.
9.—The industry or business in which the work was done.
10.—The month and year the deceased last worked at the occupation.
11.—The number of years the deceased followed the occupation.

In stating the occupation, avoid the use of such indefinite terms as "employee," "worker," "operative," etc. Find out the particular kind of work done and return that, as *spinner, weaver*, etc.

In stating the industry or business, avoid the use of such general terms as "store," "factory," "mill," etc. State the particular kind of store, factory, mill, etc., as *grocery store, soap factory, cotton mill*, etc.

Distinguish carefully the different kinds of engineers by stating the full descriptive titles, as *civil engineer, mechanical engineer, mining engineer, stationary engineer*, etc. Avoid the term "laborer" when a more precise statement of the occupation can be secured. Do not use the word "mechanic," but give the exact occupation, as *carpenter, painter, machinist*, etc. Distinguish carefully between *retail merchants* and *wholesale merchants*. A person who sells goods should be called a *salesman* and not a *clerk*.

Statement of cause of death.—Cause of death means the disease, injury, or complication which causes death, not the mode of dying, *e. g.*, heart failure, asphyxia, asthenia, etc. As principal cause name the disease or injury causing death. As related causes, name earlier morbid conditions, if any, related to the principal cause and any important complication of the principal cause. Under contributory causes of importance not related to principal cause, name other important diseases or injuries. Examples:

Example I		Example II	
The **principal cause of death** and related causes of importance in order of onset were as follows:	Date of onset	The **principal cause of death** and related causes of importance in order of onset were as follows:	Date of onset
Arteriosclerosis	*1915*	*Attack of epilepsy*	*1 week ago*
Chronic interstitial nephritis	*1921*	*Run over by street car*	*1 week ago*
Cerebral hemorrhage	*July 5, 1927*	*Peritonitis*	*3 days ago*
Contributory causes of importance not related to principal cause:		**Contributory causes** of importance not related to principal cause:	
Fracture of arm		*Influenza*	*6 weeks ago*
Automobile accident	*May 5, 1927*		

In a group of causes containing the principal cause and related causes, the causes should be given in the order of onset, so that in a group of three causes the principal cause may appear in either first, second, or third position. The principal cause in each of the above examples happens to be the second cause given.

ADDITIONAL SPACE FOR FURTHER STATEMENTS BY PHYSICIAN

in regard to primary, subsequent, and contributory causes; with the occupational relationships to death; and with the more scientific or objective support of cause of death by postmortem, operative, or laboratory evidence—chemical, bacteriological, or biological. These improved certificates may reasonably be expected, with the passage of time, to furnish a mass of data on mortality of presumably superior accuracy as to cause of death, as compared with anything hitherto available.

In some localities a special certificate is used to record still-births on the ground that "there is merit in requiring a birth and death certificate for children born alive and soon dying, and a different form to record viable products of conception stillborn." The form of still-birth certificate used in New York City is as follows:

5-H   25-2696 25-B

Department of Health of The City of New York
BUREAU OF RECORDS

No fœtus of any period of uterine gestation should be interred or disposed of in any other manner without a permit therefor having been obtained from the Department of Health, such permit to be granted upon the presentation of a proper return.

Persons who are unable or unwilling for any reason to bury a fœtus should immediately notify the Department of Health, which Department will see that the fœtus is properly and promptly buried in the City Cemetery.

CERTIFICATE OF A STILL-BIRTH

The death of an infant that has breathed must not be reported as a still-birth; such cases must be reported by filing a certificate of birth and a certificate of death.

Borough of.. Registered No...............

No...St.   Character of premises, whether tenement, private, hotel, hospital or other place, etc...............

Sex.....................Color or Race...................Date of Still-Birth...........................192......
                                                           (Month)   (Day)

	Father	Mother
Name		
Residence		
Birthplace		
Age	Color or Race	Color or Race
Occupation		
Period of Utero Gestation	Number of Previous Pregnancies	Number of Living Births

I hereby certify that the foregoing particulars are correct as near as the same can be ascertained, and I further certify that I attended at this still-birth; that the still-birth occurred on the .... day of........................ 192...., that the actual cause of the death of this fœtus was

....

....and that said death of fœtus occurred* before during *labor.*

Predisposing cause ....

Witness my hand this........day of.................. 192...   Signature.............................M. D.

Filed..   Address...............................

Place of Burial		Date of Burial	
Undertaker		Address	

Margin reserved for binding — *NO MUTILATED CERTIFICATE WILL BE RECEIVED* — *The particulars required on this Certificate must be written with black ink*

STILL-BIRTH PROCEDURE FOR MIDWIVES

Should the child not breathe after birth, the midwife must report the fact at once, by telephone or messenger, to the Department of Health, when an inspector will visit the case and countersign the still-birth certificate which the midwife must leave at the home.

The fœtus must not be removed from the premises until this certificate has been approved by the inspector from the Department of Health and a permit has been issued by the Bureau of Records.

I hereby certify that I have been employed as undertaker by.............................

the.......................of deceased. This statement is made to obtain a permit for the
   (relationship)

burial or cremation of the remains of deceased..

Signature..

THE INTERNATIONAL LIST OF THE CAUSES OF DEATH

If the statistics of mortality are to be comparable from locality to locality, it is plain that a uniform system of nomenclature of the causes of death must everywhere be used. Similarly, if hospital records are to be comparable, a uniform system of nomenclature of morbid conditions and of treatments and results must be in operation.

The science of nosology, or the classification of disease, attracted a great deal more attention from medical men a century ago than it does now. The predominant system in vogue for a long time was
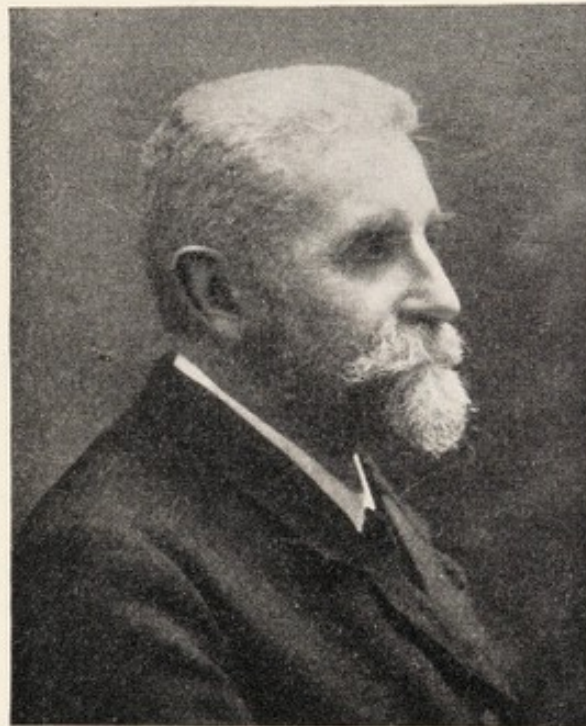
Fig. 12.—Portrait of Dr. Jacques Bertillon. (Reproduced through the kindness of Dr. Frederick L. Hoffman, to whom the original belongs, and Brig.-Gen. Robert E. Noble, Librarian of the Surgeon-General's office.)

due to Cullen. The first attempt to adapt it specifically to statistical uses was due to William Farr. In the First Annual Report of the Registrar-General of England and Wales Farr said:

"The advantages of a uniform statistical nomenclature, however imperfect, are so obvious that it is surprising no attention has been paid to its enforcement in Bills of Mortality. Each disease has in many instances been denoted by three or four terms, and each term has been applied to as many different diseases; vague, inconvenient names have been employed, or complications have been registered instead of primary diseases. The nomenclature is of as much importance in this department of inquiry as weights and measures in the physical sciences, and should be settled without delay."

The First Statistical Congress, held in Brussels in 1853, selected Farr and Marc d'Espine of Geneva to draw up a report upon a classification adapted to international use. It is interesting to note that the resolution to this end was introduced in the Congress by Dr. Achille Guillard, who was the maternal grandfather of Dr. Jacques Bertillon. In the last quarter of a century Bertillon has been perhaps more active than anyone else in perfecting and extending the use of the International Classification.

The classification prepared by Farr and d'Espine was adopted in Paris in 1855, in Vienna in 1857, and was translated into six languages. It was revised in 1864, 1874, 1880, and 1886. With further revision it was adopted by the International Statistical Institute in Chicago in 1893, and provisions were made for decennial revisions. The first of these was made in 1900, the second in 1909, the third in 1920, and the most recent one in 1929.

The present form of the International List, after its latest revision, is as follows:

INTERNATIONAL LIST OF CAUSES OF DEATH
(Fourth Decennial Revision by the International Commission, Paris, October, 1929.)

(The numbers in the list represent obligatory divisions. The subdivisions indicated by letters *a*, *b*, *c*, etc., are optional. When a cause of death is obligatorily divided among several numbers, it is essential to reserve in the tables a line for the total, relative to this cause. Example: Tuberculosis (all forms) Nos. 23 to 32.)

I. *Infectious and Parasitic Diseases*

1. Typhoid fever.
2. Paratyphoid fever.
3. Typhus fever.
4. Relapsing fever.
5. Undulant fever.
6. Smallpox:
    (*a*) Variola major.
    (*b*) Variola minor, alastrim.
    (*c*) Not specified.
7. Measles.
8. Scarlet fever.
9. Whooping cough.
10. Diphtheria.
11. Influenza:
    (*a*) With respiratory complications specified.
    (*b*) Without respiratory complications specified.

12. Cholera.
13. Dysentery:
    (a) Amebic.
    (b) Bacillary.
    (c) Unspecified or due to other causes.
14. Plague:
    (a) Bubonic.
    (b) Pneumonic.
    (c) Septicemic.
    (d) Unspecified.
15. Erysipelas.
16. Acute poliomyelitis and acute polio-encephalitis.
17. Lethargic or epidemic encephalitis.
18. Epidemic cerebrospinal meningitis.
19. Glanders.
20. Anthrax (Bacillus anthracis), malignant pustule.
21. Rabies.
22. Tetanus.
23. Tuberculosis of the respiratory system.
24. Tuberculosis of the meninges and central nervous system.
25. Tuberculosis of the intestines and peritoneum (including the mesenteric and retroperitoneal glands).
26. Tuberculosis of the vertebral column.
27. Tuberculosis of the bones and joints (vertebral column excepted):
    (a) Bones.
    (b) Joints.
28. Tuberculosis of the skin and subcutaneous cellular tissue.
29. Tuberculosis of the lymphatic system (bronchial, mesenteric, and retroperitoneal glands excepted).
30. Tuberculosis of the genito-urinary system.
31. Tuberculosis of other organs.
32. Disseminated tuberculosis:
    (a) Acute.
    (b) Chronic.
    (c) Unspecified.
33. Leprosy.
34. Syphilis:
    (a) Congenital.
    (b) Acquired.
    (c) Unspecified.
35. Gonococcus infection and other venereal diseases.
36. Purulent infection, septicemia, non-puerperal:
    (a) Septicemia.
    (b) Pyemia or pyohemia.
    (c) Gas gangrene.
37. Yellow fever.
38. Malaria:
    (a) Malarial fever
    (b) Malarial cachexia.

39. Other diseases due to protozoal parasites.
40. Ancylostomiasis.
41. Hydatid cysts:
    (a) Of the liver.
    (b) Of other organs.
42. Other diseases caused by helminths.
43. Mycoses.
44. Other infectious and parasitic diseases:
    (a) Chicken-pox.
    (b) German measles.
    (c) Others under this title.

II. *Cancers and Other Tumors*

45. Cancer and other malignant tumors of the buccal cavity and pharynx:
    (a) Lip.
    (b) Tongue.
    (c) Mouth.
    (d) Jaw.
    (e) Other and unspecified parts of buccal cavity.
    (f) Pharynx.
46. Cancer and other malignant tumors of the digestive tract and peritoneum:
    (a) Esophagus.
    (b) Stomach and duodenum
    (c) Intestine (except rectum and anus).
    (d) Rectum and anus.
    (e) Liver and biliary passages.
    (f) Pancreas.
    (g) Peritoneum.
    (h) Others.
47. Cancer and other malignant tumors of the respiratory system:
    (a) Larynx.
    (b) Lungs and pleura.
    (c) Others.
48. Cancer and other malignant tumors of the uterus.
49. Cancer and other malignant tumors of other female genital organs:
    (a) Ovary and Fallopian tube.
    (b) Vagina and vulva.
    (c) Others.
50. Cancer and other malignant tumors of the breast.
51. Cancer and other malignant tumors of the male genito-urinary organs
    (a) Kidneys and suprarenals (male).
    (b) Bladder (male).
    (c) Prostate.
    (d) Testes.
    (e) Scrotum.
    (f) Others.
52. Cancer and other malignant tumors of the skin.

53. Cancer and other malignant tumors of other or unspecified organs:
    (a) Kidneys and suprarenals (female).
    (b) Bladder (female).
    (c) Brain.
    (d) Bones (except jaw).
    (e) Others.
54. Non-malignant tumors:
    (a) Of the ovary.
    (b) Of the uterus.
    (c) Of other female genital organs.
    (d) Of the brain.
    (e) Of other organs.
55. Tumors of which the nature is not specified:
    (a) Of the ovary.
    (b) Of the uterus.
    (c) Of other female genital organs.
    (d) Of the brain.
    (e) Of other organs.

III. *Rheumatic Diseases, Nutritional Diseases, Diseases of the Endocrine Glands and Other General Diseases*

56. Acute rheumatic fever.
57. Chronic rheumatism, osteo-arthritis.
58. Gout.
59. Diabetes mellitus.
60. Scurvy:
    (a) Infantile scurvy (Barlow's disease).
    (b) Scurvy.
61. Beriberi.
62. Pellagra.
63. Rickets.
64. Osteomalacia.
65. Diseases of the pituitary body.
66. Diseases of the thyroid and parathyroid glands:
    (a) Simple goiter.
    (b) Exophthalmic goiter.
    (c) Myxedema and cretinism.
    (d) Tetany.
    (e) Others.
67. Diseases of the thymus gland.
68. Diseases of the adrenals (Addison's disease; not specified as tuberculous).
69. Other general diseases.

IV. *Diseases of the Blood and Blood-making Organs*

70. Hemorrhagic conditions:
    (a) Primary purpuras.
    (b) Hemophilia.

6

71. Anemias:
    (a) Pernicious anemia.
    (b) Others.
72. Leukemias and pseudoleukemias:
    (a) True leukemias.
    (b) Pseudoleukemias (Hodgkin's disease).
73. Diseases of the spleen.
74. Other diseases of the blood and blood-making organs.

V. *Chronic Poisonings and Intoxications*

75. Alcoholism (acute or chronic).
76. Chronic poisoning by mineral substances:
    (a) Lead.
    (b) Occupational (except lead).
    (c) Others.
77. Chronic poisoning by organic substances:
    (a) Occupational.
    (b) Others.

VI. *Diseases of the Nervous System and of the Organs of Special Sense*

78. Encephalitis (non-epidemic):
    (a) Abscess of the brain.
    (b) Others.
79. Meningitis:
    (a) Simple meningitis.
    (b) Non-epidemic cerebrospinal meningitis.
80. Progressive locomotor ataxia (tabes dorsalis).
81. Other diseases of the spinal cord.
82. Cerebral hemorrhage, cerebral embolism and thrombosis:
    (a) Cerebral hemorrhage.
    (b) Cerebral embolism and thrombosis.
    (c) Hemiplegia and causes unspecified.
83. General paralysis of the insane.
84. Dementia praecox and other psychoses:
    (a) Dementia praecox.
    (b) Other psychoses.
85. Epilepsy.
86. Convulsions (under five years of age).
87. Other diseases of the nervous system:
    (a) Softening of the brain.
    (b) Neuralgia and neuritis.
    (c) Others.
88. Diseases of the organs of vision.
89. Diseases of the ear and of the mastoid process:
    (a) Otitis.
    (b) Diseases of the mastoid process.
    (c) Others.

VII. *Diseases of the Circulatory System*

90. Pericarditis.
91. Acute endocarditis:
    (*a*) Specified as acute.
    (*b*) Unspecified (under forty-five years of age).
92. Chronic endocarditis, valvular diseases:
    (*a*) Endocarditis specified as chronic and valvular disease.
    (*b*) Endocarditis unspecified (forty-five years and over).
93. Diseases of the myocardium:
    (*a*) Acute myocarditis.
    (*b*) Myocarditis unspecified (under forty-five years of age).
    (*c*) Chronic myocarditis and myocardial degeneration.
    (*d*) Unspecified.
94. Diseases of the coronary arteries and angina pectoris:
    (*a*) Angina pectoris.
    (*b*) Diseases of the coronary arteries.
95. Other diseases of the heart:
    (*a*) Functional diseases of the heart.
    (*b*) Other and unspecified.
96. Aneurysm (except of the heart).
97. Arteriosclerosis (diseases of the coronary arteries excepted).
98. Gangrene (not gas gangrene, see 36*c*):
    (*a*) Senile.
    (*b*) Others.
99. Other diseases of the arteries.
100. Diseases of the veins (varices, hemorrhoids, phlebitis, etc.; not phlegmasia alba dolens, see 148*a*).
101. Diseases of the lymphatic system (lymphangitis, etc.).
102. Idiopathic anomalies of the blood pressure.
103. Other diseases of the circulatory system.

VIII. *Diseases of the Respiratory System*

104. Diseases of the nasal fossæ and annexa.
105. Diseases of the larynx.
106. Bronchitis:
    (*a*) Acute.
    (*b*) Chronic.
    (*c*) Unspecified.
107. Bronchopneumonia (including capillary bronchitis):
    (*a*) Bronchopneumonia.
    (*b*) Capillary bronchitis.
108. Lobar pneumonia.
109. Pneumonia, unspecified.
110. Pleurisy.
111. Congestion, edema, embolism, hemorrhagic infarct, and thrombosis of the lungs:
    (*a*) Pulmonary embolism and thrombosis.
    (*b*) Others.

112. Asthma.
113. Pulmonary emphysema.
114. Other diseases of the respiratory system (tuberculosis excepted):
    (a) Chronic interstitial pneumonia including occupational diseases of the respiratory system.
    (b) Others, including gangrene of the lung.

IX. *Diseases of the Digestive System*

115. Diseases of the buccal cavity and annexa and of the pharynx and tonsils (including adenoid vegetations):
    (a) Pharynx and tonsils.
    (b) Others.
116. Diseases of the esophagus.
117. Ulcer of the stomach and duodenum:
    (a) Stomach.
    (b) Duodenum.
118. Other diseases of the stomach (cancer excepted).
119. Diarrhea and enteritis (under two years of age).
120. Diarrhea, enteritis, and ulceration of intestines (two years and over):
    (a) Diarrhea, enteritis.
    (b) Ulceration of intestines.
121. Appendicitis.
122. Hernia, intestinal obstruction:
    (a) Hernia.
    (b) Intestinal obstruction.
123. Other diseases of the intestines.
124. Cirrhosis of the liver:
    (a) Specified as alcoholic.
    (b) Not specified as alcoholic.
125. Other diseases of the liver (including yellow atrophy of liver):
    (a) Yellow atrophy of liver.
    (b) Others.
126. Biliary calculi.
127. Other diseases of the gall-bladder and biliary passages.
128. Diseases of the pancreas.
129. Peritonitis, cause not specified.

X. *Diseases of the Genito-urinary System*

130. Acute nephritis.
131. Chronic nephritis.
132. Nephritis, unspecified.
133. Other diseases of the kidneys and ureters (puerperal diseases excepted):
    (a) Pyelitis.
    (b) Others.
134. Calculi of the urinary passages:
    (a) Calculi of the kidneys and ureters.
    (b) Calculi of the bladder.
    (c) Other and unspecified.

135. Diseases of the bladder (tumors excepted):
    (*a*) Cystitis.
    (*b*) Others.
136. Diseases of the urethra, urinary abscess, etc.:
    (*a*) Stricture of the urethra.
    (*b*) Others.
137. Diseases of the prostate.
138. Diseases of the male genital organs—not specified as venereal.
139. Diseases of the female genital organs—not specified as venereal:
    (*a*) Ovaries, tubes, and parametrium.
    (*b*) Uterus.
    (*c*) Breast.
    (*d*) Others.

XI. *Diseases of Pregnancy, Childbirth, and the Puerperal State*

140. Abortion with septic conditions.
141. Abortion without mention of septic conditions (to include hemorrhages).
142. Ectopic gestation.
143. Other accidents of pregnancy (not to include hemorrhages).
144. Puerperal hemorrhage:
    (*a*) Placenta previa.
    (*b*) Other hemorrhages.
145. Puerperal septicemia (not specified as due to abortion):
    (*a*) Puerperal septicemia and pyemia.
    (*b*) Puerperal tetanus.
146. Puerperal albuminuria and eclampsia.
147. Other toxemias of pregnancy.
148. Puerperal phlegmasia alba dolens, embolus, sudden death (not specified as septic):
    (*a*) Phlegmasia alba dolens.
    (*b*) Embolism and thrombosis.
149. Other accidents of childbirth:
    (*a*) Cesarean operation.
    (*b*) Others.
150. Other and unspecified conditions of the puerperal state.

XII. *Diseases of the Skin and Cellular Tissue*

151. Furuncle, carbuncle.
152. Phlegmon, acute abscess.
153. Other diseases of the skin and annexa, and of the cellular tissue.

XIII. *Diseases of the Bones and Organs of Locomotion*

154. Osteomyelitis.
155. Other diseases of the bones (tuberculosis excepted).
156. Diseases of the joints and other organs of locomotion:
    (*a*) Joints (tuberculosis and rheumatism excepted).
    (*b*) Other organs of locomotion.

XIV. *Congenital Malformations*

157. Congenital malformation (still-births not included):
 (*a*) Congenital hydrocephalus.
 (*b*) Spina bifida and meningocele.
 (*c*) Congenital malformation of the heart.
 (*d*) Monstrosities.
 (*e*) Others.

XV. *Diseases of Early Infancy*

158. Congenital debility.
159. Premature birth.
160. Injury at birth:
 (*a*) Cesarean operation.
 (*b*) Without Cesarean operation.
161. Other diseases peculiar to early infancy:
 (*a*) Atelectasis.
 (*b*) Icterus of the newborn.
 (*c*) Sclerema.
 (*d*) Others.

XVI. *Senility*

162. Senility:
 (*a*) With senile dementia.
 (*b*) Without senile dementia.

XVII. *Violent and Accidental Deaths*

All violent or accidental deaths should be included under the headings 163 to 198 so that all deaths without exception are included under one or other of the 200 rubrics of the list.

For the deaths included in numbers 176 to 195, a second independent tabulation under the following headings is obligatory:

1. Accidents in mines and quarries.
2. Accidents caused by machinery.
3. Accidents by means of transportation:
 (*a*) Railroads and street cars.
 (*b*) Automobiles, motorcycles.
 (*c*) Other means of transportation by land.
 (*d*) Transportation by water.
 (*e*) Transportation by air.

163. Suicide by solid or liquid poisons or by absorption of corrosive substances:
 (*a*) Arsenic.
 (*b*) Hydrocyanic acid.
 (*c*) Opium, morphin, laudanum.
 (*d*) Strychnin.
 (*e*) Corrosive sublimate.
 (*f*) Carbolic acid.
 (*g*) Lysol.
 (*h*) Other poisons or kind not stated.

164. Suicide by poisonous gas.

165. Suicide by hanging or strangulation.
166. Suicide by drowning.
167. Suicide by firearms.
168. Suicide by cutting or piercing instruments.
169. Suicide by jumping from high places.
170. Suicide by crushing.
171. Suicide by other means.
172. Infanticide (murder of infants under one year):
    (a) Immediately after birth.
    (b) Others, under one year.
173. Homicide by firearms (persons one year and over).
174. Homicide by cutting or piercing instruments (persons one year and over).
175. Other homicides of persons one year and over.
176. Attack by venomous animals.
177. Poisoning by food.
178. Accidental absorption of poisonous gas.
179. Other acute accidental poisonings (gas excepted):
    (a) Wood alcohol.
    (b) Denatured alcohol.
    (c) Carbolic acid.
    (d) Opium, morphin, laudanum.
    (e) Strychnin.
    (f) Other poisons or kind not stated.
180. Conflagration.
181. Accidental burns (conflagration excepted).
182. Accidental mechanical suffocation.
183. Accidental drowning.
184. Accidental traumatism by firearms (wounds of war excepted).
185. Accidental traumatism by cutting or piercing instruments (wounds of war excepted).
186. Accidental traumatism by fall, crushing, landslide:
    (a) Fall down stairs.
    (b) Fall in building operations.
    (c) Other falls.
    (d) Crushing, landslide.
187. Cataclysm (all deaths attributed to a cataclysm regardless of their nature)
188. Injuries by animals.
189. Hunger or thirst.
190. Excessive cold.
191. Excessive heat.
192. Lightning.
193. Accidents due to electric currents.
194. Other accidents:
    (a) Foreign body.
    (b) Others.
195. Violent deaths of which the nature (accident, suicide, homicide) is unknown.
196. Wounds of war.
197. Execution of civilians by belligerent armies.
198. Legal executions.

XVIII. *Ill-defined Causes of Death*

199. Sudden death.
200. Cause of death not specified or ill-defined:
    (*a*) Ill-defined.
    (*b*) Not specified or unknown.

In addition to the detailed list as given above the Commission recommended, in its 1929 revision, two other briefer lists. The first of these, called the *Intermediate List*, contains 85 titles. The second, called the *Abridged List*, contains 43 titles.

These lists are as follows:

INTERMEDIATE LIST

(The numbers in parentheses are those of the detailed list given above.)

I. *Infectious and Parasitic Diseases*

1. Typhoid fever and paratyphoid fever (1 and 2).
2. Typhus fever (3).
3. Smallpox (6).
4. Measles (7).
5. Scarlet fever (8).
6. Whooping-cough (9).
7. Diphtheria (10).
8. Influenza (11).
9. Dysentery (13).
10. Plague (14).
11. Tuberculosis of the respiratory system (23).
12. All other tuberculosis (24 to 32 inclusive).
13. Syphilis (34).
14. Purulent infection, septicemia, non-puerperal (36).
15. Malaria (38).
16. Other diseases due to protozoa or helminths (39 to 42 inclusive).
17. Other infectious and parasitic diseases* (4, 5, 12, 15 to 22 inclusive, 33, 35, 37, 43, and 44).

II. *Cancers and Other Tumors*

18. Cancer and other malignant tumors (45 to 53 inclusive).
19. Non-malignant tumors (or of which the nature is not specified) (54 and 55).

III. *Rheumatic Diseases, Nutritional Diseases, Diseases of the Endocrine Glands, and Other General Diseases*

20. Acute rheumatic fever (56).
21. Chronic rheumatism and gout (57 and 58).

* The other infectious diseases should be specified when they cause an appreciable mortality, and certain of them (cholera, yellow fever, leprosy) should be specified even if they cause only a single death.

22. Diabetes mellitus (59).
23. Diseases due to vitamin deficiencies (60 to 64 inclusive).
24. Diseases of the thyroid and parathyroid glands (66).
25. Other general diseases (65, 67 to 69 inclusive).

IV. *Diseases of the Blood and Blood-making Organs*

26. Pernicious and other anemias (71).
27. Leukemias, pseudoleukemias, and other diseases of the blood and blood-making organs (70, 72 to 74).

V. *Chronic Poisonings and Intoxications*

28. Alcoholism (acute or chronic) (75).
29. Chronic poisoning (76 and 77).

VI. *Diseases of the Nervous System and of the Organs of Special Sense*

30. Meningitis (79).
31. Progressive locomotor ataxia (80).
32. Cerebral hemorrhage, cerebral embolism and thrombosis (82).
33. General paralysis of the insane (83).
34. Dementia praecox and other psychoses (84).
35. Epilepsy (85).
36. Other diseases of the nervous system (78, 81, 86, and 87).
37. Diseases of the eye, the ear, and the annexa (88 and 89).

VII. *Diseases of the Circulatory System*

38. Pericarditis (90).
39. Acute endocarditis (91).
40. Chronic endocarditis, valvular diseases (92).
41. Diseases of the myocardium (93).
42. Diseases of the coronary arteries, and angina pectoris (94).
43 Other diseases of the heart (95).
44. Aneurysm (except of the heart) (96).
45. Arteriosclerosis and gangrene (97 and 98).
46. Other diseases of the circulatory system (99 to 103 inclusive).

VIII. *Diseases of the Respiratory System*

47. Bronchitis (106).
48. Pneumonia (107 to 109 inclusive).
49. Pleurisy (110).
50. Other diseases of the respiratory system (tuberculosis excepted) (104 and 105, 111 to 114 inclusive).

IX. *Diseases of the Digestive System*

51. Ulcer of the stomach and duodenum (117).
52. Diarrhea and enteritis (under two years of age) (119).
53. Diarrhea, enteritis, and ulceration of intestines (two years and over) (120).
54. Appendicitis (121).

55. Hernia, intestinal obstruction (122).
56. Cirrhosis of the liver (124).
57. Other diseases of the liver and biliary passages (including biliary calculi) (125 to 127 inclusive).
58. Other diseases of the digestive system (115, 116, 118, 123, 128, and 129).

X. *Diseases of the Genito-urinary System*

59. Nephritis (130 to 132 inclusive).
60. Other diseases of the kidneys and ureters (puerperal diseases excepted) (133).
61. Calculi of the urinary passages (134).
62. Diseases of the bladder (tumors excepted) (135).
63. Diseases of the urethra, urinary abscess, etc. (136).
64. Diseases of the prostate (137).
65. Diseases of the genital organs—not specified as venereal (138 and 139).

XI. *Diseases of Pregnancy, Childbirth, and the Puerperal State*

66. Accidents of pregnancy (141, 142, 143).
67. Puerperal hemorrhage (144).
68. Septicemia and puerperal infection (140, 145).
69. Toxemias of pregnancy (albuminuria and eclampsia) (146 and 147).
70. Other puerperal diseases (148 to 150 inclusive).

XII. *Diseases of the Skin and Cellular Tissue*

71. Diseases of the skin and cellular tissue (151 to 153 inclusive).

XIII. *Diseases of the Bones and Organs of Locomotion*

72. Diseases of the bones and of the organs of locomotion (tuberculosis and rheumatism excepted) (154 to 156 inclusive).

XIV. *Congenital Malformations*

73. Congenital malformations (still-births not included) (157).

XV. *Diseases of Early Infancy*

74. Congenital debility (158).
75. Premature birth (159).
76. Injury at birth (160).
77. Other diseases peculiar to early infancy (161).

XVI. *Senility*

78. Senility (162).

XVII. *Violent and Accidental Deaths*

79. Suicide (163 to 171 inclusive).
80. Homicide (172 to 175 inclusive).
81. Accident (176 to 194 inclusive).

82. Violent deaths of which the nature (accident, suicide, homicide) is unknown (195).
83. Wounds of war (including execution of civilians by belligerent armies) (196 and 197).
84. Legal executions.

XVIII. *Ill-defined Causes of Death*

85. Cause of death not specified, or ill defined (199 and 200).

ABRIDGED LIST

(The numbers in parentheses are those of the Detailed List.)

I

1. Typhoid fever and paratyphoid fever (1 and 2).
2. Typhus fever (3).
3. Smallpox (6).
4. Measles (7).
5. Scarlet fever (8).
6. Whooping-cough (9).
7. Diphtheria (10).
8. Influenza (11).
9. Plague (14).
10. Tuberculosis of the respiratory system (23).
11. All other tuberculosis (24 to 32 inclusive).
12. Syphilis (34).
13. Malaria (38).
14. Other infectious and parasitic diseases (4, 5, 12, 13, 15 to 22 inclusive, 33, 35 to 37 inclusive, 39 to 44 inclusive). (See note to corresponding item in Intermediate List.)

II

15. Cancer and other malignant tumors (45 to 53 inclusive).
16. Non-malignant tumors (or of which the nature is not specified) (54 and 55).

III, IV, V, and VI

17. Chronic rheumatism and gout (57 and 58).
18. Diabetes mellitus (59).
19. Alcoholism (acute or chronic) (75).
20. Other general diseases and chronic poisonings (56, 60 to 74 inclusive, 76 and 77).
21. Progressive locomotor ataxia and general paralysis of the insane (80, 83).
22. Cerebral hemorrhage, cerebral embolism and thrombosis (82).
23. Other diseases of the nervous system and of the organs of special sense (78, 79, 81, 84 to 89 inclusive).

VII

24. Diseases of the heart (90 to 95 inclusive).
25. Other diseases of the circulatory system (96 to 103 inclusive).

VIII

26. Bronchitis (106).
27. Pneumonia (all forms) (107 to 109 inclusive).
28. Other diseases of the respiratory system (tuberculosis excepted) (104 and 105, 110 to 114 inclusive).

IX

29. Diarrhea and enteritis (119 and 120).
30. Appendicitis (121).
31. Diseases of the liver and biliary passages (124 to 127 inclusive).
32. Other diseases of the digestive system (115 to 118 inclusive, 122, 123, 128, and 129).

X

33. Nephritis (130 to 132).
34. Other diseases of the genito-urinary system (133 to 139 inclusive).

XI

35. Septicemia and puerperal infection (140 and 145).
36. Other diseases of pregnancy, of childbirth, and of the puerperal state (141 to 144 inclusive, 146 to 150 inclusive).

XII and XIII

37. Diseases of the skin, of the cellular tissue, of the bones, and of the organs of locomotion (151 to 156 inclusive).

XIV and XV

38. Congenital debility, congenital abnormalities, premature birth, etc. (157 to 161 inclusive).

XVI

39. Senility (162).

XVII

40. Suicide (163 to 171 inclusive).
41. Homicide (172 to 175 inclusive).
42. Violent or accidental death (except suicide and homicide) (176 to 198 inclusive).

XVIII

43. Cause of death not specified, or ill defined (199 and 200).

RECOMMENDATIONS OF THE INTERNATIONAL COMMISSION

Certain recommendations made by the Commission in connection with the 1929 revision are of general interest. The following notes are free renderings of and running comments upon the sense of certain items in the *Procès Verbaux*, made in advance of their definitive publication, and therefore not official.

The Commission regards it as a matter of first importance that serious efforts should be made in every country to give special instruction to practitioners and students of medicine regarding the principles according to which death certificates should be filled out. This is a sound recommendation. It can probably be regarded as certain that if the medical schools in this country gave attention seriously to this matter the quality of our vital statistics in respect of the causes of death would be measurably improved within a decade.

In regard to the use of the International List the Commission recommends that countries not in a position to apply the Detailed List in all its subdivisions should nevertheless adhere to the convention and furnish figures for groups of causes of deaths, which should not be more condensed than those of the Intermediate List.

It is recommended that death certificates of persons dying after a surgical operation contain statements as to the morbid condition leading to surgical intervention and the nature of the surgical operation performed. Such procedure would certainly enhance the value of the returns for the student.

In view of the great sociological interest and importance of industrial (occupational) accident mortality two recommendations are made: (1) That the death certificate shall state, with the maximum attainable precision, the *last* occupation followed by the deceased, and (2) that governments should consider whether, in addition to the information now given on death certificates, there should not also be a specific statement as to whether the accident leading to death is, or is not, to be regarded as occupational, at least in the case of the principal rubrics under accidental deaths.

STILL-BIRTHS AND MORBIDITY

In connection with the 1929 revision the International Commission prepared a brief list of causes of still-births (*mortinatalité*), as follows:

I. *Death of the Fetus During Gestation*

1. Syphilis and other chronic diseases.
2. Toxemia of pregnancy (eclampsia, albuminuria, retroplacental hemorrhage).
3. Malformation incompatible with life.
4. Other causes and causes not specified.

II. *Deaths from Premature Birth*

5. Maternal overwork.
6. Traumatism producing premature labor.
7. Placenta previa.
8. Acute infection.
9. Chronic infection, particularly syphilis.
10. Other causes, and causes not specified.

III. *Death of the Fetus During Parturition*

11. Abnormal presentations.
12. Obstacles to the expulsion of the child.
13. Other causes and causes not specified.

The Commission has made another departure in codifying the names of diseases, as distinguished from the nomenclature of the causes of death. The French text follows:

NOMENCLATURE DES MALADIES

La nomenclature des maladies ne diffère de la nomenclature des causes de décès que par la subdivision de quelques rubriques, désignées par des lettres capitales, A, B, C, etc.

On ne reproduira ici que les rubriques ainsi subdivisées.

34. Syphilis.
    A. Congénitale.
    B. Acquise.
        1. Primaire.
        2. Secondaire.
        3. Tertiaire.
    C. Non-spécifiée.
35. Gonococcie et autres maladies vénériennes.
    A. Infections gonococciques (excepté ophtalmie).
    B. Ophtalmie gonococcique.
    C. Autres maladies vénériennes.
43. Mycoses.
    A. Teignes, trichophytie et favus.
    B. Autres mycoses.
88. Maladies des organes de la vision.
    A. Conjonctivite.
    B. Kératite.
    C. Iritis.
    D. Cataracte.
    E. Rétinite.
    F. Glaucome.
    G. Autres.
115. Maladies de la cavité buccale.
    A. Maladies des dents ou des gencives.
    B. Autres.

149. Autres accidents de l'accouchement.

> Bien qu'il ne s'agisse pas de maladie, une rubrique "accouchement normal" est nécessaire pour la statistique des personnes présentes dans les hôpitaux, maternités, etc.

    A. Accouchement normal.

    B. Accidents de l'accouchement.

153. Autres maladies de la peau, de ses annexes et du tissu cellulaire.

    A. Pelade.

    B. Autres maladies.

158. Débilité congénitale.

> Bien qu'il ne s'agisse pas de malades, une rubrique "nouveau-nés sortis de l'hôpital ou de la maternité sans avoir été malades" est nécessaire pour le statistique des personnes présentes dans les hôpitaux, maternités, etc.

    A. Nourrissons sortis de l'hôpital sans avoir été malades.

    B. Débilité congénitale.

194. Autres accidents.

    A. Corps étranger.

    B. Luxation.

    C. Entorse.

    D. Fracture (sans autre indication).

    E. Plaie.

    F. Autres.

200. Causes non spécifiées ou mal définies.

    A. Causes non spécifiées ou mal définies.

    B. Surmenage.

    C. Simulation, malade en observation.

> Bien qu'il ne s'agisse pas de maladie véritable, une rubrique "simulation" est nécessaire pour la statistique des personnes ayant séjourné dans un hôpital, une maison de santé, etc.

THE OFFICIAL STATISTICAL TREATMENT OF JOINT CAUSES OF DEATH

Few persons not professional vital statisticians understand the real meaning of mortality statistics tabled under the International Classification. The official charged with compiling such statistics has to work under a set of essentially arbitrary rules. Otherwise he never could make an intelligent compilation, because of two important facts:

1. Some physicians all the time, and all physicians some of the time, will use their own terminology instead of that of the International Classification in reporting the cause of death on the original death certificate.

2. Physicians will, quite properly, report more than one morbid condition as a causal factor in the death.

What shall the vital statistician do under such premises? What he actually does do is so important for a right understanding of what official vital statistics of the present day really mean *medically*, that it seems desirable to reproduce here, in part, the excellent discussion of the matter contained in the last issued "Manual of the International List." This discussion shows the general principles according to which causes of death are handled in modern statistical offices. From time to time some slight modifications in respect of details are made. Discussions of these modifications and accounts of the procedure under the rules are embodied each year in the textual matter of the annual volumes of Mortality Statistics from the Census Bureau. Here we are only concerned with general principles.

The expression "joint causes of death" is a convenient one for those cases in which the physician reports two or more causes or conditions upon the certificate of death of an individual. According to the general practice of statistical compilation only one cause can be tabulated for each death, consequently a process of selection is necessary. The method employed for this purpose may have a very considerable influence upon the resulting statistics. Dr. Julius J. Pikler* has very forcefully directed attention to the importance of the study of contributory causes of death that usually are lost entirely in compilation, but the full statement of such causes would be difficult, especially for related tables and a detailed classification, in a report dealing with large numbers of returns.

The International Commission did not give special consideration to this subject in 1909, but at the suggestion of Dr. Bertillon it was agreed that the rules employed since 1900 should be continued in force and a special committee was appointed to report on the subject. Following are the rules in question as given in the French edition of 1903:

1. If one of the two diseases is an *immediate* and *frequent* complication of the other, the death should be classified under the head of the primary disease. Examples:

Infantile diarrhea and *convulsions*, classify as *infantile diarrhea*.
Measles and *bronchopneumonia*, classify as *measles*.
Scarlet fever and *diphtheria*, classify as *scarlet fever*.
Scarlet fever and *nephritis*, classify as *scarlet fever*.

* Das Budapester System der Todesursachenstatistik, 1909.

2. If the preceding rule is not applicable, the following should be used: If one of the diseases is *surely* fatal* and the other is of less gravity, the former should be selected as the cause of death.   Examples:

    Cancer and *bronchopneumonia*, classify as *cancer*.

    Pulmonary tuberculosis and *puerperal septicemia*, classify as *tuberculosis*.

    Icterus gravis and *pericarditis*, classify as *icterus gravis*.

3. If neither of the above rules is applicable, then the following: If one of the diseases is *epidemic* and the other is not, choose the epidemic disease.   Examples:

    Typhoid fever and *saturnism*, classify as *typhoid fever*.

    Measles and *biliary calculi*, classify as *measles*.

4. If none of the three preceding rules is applicable, the following may be used: If one of the diseases is *much more frequently fatal* than the other, then it should be selected as the cause of death.   Examples:

    Rheumatism (*without metastasis*) and *salpingitis*, classify as *salpingitis*.

    Pericarditis and *appendicitis*, classify as *pericarditis*.

5. If none of the four preceding rules applies, then the following: If one of the diseases is of *rapid development* and the other is of slow development, the disease of rapid development should be taken.   Examples:

    Diabetes and *icterus gravis*, classify as *icterus gravis*.

    Cirrhosis and *angina pectoris*, classify as *angina pectoris*.

    Pleurisy and *senile debility*, classify as *pleurisy*.

6. If none of the above five rules applies, then the diagnosis should be selected that best characterizes the case.   Example:

    Saturnism and *peritonitis*, classify as *saturnism*.

Precise diagnoses should be given the preference over vague and indeterminate ones, such as "Hemorrhage," "Encephalitis," etc.   Arbitrary decisions should be avoided as much as possible by the use of the preceding rules.   None of them is absolute, but all are subject to exceptions which may vary according to local usages.†
In practice the first rule, which is the most logical of all, is the one of most frequent application.   The others have been formulated only to prepare for all cases and to treat them with system and uniformity.

These rules differ but slightly from those given in the Manual of 1902, which were based upon the French edition of 1900.   They are a development of practical experience, as shown by the forms in which they have appeared in various editions of the International

---

* Apart from all treatment.   This provision is necessary to assure stability in the application of the rules.   Otherwise a therapeutic discovery, for example, that of the antidiphtheric serum, would modify the tables and injure the comparability of the statistics.

† Particularly we should note the impropriety of certain expressions.   For example, if a physician writes *Typhoid fever, chronic nephritis*, it is almost certain that he intended to indicate typhoid fever complicated with albuminuria and not a patient with Bright's disease attacked with typhoid fever.

When a disease ordinarily rare or absent undergoes a large extension (*e. g.*, cholera, yellow fever, etc.) the total deaths should be noted without any exception whatever. For such cases it is necessary to waive all ordinary rules.

7

Classification, and may be compared with the rules given in the introductory text of the Alphabetische Liste von Krankheiten und Todesursachen, Kaiserliches Gesundheitsamt, Germany, 1905:

When several diseases are reported as causes of death, the following rules should be observed:

1. The death is, as a rule, to be assigned to that number which represents the probable primary cause (Grundleiden). For example, when nephritis and valvular heart disease are returned, the death should be classified under the heart disease as the probable primary cause. Only when the primary cause is not a real disease may it be disregarded. For example, with "senile debility and bronchitis" or "debility and intestinal catarrh," the deaths should be classified not as senile debility or congenital debility, but as chronic bronchitis and as intestinal catarrh.

2. With two independent diseases, the more severe should be chosen.

3. With an infectious disease and a non-infectious disease, the former should be chosen. Example: Insanity and typhoid fever, classify as typhoid fever.

4. If acute diseases are reported with chronic diseases, the acute diseases are to be preferred. Example: Gastric ulcer and croupous pneumonia, classify as croupous pneumonia.

5. If two infectious diseases are reported as causes of death, then smallpox, scarlet fever, measles, typhus fever, diphtheria and croup, whooping-cough, croupous pneumonia, influenza, typhoid fever, paratyphoid fever, Weil's disease, relapsing fever, cerebrospinal fever, erysipelas, tetanus, septicemia, puerperal fever, plague, Asiatic cholera, dysentery, anthrax, glanders, rabies, and trichiniasis should have the preference over tuberculosis, malaria, or a venereal disease.

6. Causes of death from violence are usually preferred.

7. Such returns as heart weakness ["heart failure"], cardiac paralysis, paralysis of the lungs, pulmonary edema, coma, and the like, should be disregarded if other causes are named.

8. With tuberculosis of several organs, including that of the lungs, tuberculosis of the lungs should be selected.

It will be interesting also to compare the rules published by the Society of Medical Officers of Health of England*:

RULES AS TO CLASSIFICATION OF CAUSES OF DEATH

With the following exceptions the general rule should be to select from several diseases mentioned in the certificate the *disease of the longest duration*. In the event of no duration being specified, the disease standing first in order should be assumed to be the disease of longest duration.

Exceptions to the Above Rule

Any one of the *chief infective diseases* should be selected in preference to any other cause of death. If two infective diseases in succession be specified, the disease of *longer* duration should be selected.

* The New Tables Issued by the Local Government Board and the Schedules of Causes of Death issued by The Incorporated Society of Medical Officers of Health, London, 1901.

Thus scarlet fever should be selected in preference to bronchopneumonia, and phthisis in preference to bronchitis.

Definite diseases, ordinarily known as *constitutional diseases*, should have preference over those known as local diseases.

Thus cancer should be selected in preference to pneumonia, and diabetes in preference to heart disease.

When *apoplexy* occurs in conjunction with definite *disease of the heart* or *kidneys*, the heart disease or the kidney disease, as the case may be, should be preferred.

When *hemiplegia* is mentioned in connection with *embolism*, the *embolism* should be selected.

When *embolism* occurs in connection with *childbirth*, the death should be referred to *accidents of childbirth*.

In calculating the death-rate from "diarrhea," deaths certified as due to *diarrhea*, either alone or coupled with some ill-defined cause (such as "atrophy," "debility," "marasmus," "thrush," "convulsions," "teething," "old age," or "senile decay"), *epidemic* or *summer diarrhea*, *epidemic* or *zymotic enteritis*, *intestinal* or *enteric catarrh*, *gastro-intestinal* or *gastro-enteric catarrh*, *dysentery* or *dysenteric diarrhea*, *cholera* (not being "Asiatic cholera"), *cholera nostras*, *cholera infantum*, and *choleraic diarrhea* should be included.

The following miscellaneous examples are given as indicating the method of classification in cases of difficulty that frequently arise:

Causes of Death in Order Given in Death Certificate	*To be Classified Under—*
Whooping-cough, bronchopneumonia, scarlet fever.	Whooping-cough, i of longer duration than scarlet fever.
Scarlet fever six months, otitis media, abscess of brain.	Scarlet fever.
Laryngeal and pulmonary phthisis.	Phthisis.
Pneumonia, old age.	Pneumonia.
Old age, bronchitis.	Bronchitis.
Phthisis, diabetes mellitus.	Select disease of longest duration.
Diphtheria nine months, paralysis.	Diphtheria.
Puerperal perimetritis.	Puerperal fever.
Cerebral embolism.	Embolism.
Spasmodic croup.	Laryngismus stridulus.
Acute hydrocephalus.	Tubercular meningitis.
Bronchitis, phthisis.	Phthisis.

Through the kindness of Dr. John Tatham, formerly Medical Superintendent of the Registrar-General's office, England, a copy of the Instructions to Abstractors, as employed in that office in 1909, was supplied to the Bureau of the Census. Certain decisions of special interest are taken therefrom:

1. Any general disease (except pyrexia, premature birth, congenital defects, want of breast milk, teething, and chronic rheumatism) to be taken in preference to any local disease except aneurysm and strangulated hernia.

2. Any of the following diseases are to be given preference over any other diseases: Aneurysm, anthrax, Asiatic cholera, cancer, carcinoma, glanders, rabies, industrial poisoning, malignant disease, opium or morphin habit, puerperal septic disease, sarcoma, smallpox, strangulated hernia, tetanus, and vaccination.

3. Any disease in this group is to be preferred over any other disease except those named in the preceding group: Cerebrospinal fever, diphtheria, dysentery, typhoid fever, German measles, malaria, measles, mumps, relapsing fever, scarlet fever, typhus fever, and whooping-cough.

4. The following diseases to be preferred except for those named in the two preceding lists: Acute hydrocephalus, alcoholism, influenza, lupus, phthisis, pulmonary tuberculosis, rheumatic fever (acute and subacute rheumatism), scrofula, syphilis, tabes mesenterica, tuberculous meningitis, tuberculous peritonitis, tuberculosis of other organs, and general tuberculosis.

5. For the following list, prefer the disease of longer duration or the disease first written: Carbuncle (not anthrax), diabetes mellitus, epidemic diarrhea, epidemic enteritis, enteritis, diarrhea due to food, erysipelas, gout, hemophilia, infective endocarditis, infective enteritis, pernicious anemia, phagedena, phlegmon (not anthrax), pneumonia (all forms), purpura hæmorrhagica, pyemia (not puerperal), rheumatoid arthritis, rheumatic gout, rheumatism of heart, rickets, scurvy, septicemia, other septic diseases, septic infections, starvation, and varicella.

6. Premature birth and congenital defects (malformations) to be preferred for decedents under three months of age to other causes except those of groups 2 and 3.

7. Chlorosis and anemia (not pernicious) only when alone.

8. For combinations of local diseases, usually select disease of longer duration or that first written.

9. Any definite disease accelerated by violence is to be classed to the disease.

10. Tetanus, septicemia, blood-poisoning, pyemia, or erysipelas following violence to be classed to tetanus or the septic disease if the injury is slight; but if severe enough to kill by itself, the death should be classed to the form of violence.

For returns upon the Standard Certificate of Death, and especially for those returns in which the instructions have been regarded by the reporting physicians, the following suggestions are made by the United States Bureau of the Census:

1. Select the primary cause, that is, the real or underlying *cause of death*. This is usually—
   (a) The cause first in order.
   (b) The cause of longer duration. If the physician writes the cause of shorter duration first, inquiry may be made whether it is not a mere symptom, complication, or terminal condition.
   (c) The cause of which the contributory (secondary) cause is a frequent complication.
   (d) The physician may indicate the relation of the causes by words, although this is a departure from the way in which the blank was intended to be filled out. For example, "Bronchopneumonia *following* measles" (primary cause last) or "Measles *followed by* bronchopneumonia" (primary cause first).

2. If the relation of primary and secondary is not clear, prefer general diseases, and especially dangerous infective or epidemic diseases, to local diseases.

3. Prefer severe or usually fatal diseases to mild diseases.

4. Disregard ill-defined causes (Class XVIII), and also indefinite and ill-defined terms (*e. g.*, "debility," "atrophy") in Classes XIV and XV that are referred, for certain ages, to Class XVIII, as compared with definite causes. Neglect mere modes of death (failure of heart or respiration) and terminal symptoms or conditions (*e. g.*, hypostatic congestion of lungs).

5. Select homicide and suicide in preference to any consequences, and severe accidental injuries, sufficient in themselves to cause death, to all ordinary consequences. Tetanus is preferred to any accidental injury, and erysipelas, septicemia, pyemia, peritonitis, etc., are preferred to less serious accidental injuries. Prefer definite means of accidental injury (*e. g.*, railway accident, explosion in coal mine, etc.) to vague statements or statement of the nature of the injury only (*e. g.*, accident, fracture of skull).

6. Physical diseases (*e. g.*, tuberculosis of lungs, diabetes) are preferred to mental diseases as causes of death (*e. g.*, manic depressive psychosis), but general paralysis of the insane is a preferred term.

7. Prefer puerperal causes except when a serious disease (*e. g.*, cancer, chronic Bright's disease) was the independent cause.

8. Disregard indefinite terms and titles generally in favor of definite terms and titles. The precise line of demarcation is difficult to lay down, but may be indicated broadly by the kinds of type employed in the International List in the form distributed by the Census to all physicians in the United States.*

From these suggestions and from the instructions employed in various offices it will be apparent that there is a considerable factor of uncertainty in the results when a large proportion of joint causes is involved. No rules yet formulated will insure absolutely identical compilations from the same material, and the methods employed in the same office may vary from year to year. The most efficient editor is not the one who follows any set of listed arbitrary decisions, but rather the one who is constantly on the lookout for cases in which it should not be followed, and who calls attention to such cases. A list of this kind cannot incorporate considerations of duration, sex, place of death, age, occupation, etc., any or all of which may have an important bearing upon the classification of deaths, and in individual cases such data on transcripts often indicate an assignment contrary to the listed one.

The whole subject of joint causes is a difficult one, and there is still no international agreement about the matter. The International Commission, however, refuses to recommend the formation

* See Physicians' Pocket Reference to the International List of Causes of Death.

of any general, uniform, international code for joint causes. This position is based upon arguments which seem to many competent vital statisticians singularly narrow and lacking in any deep understanding of the basic philosophy and psychology underlying the recording and tabulating of the causes of death. It is generally admitted that the work of the United States Census Bureau is the most advanced, on this matter, of any country in the world. And yet, as the late Dr. William H. Davis has said, "the treatment of joint causes of death has never been adequately discussed by international conferences, nor been adequately treated by anybody."

RELIABILITY OF STATISTICS OF SEPARATE CAUSES OF DEATH

Philosophically considered a true determination of the "cause of death" is in a great many cases, indeed the majority probably of all cases, an extraordinarily difficult matter. This every pathologic anatomist knows. The difficulty arises from many different circumstances. Some illustrations will perhaps make the point clear. A woman has cancer of the breast, is operated upon in the hope of curing this disease, develops a postoperative pneumonia, and dies. Now if the women had not had the *cancer* and had therefore not been operated on for its relief, this train of circumstances would not have got under way. This way of looking at the matter plainly suggests that the cancer is fundamentally the cause of this death. But, on the other hand, if she had not been operated on, even though she still had the cancer, she would not have died *when* she did, but at some later time. This view rather tends to make the *operation* the cause of death, at least at the particular time and place at which it occurred. Again, suppose she had been operated on, and had *not* developed the postoperative pneumonia. Then she might have been permanently cured of the cancer (some are) and lived to a ripe old age. This view of the case truly makes the *pneumonia* the cause of death. Which of the three things— cancer, operation, or pneumonia—is to be charged as the primary cause of death plainly depends upon the point of view, or, put in another way, upon what definitions or rules are set up as to what shall be called the cause of death.

As has already been shown, official vital statistics operate under

such a set of rules. In the case cited, cancer would be given as the primary cause of death, and the postoperative pneumonia as the secondary or complicating cause. To the philosophic mind this is probably the least satisfactory solution of the three. Why it is the officially chosen one is because of an often overlooked, and in some of its aspects quite vicious, underlying concept in official vital statistics. *There is ever present in vital statistics, and from the beginning always has been, an attempt to make the incidence of mortality a measure or index of the incidence of morbidity.* Mortality is not and never can be a good index of morbidity, generally speaking. What actually is done is to weaken and impair the value of the statistics for the study of *mortality* in the hope to make them a little better indices of *morbidity*. This tendency is apparent in the illustration given above. It is thought desirable to get as complete records as possible of the *prevalence* of cancer in the population, as a disease. Therefore, the rule is that, in general, if a person dies who is known to have had cancer prior to death, the death is to be charged to cancer. In consequence, it results that no one can get from the official statistics an accurate answer to the question: "How many persons per 1000 living did cancer kill in 1920?" Instead, what he gets is information as to how many persons died per 1000 living in 1920, who had had cancer before they died, assuming that the diagnosis is correctly made in every case. The latter information, as anyone with a logical mind will at once perceive, is quite different from the former.

Now if all secondary and complicating conditions were accurately reported and compiled, the case would be far better in respect of the objection just discussed. But this is an unattainable counsel of perfection. Even if it were accomplished there would still remain a large source of error in statistics of the causes of death. This arises from the fact that all physicians are not equally intelligent or clever diagnosticians. Clinical diagnosis is not yet an exact science. A person dies: the attending physician quite honestly thinks he knows what this patient died of, and registers his conviction on the death certificate. Actually, the physician may have been mistaken in his diagnosis, too often grossly so. But his error gets embalmed in the official vital statistics.

This phase of the problem has been the subject of careful study by a committee of the American Public Health Association.[5] Every student of vital statistics should study and ponder over this committee's report. He will be bound to reach the conclusion that there are but few indeed of the rubrics of the International List whose figures can be unreservedly accepted at their face value.

The following classes of official vital statistics alone can, in the writer's opinion, be subjected to analysis as *scientifically accurate* records of natural phenomena:

1. Deaths from all causes (either for all ages together or for separate age groups, as, for example, "infant mortality" (deaths under one year of age).

2. Traumatism (Rubrics 178, 180 to 188 inclusive, 192 to 194 inclusive, and 196 to 198 inclusive).

3. Homicide (Rubrics 172 to 175 inclusive).

This is neither a long nor, except in its first item, a specially important list. But when we deal with other rubrics we are dealing with mixtures of unknown composition, and with data of a wholly different order of accuracy than those, for example, of the physicist or the chemist. We are forced, of course, in the practical conduct of a statistical business to deal with other rubrics, but, at any rate, one should, when so doing, always remember that the material is fundamentally of a dubious character. In the discussion of this point in an earlier edition of this book suicide was included as a fourth rubric in the short list above. It cannot withstand criticism, however, as undoubtedly the fact of suicide is sometimes concealed by surviving members of the family; how often no one knows. A similar objection may be made about homicide. But probably in case of both suicide and homicide the error in the returns so produced is statistically negligible.

Professor Haven Emerson, of Columbia University, who is one of the foremost authorities in the world on the accuracy of certified causes of death, suggests (*in litt.*) that

"certain rubrics of the International List may be considered reasonably accurate if they deal with neoplasms, lesions, conditions verifiable by direct observation of the surface of the body or its interior where accessible by inspection through body orifices or where operative procedures have made the interior tissues visible by direct inspec-

tion, or where there is a specific test or organism determining the cause of death, as the diphtheria bacillus, the Plasmodium malariæ, lead or alcohol poisoning."

This is undoubtedly true where the interest and painstaking care of the physician in filling out death certificates can be counted on. But the testimony of registrars is that these qualities are by no means universal. Dr. Emerson's suggestion defines a set of conditions which *permit* the reasonably accurate recording of the cause of death, but they do not *compel* it, and these would seem to be two different things.

There are still other factors in the case, as Dr. Emerson goes on to point out:

"Where, as in France, the report is made by the head of the household for purposes of the *état civil*, there is a wider error than where, as in Switzerland, there is a separation of the factor identity of a death from the pathological report by a physician of the cause of death.

"As I recall it the late Dr. Ney of Switzerland found that the installation of the present system where the confidential character of the certificate of cause of death was scrupulously maintained, and thus the physician was protected against civil damage suits by the family of the deceased, resulted in an increase in some Cantons of from 50 to 70 per cent. in the certification of syphilis as a cause of death.

"I believe the Swiss system has been for the past ten to fifteen years superior in principle and practice to that of any other country."

At this point the discussion of the interesting problem of the accuracy of the statistical recording of causes of death must be dropped, because of considerations of space. It should be emphasized, in conclusion, that there is no subject of greater importance in the whole field of vital statistics, and there is no subject on which further research is more needed and will be more permanently profitable and useful.

SUGGESTED READING

1. Rossiter, W. S.: A Century of Population Growth, Washington (Bureau of the Census), 1909. (For discussion of development of census methods.)
2. Hooker, R. H.: Modes of Census Taking in the British Dominions, Jour. Roy. Stat. Soc., vol. 57, pp. 298–358, 1894.
3. Manual of the International List of Causes of Death. Based on the Second Decennial Revision by the International Commission, Paris, July 1 to 3, 1909, Washington (Bureau of the Census), 1911. (Every student of vital statistics should thoroughly study this or later manuals. It alone really gives an understanding of the basic content of official vital statistics.)

4. Pearl, R., and Bacon, A. L.: Biometrical Studies in Pathology. I. The Quantitative Relations of Certain Viscera in Tuberculosis, The Johns Hopkins Hospital Reports, vol. 21, Fasc. III, pp. 157–230, 1922. (This paper illustrates in the specific case of tuberculosis the inaccuracy of recorded causes of death. See particularly pp. 211–226.)

5. The Accuracy of Certified Causes of Death, Public Health Reports, vol. 32, pp. 1557–1632, September 28, 1917.

6. Pearl, R.: The Biology of Death, Philadelphia (J. B. Lippincott Co.), 1922. (Chaps. IV and V.)

7. Pearl, R.: Studies in Human Biology, Baltimore (Williams & Wilkins Co.), 1924. (Chaps. V, VII, and XI.) (The readings suggested under 6 and 7 give some specific examples of the constructive uses of the International List of Causes of Death in attempts to solve problems of human biology.)

8. Dudfield, R.: A Critical Examination of the Methods of Recording and Publishing Statistical Data Bearing on Public Health; and Suggestions for the Improvement of Such Methods, Jour. Roy. Stat. Soc., vol. 68, pp. 1–40, 1905. (This paper gives much valuable information about English registration methods, the definition of the various areas used in English official statistics, etc. The non-English reader will find it very useful in gaining a correct idea of just what is the content of English statistics.)

9. League of Nations Health Organization: Statistical Handbooks Series. (Prepared by Prof. Major Greenwood, F. R. S., F. R. C. P., and Major P. Granville Edge, with the exception of 12, which was prepared by Drs. Marcel Ney and H. Carrière.)

    (1) The Official Vital Statistics of the Kingdom of the Netherlands, 1924, 77 pp.

    (2) The Official Vital Statistics of the Kingdom of Belgium, 1924, 84 pp.

    (3) The Official Vital Statistics of England and Wales, 1925, 115 pp.

    (4) The Official Vital Statistics of the Kingdom of Spain, 1925, 59 pp.

    (5) The Official Vital Statistics of the Austrian Republic, 1925, 68 pp.

    (6) The Official Vital Statistics of the Scandinavian Countries and the Baltic Republics, 1926, 107 pp.

    (7) The Official Vital Statistics of the Republic of Portugal, 1926, 59 pp.

    (8) The Official Vital Statistics of the Republic of Czechoslovakia, 1927, 71 pp.

    (9) The Official Vital Statistics of the French Republic, 1927, 115 pp.

    (10) The Official Vital Statistics of the Kingdom of Hungary, 1927, 78 pp.

    (11) The Official Vital Statistics of Ireland, The Irish Free State and Northern Ireland, 1929, 112 pp.

    (12) The Official Vital Statistics of Switzerland, 1928, 88 pp.

    (13) The Official Vital Statistics of the Kingdom of Scotland, 1929.

    (These treatises are invaluable to the student of vital statistics. They give, in great detail and with the most painstaking, critical care, the statistical procedures in the different countries. From them the student can form a judgment of the *meaning* of the vital statistics in the different countries. It is greatly to be hoped that the series will be continued until all the countries of the world have been included.)

CHAPTER IV

TABULAR PRESENTATION OF STATISTICAL DATA

THE raw material of statistics consists of individual observations of phenomena. The simplest way to tabulate such material is, of course, to make a *list* of the observations, in which each single one constitutes an item of the table. But this can scarcely be called tabulation, because it does not perform the essential function of that operation.

The purpose of tabulation is so to arrange observations that like cases shall be put together and their frequency of occurrence in the whole group thus be made apparent.

The degree of likeness of the cases to be put together may be defined quantitatively in any way one likes. For example, it may be decided for purposes of tabulation to call all men whose stature falls anywhere between 65.00 and 65.99 inches, *alike* in stature, and put them in the same class. Evidently, then, the first necessary step in tabulating observations after they have been collected is to *classify* them, quantitatively if possible.

DICHOTOMOUS CLASSIFICATION

Logically considered, *classification is the process of partitioning a universe into mutually exclusive categories or compartments.* The number of such compartments may be anything from two up. If it is exactly two, the classification is called *dichotomous*. This is the alternative category type of classification. At the moment of this writing:

Every living person in the world $\begin{cases} \text{Either } has \text{ smallpox} \\ \text{Or } does \ not \text{ have smallpox} \end{cases}$

So then it is possible to put every person into his proper compartment relative to this classification.

But this process can be continued indefinitely:

$$\text{Every living person either} \begin{cases} \text{Has smallpox and} \begin{cases} \text{Has a fever and} \begin{cases} \text{Has an automobile, } n_1 \\ \text{or} \\ \text{Has no automobile, } n_2 \end{cases} \\ \text{or} \\ \text{Has no fever and} \begin{cases} \text{Has an automobile, } n_3 \\ \text{or} \\ \text{Has no automobile, } n_4 \end{cases} \end{cases} \\ \text{or} \\ \text{Does not have smallpox and} \begin{cases} \text{Has a fever and} \begin{cases} \text{Has an automobile, } n_5 \\ \text{or} \\ \text{Has no automobile, } n_6 \end{cases} \\ \text{or} \\ \text{Has no fever and} \begin{cases} \text{Has an automobile, } n_7 \\ \text{or} \\ \text{Has no automobile, } n_8 \end{cases} \end{cases} \end{cases}$$

If at the end of such a process of dichotomizing the number of cases in each of the final classes be counted, we shall have the frequency of occurrence of individuals alike in the respects indicated by the line of the classification back to the start. Thus in the example given above we may contrast the n_1 persons in the condition of having smallpox, *and* fever, *and* an automobile, with the n_8 individuals who have wholly escaped this concatenation of disasters.

An example of a table of this sort is presented as Table 2. It is based upon data collected to determine the incidence of influenza among tuberculous and non-tuberculous persons in the same family during the influenza pandemic of 1918 (cf. Pearl[3]).

TABLE 2

SHOWING THE INCIDENCE OF INFLUENZA AMONG TUBERCULOUS AND NON-TUBER-CULOUS WHITE INDIVIDUALS, ARRANGED BY PRESENCE OR ABSENCE OF OTHER CASES OF INFLUENZA

Tuberculous, 2375.				Not tuberculous, 8820.			
Influenza, 595.		No influenza, 1780.		Influenza, 1971.		No influenza, 6849.	
Other cases in household, 460	No other cases in household, 135	Other cases in household, 533	No other cases in household, 1247	Other cases in household, 1788	No other cases in household, 183	Other cases in household, 2568	No other cases in household, 4281

From Table 2 we note that of the 2375 tuberculous persons, 595, or 25 per cent., had influenza, while 1780, or 75 per cent., did not have this disease during the epidemic. Of the 8820 non-tuberculous individuals living in the same households as the tuberculous, 1971, or 22.3 per cent., had influenza, and 6849, or 77.7 per

cent., did not have it.   It therefore appears that, under the same environmental conditions of living, only 2.7 per cent. more of the tuberculous individuals than of the non-tuberculous contracted influenza during the epidemic.

Of the 595 tuberculous persons who had influenza, 460, or 77.3 per cent., were in households where at least one other person also had influenza during the epidemic. Of the 1971 non-tuberculous persons who had influenza, on the other hand, 1788, or 90.7 per cent., were in households where at least one other person also had influenza. Or, in other words, 22.7 per cent. of the tuberculous who had influenza were the only cases of the latter disease in their households, while only 9.3 per cent. of the non-tuberculous who had influenza were the sole cases in the household.

Of 1780 tuberculous persons who did not have influenza during the epidemic, only 533, or 29.9 per cent., were exposed to influenza infection in the household, whereas of the 6849 non-tuberculous persons who did not have influenza, 2568, or 37.5 per cent., were exposed to infection within the household.

These examples will suffice to show how a simple dichotomous statistical table is to be read.

Now instead of dividing the residual universe into just two parts each time we may equally well divide it into a number of parts.   This leads to some sort of *linear* classification.

Such a linear classification and tabulation based thereon may be combined terminally with a preceding dichotomous table, and this often furnishes a useful form of statistical tabulation. An example is given in Table 3, which is an expansion of Table 2. The linear classification in this case is relative to the number of persons in the household, proceeding from 1 to 15, down the first or left-hand column of the table.

It will be noted at once that this expansion by size of household throws interesting and significant light upon the results stated above from the more meager distributions of Table 2. The manner in which this is accomplished may be left to the reader to work out for himself as a useful exercise in getting familiar with the reading of statistics.

The linear classification of Table 3, by number of persons in

TABLE 3

SHOWING THE INCIDENCE OF INFLUENZA AMONG TUBERCULOUS AND NON-TUBERCULOUS WHITE INDIVIDUALS, ARRANGED (*A*) BY NUMBER OF PERSONS IN HOUSEHOLD, AND (*B*) BY PRESENCE OR ABSENCE OF OTHER CASES OF INFLUENZA

Number in household.	Tuberculous.				Not tuberculous.			
	Influenza.		No influenza.		Influenza.		No influenza.	
	Other cases in household.	No other cases in household.	Other cases in household.	No other cases in household.	Other cases in household.	No other cases in household.	Other cases in household.	No other cases in household.
1......	..	..	..	14	..	..	..	..
2......	4	10	12	108	4	15	7	100
3......	46	39	38	161	76	22	118	292
4......	72	28	81	255	168	37	243	696
5......	89	27	78	221	262	21	363	749
6......	73	16	96	210	303	29	419	822
7......	71	9	83	123	358	18	480	636
8......	51	2	68	82	257	24	414	446
9......	22	3	40	33	117	8	188	246
10......	18	1	16	20	114	3	138	170
11......	8	..	12	12	49	5	91	43
12......	3	..	5	5	36	1	63	43
13......	2	..	3	2	32	..	28	24
14......	..	..	..	..	..	..	..	..
15......	1	..	1	1	12	..	16	14
Totals.	460	135	533	1247	1788	183	2568	4281
	595		1780		1971		6849	
	2375				8820			

the household, presents no problem for decision as to where each case belongs in making the table. There are no fractional components of a household. Each will have 2, or 4, or some other definite and simple whole number of individuals in it. When, however, we deal with things which are *measured*, instead of *counted*, a new element enters the tabulating situation. This may be illustrated by Table 4, which is a simple statistical table based upon a linear classification.

In Table 4 the observed systolic blood-pressures are divided into seven mutually exclusive *classes*. Each class includes an elemental range of 20 mm. pressure. This classification says that systolic pressures of between say 130 and 150 mm. are to be regarded for practical purposes as alike. The correct way to state class

TABLE 4

FREQUENCY DISTRIBUTION OF SYSTOLIC BLOOD-PRESSURES IN 102 MEN AGED SEVENTY-
FIVE AND OVER.   (From Thompson and Todd, Lancet, 1922, II, 503.)

Systolic pressure (mm. Hg).	Absolute frequency.
110–129	18
130–149	31
150–169	23
170–189	20
190–209	7
210–229	1
230–249	2
Total	102

limits in setting up a frequency table is that followed in Table 4. The class range 110–129 means theoretically that all pressures are included which are equal to or *greater* than 110.0000 . . . and are equal to or *less* than 129.9999. . . .

In grouping observations in this way we are doing essentially the same thing that is done in measuring when the graduations on the measuring scale have a defined degree of coarseness. Suppose, for example, each one of a group of men to be measured as to height with a stick graduated only to inches. Some few men in the group will be of a height which exactly coincides with one of the inch markings on the stick and their height will be that exact number of inches. More of them, however, will have a height falling somewhere in between two consecutive inch marks on the stick. Say there are four men whose height falls between the 72-inch and 73-inch mark. These four men differ from each other in height by less than an inch. If we have agreed at the start to measure only to the fineness of 1 inch, this is equivalent to saying that we propose to regard individuals differing from each other by less than an inch, as being of the same height.

Scales can be read in two different ways. Thus an individual whose *actual* height is 72.2 inches may be said to be, on the basis of measuring with a scale divided into inches only:

either *a*, more than 72 inches in height but less than 73 inches;

or *b*, nearer 72 inches than 73 inches in height.

In practical statistical work it makes some difference which of these methods of recording scale readings is adopted. A group of individuals in our statistics recorded as 72 inches in height according

to the first method (*a*) of reading and recording will include individuals ranging in height between 72.0000 . . . 01 inches and 72.9999 . . . 9 inches. On this method of reading the central point of the group will be, to a practical degree of approximation, 72.5 inches. But a group of individuals recorded as 72 inches in height, on the second (*b*) method of reading and recording, will include individuals ranging in height from 71.50000 . . . 01 inches and 72.4999 . . . 9 inches. The central point of the group, read and recorded in this way, will be, again to a practical degree of approximation, 72 inches.

In vital statistics this point is perhaps of greatest practical importance relative to the recording of age. Here method (*a*) records "age at last birthday," while method (*b*) records "age at nearest birthday." After a somewhat vacillating policy in the past, official vital statisticians (for example, the Census Bureau) have now adopted method (*a*) as a definite policy in their work. This obviously makes for greater accuracy. If a person at a given moment is near the half way point between two consecutive birthdays it is not easy, without careful figuring, to say which of the two is *nearer*. But if he knows his age at all he can instantly say what it was at his *last* birthday. When really accurate work with statistical data is attempted it is always necessary to be sure whether method (*a*) or method (*b*) was used in the original records.

DOUBLE DICHOTOMOUS TABLES

The principle of dichotomous classification, with expansion of terminal classes linearly, may be applied to both sides of a table. There will then result what may be called a *double dichotomous table*, which is one of the most useful forms of tabulation for raw, basic statistical data. Why it is so is because it permits the greatest freedom and variety in the subsequent constructive and derivative use of the material.

Table 5 is a simple example of a double dichotomous table. This table presents certain information derived from the autopsy protocols of 358 persons found at autopsy in the Johns Hopkins Hospital to have miliary tuberculosis of some organ or organs of the body. There are $8 \times 12 = 96$ elemental cells in this table.

TABLE 5

Original Data on Color, Sex, Age, and Location of Lesions of 358 Persons Found at Autopsy to Have Miliary Tuberculosis

Tuberculous lesions — lungs	heart	kidneys	White Males Under 20	White Males 20 to 49	White Males 50 and over	White Females Under 20	White Females 20 to 49	White Females 50 and over	Colored Males Under 20	Colored Males 20 to 49	Colored Males 50 and over	Colored Females Under 20	Colored Females 20 to 49	Colored Females 50 and over	Totals
Not present in lungs	Not present in heart	Not present in kidneys	1	3	..	1	..	..	12	6	1	4	3	2	33
Not present in lungs	Not present in heart	Present in kidneys	1	1	1	..	1	..	6	8	3	1	..	..	22
Not present in lungs	Present in heart	Not present in kidneys	..	1	..	..	2	..	..	1	..	..	..	..	4
Not present in lungs	Present in heart	Present in kidneys	..	..	..	..	..	..	..	..	..	..	..	..	0
Present in lungs	Not present in heart	Not present in kidneys	8	25	7	16	9	..	16	38	6	32	13	4	174
Present in lungs	Not present in heart	Present in kidneys	4	17	5	7	5	1	7	28	6	12	3	..	95
Present in lungs	Present in heart	Not present in kidneys	..	2	..	..	..	..	3	5	..	2	..	..	12
Present in lungs	Present in heart	Present in kidneys	..	..	2	1	2	..	2	6	3	1	1	..	18
Totals			14	49	15	25	19	1	46	92	19	52	20	6	

Subtotals (Totals section):

Group	Not present in heart	Present in heart	Lung subtotal
Not present in lungs	55	4	59
Present in lungs	269	30	299

Column totals: White Males 78; White Females 45; White 123. Colored Males 157; Colored Females 78; Colored 235.

Grand Total 358

Each cell tells the number (*i. e.*, the *frequency*) of individuals in the total universe of 358 who were alike in the following respects:

1. Color.
2. Sex.
3. Age (in broad classes).
4. Presence (or absence) of tuberculous lesions in lungs.
5. Presence (or absence) of tuberculous lesions in heart.
6. Presence (or absence) of tuberculous lesions in kidneys.

Furthermore, the frequency of every possible *combination* of these categories is stated in Table 5.

This table will repay careful and detailed study from the standpoint of statistical methodology.   First, let us see by some examples how it is to be read.

(*Single cell reading*).   There was 1 colored male with miliary tuberculosis, falling in the age class twenty to forty-nine years, who had no tuberculous lesions in either kidneys or lungs, but did have a tuberculous lesion of the heart.

(*Primary subtotal reading*).   There were 15 white males aged fifty or over among the 358 persons who had miliary tuberculosis.

(*Secondary subtotal reading*).   There were but 4 persons, in the 358 who had miliary tuberculosis, who had a tuberculous lesion of the heart, but at the same time lacked any such lesion of the lungs.

(*Tertiary subtotal reading*).   There were 123 white and 235 colored persons in this experience of miliary tuberculosis.

It is obvious that this form of table may be expanded to any desired degree.   The double dichotomous type of table leads up to and exemplifies the theoretical *ideals* of statistical tabulation. These ideals always to be kept in mind in tabulating raw statistical data as a matter for reference and possible future synthetic or derivative use are:

1. Make the information in each cell *exclusive* relative to as many different categories as is possible, while still conforming to the ideal of

2. Making a *tabulation*, not a mere list.

The first of these ideals perhaps needs a little further illustration to make its meaning entirely clear.   The records of the Baltimore

Health Department for 1917 show that in that year there died 223 bookkeepers and clerks and 124 drivers and hostlers.

The same records also show that in the same year there died 1213 persons of tuberculosis of the lungs.

But it is impossible to determine from the records how many of the bookkeepers or of the hostlers died of tuberculosis of the lungs. Some part surely of the 223 bookkeepers and the 124 drivers and hostlers had tuberculosis. Why it is impossible from the published tabulations to find out how many were in this part, is that the elemental cells of each of the published tables are too *inclusive*. Two hundred and twenty-three and 124 are elemental cell frequencies of the published table of deaths by occupations, and 1213 is an elemental cell frequency in the published table of deaths by causes. But the 223 persons of the first mentioned cell are *alike in only one respect*, namely, that they were all either clerks or bookkeepers. They *included* males and females, whites and colored, persons dying of tuberculosis, cancer, etc. In short, the information is *exclusive* relative only to one single category. This may be satisfactory or desirable in derivative tables of constants and the like, but it is eminently unsatisfactory in original tables of the raw statistical material.

CORRELATION TABLES

A table of double entry in which the condition or status of each individual is entered relative to two characteristics, or attributes simultaneously is called a *correlation table*. This type of table is one of the most important in statistical work. An example of a correlation table is shown in Table 6. This table correlates the relative cell volume of the blood (volume of corpuscles as percentage of total volume) with body-weight, in 449 males having active tuberculosis.*

As an illustration of the manner in which correlation tables are to be read, it is seen from Table 6 that, in this experience, there were 20 males whose body weight fell somewhere within the scale range

* Pearl, R., and Miner, J. R.: A Biometric Study of the Relative Cell Volume of Human Blood in Normal and Tuberculous Males, Bull. Johns Hopkins Hosp., vol. 40, pp. 3–32, 1927. Table on p. 26.

TABLE 6

CORRELATION OF RELATIVE CELL VOLUME WITH BODY-WEIGHT AMONG ACTIVELY TUBERCULOUS MALES

BODY-WEIGHT	RELATIVE CELL VOLUME														Totals
	25.5–27.4 per cent	27.5–29.4 per cent	29.5–31.4 per cent	31.5–33.4 per cent	33.5–35.4 per cent	35.5–37.4 per cent	37.5–39.4 per cent	39.5–41.4 per cent	41.5–43.4 per cent	43.5–45.4 per cent	45.5–47.4 per cent	47.5–49.4 per cent	49.5–51.4 per cent	51.5–53.4 per cent	
pounds															
89.5– 99.4	–	1	–	–	1	–	3	1	–	–	–	–	–	–	6
99.5–109.4	1	1	1	1	–	3	4	2	4	3	–	–	–	1	21
109.5–119.4	–	–	–	3	–	5	4	7	8	4	3	3	1	2	40
119.5–129.4	2	–	–	1	2	6	8	10	10	16	15	7	5	2	84
129.5–139.4	–	–	2	3	2	7	6	9	20	19	15	14	6	3	106
139.5–149.4	–	–	–	–	1	–	3	13	17	22	22	10	6	1	95
149.5–159.4	–	–	–	2	1	–	2	3	4	9	17	3	4	1	46
159.5–169.4	–	–	–	1	1	1	–	2	8	3	5	4	1	–	26
169.5–179.4	–	–	–	–	–	1	–	1	–	–	4	1	4	–	11
179.5–189.4	–	–	–	–	–	1	–	–	2	2	3	1	1	1	11
189.5–199.4	–	–	–	–	–	–	–	–	–	–	–	–	1	1	2
199.5–209.4	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
209.5–219.4	–	–	–	–	–	–	–	–	1	–	–	–	–	–	1
Totals......	3	2	3	11	8	24	30	48	74	78	84	43	29	12	449

extending from 129.5 pounds to just under 139.5 pounds (denoted in the left marginal rubrics as 129.5–139.4); and these same 20 males exhibited relative cell volumes of their blood falling within the scale range extending from 41.5 to just under 43.5 per cent. (denoted in the marginal rubrics along the top as 41.5–43.4 per cent.).

ARRANGEMENT OF STATISTICAL TABLES

Much of the cogency and force of statistical tables, otherwise correct, depends upon their *arrangement*. This is a subject about which it is difficult, if not wholly impossible, to state general principles, yet in no other respect is it easier to distinguish the performance of the experienced professional statistician from that of the amateur. One may say: "Make a clear, concise, easily read table, which bears directly upon the subject under discussion, and upon no other subject," but obviously this counsel is rich in why-ness and poor in how-ness. Perhaps an illustration may be helpful.

In the excellent paper by Dr. Huntington Williams on "Epidemic Jaundice in New York State, 1921–1922,"* the table here reproduced as Table 7 appears.

Now let us examine the first purpose of this table. It is stated in the original that: "Each of eighteen common symptoms is recorded in Table 1 (Table 7 here) for every case in the series of 700 that were studied. Symptoms are reported [on the physician's original case reports presumably] positive, negative, or not re-

TABLE 7

ORIGINAL FORM OF TABLE ON SYMPTOMATOLOGY OF EPIDEMIC JAUNDICE

Symptom.	Cases positive.		Cases negative.		Not recorded.	
	Number.	Per cent.	Number.	Per cent.	Number.	Per cent.
Jaundice...............	647	92.4	11	1.6	42	6.0
Anorexia...............	574	82.0	68	9.7	58	8.3
Nausea................	619	88.4	46	6.6	35	5.0
Vomiting...............	503	71.9	169	24.1	28	4.0
Headache..............	488	69.7	139	19.9	73	10.4
Constipation...........	463	66.1	110	15.7	127	18.2
Prostration............	211	30.1	81	11.6	408	58.3
Clay-colored stools.......	558	79.7	46	6.6	96	13.7
Bile-stained urine........	617	88.2	10	1.4	73	10.4
Abdominal pain.........	417	59.6	211	30.1	72	10.3
Fever..................	524	74.9	105	15.0	71	10.1
Chills.................	334	47.7	293	41.9	73	10.4
Limb pains.............	235	33.6	297	42.4	168	24.0
Diarrhea...............	106	15.2	442	63.1	152	21.7
Conjunctival congestion...	66	9.4	103	14.7	531	75.9
Epistaxis...............	61	8.7	525	75.0	114	16.3
Herpes................	28	4.0	536	76.6	136	19.4
Hiccup................	98	14.0	478	68.3	124	17.7
Unusual prevalence of rats on premises............	167	23.9	262	37.4	271	38.7

corded." Now, plainly, the purpose of the tabulation is to show the relative and absolute frequency of each of the symptoms taken by itself. But, plainly, "not recorded" furnishes no information about symptoms. It only tells the reader that no record was made of symptoms. Hence its inclusion in a table which only purports to tell us about *symptoms* is superfluous and wholly beside the point. But since the "not recorded" cases are included in the percentages

* Jour. Amer. Med. Assoc., vol. 80, pp. 532–534, 1923.

(which add to 100 across the table, and therefore include the whole of each universe), the percentages defeat the main purpose of the table, which is to inform us as to which symptoms are relatively most frequent.   Furthermore, even if this difficulty were corrected, we should still have to search laboriously down the list to find which was the most frequent symptom, the next most frequent, and so on, owing to the fact that no attention is paid to the order of arrangement of the symptoms.

TABLE 8

SHOWING THE ABSOLUTE AND RELATIVE FREQUENCY OF OCCURRENCE OF DIFFERENT SYMPTOMS IN SO MANY OF 700 CASES OF EPIDEMIC JAUNDICE AS FURNISHED DEFINITE RECORDS OF PRESENCE OR ABSENCE OF EACH OF THE INDICATED SYMPTOMS

Order.	Symptom.	Symptom present.		Symptom absent.		Total cases with any record about this symptom.
		No.	Per cent.	No.	Per cent.	
1	Jaundice..................	647	98	11	2	658
2	Bile-stained urine..........	617	98	10	2	627
3	Nausea...................	619	93	46	7	665
4	Clay-colored stools.........	558	92	46	8	604
5	Anorexia.................	574	89	68	11	642
6	Fever....................	524	83	105	17	629
7	Constipation..............	463	81	110	19	573
8	Headache.................	488	78	139	22	627
9	Vomiting.................	503	75	169	25	672
10	Prostration...............	211	72	81	28	292
11	Abdominal pain............	417	66	211	34	628
12	Chills....................	334	53	293	47	627
13	Limb pains...............	235	44	297	56	532
14	Conjunctival congestion.....	66	39	103	61	169
15	Unusual prevalence of rats on premises..............	167	39	262	61	429
16	Diarrhea.................	106	19	442	81	548
17	Hiccup...................	98	17	478	83	576
18	Epistaxis.................	61	10	525	90	586
19	Herpes...................	28	5	536	95	564

Let us then examine the table (now Table 8) in rearranged form, to fulfil in maximum degree possible from the published data the fundamental purpose for which it was tabulated.

Table 8 tells the story of symptomatology much more simply, directly, and accurately than does Table 7, of which it is merely a rearrangement.   It is seen at a glance, for example, that more

than 90 per cent. of the cases about which anything definite as to
the symptoms was known, exhibited at least one of the four following
symptoms: jaundice, bile-stained urine, nausea, clay-colored stools.
Fewer than 20 per cent. of the cases had either diarrhea or hiccup,
or epistaxis, or herpes, each taken by itself.

In making this rearrangement three changes were made from
the original table:

(*a*) The percentages were calculated on the basis of the *known*
universe of discourse. To do otherwise in this case makes the
percentages virtually meaningless.

(*b*) Percentages were tabled only in *whole numbers*. No
derivative calculations will be made from these percentages. Their
sole purpose is quickly and simply to inform the reader of the
relative frequencies of certain conditions. Decimals are only an
annoyance under such circumstances.

(*c*) The symptoms are arranged in *descending order* of relative
frequency. This makes rapid and intelligent reading, and evalua-
tion of the table as a whole, easy of accomplishment. What could
be more desirable if the author wishes to instruct and entertain
his reader?

The percentage figures of Table 8 are shown graphically in
Fig. 35 of Chapter VI on p. 168.

It will be good practice for the reader, in developing for himself
skill in the planning and arrangement of tables, mentally to criticize
statistical tables as he encounters them in his general medical read-
ing, and try whether he could re-arrange the same data into more
accurate, intelligible, or simple form. This particular process will be
materially aided, to say nothing of the general training in accuracy
and precision of mental processes which will incidentally accrue, if
one approaches a statistical table in some such manner as this:

What is the *purpose* of this table? What is it *supposed* to
accomplish in the mind of the reader?

Does it? Well? Indifferently? Badly? Not at all?

Wherein does its failure of attainment fall?

When this last question has been analyzed and settled, the
process of making a satisfactory table to accomplish the purpose is
much more than half finished.

SUGGESTED READING

1. Yule, G. U.: Introduction to the Theory of Statistics. Chapters I–V inclusive. (A detailed and important treatment of the statistical consequences which flow from dichotomous and other forms of classification. The student should work through the practical exercises given at the end of each of these chapters in Yule.)

2. Crum, W. L., and Patton, A. C.: An Introduction to the Methods of Economic Statistics, Chicago (A. W. Shaw Co.), 1925. Chapters V and VI.

3. Pearl, R.: Preliminary Note on the Incidence of Epidemic Influenza Among the Actively Tuberculous, Quart. Publ. Amer. Stat. Assoc., vol. 16, pp. 536–540, 1919.

4. Watkins, G. P.: Theory of Statistical Tabulation, Quart. Publ. Amer. Stat. Assoc., vol. 14, pp. 742–757, 1915.

CHAPTER V

ORIGINAL SCIENTIFIC RECORDS AND THEIR TRANSLATION TO TABULAR FORM

Up to this point in the discussion original statistical data have been tacitly assumed to be given. The reader has not been required to undertake any responsibility regarding their collection. We have, to be sure, examined with some care the methods by which official vital statistics are obtained (Chapter III). But this was a special study of how an official government body—the Census Bureau—goes about furnishing vital statisticians with their basal sustenance, so to speak.

It is only a small fraction of the scientific data to which statistical methods of research may be usefully applied that governments take the trouble to furnish to students. Mostly the student of any subject has to collect his own data, by means of observation and experiment, from the phenomenal world around him. In this chapter the attempt will be made to discuss briefly some general principles underlying the collecting, recording, and putting into tabular form of original observational data in the manner most convenient and useful for subsequent statistical treatment.

THE COLLECTION OF SCIENTIFIC DATA

All scientific data are answers to specific questions put to Nature by the investigator. The scope of both the question and the answer is necessarily sharply and narrowly delimited in each particular case. For example, if an investigator starts collecting data on the length of the human skull, what he does is to measure, with appropriate instruments and with the greatest attainable accuracy, the length of each one of a series of skulls. When he sets down in his record book that the length of skull No. 1 was 137 mm., the record 137 mm. is *the answer to the implied question "what is the length of skull No. 1?"* Wherever possible science asks this simple

type of question, and records the answers as numerical statements of quantity. This is not always possible, however, basically for the reason that there are a great many things that are interesting which no one has yet found a way to measure, or express quantitatively. With the progress of science the number of such things is, happily, all the time getting smaller. But it is still undeniably large. Where the simple straightforward numerical answer is impossible, the record has to be of a more complex character. For example, when a physician makes a stethoscopic examination of the lungs and writes down as a part of his record "moist râles at left base," the answer is enormously more complicated in its implications than is the craniologist's precise figure of skull length.

If the primary business of science is to ask questions and set down the answers to them, then the questionnaire may be regarded as the canonical form of scientific record. This is employing the term "questionnaire" in a broader and more inclusive sense than is usual. It is so used here to emphasize the essential nature of original scientific records, namely, that *they are individual answers to specific questions*. If this concept is once clearly and firmly grasped by the mind, it will greatly help the student in wrestling with the never-ending problems of methodology which will keep on arising so long as he attempts to do any original work in any branch of science. For it will enable him to see that there are really only two great methodological problems in scientific research. The first is: "What will be the most effective and useful way to ask the question?" The second is: "How may it best be assured that the answer shall be correct, precise, clear, and without ambiguity?" Obviously these two questions are not wholly independent. A substantial part of the second is implied in the first. In the opinion of many competent investigators it is the first methodological problem that is the most important and the most difficult. They hold that the successful and fruitful outcome of original investigation depends most upon the *Fragestellung*—how the question is put to Nature.

Anything like adequate didactic instruction to the beginner on this important matter, if not wholly impossible in the nature of the case, is certainly too vast an undertaking for the scope of this book as well as for the limited competency of its author. The best prac-

tical help here which can be offered the student is to suggest that he read widely and deeply in the history of science. Reference No. 16 in the reading list at the end of the chapter was especially prepared to give the student a guided start in this direction. The history of science tells us at least how the great investigators, whose efforts have brought about the achievement of so much knowledge of Nature and her laws as we now possess, put their questions. And example is not a bad pedagogical technique.

SOME ESSENTIAL IDEALS IN THE MAKING OF SCIENTIFIC RECORDS

Let us turn now to some consideration of the relatively simpler problem of the recording of the answers to questions scientifically asked. This may be, and often is, regarded as merely a sordid, mechanical business, not worthy really serious attention. But surely it is a pity to increase the labor and strain of scientific work by requiring it to be done with illegible, incomprehensible, or ambiguous original records. It requires a tremendous amount of labor, in the aggregate, to collect accurate, scientific data. Surely it is not much to ask that the original record of them be so made as to be permanently clear, precise, and useful for whatever subsequent purposes it may be desired to put it to.

The following list of desirable characteristics of original scientific records makes no effort toward pedantic completeness. But perhaps it may stimulate the student himself to think a little when he is engaged in the dull spade work of measuring, counting, taking case histories, or making physical diagnoses.

1. *Accuracy.*—This must, of course, come first in the making of scientific records. It is attained, more than in any other way, by the exercise of two rather rare human qualities, at least in their native, uncultivated state. These are *carefulness* and *attentiveness*. Most mistakes in the recording of the results of experiments or measurements are due to careless wandering of the attention, momentarily or longer, from the business immediately in hand. Unfortunately there is no general, infallible, mechanical apparatus adequate to obviate this difficulty. Perhaps the best suggestion is that a habit be formed to check each individual record directly after it is set down on paper, there and then, against

the observed object, while the latter is still at hand. Of course, in the nature of things, this cannot always be done. But, anyhow, it will not be a bad habit to try to form in this imperfect world.

2. *Altruism.*—Every scientific worker should always keep before his mind that somebody else may sometime want to make use of his original records. An unexpected disabling illness, or death, or even the getting of a better job may lead to such a situation. Whence it follows that every page, every line, and indeed every word and figure of the record should be absolutely clear as to its meaning, in a proximate and immediate sense. Everyone is prone to abbreviate and condense in the tedious work of making records. This is obviously good sense, and unobjectionable in principle. But a meticulously detailed account of the abbreviations, the manner of condensation, should go along with the records. Any considerable experience of working with records collected by other people engenders strong views on this point. It is perfectly natural for anyone to feel that *he* understands his system of abbreviating his notes, and to forget that what is so clear to him at the time will not be clear to others, or incidentally to himself after a sufficient lapse of time, *unless he takes the trouble to set down the explanation at the time along with the record.*

3. *Neatness and Legibility.*—A scientific record which no one, even its maker, can *certainly* read, may be accurately defined as a total loss. If it is difficult to read it is a nuisance. Taking pains to make figures and writing neat, plain, and legible pays extremely well in subsequent saving of time. Furthermore neatness in the arrangement of the record is important.

4. *Permanence.*—Original records should be made on (*a*) a good quality of paper, cut in sheets of uniform size, and either bound in a book at the start, or, if loose, the sheets should be bound as soon as the particular set of records is completed; or (*b*) on card forms, cut to uniform size from good stock. For many years it has been the practice in the writer's laboratory to use uniform paper, of a standard size, ruled to suit our requirements, for all records, computations, and preparation of manuscript. At the end of a particular study, after the results have been published, all of the papers connected with it are bound neatly together in heavy board covers

by the laboratory *Diener* at a cost of from 15 to 25 cents a volume, and filed for permanent record.

Original records should be made in ink wherever possible. It is recognized that this is a counsel of perfection. Some people are bound to use pencils. If the pencil urge is uncontrollable, it should be satisfied with either indelible pencils, or ordinary pencils with hard leads. Any other sort will, in time, lead to blurred and illegible records.

5. *Comprehensiveness.*—Nothing is more annoying in working with statistical records, or indeed any other kind of records, than to find no statement whatever made about some particular point, which certainly was observed at the time. Such omissions arise chiefly in one or another of three ways: (*a*) The point was observed, but through carelessness was not recorded in every case; (*b*) the point was observed to be "normal" in some cases, and on that account thought not worth recording; (*c*) in planning the investigation no place was made in the scheme for observing the point at all. Such gaps in the records may be avoided by two relatively simple means. The first is to plan the investigation in advance with sufficient care to ensure that all pertinent data, so far as it is possible to envisage them in the then existing state of knowledge, shall be included in the plan of the records. The second is to make it an unfailing rule to record something regarding every item in the record plan in every case. This "something" may be a positive or a negative finding; the situation may be "normal" or interestingly abnormal; but in any case something about it should go into the record. This matter of comprehensiveness of records will be discussed further in a later section of this chapter.

6. *Minimal Errors of Personal Equation.*—It is a well-established fact that observations are influenced by the unconscious bias or so-called "personal equation" of the observer. In astronomy, and to some extent in the other physical sciences, careful attention is paid to personal equation. Very little, though, is given to it in the biological sciences or medicine. It can, however, lead to considerable errors, greater in magnitude indeed than the errors of random sampling, to which the statistician pays so much attention. The student should read the important papers of Pearson[1] and Yule[2] on the

subject. An example may be given briefly here to indicate how divergent the results of thoroughly trained, competent biologists may be, when observing identical things. Further details regarding the experimental results may be found in the original publication.[3]

When a yellow starchy (flint) variety of maize was crossbred with a white sweet (sugar) variety there was produced in the first generation uniformly yellow starchy progeny, in accordance with Mendel's law of dominance. When these first generation crossbred kernels were planted they gave rise to a second crossbred generation in which each ear bore four different kinds of kernels, in approximately the following proportions: 9 yellow starchy; 3 yellow sweet; 3 white starchy; 1 white sweet.

Fifteen trained observers were asked each to sort into these four classes and count independently the kernels on each of a number of second generation ears. The results of the count on one such ear are shown in Table 9. The fifteen observers included two plant pathologists, two professors of agronomy, one professor of philosophy (originally trained as a biologist), four biologists, one computer, one practical corn breeder, and one professor and three assistants in plant physiology. The following remarks about the group are pertinent in judging the results.

"In the first place it is obvious that any one of them (with the possible exception of X) might in the ordinary course of his work carry on a Mendelian experiment with maize, either independently or in co-operation with someone else. If this were done and the results published they would certainly be accepted by the biological public as a precise and true statement of the facts regarding the material which was in the experimenter's hands. That is, if any worker in this list published a statement that a Mendelian experiment which he had conducted with maize led to a ratio of, for example 759 : 234 : 252 : 90 this statement would not be doubted or questioned. In the second place it is worth while to consider the training, or lines of work with which these 15 observers have had to do. Of six (Nos. I, II, XI, XII, XIII, XIV) the training and work has been primarily *botanical*. Four of these (the Danish group, Nos. XI to XIV inclusive) have had particularly to do with the data of experimental plant breeding, in connection with the brilliant and fundamental researches of Professor Johannsen. The training and special field of work of five (Nos. V, VI, VII, VIII, and XV) of the observers has been *zoölogical*. Of these five three (Nos. VI, VII, and VIII) have had experience with the data and methods of investigation in experimental breeding. Another of the five (No. V) adds to the special training of the zoölogist that of philosopher and psychologist, which by traditional standards, at least, ought to aid in the development of a discriminative judgment. The training of two of the

observers (Nos. III and IV) has been agricultural.   Further, both of these men belong by birth, early life, and education to the "corn belt" section of the country, and are thoroughly and intimately familiar with maize.   They have had experience in corn judging, which demands the appreciation of very small differences in ear characters.   Observer No. X, while not a scientific student of breeding, has had successful experience in corn breeding, and is a careful observer.   Observer No. IX has been especially trained in biometric work in the writer's laboratory and has had considerable experience in measuring, sorting small variations out of mixed material, and similar work."

TABLE 9

SHOWING THE CLASSIFICATION OF THE KERNELS OF EAR NO. 8 BY THE DIFFERENT OBSERVERS

Observer.	Classes of Kernels.					
	Yellow starchy.	Yellow sweet.	White starchy.	White sweet.	Total starchy.	Total sweet.
Mendelian Expectation.	299.25	99.75	99.75	33.25	399.00	133.00
I.	352	102	52	26	404	128
II.	322	49	82	79	404	128
III.	298	75	108	51	406	126
IV.	332	101	71	28	403	129
V.	305	101	86	40	391	141
VI.	313	100	90	29	403	129
VII.	308	86	95	43	403	129
VIII.	311	101	92	28	403	129
IX.	327	101	78	26	405	127
X.	308	92	95	37	403	129
XI.	311	97	92	32	403	129
XII.	313	99	91	29	404	128
XIII.	308	97	95	32	403	129
XIV.	312	104	91	25	403	129
XV.	333	97	73	29	406	126
Totals.	4753	1402	1291	534	6044	1936
Means.	316.87	93.47	86.07	35.60	402.93	129.07

The results shown in Table 9 and Fig. 13 seem to have real significance relative to the methodology of biology in general, apart from specifically genetic problems.   It must be remembered that each individual handled, sorted, and counted *the same identical kernels of corn*.   They were required to discriminate only with reference to the color and the form of each kernel.   *Yet no two of the fifteen highly trained and competent observers agreed* as to the distribution of these 532 kernels.   When it is recalled that pathologists, clinicians, and anthropologists have to make fine distinctions

relative to color and form regularly in the course of their work, the thought suggests itself that perhaps their records of observation on man, a more complicated entity than a maize kernel, may not have that absolute and ultimate verity that some naïve persons perhaps suppose they have.
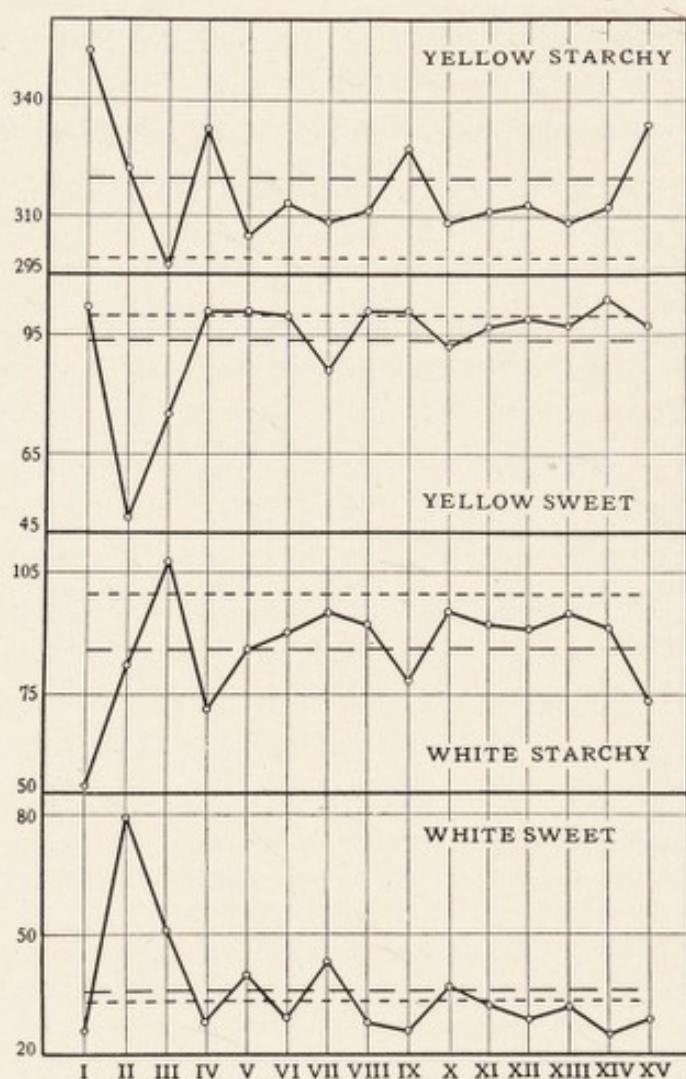
Fig. 13.—Diagram showing the counts for ear No. 8 by each of the different observers. The horizontal dotted line gives the Mendelian expectation, and the horizontal dash line the average of the counts of all 15 observers.

It should be recognized as a general principle, and kept always in mind, in measuring and recording, that every individual has bias or "personal equation" in his observing and measuring. There is no way completely to eliminate its effects. The most that can be done is to minimize them. The first step toward this is for an

individual to find out by preliminary observation approximately what the trend and amount of his personal equation is relative to the particular thing being measured or observed. Then, at least, he knows where he needs guarding against himself, and can make allowances and use extra care.

Space cannot be spared for further discussion of the matter here. But before leaving it entirely the student who is interested in philosophy or metaphysics is asked to contemplate Table 9 from the point of view of the statistical method of acquiring knowledge. It may fairly be said that Ear No. 8 carried 532 kernels. The testimony of fifteen independent witnesses agrees to this. Perhaps with as great warrant as is ever attainable we may say that we *know* that Ear No. 8 had 532 kernels. But how many white starchy kernels did it have? I mean how many did it *really* have? There must have been some determinate number because it is certainly *known* that *some* of the 532 kernels were white starchy. But *how many?* It seems a simple problem. One only has to count them. They do not run away or change. But still I should like to *know* how many of them there were on this ear. And still more I should like to know some method by which *definite and certain knowledge* on the point *could possibly be obtained*, by the use only of visual observation of the kernels themselves and the process of counting. Examine the fourth column of Table 9 and think it over.

7. *Purposeful Adaptation.*—Original record forms should be carefully planned in advance so that the orderly arrangement of the individual items will most effectively conduce to speed and accuracy, first in the recording of the original observations, and second, in their subsequent tabulation. For example, in planning a blank form for recording anthropometric measurements all those measurements which are taken with one instrument, for instance the heights measured with the anthropometer, may conveniently follow each other consecutively in one group.

8. *Inclusiveness.*—All observations made should be included in the original records, as they are made. If some particular observations are suspected of being bad, a note should be made saying so. But they should not be thrown away. To do so is to implant permanently in the observer's hitherto pure mind a horrid (however

9

small) conviction of the sin of picking and choosing only favorable cases. No honest man would, of course, be guilty of intentionally doing this. But the only way to avoid it is never to take the first step. Put down on the record everything that Nature offers. Later on the record can be looked over and studied, and a calm and reasoned attempt can be made to see what it all means. But if part of what was actually observed has been omitted from the record nothing further can ever be honestly done about that fact.

9. *Absence of Ambiguity.*—A record which is capable of being read in either of two ways is a thorn in the scientific flesh. Unless care is given to the point such records turn up with curious frequency. Let an example point the precept. It is the prevailing custom in America to write dates in the form April 2, 1930. It is a common custom in Europe to write dates in the form 2 April, 1930. In both America and Europe scientific men have a habit of using a numerical shorthand in dating their records. But who is to be *sure* whether a record of 4/2/1930, considered as record, means April 2, 1930, or February 4, 1930? It has long been a rule of the writer's laboratory that all dates should be of the form Apr. 2, 1930, as a maximum concession to the shorthand urge. But in going over our old records it is amazing to see how many have risen superior to any such attempted restraint upon free self-expression. And of what conceivable use is a date record 4/2 say n years after the actual year of its making? Other examples of ambiguous records might be given.

MEDICAL RECORDS

It will perhaps be useful at this point to transfer the discussion from the general to the specific, and consider medical records as examples pertinent to the interest of the class of readers for which this book is especially intended.

The fundamental medical record is the individual case history. Upon it depends any and all useful information, whether statistical or otherwise in character, which may be wanted for any purpose whatever. It is, therefore, of the highest importance that case histories conform to the best standards of scientific record making, on the one hand, and of modern business office practice on the other

hand.   There seem to be relatively few hospitals where the highest standards in either of these respects are even approximated.

From the standpoint of scientific record taking, case histories are most often defective in what they *fail* to record about the patient.   It is by no means impossible to find case histories that fail to record the sex of the patient; while any indication of what *kind* of person he was, in the common sense of the word, whether fat or lean, white or colored, rich or poor, young or old, etc., is all too frequently kept a deep secret from any subsequent reader of the history.   Again, even in the special medical portions of the history the writer forgets, with almost unbelievable frequency, to make any record of highly important facts.

The root of the difficulty apparently lies in the method by which case histories are written.   The general scheme or outline which a history is to follow seems often to reside in the head of the particular writer, and there only.   And heads, especially of human beings, do vary so!   There is a simple procedure which will help to remedy the difficulty.   It is, as a first step, to draw up and have printed a series of *standard* history forms, which will cover not merely general routine facts common to all diseased conditions, but special forms as well, for at least all of the more frequently occurring conditions. These blank forms will contain definitely indicated spaces in which some statement of fact, either positive or negative, *absolutely must be recorded in every single case.*   If on the case record form for gall-stone cases, for example, there is printed the question, "Did this patient ever have typhoid?" or the equivalent of this question, one or another of three answers may be definitely recorded, either "yes," or "no," or "nobody knows."   If, furthermore, every worker in the service clearly understands that any history for which he is responsible that comes into the history department, with any blank spaces in its standardized portion, will not be accepted for filing, but will be forthwith returned to him for completion, future students will be able to compile comprehensively and definitely the teachings of the experience of that hospital relative to the etiologic relations between typhoid and biliary calculi.

It is, of course, to be understood that no blank form, however carefully it may be devised, can ever suffice for the recording of the

whole history. There must be some portions written or dictated with entire freedom from Procrustean rigidities. The reason why this is so is plain. One of the chief characteristics of living things, whether men or mice, is that they vary individually. But formal blanks do not vary. An invariable phenomenon cannot accommodate itself to a variable one. But this is no valid argument against having certain essential parts of the history recorded in standardized form. There are some facts that everyone will agree ought to form a part of every case history which is to be permanently preserved. It is that class of facts which should be recorded upon standardized formalized sheet or sheets incorporated into each history. Then, *in addition*, the clinician may write or dictate as much more as he likes in an entirely free untrammeled style. The formalized portion merely serves as the schema of the whole, to make sure that no point of importance for future students is left out, because forgotten, in the greater present interest of other more immediately exciting features of the case.

It is particularly important that a definite statement or record be made that a structure or function is *normal* when it is so. In the minds of many persons, perhaps particularly in the field of medicine, there has grown up the notion that what is normal is of no interest and, therefore, nothing needs to be said about it in the record. Later on someone comes to study the record. Let us say, to take a concrete example, that this subsequent student wants to know definitely whether the tonsils in this particular case were diseased or not. No mention of tonsils can be found. Two alternatives then present themselves to the second student:

1. The tonsils were not diseased, and on that account the original recorder said nothing about them.
2. The original recorder forgot to look at the tonsils or forgot to make a record of his findings.

Either horn of the dilemma is equally unfortunate. "No information" is the sad, but only possible, conclusion which can be regarded as accurate.

EXAMPLES OF BLANK FORMS

In this section it is proposed to give some examples of record forms, which have been successfully and satisfactorily used in

actual investigations over a number of years. These will help to illustrate some of the general principles which have been discussed above. It should be understood that these examples are *not* put forward as models. Perfection is rarely attained in this rapidly moving world, and almost never by statisticians. The examples of blank forms here presented are open to criticism from a number of viewpoints. But, even with their recognized imperfections, they have worked well in practice. Every time a new batch of these blanks is printed some defects which experience has brought to light are corrected, and some new items or improvements added. This somewhat easy-going attitude represents the sum-total of our striving toward absolute perfection in record forms.

If any reader should wish to make trial of such record forms as are here illustrated, he should *not* copy slavishly these particular forms. Instead he should draw up his own, designing them to meet specifically his particular needs, just as these were drawn up to meet our special requirements. The examples given here can, and should, at best, serve only as suggestions.

In Figs. 14–21 inclusive are a part of the forms used in the investigation of constitutional factors in disease which has been in progress for the past five years in the Institute for Biological Research of the Johns Hopkins University.[4] These eight forms have to do with the personal and medical history, and the physical examination, of a patient being studied constitutionally. The problem of primary interest in our constitutional work has been that presented by the clinical picture commonly called "essential hypertension," and the blanks have all been constructed with that problem in mind. This will account for some of the obvious omissions, were these to be regarded as general medical history blanks.

All the forms are printed on sheets 11 x 8.5 inches, of a good grade of bond paper. A binding margin, 1 inch wide, is left on the left-hand long side of each sheet. All forms are printed on one side of the paper only, leaving the back available for additional notes.

Forms A 1–8 are used exclusively for data pertaining to the patient (and when possible to other members of the family who consent to have complete examinations made), and are filled in by the examining physician, who is especially trained in the technic.

In planning the investigation it was constantly borne in mind that the data collected should comply with the following criteria.

Fig. 14.—Constitutional Form A-1 in reduced facsimile.

The questions asked must be (1) strictly relevant to the purpose of the investigation; (2) inclusive, *i. e.*, must apply to all individuals, allowing for differences of sex and age; (3) systematic, *i. e.*, grouped

in logical order; (4) specific, *i. e.*, definite and unambiguous; (5) comprehensive, *i. e.*, covering the fields the investigators were cap-

Fig. 15.—Constitutional Form A-2 in reduced facsimile.

able of exploring; (6) objective, *i. e.*, avoiding individual bias as far as possible; and (7) conveniently arranged for recording and analyzing.

The forms set up for recording the data cover with reasonable completeness, subject to the implications of criterion 5, all fields

Constitutional Form A-3

Family Name: Family No.

Name of Person:

Teeth: No. extracted 1, 2, 3, 4, 5, 6, 7. Why? pyorrhœa, abscess, toothache.
    Gums bleed occasionally, frequently.

Goitre: No, yes (specify)

Cardio Resp: Continually clearing throat, chronic cough, morning cough
    Expectoration: none, occasional, frequent; hard to raise, easy to raise, white, grey, yellow, blood streaked 1, 2, 3, 4, 5, 6 times, large amount of blood 1, 2, 3, 4, 5 times
    Severe pain on breathing: rt. side, lt. side
    Night sweats: occasional, nightly
    Shortness of breath: none, running for a street car, walking up one flight of steps, walking along level street, sudden attacks requiring pt. to sit up in bed
    Consciousness of rapid beating of heart: none, with excitement, after eating, with exercise, sudden onset while sitting still, sudden onset while lying in bed
    Does pt. have attacks of pain in chest (over heart or in lt. arm) which come on suddenly: with, without exercise, causing fear of death, or cessation of all activity for the moment? Yes, no, occasional, frequent, daily.
    Other history of heart or lung diseases: (specify)

Gastro-intestinal: Eats: heartily, average, sparingly, always hungry but afraid to eat, never hungry
    Bowel movements: regular, irregular, 1, 2, 3, 4, 5 daily, weekly, easy, difficult, loose
    Piles: none, itch, protrude, painful, bleed.
    Abdominal pain: none, upper half, lower half, relieved by eating, increased by eating, dull, severe, shooting
    Hernia: no, yes (specify kind)
    Other history of abdominal disease (specify)

Genito-urinary: Gets up at night to pass urine 0, 1, 2, 3, 4, 5 times
    Age at onset of nocturia   (yrs.)
    Gonorrhea: attacks 0, 1, 2, 3, age at first attack   (yrs.) Epididymitis, prostatitis
    Syphilis: no, yes, chancre, secondary rash, hereditary, other manifestations: (specify)
        Age at onset   (yrs.)

    Treatment: (specify and state physician or dispensary)

    Other history of genito-urinary disease

Menstrual History: Age at onset   (yrs.) Painful, painless, profuse, average, scanty
    Regular: interval   days, duration   days
    Irregular: usual interval   days; shortest interval   days: longest interval (unassoc. with pregnancy)   mos. duration   days
    Completely irregular: shortest interval   days, longest   days.
    Duration: usual   days; shortest   days; longest   days.
    Last time: month   year

Do Not Write In This Space

Fig. 16.—Constitutional Form A-3 in reduced facsimile.

of general medicine. They do not include such specialties as psychiatry, obstetrics, and genito-urinary diseases of males and females, because these and certain other branches of medicine are

at present outside the scope of our investigations; but for them supplementary forms along similar lines could readily be constructed,

---

Constitutional Form A-4

Family Name: Family No.

Name of Person:

Marital History: Number of pregnancies (or wife's) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Number of miscarriages
Month of pregnancy

1	2	3	4	5	6	7	8		

Number of premature living born
Month of pregnancy

1	2	3	4	5	6	7	8		

Number of term living born 1, 2, 3 ,4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14

Number of born dead 1, 2, 3, 4, 5, 6

Sexual Habits: Frequency 1, 2, 3, 4, 5, 6, 7 per week   month   year

Birth Control: Not used   used (specify methods)

Allergic History: Hay fever, asthma, angioneurotic oedema, urticaria, food, other protein, other
    symptoms: (specify)

Skin: (History of any skin disease)

Congenital malformations, tumors, etc. (specify)

Bleeding: (History suggestive of hemophilia): No   yes (specify)

Nervous system: Frequency of attacks of dizziness: none, daily, weekly, monthly, yearly
    Patient has fallen in attacks, has not fallen
    Frequency of convulsions: General: none, daily, weekly, monthly, yearly
                        Localized: none, daily, weekly, monthly, yearly
    Other history of nervous disorders: (specify)

Locomotor System: History of disease of joints, difficulty in walking, etc. (specify)

Weight: Has there ever been any rapid change in weight during adult life?
    Gained   lbs.   yr.   Why?
    Lost   lbs.   yr.   Why?

Habits: Sleeping: Average hours 5, 6, 7, 8, 9, 10 sleep unbroken, broken, 1, 2, 3, 4, 5 times
    Why?   Awakens rested, tired.

    Alcohol used: no, yes (get as full details as possible throughout life)

    Wine?   Beer?

    Whiskey or other spirits?

    Tobacco used: no, yes (get as full details as possible throughout life)

    Pipe?   Cigars?

    Cigarettes?   Chewing?

    Snuff?

Do Not Write In This Space

Fig. 17.—Constitutional Form A-4 in reduced facsimile.

and in certain cases, notably in the study of Friedreich's ataxia, we have made such supplementary record forms.

Forms A 1–4, shown in Figs. 14–17 inclusive, deal with the essen-

tial facts of the *patient's personal and medical history*, as obtained by direct questions asked by the physician. The questions are simple

---

Constitutional Form A-5

Family Name: Family No.

Name of Person:

PHYSICAL EXAMINATION

Date

Body temperature: Hour when taken   Examined by

Head: Mucous membrane; normal color, pale, cyanotic

Eyes: deep set, average, prominent, very prominent. Upper lids: normal, puffy.
 Lower lids: normal, puffy, very puffy. Arcus senilis: none, slight, marked.
 Pupils: regular, irregular, equal, unequal, lt. larger, rt. larger, react to light, react on accommodation.
 E. O. M.: normal, abnormal (specify)

Nose and throat: Nasal obstruction: none, moderate, marked, left, right.
 Tonsils present, absent, normal, obviously diseased.

Teeth: gums and mouth: pyorrhoea, none, mild, moderate, severe, other abnormality (specify)

Upper teeth   Lower teeth

(Strike out those that are missing; C – crown. B – bridge. F – filled. G – nonvital teeth. H – obvious cavities unfilled. R – roots only.)

Ears: Hearing: hears watch tick Rt. ear yes no Lt. ear yes no
     hears normal voice Rt. ear yes no Lt. ear yes no
     hears loud voice Rt. ear yes no Lt. ear yes no
 Discharge: Lt. ear yes no.   Rt. ear yes no

Neck: Thyroid is: not felt, isthmus palpable, lobe palpable, right, left, smooth, nodular, markedly enlarged.
 Trachea is: in midline, deviated to left, right

Lymph nodes: Angles of jaw: small, large, tender, soft, firm, not felt
     Cervical: small, large, tender, soft, firm, not felt
     Epitrochlears: small, large, tender, soft, firm, not felt
     Axillary: small, large, tender, soft, firm, not felt
     Inguinal: small, large, tender, soft, firm, not felt

Chest: 1. General description:
 Clavicles: invisible, visible, prominent
 Chest shape: normal, flat, chicken breasted, barrel shaped
 Harrison's groove: present, absent
 Costal margin: normal, flaring
 Movement of chest wall: rt. equals lt., rt. more, rt. less

Respiration rate:

Do Not Write In This Space

Fig. 18.—Constitutional Form A-5 in reduced facsimile.

and direct, but comprehensive withal, and are restricted to circumstances, events, bodily functions and variations, symptoms and diseases, concerning which a person of ordinary intelligence can

usually give reliable answers, or, if the patient is a child, which the mother can answer. They differ little from those covered in the

Constitutional Form A-6

Family Name: Family No.

Name of Person:

2. Lungs: Distance of lung bases at rest below 7th cerv. spine   cm.
          if not at equal levels rt.   cm., lt.   cm.
      Condition of lungs on exam.: normal, abnormal (specify pathology)

3. Heart Dorsal, upright
        P. M. I., seen, not seen; localized, diffuse, weak, forceful
        Distance from midline   cm., interspace 4, 5, 6, 7;
        Distance from suprasternal notch   cm.
        Cardiac dullness: cm. to left M. S. L.;   cm. to right M. S. L.
        Heart rhythm: regular, presystolic gallop, protodiastolic gallop, extrasystolic, fi-
                     brillating
        Murmurs: none, yes (specify)
4. Pulse: Dorsal, upright. Rate   Blood pressure: S.   D.
        Rhythm: regular, extrasystole, none, occasional, frequent, completely irregular

Arteries: Temporal: Vessel wall seen, not seen, tortuous, very tortuous
        Radial: Rt. pulse equals lt., rt. larger, rt. smaller.
                Vessel walls: not felt, felt, soft, hard, diffuse thickening, nodular, straight, tor-
                     tuous, very tortuous
        Brachial: vessel walls (*elbow straight*): not seen, seen, not felt, felt, soft, hard, diffuse
                thickening, nodular, straight, tortuous, very tortuous.
        Posterior tibial: Pulse present, absent, rt. equals lt., rt. larger, rt. smaller
        Dorsalis pedis: Pulse present, absent, rt. equals lt., rt. larger, rt. smaller

Abdomen: Abdominal walls: fatty layer: thin, medium, thick. Muscle wall: relaxed, not relaxed,
                flabby, firm: Muscle spasm localized in U. R. Q., U. L. Q., L. R. Q., L. L. Q.
        Liver: not felt, felt   cm. below costal margin, smooth, not smooth
        Spleen: not felt, felt   cm. below costal margin, smooth, not smooth
        Kidney rt.: not felt, felt, normal size, enlarged, tender
        Kidney lt.: not felt, felt, normal size, enlarged, tender
        Hernia: no, yes, rt., lt., direct, indirect, inguinal, umbilical, femoral
        Additional Note:

        Reflexes: Biceps: Rt. absent, average, diminished, exaggerated.
                              Lt. absent, average, diminished, exaggerated.
                 Knee Kick: Rt. absent, average, diminished, exaggerated.
                              Lt. absent, average, diminished, exaggerated.
                 Ankle Jerk: Rt. absent, average, diminished, exaggerated.
                              Lt. absent, average, diminished, exaggerated.
                 Abdominal: Rt. absent, average, diminished, exaggerated.
                              Lt. absent, average, diminished, exaggerated.
                 Romberg: Absent, slight, marked.

Extremities: Tremor of extended fingers: none, fine, coarse. Additional note:

Malformations: (specify)

Do Not Write In This Space

Fig. 19.—Constitutional Form A-6 in reduced facsimile.

history taken in a properly conducted office consultation. The only important distinction lies in the fact that in this work *all the data*

are recorded for every patient. We permit no blank spaces or missing records.

Constitutional Form A-7

Family Name: Family No.

Name of Person:

LABORATORY STUDIES

Urine Examination

Date	Appearance	Reaction	Sp. gr.	Alb.	Sugar	Microscopic

Kidney function (phenolsulphonephthalein eliminated)

Date	During first hour		During second hour		Total	
	Amount	Per cent	Amount	Per cent	Amount	Per cent

Wasserman Reaction on Blood Serum   Spinal Fluid Examination

Date	Result	Date	Result

Basal metabolism: (give date and result)

Do Not Write In This Space

Fig. 20.—Constitutional Form A-7 in reduced facsimile.

Taking up the forms in order, it will be noted that in *Form A* 1, stress is laid first on the general conditions of the patient's life, as regards urban or rural residence, the kind of work done and its

demands on energy output, and second on his present illness.   If
the patient has ever been treated before in the dispensary or hos-

Constitutional Form A-8

Family Name: Family No.

Name of Person: Date:

EXAMINATION OF EYE GROUNDS

Right eye: normal, abnormal (specify below)

Arteries: Irregular diameter   Slight, medium, marked
           Reduced diameter   Slight, medium, marked
           Tortuous   Slight, medium, marked
           Pulsating.   Slight, medium, marked

    Increased light reflex   Punctate, uniform
    Translucency   Diminished, absent

Veins: Increased diameter   Slight, medium, marked
       Compressed by arteries   Slight, medium, marked

Retina: Oedema   Slight, medium, marked
        Exudate   Slight, medium, marked
        Haemorrhages   Fresh, organized

Optic disc: (Describe abnormality)

Remarks:

Left eye: normal, abnormal (specify below)

Arteries: Irregular diameter   Slight, medium, marked
           Reduced diameter   Slight, medium, marked
           Tortuous   Slight, medium, marked
           Pulsating   Slight, medium, marked

    Increased light reflex   Punctate, uniform
    Translucency   Diminished, absent

Veins: Increased diameter   Slight, medium, marked
       Compressed by arteries   Slight, medium, marked

Retina: Oedema   Slight, medium, marked
        Exudate   Slight, medium, marked
        Haemorrhages   Fresh, organized

Optic disc: (Describe abnormality)

Remarks:

Do Not Write In This Space

Fig. 21.—Constitutional Form A-8 in reduced facsimile.

pital full abstracts are made on other intercalated sheets of the
symptoms, objective findings, diagnosis, operations, etc., if any,
as recorded in these histories, and these are filed with the examina-

tion forms. In many cases there are elaborate histories, involving several departments and extending over many years.

Form A 2 covers the history of surgical operations, hospital admissions, the occurrence of various common infections, and some important chronic organic diseases. Gonorrhea and syphilis, which evidently deserve greater detail, are taken up under the genito-urinary system, on *Form A* 3.

The remainder of *Form A* 2 and *Forms A* 3 and 4 are devoted to the head (including the hair, headaches, vision, hearing, the nose, the teeth, the thyroid), the cardio-respiratory, the gastro-intestinal, and the genito-urinary systems, and to the menstrual history.

On *Form A* 4 one third of the space is devoted to the important and usually neglected subject of sexual history. In addition to the total number of pregnancies, provision is made for the recording of not only the number of living and dead born full-term children but the number of miscarriages (including abortions) and of premature living born, with the periods of gestation by months.

The frequency of sexual intercourse, by week, month or year (by five- or ten-year periods of life when possible), and the use or non-use of what were, or were believed to be, contraceptive methods (including induced abortions) are set down under these headings with as much detail as it is possible to obtain.

In training the clinicians who take the histories the following memorandum of instructions is used, relative to the questions regarding the reproductive history. It is included here as an example of the sort of instructions which should be worked out for complicated or difficult points in any large, systematic, record-taking enterprise.

Further knowledge concerning the correlation between frequency of intercourse and frequency of conception is urgently needed. Birth control, conscious and designed, is affected by (*a*) abstinence in varying degree, or (*b*) contraceptive measures of various sorts. To these must be added, in the world as it actually exists, the causing of the expulsion of fertilized ova, or induced abortion. The nullifying effects of these measures upon the consequences of frequent sexual intercourse is obvious.

Various diseases, such as syphilis, pneumonia, influenza, malaria, variola, by killing or causing the expulsion of fertilized ova, cut down the proportion of live births per marriage. Of these the most important is thought to be syphilis. However, the connection between the others—particularly pneumonia and influenza—and abortions and miscarriages is to be inquired into carefully in each case.

The above considerations are sufficient to indicate the importance of careful and painstaking efforts to obtain full information on these two questions of sexual habits and birth control.

In regard to the first of these, detailed information concerning the actual frequency of sexual intercourse during the individual's entire sexual life is desired. To record once or twice per month for a man or a woman aged forty-five, fifty or sixty years, while possibly or even probably correct for these particular ages, fails by far to give the true history of this physiological activity in this person. The other ages are of equal or of even greater importance.

It is suggested that in the case of married women the investigator start with the present time and then trace back by ten-year periods in even decades (i. e., 20–29, 30–39, etc.) to the beginning of married life. In the case of widows the questioning should begin with the period immediately previous to the husband's death. Multiple marriages are to be treated, of course, on the same plan as outlined above. In the case of unmarried women who acknowledge having been pregnant, begin with the particular set of sexual activities leading to the first impregnation, and follow the lead backward and forward from that.

The natural sequence for these questions (in the case of women) is that of the history sheets—i. e., after the information regarding pregnancies. In some cases the investigator may judge that it is best to postpone this part of the history until after the patient's confidence has been gained more securely.

The question of birth control is best approached indirectly and without the use of this term. For instance, a woman of thirty-five, married fifteen years, childless or with one or two pregnancies only, should be asked first to what she attributes her failure to have more children, and the question of preventive measures will follow naturally. Similarly a woman of many pregnancies may be asked very naturally if she had never been advised or tempted to try preventive measures to avoid such frequent conceptions.

Frequent abortions or miscarriages, particularly in the absence of some disease of the generative organs, or of those previously mentioned, are suggestive of contraception, and the woman should be questioned carefully concerning them, especially as to whether any of them followed taking medicine or the use of mechanical measures.

Failure to conceive is frequently, though by no means always, associated with gonorrhea, puerperal salpingitis, uterine displacements, etc.

Men are to be questioned with the same tact, but with a somewhat wider latitude, especially in regard to premarital intercourse and extramarital intercourse; also in regard to impotence.

Checking the statements of man and wife independently is of great importance.

When contraceptive measures are acknowledged to have been used, find out exactly and record what they were, and the patient's opinion of their effectiveness. This applies, of course, to persons of both sexes. It is probable that with women the questions of frequency of sexual intercourse and birth control will be approached best in connection with the number of pregnancies, or with failure to conceive, as the case may be.

It is important here, as throughout the examination, to impress the patient with the idea that these questions may have a bearing upon his or her particular illness.

Inquiry into the state and functioning of the bodily organs con-

cludes with affections of the skin, the nervous, and the locomotor systems.

Anomalies are covered under the headings allergic history, congenital malformations, tumors, and bleeding.

This anamnesis concludes with detailed information regarding gain and loss of weight, duration and quality of sleep, and the use of tobacco and alcohol.

The *physical examination* of the patient (*Forms A* 5 to *A* 8 inclusive) covers the systematic inquiry into the normal and pathological anatomy of most of the organs and other structures (the larynx and the genito-urinary organs excepted) which are readily susceptible to examination in a medical clinic by inspection, palpation, percussion, and auscultation, with the aid, where necessary, of such relatively simple instruments of precision as the thermometer, the watch, the reflex hammer, the stethoscope, the sphygmomanometer, and the ophthalmoscope.

It will be noted on *Forms A* 5 and *A* 6 that, in respect of the scope and order of subject matter, this examination follows closely the clinical history recorded on Forms A 1–4.

The relatively considerable space and detail devoted to the heart and the arteries are demanded. both by the comparative importance of these organs and by the facility with which their anatomy and physiology may be investigated by simple means. Results of laboratory examinations (routine for urine and blood Wassermann tests) are recorded on *Form A* 7. In addition to those listed, other tests applicable to special cases, as for instance spinal fluid tests in syphilitics, blood clotting time and blood cell counts for hemophiliacs, etc., are recorded here.

On account of the unique opportunity afforded by the retinal arteries for observing the state of minute arterial vessels, *Form A* 8, covering the examination of the eye grounds, was incorporated.

Particular attention is directed to the fact that in all these forms for recording the data of both clinical history and physical examination, provision is made for indicating answers to questions in the most simple, direct, and unambiguous manner. Wherever it is practicable the record is made by simply drawing a circle about "yes" or "no," or about one out of several alternatives.

A much more detailed and elaborate system of blanks, on a similar general plan to those illustrated above, has been devised and perfected by Dr. Halbert L. Dunn,[5, 6] for use in making clinical records in hospitals or in private practice. His system is also developed thoroughly for mechanical tabulation of the records by the Hollerith machine described later in this chapter. The reader interested in such medical record forms should also consult the account of the successful installation of such a system in the Cincinnati Children's Hospital given by Hoyer and Mitchell.[7] It is interesting to note that these latter authors do not recommend the routine installation of a mechanical tabulating system for all hospital records, although their original record forms are so drawn as to facilitate transference to punch cards at any future time if this seems desirable for particular investigations.

It is to be hoped that the student will realize that the idea of such blank forms for medical records as have been illustrated here, or have been devised by Dunn, is not a new one. If he should harbor any such erroneous idea he may well read, as a single corrective example, Chapter VII "On the Mode of Investigations and Recording Cases" in a treatise on the diseases of the ovaries, published in 1873, by the distinguished gynecologist and obstetrician, T. Spencer Wells.* There he will find illustrated an excellent set of blanks, embodying all of the essential principles discussed in this chapter of the present book.

The blanks so far discussed have been of the type for which it is intended that the blank shall be filled out by an expert (in the forms discussed, by a physician) at the time of, and as the result of a personal conference with the subject. There will now be given an example of a different type of record form, to be filled out by the subject himself. In the nature of the case such record forms must always be simpler and less technical in character than those of the first type.

The record form illustrated in Figs. 22–25 inclusive has been used for some five years in an investigation of human longevity.[8] It has been circulated chiefly to *living* persons ninety-five or more

* Wells, T. Spencer: Diseases of the Ovaries: Their Diagnosis and Treatment, New York (Appleton), 1873.

years of age, and to a smaller extent to living persons between ninety and ninety-five years of age. The blank is printed on good

THE JOHNS HOPKINS UNIVERSITY
INSTITUTE FOR BIOLOGICAL RESEARCH

Investigation of Longevity

By filling in the information asked for on this form, you will be greatly aiding the program of our investigation as to the factors which influence longevity. If for any reason you are yourself unable to write in the information desired, will you not please get someone in your household to fill it in for you. This form, after filling out, should be returned in the addressed stamped envelope enclosed herewith to DR. RAYMOND PEARL, Institute for Biological Research, 1901 East Madison Street, Baltimore, Maryland.

NAME

ADDRESS

WHERE WERE YOU BORN?   WHEN WERE YOU BORN?

If born abroad in what year did you COME TO THIS COUNTRY?

How OLD were you when you came?

How many BROTHERS did you have?   How many SISTERS did you have?

Are any of your brothers and sisters alive now?

If so, give name and address.

How MANY TIMES have you been MARRIED?   What was your AGE when MARRIED?

DATE of MARRIAGES?

Give NAME of your first husband - wife.

How old was he - she at death?

When did he - she die (date)?

Give NAME of your second husband - wife.

How old was he - she at death?

When did he - she die (date)?

Was your HUSBAND'S - WIFE'S FAMILY especially LONG-LIVED?
          (Give any particulars that you know of.)

How many CHILDREN have you had?   BOYS?   GIRLS?

If you were married more than once specify how many CHILDREN BY EACH HUSBAND - WIFE

How many of your CHILDREN are NOW LIVING?

How many GRANDCHILDREN have you had?

How many of your GRANDCHILDREN are NOW LIVING?

How many GREAT-GRANDCHILDREN have you had?

PLEASE TURN OVER

Do Not Write on This Side of Vertical Line

Fig. 22.—First page of longevity record form.   Reduced facsimile.

quality bond paper, on both sides of two sheets, making a four-page leaflet, $8\frac{1}{2}$ x 11 inches in size.

The greatest *desideratum* in a blank to be filled out by the subject himself is *clearness*, so that there can be no mistaking the

YOUR FAMILY		
RELATIVE	Put in this column the AGE OF THE RELATIVE AT DEATH, OR if LIVING, the AGE NOW.	Put in this column the CAUSE OF DEATH so far as you know it. If person is not dead, write "LIVING"
Your FATHER		
Your MOTHER		
Your FATHER'S FATHER		
Your MOTHER'S FATHER		
Your FATHER'S MOTHER		
Your MOTHER'S MOTHER		
Your CHILDREN (list each child separately by NAME) here	Put in this column the AGE OF EACH CHILD AT DEATH, or if LIVING, the AGE NOW.	Put in this column the CAUSE OF THE CHILD'S DEATH, so far as you know it. If child is not dead, write "LIVING".
1st		
2nd		
3rd		
4th		
5th		
6th		
7th		
8th		
9th		
10th		
11th		
12th		
Your BROTHERS and SISTERS (list each one separately) NAME here	Put in this column the AGE OF EACH BROTHER AND SISTER AT DEATH, or if LIVING, the AGE NOW.	Put in this column the CAUSE OF THE BROTHER'S OR SISTER'S DEATH. For those not dead, write "LIVING".

Do Not Write on This Side of Vertical Line

Fig. 23.—Second page of longevity record form.   Reduced facsimile.

meaning of the question to which an answer is wanted.   Another point to be kept in mind is to ask for some of the same information, in diverse ways, in different parts of the blank, in order to have a

check on the care and reliability with which the blank has been filled out.

PERSONAL HABITS AND HEALTH

To what extent have you USED ALCOHOLIC BEVERAGES during your life?

WINE?   BEER?

WHISKEY or other SPIRITS?

To what extent and in what form have you USED TOBACCO?

PIPE?   CIGARS?

CIGARETTES?   CHEWING?

SNUFF?

How has your HEALTH been generally throughout life?

Have you ever had MEASLES?   SCARLET FEVER?   WHOOPING COUGH?

TYPHOID FEVER?   MALARIA?   SMALL POX?   PNEUMONIA?

DIPHTHERIA?   GOITER?   OTHER SERIOUS ILLNESS?

Have you ever undergone a SURGICAL OPERATION?

If so, please state its NATURE, and the DISEASE for which it was undertaken?

WHAT WAS YOUR AGE AT THE TIME?

Please state any other DETAILS ABOUT YOUR HEALTH which you think might be of interest.

What have been your general HABITS during life as to EATING, DRINKING, SLEEPING and WORKING?

TO WHAT DO YOU CHIEFLY ATTRIBUTE YOUR LONG LIFE?

PLEASE TURN OVER

Do Not Write on This Side of Vertical Line

Fig. 24.—Third page of longevity record form.   Reduced facsimile.

Whenever the nature of the investigation permits it, there are certain definite advantages in having the original records made on *card* forms, rather than in record books, or loose leaves of paper.

If the records are on cards the work of subsequent tabulation of the data is greatly facilitated. Furthermore, the problem of filing the records for ready reference is simplified.

RESIDENCE, OCCUPATION, ETC.

In what PLACES have you RESIDED at different times in your life?

Have you LIVED mostly in the COUNTRY or CITY?

What OCCUPATIONS have you followed at different times during life?

To what extent have you done HARD MANUAL LABOR?

What is your RELIGIOUS FAITH?

To what RACE STOCK (English, Scotch, Irish, German, French, etc.), do you chiefly belong?

What is your HEIGHT?   AVERAGE WEIGHT?

How has your WEIGHT CHANGED since you were 25 years of age?

What, in general, has been your BUILD DURING ADULT LIFE?

    A. THIN AND LEAN?

    B. MODERATELY THICK-SET OR CHUNKY?

    C. DISTINCTLY FAT?

Color of hair at age 25?   Now?

Color of eyes?

Were you a blond or a brunette?

BY WHOM WAS THIS BLANK FILLED OUT?

WHAT IS YOUR RELATION TO

PLEASE GIVE ME THE NAME AND ADDRESS OF ANY OTHER RELATIVE WHO MIGHT BE ABLE TO FURNISH ADDITIONAL OR MISSING INFORMATION

DATE WHEN THIS BLANK WAS FILLED OUT

PLEASE TURN OVER

Do Not Write on This Side of Vertical Line

Fig. 25.—Fourth page of longevity record form. Reduced facsimile.

An example of a *card* form for original records is shown in Figs. 26 and 27. This is printed on medium weight card stock, of the

best quality. It is 5 x 7 inches in size. This form was intended to be filled out by physicians in obstetrical clinics, to provide information about normal fertility and birth control, as actually prac-

OBSTETRICAL SERVICE No.		PARA.		EDUCATION (PUT CHECK AGAINST ONE GROUP)	ILLITERATE / SECONDARY SCHOOL / HIGH SCHOOL / COLLEGE

COLOR	RACE STOCK	AGE	YEAR MARRIED	HAS PATIENT ANY DISEASE OF UROGENITAL SYSTEM? IF SO, SPECIFY.
W. C.				

REPRODUCTIVE HISTORY

HAS PATIENT EVER USED ANY METHOD FOR PREVENTION OF CONCEPTION? YES. NO.

PREGNANCY	YEAR	RESULT	
1		M. L. S.	IF SO, SPECIFY METHOD OR METHODS USED, IN AS GREAT DETAIL AS POSSIBLE.
2		M. L. S.	
3		M. L. S.	
4		M. L. S.	
5		M. L. S.	HOW LONG WAS EACH OF ABOVE SPECIFIED METHODS PRACTISED? MAKE ANSWER AS DETAILED AS POSSIBLE.
6		M. L. S.	
7		M. L. S.	
8		M. L. S.	WHAT IS PATIENT'S OPINION AS TO EFFECTIVENESS OF METHOD OR METHODS?
9		M. L. S.	
10		M. L. S.	
11		M. L. S.	HAS PATIENT EVER HAD SELF-INDUCED OR OTHER ABORTION?
12		M. L. S.	
13		M. L. S.	
14		M. L. S.	WARD OR PAY PATIENT? ECONOMIC AND SOCIAL POSITION? POOR, LOWER MIDDLE CLASS, UPPER MIDDLE CLASS, RICH.
15		M. L. S.	
16		M. L. S.	
17		M. L. S.	

M--MISCARRIAGE OR ABORTION. L--LIVE BABY. S--STILL-BORN BABY.   (OVER

Fig. 26.—Face of card record form for collecting data on fertility and contraception. Reduced facsimile.

INSTRUCTIONS. THE USUAL METHODS OF PREVENTING CONCEPTION FALL IN THE FOLLOWING GENERAL CLASSES: COITUS INTERRUPTUS (WITHDRAWAL); CONDOM; PESSARY; ABSTINENCE FROM INTERCOURSE DURING PART OF MONTH; VAGINAL DOUCHES, PLAIN OR MEDICATED; MEDICATED SUPPOSITORIES. IN FILLING OUT BLANK BE SURE TO GIVE CLEAR DETAILS AS TO WHICH OF ABOVE METHODS, OR OTHER METHOD, THE PATIENT HAS PRACTISED. UPON THE DEFINITENESS AND PRECISION OF THE INFORMATION ON THIS POINT DEPENDS THE SIGNIFICANCE OF THE DATA FOR THE RESEARCH PLANNED.

REMARKS:

Fig. 27.—Reverse of card record form shown in Fig. 26.

ticed, which might serve in some part as an independent check upon statements reported from time to time by birth control organizations.*

* The writer will be glad to furnish a supply of these card forms to the obstetrical department of any hospital willing to co-operate in the investigation, and return the blanks properly filled in.

By criticising and improving the record forms which are given as examples in this chapter the student will learn more in a practical sense of the principles involved in the construction of such blanks than can be imparted by any amount of didactic precepts.

THE PRESERVATION OF CASE HISTORIES

Turning to the question of the way case histories are handled after they are written, which is essentially a matter solely of business or office management and not of medicine or science, there are two defects in the common practice. These relate, first, to the fixation of responsibility for the recording of each item in the history, and, second, to the filing of the completed histories. From every point of view, whether of administration, research or other, it is of the highest importance that future students of a hospital's records should know who is responsible for statements appearing in a history. How often has one heard long and inconclusive debates as to what interpretation was to be put upon some statement in a history as to a clinical finding? The decision really depended upon who originally was responsible for the statement. If it were the considered verdict of the wise and experienced old professor, it was one thing; if it were the snap judgment of the latest intern, it was quite another. All this difficulty can be removed by inaugurating and practising the principle that every sheet of a history shall bear upon its face the names of the person or persons responsible for what appears upon that page. Perhaps a word of caution needs to be added lest there should be some misunderstanding. Fixation of responsibility is not to be construed as an excuse for any weakening of the rigid canons of extreme objectivity in history or protocol writing, now generally taught in all first-class medical schools.

The purpose of filing case histories is twofold: first, to preserve them, and, second, to do it in such a way as to make them most readily accessible to anyone who may in the future want to consult them. There can be no question that this latter purpose will best be served by the so-called "unit system" of case histories, in which the hospital's complete record about any one individual forms one separate and distinct volume. The advantages of this method of preserving histories over the far more common system of binding

them up in great volumes in numerical or temporal sequence, are so obvious as not to need detailed exposition. Such a method of handling the completed records is really essential to their most efficient utilization, whether for statistical, investigational, or any other purpose.

MECHANICAL TABULATION

There are certain items of information which ought to be and generally are intended to be included in every case history. Some of these routine items are:

1. Case number.
2. Service number.
3. The patient's name.
4. Diagnosis.
5. Sex.
6. Social status (single, married, widowed, divorced).
7. Age.
8. Occupation.
9. Body weight.
10. Stature.
11. Race.
12. Birthplace.
13. Service under which patient was treated.
14. Date of admission to the hospital.
15. Duration of stay in hospital.
16. Time from onset of diagnosed condition to admission to hospital.
17. Condition at admission.
18. General health of patient prior to present illness.
19. Whether there is any family history of the diagnosed disease.
20. Whether a first entry or a readmission.
21. Whether a free, a paying, or a part-paying case.
22. Condition at discharge.
23. Whether or not an autopsy was performed.
24. Autopsy number, if any.
25. Nature of treatment.
26. Complicating pathologic conditions, additional to the one diagnosed.

In an ideal system of handling such records each history should be completely cross-indexed under each one of the following items in the above list at least: 1 to 18 inclusive, 21, 22, 23, 24, 25. Of course, nothing like such complete cross-indexing as this is even attempted, not to say accomplished.

There is only one method now known, whereby in a *practical* way such an amount of cross-indexing can possibly be accomplished. That method is to handle the routine information by the modern system of *mechanical tabulating and indexing*.* On this system the original records are transferred, by means of a machine called a "key punch" (cf. Fig. 28), to cards, the record on the card appearing as a series of punched holes. Then, by means of another
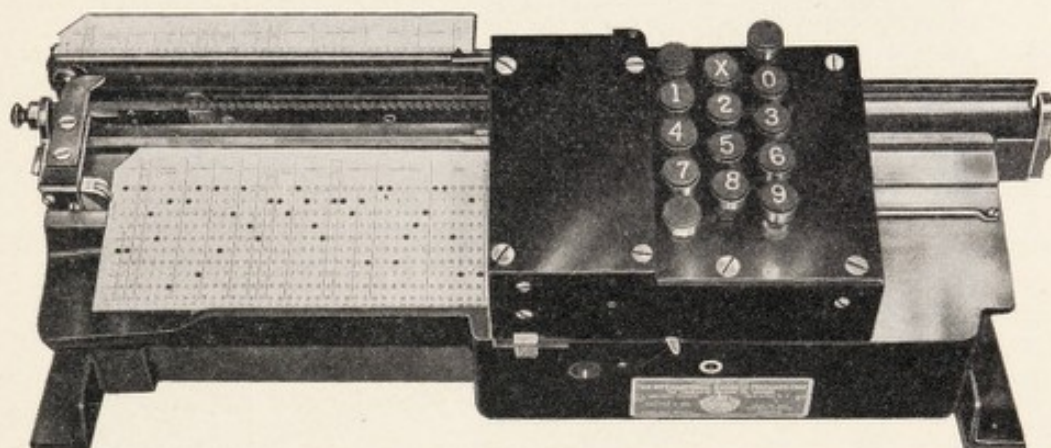

Fig. 28.—Electric key punch for transferring written records to cards to be used in mechanical tabulation and indexing.

machine, known as a "sorter" (cf. Fig. 29), the punched cards can be mechanically sorted, at a rate of about 350 to 400 cards per minute, into any desired arrangement relative to any rubric or item of information recorded upon the cards.

Let us suppose, for example, that someone wishes to assemble for study all the cases of lobar pneumonia which have been treated

* The most generally useful and flexible system of mechanical tabulation now available is that known as the Hollerith system, from its inventor, Mr. Herman Hollerith. The machines of that system are the ones illustrated here. Further information about these machines may be obtained from the manufacturers, The Tabulating Machine Company Division of the International Business Machines Corporation, 50 Broad St., New York City. It may be of interest to medical readers to know that a distinguished physician, the late Dr. John S. Billings, had a great deal to do with the initiation and early development of this invention. He was a close friend and adviser of Mr. Hollerith all through the early stages.

in the hospital. Suppose the diagnostic code number for lobar pneumonia is 102. One has then only to run the cards through the sorter relative to the field designated "diagnosis" and pick out, after the cards have been mechanically arranged in numerical order, all those bearing the punched number 102 in the diagnosis field. These 102's will all be together in one bundle, and they will be all the lobar pneumonia cases in the hospital's records. Each card
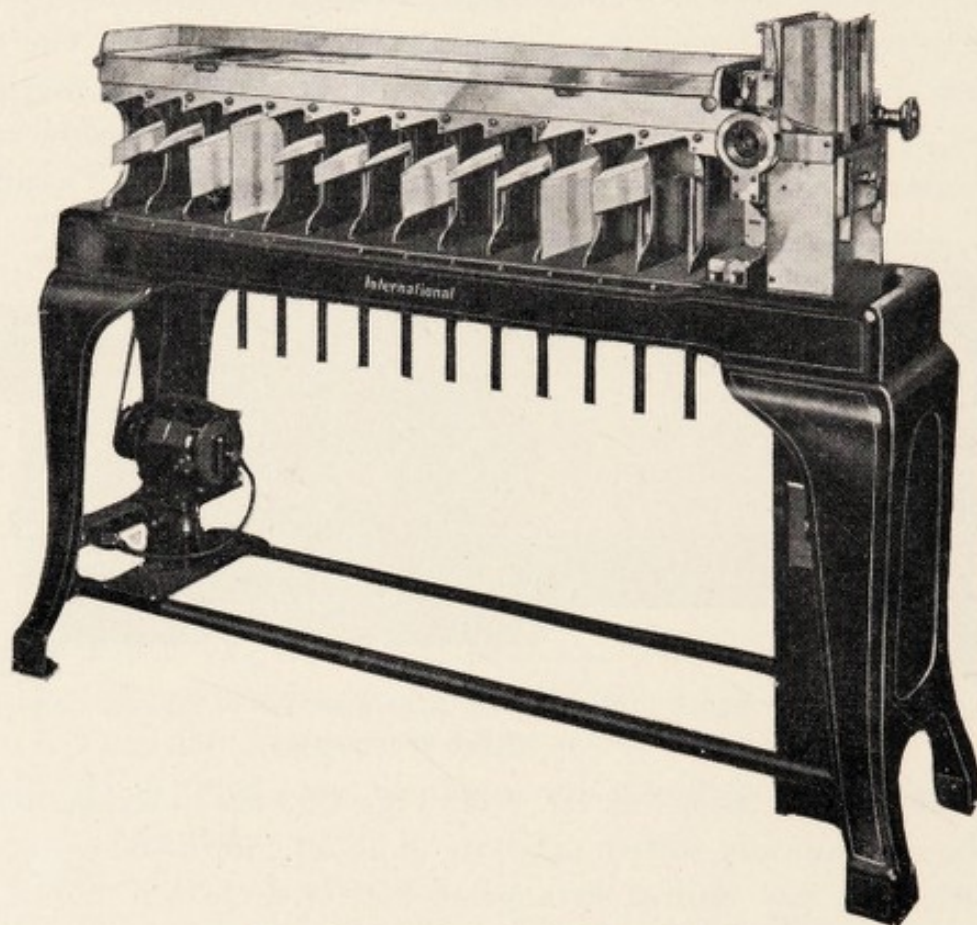
Fig. 29.—Horizontal sorting machine.

will bear the case number, from which, of course, the original histories can be consulted if one desires. If one particularly wishes to study the lobar pneumonia of negroes, he need only take his bundle of "diagnosis 102" cards, run through the sorter again relative to "race" and he will in a few moments have all the cases of this disease in negroes separated out by themselves. Suppose he is further only interested in lobar pneumonia in negro children under five years of age, say. He need only take his bundle of

negro lobar pneumonia cases and put them through the sorter again, retaining this time only those falling into ages under five. He gets his results at the rate of 350 to 400 a minute. Compare this with the laborious process that would be involved in assembling
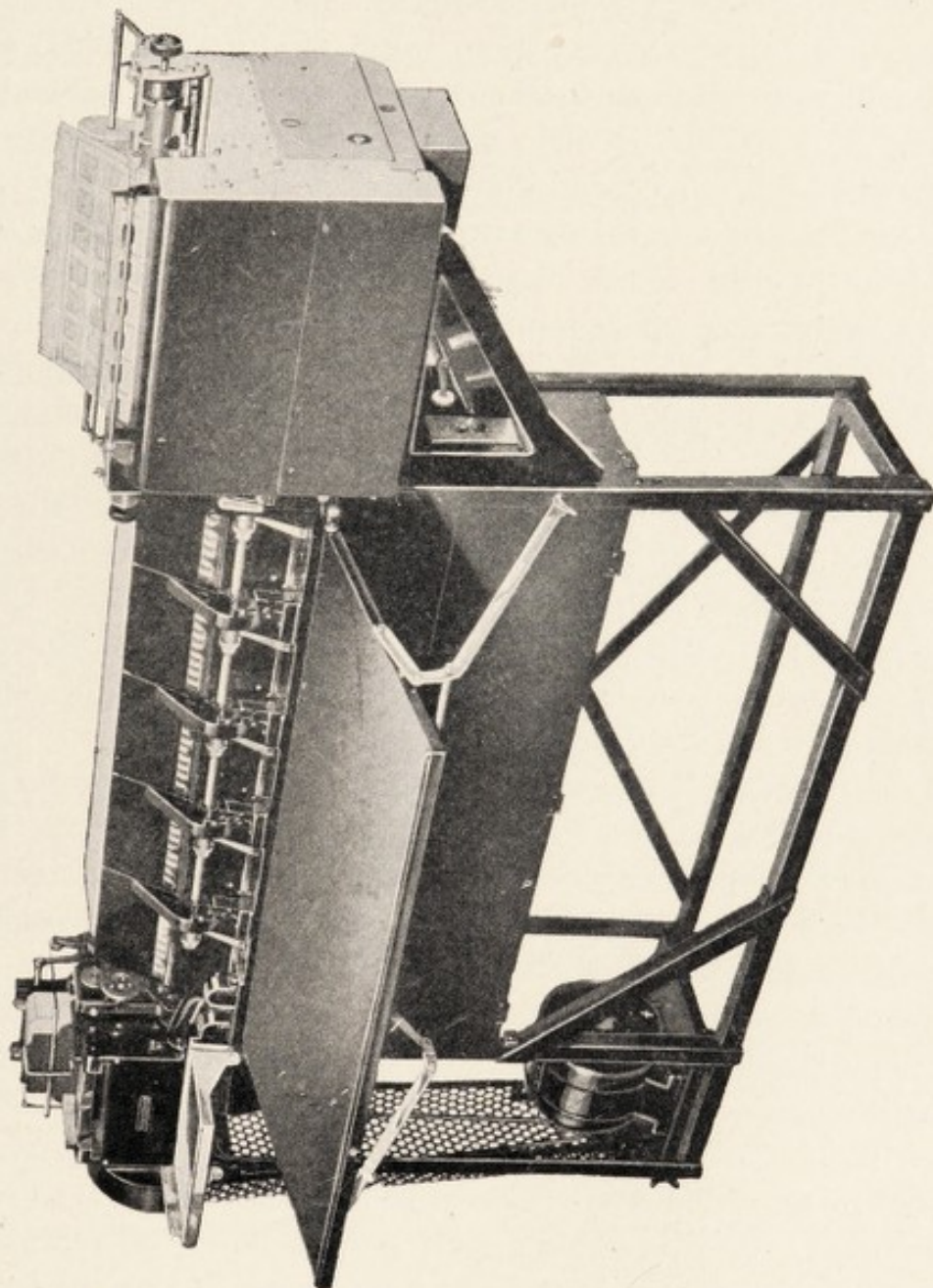

Fig. 30.—Electric accounting machine.

by hand from an ordinary card catalogue of hospital case records the case history numbers of *all* the cases of lobar pneumonia in negro children under five ever treated in the hospital. The comparison is as of hours with weeks or even months if the histories be numerous.

Again, suppose that a complete group of like case histories has been assembled by painfully laborious hand processes, and one wishes then to make a statistical tabulation of the numerical facts they contain. Weeks or months may easily be, and often are, spent upon the process. But if the records are upon punched cards, the pertinent cards, which have been mechanically assembled, need only be run again through another machine, known as a "tabulator" (cf. Fig. 30), and the results relative to any desired category of information will be mechanically counted with great rapidity and absolute accuracy, and the columns of figures will at the same time be added. At either of two stages in the process the results may be automatically taken off in printed form, from the machine, if it is desired to do so. The electric accounting machine, shown in Fig. 30, tabulates and prints the final results of the preceding operations.

It falls outside the scope of this book to go in detail into the theory and applications of mechanical tabulation. The student who wishes to become familiar with the scope and possibilities of modern mechanical tabulating will do well to apply to the Tabulating Machine Company Division of the International Business Machines Corporation, 50 Broad St., New York, for literature regarding its application in various fields.

In the statistical offices of up-to-date departments of health, and in census offices, the mechanical system of tabulating the data from birth and death certificates is employed. The economies so effected, both in time and money, are very great. The student interested in this aspect of the subject should get and study the card forms and codes used in representative health departments.

A single example of a Hollerith punch card form and its application may be given here. It is taken from Dunn and Rockwood.[6] In their paper they illustrated their original record forms by filling them out for the hypothetical case of an imaginary Mrs. H. Brown. Dr. Halbert L. Dunn has kindly given permission for the reproduction here of his discussion of the punch card form. Figure 31 represents the Hollerith card punched to represent the coded information given in the hypothetical original record form.

The first six columns indicate the case number of the chart. Case No. 1 would be punched as 000001 in columns 1, 2, 3, 4, 5, and 6, respectively. The record number

of Mrs. H. Brown, 67190, is punched as 067190 in the first six columns. The highest number which can be recorded is 999999 and allows, therefore, for a tabulation of that many separate patients.

Columns from 7 to 21 are set aside for diagnoses. Three columns, 7, 8, and 9, are allotted for the first diagnosis and two columns each are allowed for the remaining diagnoses. The first column in each diagnostic field, namely, 7, 10, 12, 14, 16, 18, and 20 can be double-punched.* In each of these first columns the blank position represents 1, the X position 2, and the zero position 3 in units of 10. It is possible by this means to indicate forty numbers in one column. The number zero will be indicated if a hole is not punched in the column, and numbers from 1 to 9 if the printed numbers from 1 to 9 are punched. If the blank position is punched, it will signify code No. 10. The numbers from 11 to 19 will be indicated by a double punch, namely, the blank position for the 1 in units of 10 and the proper unit number from 1 to 9; No. 20 by a single punch in the X position which represents 2 in units of 10; Nos. 21

Fig. 31.—An index card for the general medical examination, which has been punched for the illustrative diabetic record of Mrs. H. Brown.

to 29 by a double punch, one of which is the X position representing 2 in units of 10 and the other the proper unit number from 1 to 9; No. 30 by a single punch in the zero position which stands for 3 in units of 10 and numbers from 31 to 39 by a double punch, one of which is the zero position representing 3 in units of 10 and the other the proper unit number from 1 to 9.

The second column in each diagnostic field is punched in one position only. Eleven positions may be indicated in this column which are, respectively, the blank position, 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9. It is possible, therefore, to code into hundreds in each two-column field for any given diagnosis. The numbers would read serially as 00-blank, 000, 001, 002, 003, 004, 005, 006, 007, 008, 009, 01-blank, 010, 011, 012, etc., up to the highest number which would be 399. Counting the blanks and zeros

* In order to accord with the wiring possibilities of the printing tabulator only twenty-five columns can be double-punched. We have taken advantage of every one of these possibilities in the proposed cross-index card.

as separate numbers, this code permits the division of each diagnostic field into forty major headings in the first column with a subdivision of each of these into eleven subsidiary units in the second column. The total number of items which it is possible to list in the two-column field by this process is 440.

A survey made in several hospitals shows that usually in about 80 per cent. of the case records there is only one diagnosis. Multiple diagnoses up to four are fairly common, and over seven extremely unusual. The first diagnosis, columns 7, 8, and 9 has a third column, No. 9, which permits the subdivision of each of the 440 items in the first two columns into eleven subsidiary units allowing for 4840 items in the code of the first primary diagnosis.

If there should be multiple diagnoses of eight or more numbers for any given hospital record, the code number of these diagnoses should be written in ink on the back of the card. Any card which has been punched for seven diagnoses must be examined for written code numbers on the back. It is estimated that this event should not occur more than once in two or three hundred times.

In the record of Mrs. H. Brown, illustrated in Fig. 31, there are seven diagnoses. These diagnoses with their respective code numbers are as follows: (1) Diabetes mellitus (6 blank); (2) obesity (60); (3) diabetic acidosis (63); (4) general arteriosclerosis (80); (5) hypertension (81); (6) chronic constipation (124), and (7) hemorrhoids (126).

The diagnostic code has not been filled out to thousands and, consequently, the third column in the first diagnosis is left unpunched. The code number of diagnosis 1 (6 blank) is punched in columns 7 and 8; that of diagnosis 2 (60) in columns 10 and 11; of 3 (63) in columns 12 and 13; of 4 (80) in columns 14 and 15; of 5 (81) in columns 16 and 17; of 6 (124) in columns 18 and 19, and of 7 (126) in columns 20 and 21.

The next four columns of the punch-card from 22 to 25 represent the items of age, sex, color, outcome and civil state. Age is coded in columns 22 and 23; sex (male or female) and color (white or black) in column 24; outcome (well, improved, same, worse or dead) and civil state (single, married, widowed, divorced or separated) in column 25. For example, the age of Mrs. H. Brown is indicated as 49 in columns 22 and 23, the sex (female) and color (white) are coded as No. 2 in column 24, and civil state married and outcome improved by code No. 7 in column 25.

The main divisions of the history occupy four columns from 26 to 29; each is double-punched, using the blank key as 1, the X key as 2 and the zero key as 3 in units of ten. By this means, numbers up to forty can be indicated in the same manner as described for the first column of each diagnostic field. Column 26 represents past illnesses including abnormalities in the weight curve; 27, the respiratory, circulatory, and gastro-intestinal systems; 28, the genito-urinary and nervous system; and column 29, the routine history and the clinician's opinion of the accuracy of the history.

Only thirty-one numbers of the possible forty are used in each one of these columns. Each of the thirty-one numbers represents an abnormality or combination of abnormalities. It is possible to list five separate items or any combination of these five items by use of a combination code printed on the history. The past illness, for instance, is divided into five subdivisions. If any one of the diseases from typhoid to scarlet fever has been checked in the past illness, No. 1 is marked as positive. If the patient has had syphilis, gonorrhea or a history of abnormal weight curve,* No. 2

* The code position of the weight curve is placed with past illnesses, while its chart position naturally falls after the routine question by systems. It represents the only divergence in the arrangement by order between the index code and the printed chart.

is checked; if he has had a major operation or accident, No. 3; a nose or throat operation, No. 4; and if there is some important item in the miscellaneous, not included in the routine list of past illnesses, No. 5 is checked.

If the patient has only one of these five conditions, for instance No. 2, the coded number would be identical to the number checked on the chart. If, however, he had items 1, 2, and 4 checked, this combination would be indicated by code No. 17 in column 26 of the index card.

In Fig. 31 code No. 26 is punched in column 26 representing positive observations in the past illness in items 1, 2, 3, and 4. Likewise, code No. 17 in column 27 represents positive observations in items 1, 2, and 4.

Body height, grouped by classes, is given in column 30, and body weight, also by classes, in column 31. Mrs. H. Brown has a stature of 64 inches (162.56 cm.) represented by code No. 7 in column 30, and a body weight of 160 pounds (74.4 kg.) indicated by code No. 7 in column 31.

The physical examination occupies eight columns, from 32 to 39, each one of which is double-punched, so that it can represent numbers up to 40. All of the code numbers in the physical examination stand for abnormalities. In column 32, abnormalities of the head and face are noted; 33, of the mouth and throat; in 34, of the neck, spine and thorax; in 35, of the chest and lungs; in 36, of the heart; in 37, of the vessels and abdomen; in 38, of the extremities and neurologic symptoms and 39, of the lymph nodes, skin, genitalia, rectum, and abnormal psyche.

Abnormal laboratory observations are indicated in columns 40 to 43.

Columns 44 and 45 have been set aside for miscellaneous conditions and may be assigned in any way desired by a specific institution. We suggest a certain arrangement which may or may not be followed. In this arrangement, four conditions are coded in column 44 which might exist in any diagnosis or any case, namely, autopsy, major operations, minor operations, and previous admissions. This leaves a fifth blank space for some special interest still unassigned. Column 45 could be reserved for the indication of the principal service on which the patient was treated. Many hospitals would not desire to make such a distinction between their records.

The punch-card form illustrated in Fig. 31 is on the old standard 45-column card. There is now available for the tabulating machine equipment illustrated an 80-column card of the same dimensions as the 45-column card. The obvious advantage of this is that a much greater amount of information can be put upon a single card.

Out of some twenty years' experience with mechanical tabulation in various fields the writer may perhaps be permitted to state briefly his considered evaluation of it for scientific research purposes. Wherever the problem to be dealt with is one either of (a) cross-indexing a mass of data so that all original records falling in a particular category of manifold characteristics may be quickly picked out from a file containing a large mass of diverse, separate

records, or (*b*) tabulating a mass of *purely and inherently numerical observations*, mechanical tabulation is without any rival. It can do these two jobs more quickly, more accurately, and in every way better than any other known mode of procedure. If the investigator has problems involving either of these types of operation he would be as foolish not to employ punched cards and mechanical tabulation as he would be if he insisted on making his journeys about the country in a stage coach in preference to the railroad train or the automobile.

When, on the other hand, an investigator has to deal with data which are *inherently not numerical* in character, but have to be made so artificially by some process of coding, the case is not quite the same. The punch card then automatically destroys all qualifications, shadings or half-tones in the original written records. The system is inherently rigid. Long observation indicates that while the young investigator does not much mind this feature, the older, more critical, and perhaps wiser investigator prefers to make his tabulations of inherently qualitative, "judgment," data directly from the original record, rather than from a punch card, which can only tell him, in each particular case, the rigid category into which he, or somebody else, at some time decided that a particular observation was to be put. The real point seems to be that the investigator's point of view frequently—and rightly—changes during the course of a long investigation. As he penetrates deeper into a mass of observational material he sees meaning and relationships in it, of which he had no conception at the start. These things alter his views as to the significance and disposition of particular individual observations. But transferral of the records *via* the code route to punch cards, if the system is to display its potential smooth and accurate efficiency, must take place at the *beginning* and not at the end of the investigation. Experience shows that one becomes more and more cautious about choosing this route, and more and more inclined to adopt other devices, which, while retaining the possibility of immediate reference to the original written record about qualitative observations in each individual instance, adopt some of the features of the punch card system which facilitate rapid and accurate tabulation.

A simple form designed for precisely this purpose is shown in Fig. 32.

This card form was devised for investigations on autopsy records.[10] In each of the spaces following an organ designation the essential statements about the lesions of that organ, as given in the original protocol by the pathologist, are copied. By the use of

Fig. 32.—Reduced facsimile of card form for autopsy records.

fine handwriting a large amount of information can be put on the face of the card. If still further detail is wanted, the back of the card is available.

The card is $8\frac{1}{2}$ x 11 inches in size. The upper left-hand corner is clipped off to facilitate stacking. The cards are printed on heavy card stock of four different colors, to make easy the distinction of sex and color of the patients in tabulating. Records for white

male patients are put on white cards; for white females on pink cards; for colored males on green cards, and for colored females on yellow cards.

The purpose of the numbered cells around the edge of the card is to furnish the guides for the accurate punching of holes with a hand punch, in order to facilitate the sorting out of groups of like cards. For example, every card which records a case showing any malignant neoplasm has a hole punched where the circle is printed in cell No. 50 at the bottom of the card. This makes it possible to assemble at any time, quickly and accurately, all the cases of malignancy in the material.

SUGGESTED READING

1. Pearson, K.: On the Mathematical Theory of Errors of Judgment, with Special Reference to the Personal Equation, Phil. Trans. Roy. Soc., A, vol. 198, pp. 235–299, 1902.
2. Yule, G. U.: On the Influence of Bias and Personal Equation in Statistics of Ill-defined Qualities, Jour. Anthropol. Inst., vol. 36, pp. 325–381, 1906.
3. Pearl, R.: The Personal Equation in Breeding Experiments Involving Certain Characters of Maize, Biol. Bull., vol. 21, pp. 339–366, 1911.
4. Pearl, R., A. C. Sutton, W. T. Howard, Jr., and M. G. Rioch: Studies on Constitution, I. Methods, Human Biology, vol. 1, pp. 10–56, 1929.
5. Dunn, H. L.: The Status of Statistical Medicine, Jour. Amer. Med. Assoc., vol. 89, pp. 1273, 1274, 1927.
6. Dunn., H. L., and R. Rockwood: A Record System Suitable for Both Clinical and Statistical Medicine, Arch. Int. Med., vol. 41, pp. 499–535, 1928.
7. Hoyer, B., and A. G. Mitchell: Records and the Record System of the Children's Hospital, Cincinnati, Ohio. Methods and Problems of Med. Educ., 14th Ser., pp. 152–207, 1929.
8. Pearl, R.: Preliminary Account of an Investigation of Factors Influencing Longevity, Jour. Amer. Med. Assoc., vol. 82, pp. 259–264, 1924.
9. Pearl, R.: Modern Methods in Handling Hospital Statistics, Johns Hopkins Hosp Bull., vol. 32, pp. 184–194, 1921.
10. Pearl, R., and A. L. Bacon: Biometrical Studies in Pathology. IV. Statistical Characteristics of a Population Composed of Necropsied Persons, Arch. of Pathol. and Lab. Med., vol. 1, pp. 329–347, 1926.
11. Hollerith, H.: An Electric Tabulating System, School of Mines Quarterly (Columbia Univ.), April, 1889. (Contains an account of the plans and machines as originally developed for the 1890 census of the United States.)
12. Hollerith, H.: The Electrical Tabulating Machine, Jour. Roy. Stat. Soc., vol. 57, pp. 678–682, 1892. (An early account of the system by its inventor.)
13. Knight, F. H.: Mechanical Devices in European Statistical Work, Quart. Publ. Am. Stat. Ass., vol. 14, pp. 596–598, 1915. (A survey of the extent to which mechanical tabulation has become established in European statistical offices.)

14. Menzler, F. A. A.: The Census of 1921; Some Remarks on Tabulation, Jour. Inst. Actuaries, vol. 52, pp. 341–384, 1920–21.   (An account of the mechanical tabulation of the 1921 census of England and Wales.)

15. Health Report of the Royal Air Force for 1920, Lancet, March 25, 1922 pp. 598–601, and April 1, 1922, pp. 655–657.   (An account of the punch-card tabulation of their medical data.)

16. Pearl, R.: To Begin With.   Being Prophylaxis Against Pedantry, second edition New York (Alfred A. Knopf, Inc.), 1930.

CHAPTER VI

GRAPHIC REPRESENTATION OF STATISTICAL DATA

VALUE OF STATISTICAL DIAGRAMS

DIAGRAMS properly constructed and intelligently used constitute one of the most potent tools in the statistician's armamentarium. Even the most seductively constructed and arranged table of statistics will not convey the story which inheres in the figures with anything like the neatness and despatch attainable by graphic presentation.

The graphic side of statistical work has received a great deal of attention in recent years and there are several excellent treatises available, dealing solely with this subject (see reading list at the end of this chapter). Any detailed treatment of the subject is impossible in the space available here. The attempt will be only to set forth a few of the most elementary principles, and to introduce the reader to the more detailed literature.

GENERAL CHARACTERISTICS

Before developing the structure and uses of different types of statistical diagrams it is desirable to say a word about their underlying general characteristics.

All statistical diagrams are representations of points, lines, surfaces or solids, the positions of which in space are quantitatively defined by a system of co-ordinates.

These co-ordinates may be of various sorts. The most common sort are rectangular co-ordinates. Here a point p in a plane (Fig. 33) has its position defined (as indicated by the dotted lines) in terms of the x and y axes of reference.

The distance from o to the dotted line on the horizontal axis is known conventionally as the *abscissa* of the point p. The distance on the vertical axis from o to the dotted line is known as the *ordinate* of the point p. The horizontal or x axis is the *abscissal axis*. The vertical or y axis is the *axis of ordinates*, or the ordinal axis. Gen-

erally and usually in plotting statistical data to rectangular axes the classes of things are laid off as abscissæ, and the frequencies of these classes as ordinates. This, however, is only a convention, and not a law of nature.

Besides rectangular co-ordinates, there are sometimes used in statistical diagrams:

(a) Angular co-ordinates (as in "pie" diagrams).
(b) Polar co-ordinates.
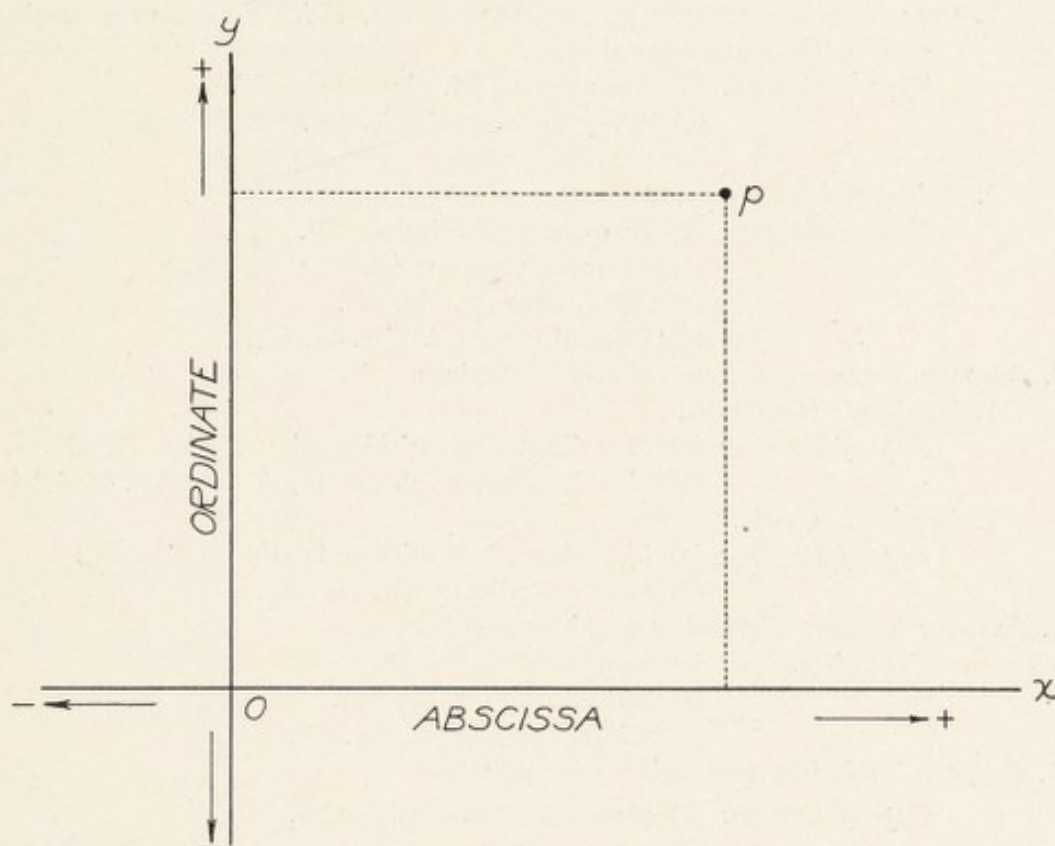
Fig. 33.—Diagram to illustrate rectangular co-ordinates. *o* is the origin. The arrows indicate the conventional directions relative to algebraic signs.

(c) "Geographical" co-ordinates (as in a statistical map, where latitude and longitude are the axes of reference, really angular co-ordinates which may become rectangular by projection to a plane).

TYPES OF DIAGRAMS

The first question which anyone should ask himself who feels an impulse to make a statistical diagram is this: What is to be

the fundamental purpose of this diagram? What is the essential
point that it is intended to convey to the viewer? The answer to
this question virtually settles the type of diagram to be employed,
because there is a rather definite adaptation of diagram types.
Some types of diagrams are much better fitted than others to the
telling of particular kinds of statistical stories.

Consider the following scheme:

A. *Purpose:* To represent *frequencies* of things or events.
    1. Categories or attributes of qualitative things, which do not vary continu-
        ously in the mathematical sense.
        Type of diagram: (*a*) Bar diagram (cf. Figs. 34 and 35).
                    (*b*) "Pie" diagram (cf. Fig. 36).
                    (*c*) Frequency polygon (Figs. 40, 41).*
    2. Things which vary continuously.
        Type of diagram: (*a*) Histogram (cf. Figs. 37–39).
                    (*b*) Frequency polygon (cf. Figs. 40, 41).
                    (*c*) Ogive curve (cf. Fig. 42).
                    (*d*) Integral curve (cf. Figs. 43, 44).
B. *Purpose:* To represent *trends* of events or things.
    1. In Time. Non-cyclic.
        Type of diagram: (*a*) Line diagram on arithlog grid (cf. Figs. 47, 48).
                    (*b*) Line diagram on arithmetic grid (cf. Figs. 45–47).†
    2. In Time. Cyclic.
        Type of diagram: (*a*) Line diagram on arithmetic grid (cf. Fig. 49).
                    (*b*) Polar co-ordinates (cf. Fig. 50).
C. *Purpose:* To show *distribution* of things or events.
        Type of diagram: (*a*) Spot map (cf. Fig. 51).
                    (*b*) Shaded map (cf. Fig. 52).
                    (*c*) Scatter diagram (cf. Fig. 53).
D. *Purpose:* To facilitate or replace computation.
        Type of diagram: (*a*) Nomogram (cf. Figs. 54–56).

Bar Diagrams

The bar diagram is the simplest possible picture of a statistical
situation. Figure 34 is a bar diagram‡ showing the proportion

    * Strictly speaking, the frequency polygon belongs here rather than under 2*b*,
where it is also listed. The rigid statistical purist will use the frequency polygon only
to depict discontinuous variation. But in actual statistical practice it always has
been, and probably will be, usefully employed as a substitute or alternate for the
histogram, especially where it is desired to compare graphically several frequency dis-
tributions in the same diagram.

    † The logistic curve (cf. Chapter XVII) is a special case of this type of diagram.

    ‡ From R. Pearl, The Nation's Food, Philadelphia, 1920, p. 237. Data on which
diagram is based are there given.

which each of the more important foods contributes to the total protein consumed in the United States by human beings.

PERCENTAGE CONTRIBUTION TO TOTAL PROTEIN CONSUMED
PER CENT
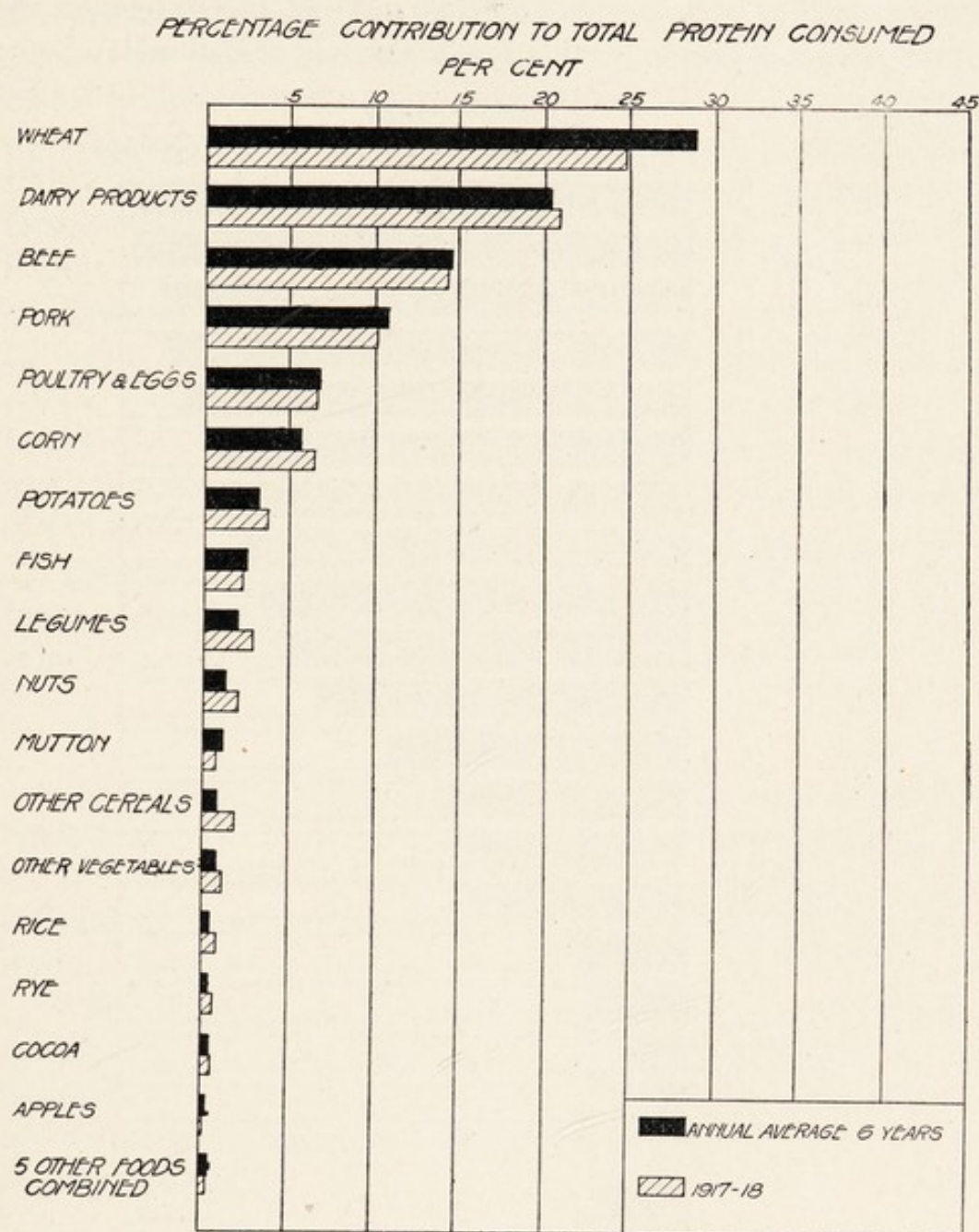
Fig. 34.—Diagram showing the percentage of the total protein consumed in the United States contributed by each of 23 commodities. The solid bars denote the average consumption in the six years preceding our entry into the war. The cross-hatched bars denote the consumption in 1917 and 1918.

From this diagram one sees at a glance the relative significance of the great staple foods in furnishing protein for human consump-

tion. Wheat stands first.   Beef contributes roughly one-half as much
protein to the national dietary as wheat, and poultry and eggs about
half as much as beef, etc.   The whole story of the sources of the
protein we, as a people, consume is accurately visualized.
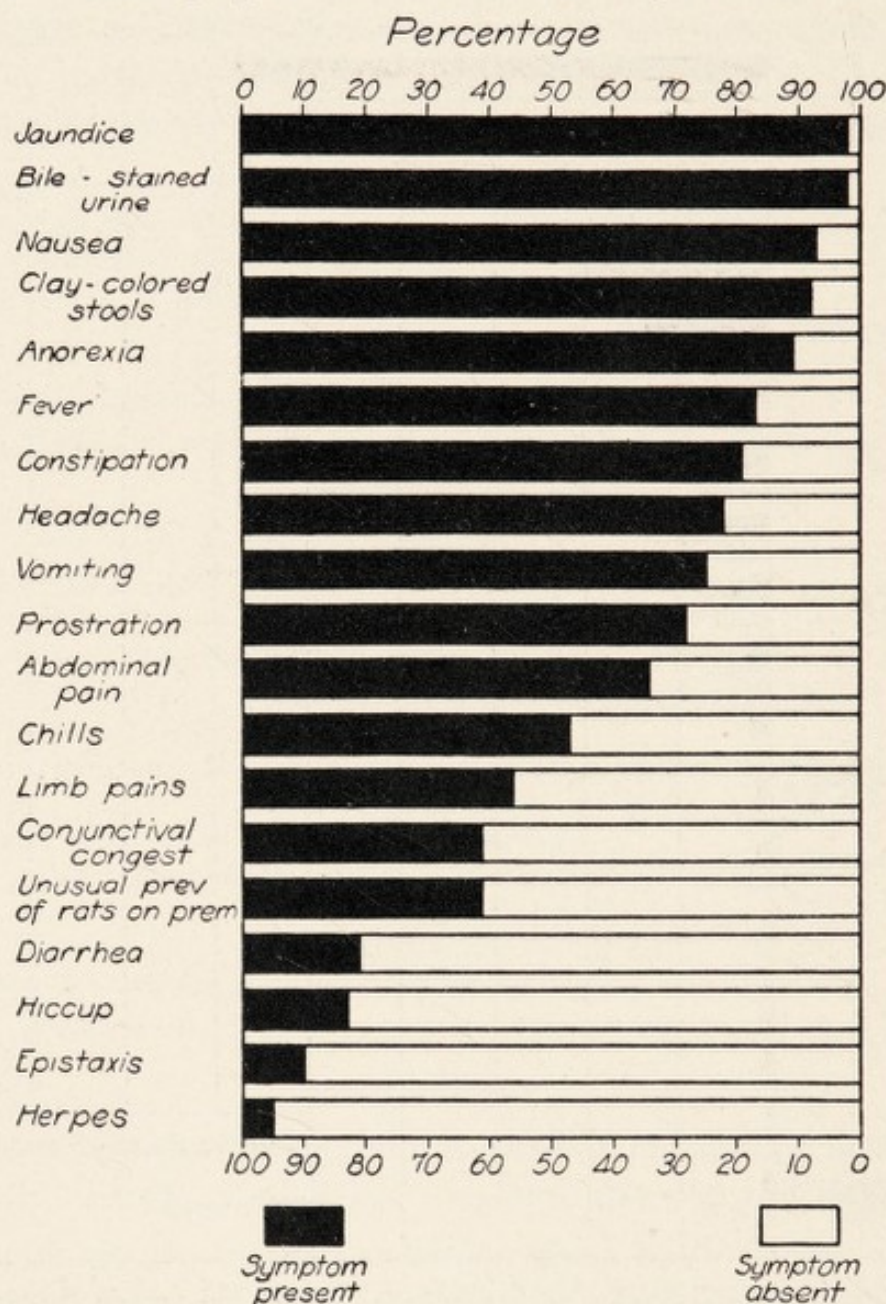
Fig. 35.—Bar diagram based upon data of Table 8, Chapter IV, showing the relative
frequency of different symptoms in epidemic jaundice.

The percentage columns of Table 8 in Chapter IV make the bar
diagram shown in Fig. 35.   This is a slightly different form of bar
diagram from that shown in Fig. 34.

Bar diagrams find perhaps their most appropriate field of usefulness in the graphic representation of *discontinuous* variates, as is illustrated in the two examples here given. Wheat and dairy products are discontinuous, discrete entities; one cannot start from wheat and by a series of minute continuous steps or gradations pass to dairy products. Similarly, jaundice and nausea are physically discontinuous phenomena. Hence it is appropriate to represent them graphically by physically separate bars. The case is quite different with continuous variates. It is possible to pass continuously by successive, unbroken small steps from a height of 60 inches say to a height of 65 inches. Hence it is proper to represent such phenomena graphically by continuous lines. One frequently sees bar diagrams in which each bar represents a physically discrete phenomenon or entity, but in the diagram the ends of the bars have been connected by a line. This is bad practice. Its absurdity is evident if one tries to read a point on the line in terms of abscissal or ordinal units. What is the meaning of something half-way between wheat and dairy products?

"Pie" Diagrams

For a reason which will be perfectly obvious to all American readers, and which foreign readers have no occasion to be interested in, sector diagrams plotted to angular co-ordinates are called colloquially "pie" diagrams. An example of such a diagram is seen in Fig. 36.

While this form of diagram is extremely popular, especially in exhibit work, I agree entirely with Brinton that it is a far less desirable type than the simple bar diagram. Its use should probably be confined strictly to popular presentation, as in exhibit and propaganda work.

Histograms, Frequency Polygons, Ogives.

It will be desirable to consider this group of graphic forms together, and because of their importance and frequent use the methods of their construction from the original data will be treated in detail. As material for this study of graphic representation the data of Table 10 may be used. This table gives the head heights

in millimeters of 68 male inmates of the Haddington District Asylum in Scotland, as reported by Tocher* (p. 39).

The data of Table 10 (p. 171) are simply a list of observations just as originally presented by Tocher. To make them into usable statistics they must first be converted into a frequency distribution in which like head heights will be brought together. This is done in Table 11 (p. 172).
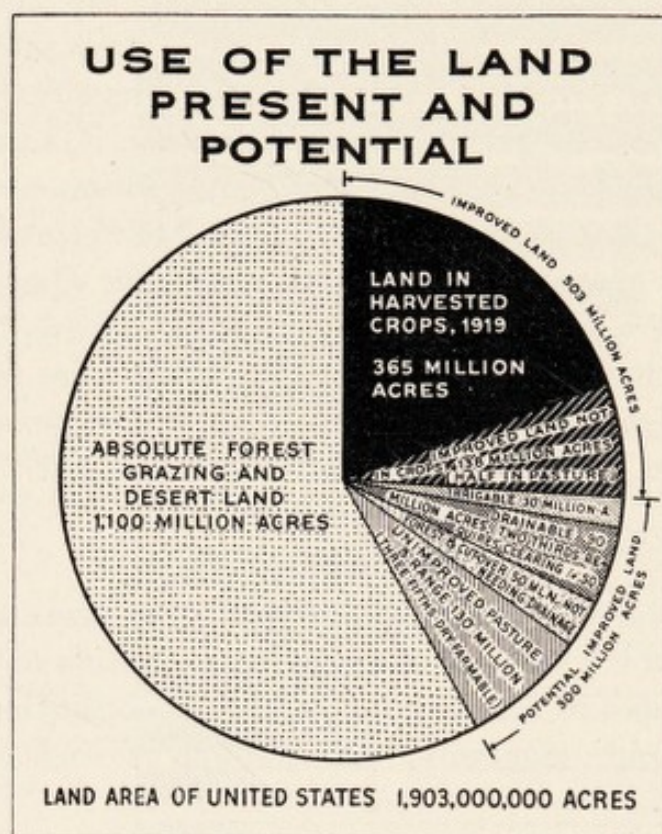
Fig. 36.—Example of diagram to angular co-ordinates. (Reproduced by permission of Dr. O. E. Baker and the editor of the Geographical Review from an article by Dr. Baker entitled "Land Utilization in the United States: Geographical Aspects of the Problem," published in the Geographical Review, vol. 13, January, 1923.)

It is evident that the extent of variation is so great in this character height of head that a class unit of 1 mm. is too fine. It is necessary to group the material into larger class units. This is done in the third column of the table, headed "Frequencies grouped in 5 mm. classes." The class limits are taken to begin on the even 5 and 10 mm. points.

* Tocher, J. F.: Anthropometric Survey of the Inmates of Asylums in Scotland. Henderson Trust Reports, vol. i, Edinburgh, 1905.

"Histogram" is the name given by Pearson to the correct graphical representation of frequency distributions. In these diagrams the class limits are laid off on the abscissal axis, and the frequencies over each abscissal element are given as the *areas* of

TABLE 10

TOCHER'S DATA ON HEAD HEIGHT OF MALE INMATES OF HADDINGTON DISTRICT ASYLUM

Patient No.	Head height, mm.	Patient No.	Head height, mm.
1	137	35	142
2	144	36	139
3	132	37	138
4	131	38	129
5	131	39	139
6	144	40	137
7	145	41	139
8	155	42	126
9	125	43	145
10	146	44	143
11	143	45	133
12	152	46	137
13	137	47	143
14	134	48	125
15	140	49	139
16	137	50	131
17	142	51	119
18	138	52	134
19	150	53	143
20	141	54	149
21	129	55	136
22	137	56	150
23	129	57	141
24	140	58	131
25	130	59	143
26	143	60	129
27	141	61	131
28	126	62	145
29	134	63	133
30	138	64	134
31	139	65	125
32	144	66	138
33	128	67	130
34	138	68	134

rectangles erected on these base elements. So long as the base elements (that is, sizes of the classes into which the material is grouped) are all equal, then obviously the *heights* of the rectangles will be proportionate to the frequency.

Suppose now we plot as a histogram the data of the first (ungrouped) half of Table 11. The result will be that shown in Fig. 37. Now it is at once evident that Fig. 37 is an inadequate and

TABLE 11

FREQUENCY DISTRIBUTION OF HEAD HEIGHTS FROM TABLE 10

Head heights, mm.	Ungrouped frequencies.	Frequencies grouped in 5 mm. classes.	Class limits for group frequencies, mm.
119	1	1	115–119
120			
121			
122		....	120–124
123			
124			
125	3		
126	2		
127		10	125–129
128	1		
129	4		
130	2		
131	5		
132	1	15	130–134
133	2		
134	5		
135			
136	1		
137	6	17	135–139
138	5		
139	5		
140	2		
141	3		
142	2	16	140–144
143	6		
144	3		
145	3		
146	1		
147		5	145–149
148			
149	1		
150	2		
151			
152	1	3	150–154
153			
154			
155	1	1	155–159
Totals	68	68	—

misleading graphical representation of the important facts about variation in head height in this group of people. It is a long, flat thing with many gaps and only roughly indicates what general

sorts of head heights occur most frequently.   The grouping, in
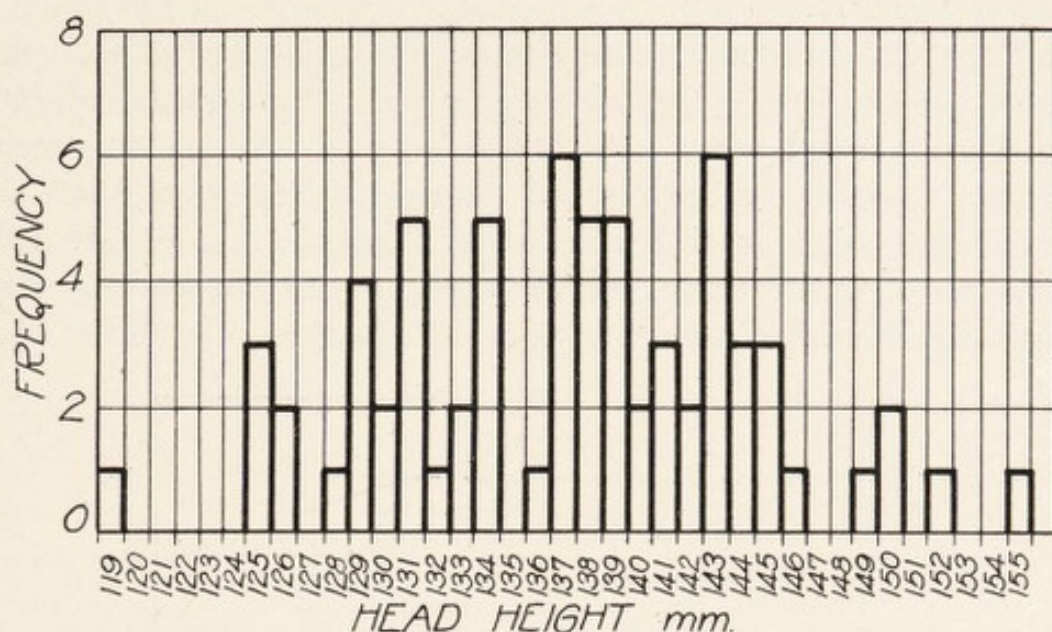
Fig. 37.—Histogram of ungrouped frequencies of head height from Table 11.
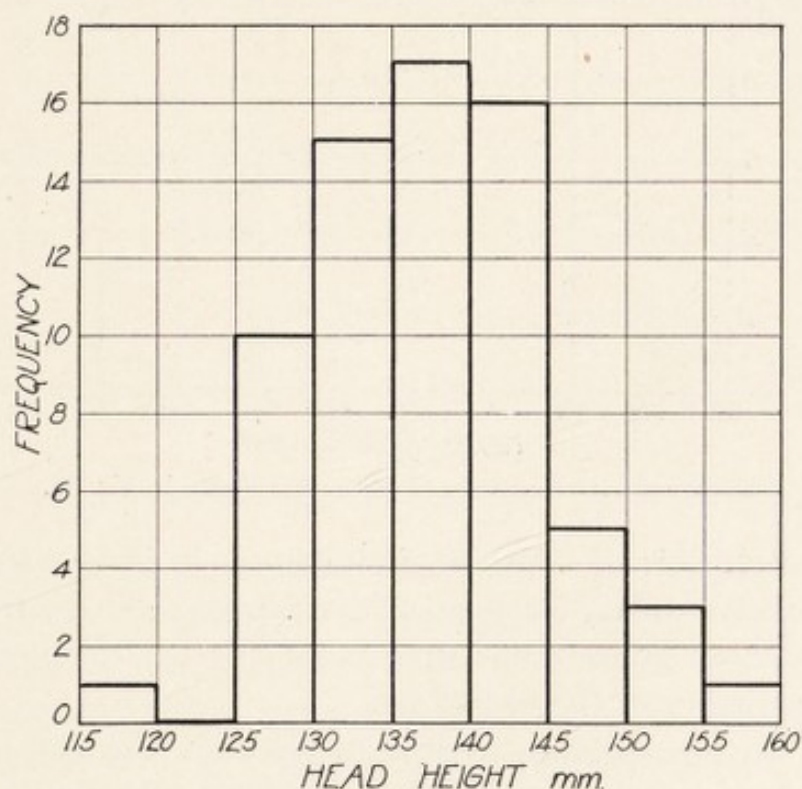
Fig. 38.—Histogram of grouped frequencies of head height from Table 11.

short, is too fine for so small a sample as 68.   A much clearer and more adequate idea of the real state of the case is given in Fig. 38,

which is a histogram plotted from the grouped data of the latter half of Table 11.

From this diagram an adequate picture is obtained of the real distribution of head heights in this group. The skewness of the distribution is apparent. Another example of a histogram is seen in Fig. 78 *infra*. A method of drawing a histogram which is preferred by some statisticians is that shown in Fig. 39. It will be seen to consist simply in the omission of that part of the vertical grid work of the drawing which lies below the top of the
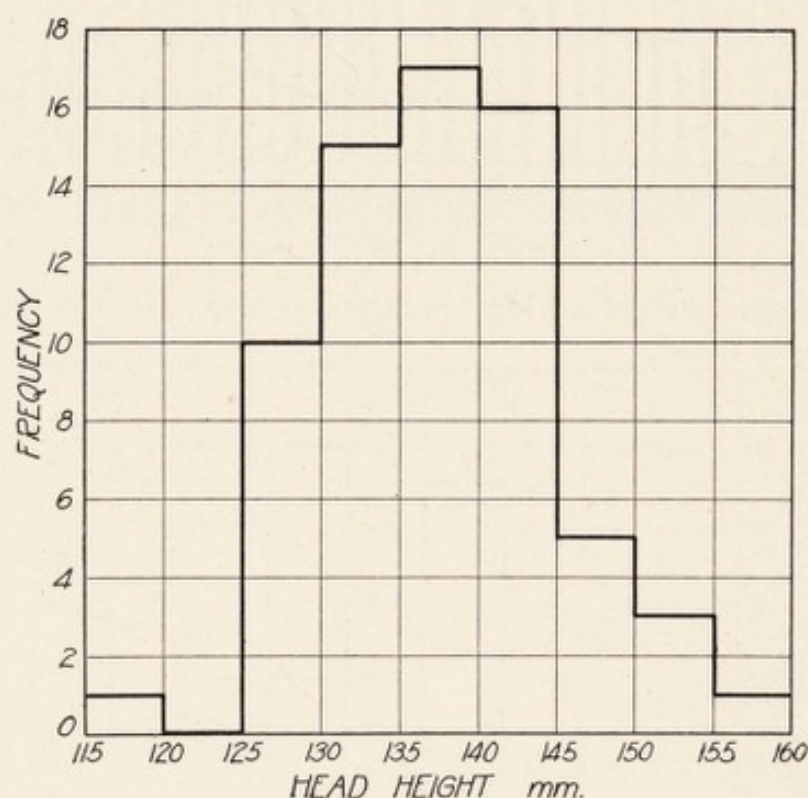
Fig. 39.—Alternative form of histogram shown in Fig. 38.

lower of each pair of adjacent rectangles. It is an attempt to realize the advantages, for comparative purposes, of the frequency polygon without at the same time sacrificing the complete mathematical accuracy of the histogram.

While the histogram is, on theoretic grounds, the most accurate method of graphically representing frequency distributions, it is sometimes more practically useful to represent them as frequency polygons.

A *frequency polygon* is the result that one gets by assuming

that the total frequency in any given class is concentrated at the center of that class, and then plotting ordinates of height proportionate to the frequencies supposed concentrated at those midpoints. The histogram of Fig. 38 is shown plotted as a frequency polygon in Fig. 40.

The frequency polygon is less accurate than the histogram because it does not truly represent the frequency areas over the base elements. But it is an extremely useful form of frequency diagram for *comparative* purposes. It may be employed freely
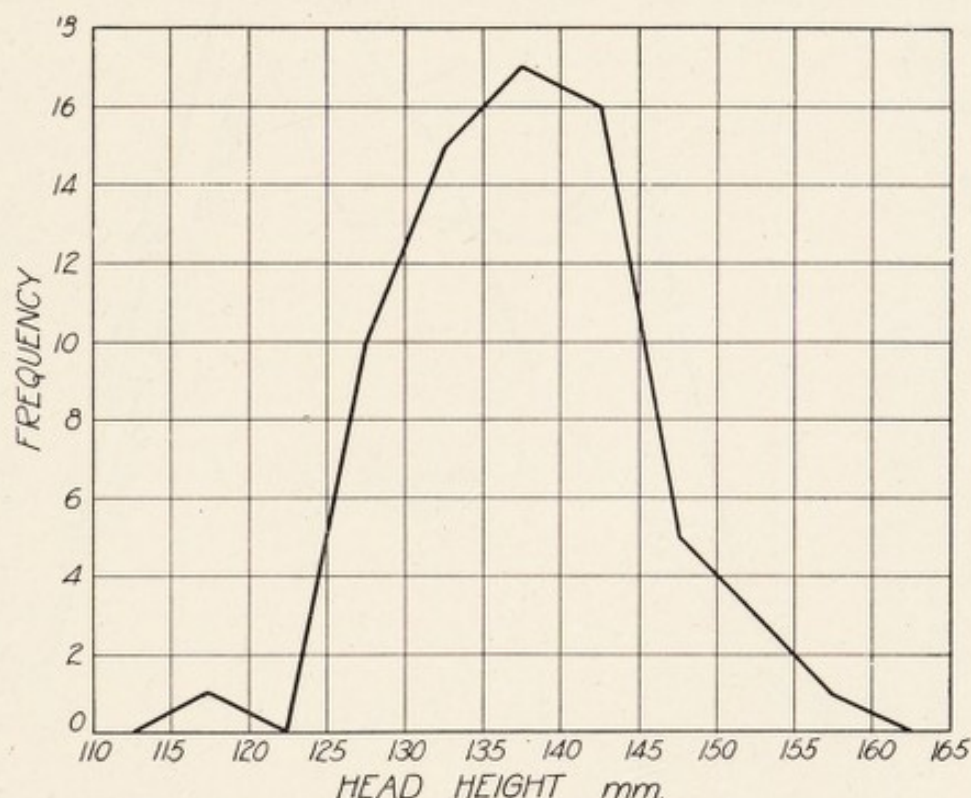
Fig. 40.—Frequency polygon of grouped frequencies of head heights from Table 11.

in place of the histogram where the only object is to give a general picture to the eye of a series of overlapping frequency distributions. An example of such comparative use is shown in Fig. 41.

Another method of representing frequency distributions graphically was devised by Galton, and the resulting type of curve was called by him the "ogive." It is the sort of curve which would be got if 1000 men taken at random were arranged in a row in order of their heights, beginning with the shortest at one end, and ending with the tallest at the other. If now a smooth line be

imagined just touching the top of the head of each man in the row, this line would be an ogive curve, in Galton's sense. The data of Table 11 are plotted as an ogive curve in Fig. 42.

It is seen that in this curve the head heights in millimeters are now taken as ordinates, and at equal intervals along the abscissal axis there is erected an ordinate for each of the 68 individuals.
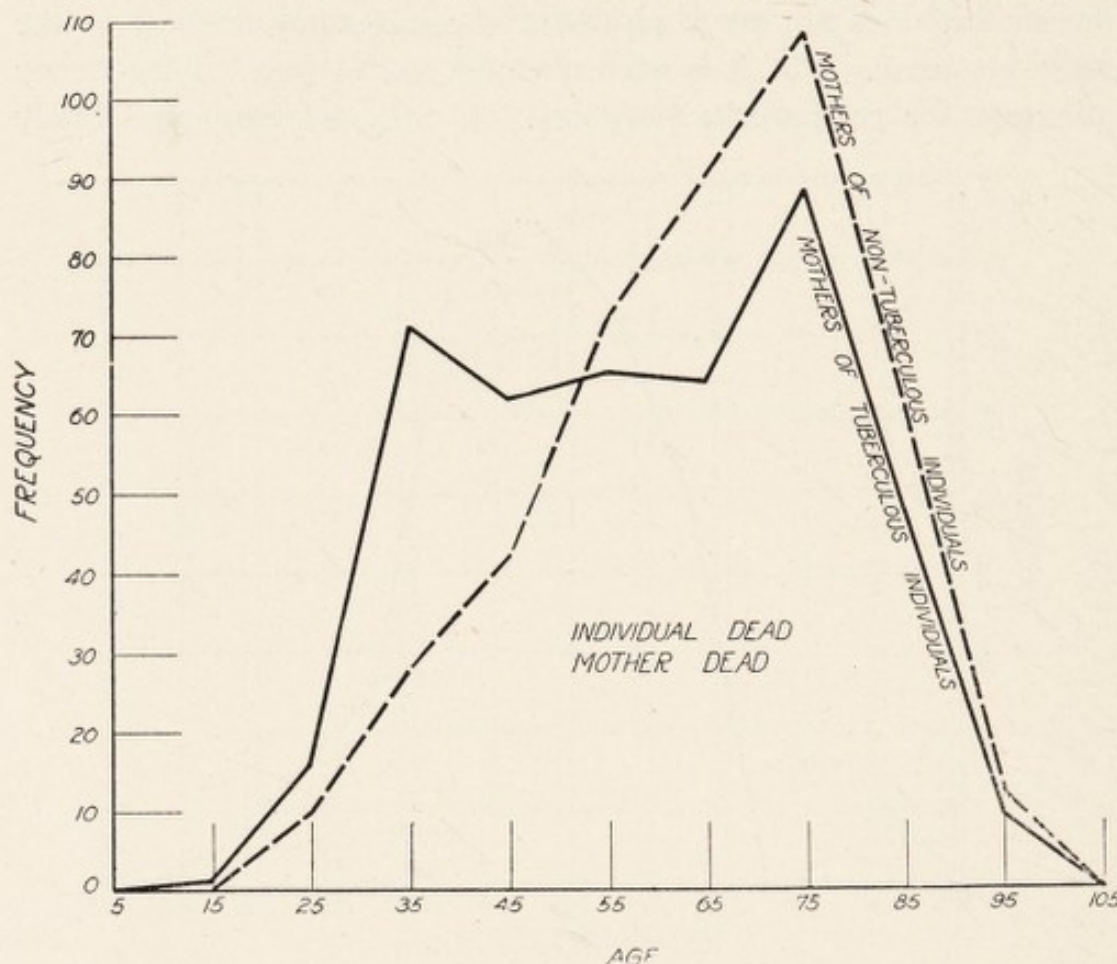
Fig. 41.—Frequency polygons showing the age distribution of dead mothers of dead (a) tuberculous (solid line) and (b) non-tuberculous (broken line) individuals. (Reproduced from Pearl, R., "The Age at Death of the Parents of the Tuberculous and the Cancerous," Amer. Jour. Hygiene, vol. 3, pp. 71–89, 1923.)

If a larger number of individuals were involved the curve would be smoother. The curve is seen to be like the mirror image of an enormously stretched out and elongated S, or an integral sign, lying on its back.

Integral or Cumulated Frequency Diagrams

So far in the discussion of the graphic representation of frequencies, we have plotted the value of each single frequency, by

itself, against its proper abscissa. Let us consider now the *integral* or accumulated diagram of frequency. In this case the frequency is successively *accumulated*, class by class, from the lower range

TABLE 12

CUMULATED FREQUENCY DISTRIBUTIONS, ABSOLUTE AND PERCENTAGE, OF THE HEAD HEIGHTS FROM TABLE 11

Head height, mm.	Cumulated frequencies.	
	Observed.	Percentage.
119	1	1.5
120	1	1.5
121	1	1.5
122	1	1.5
123	1	1.5
124	1	1.5
125	4	5.9
126	6	8.8
127	6	8.8
128	7	10.3
129	11	16.2
130	13	19.1
131	18	26.5
132	19	27.9
133	21	30.9
134	26	38.2
135	26	38.2
136	27	39.7
137	33	48.5
138	38	55.9
139	43	63.2
140	45	66.1
141	48	70.6
142	50	73.5
143	56	82.3
144	59	86.7
145	62	91.1
146	63	92.6
147	63	92.6
148	63	92.6
149	64	94.1
150	66	97.0
151	66	97.0
152	67	98.5
153	67	98.5
154	67	98.5
155	68	100.0

end on. The data of Table 11 are put in this form in Table 12. The integral curve plotted from the data of Table 12 is shown in Fig. 43.
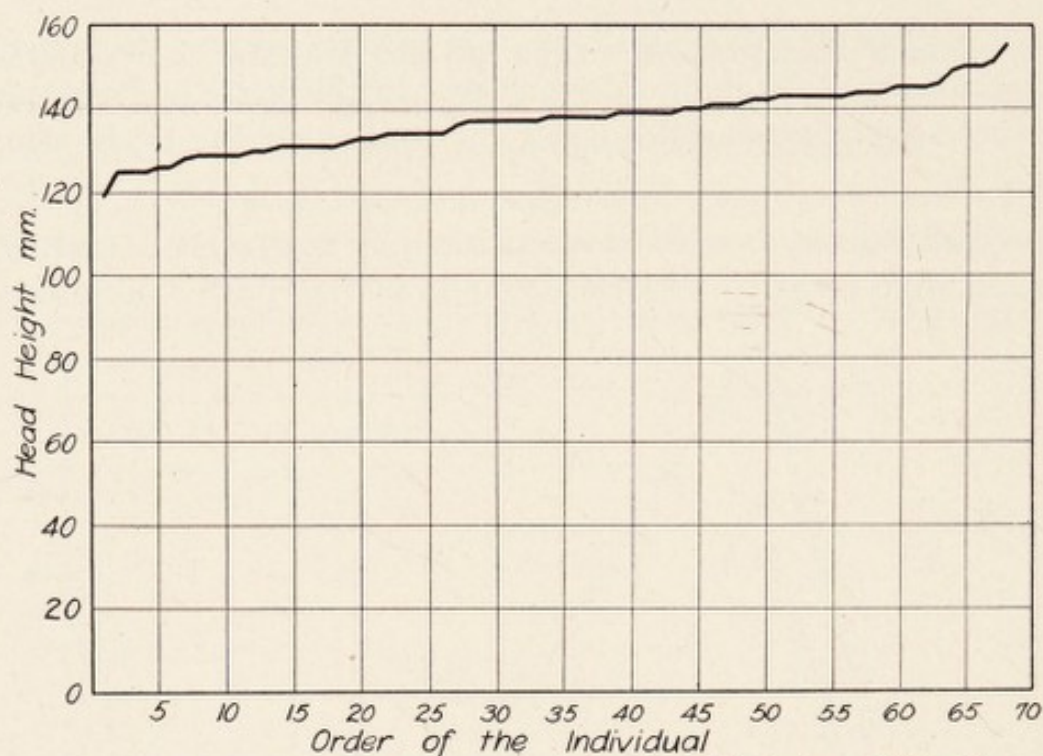
Fig. 42.—Ogive of ungrouped frequencies of head height, from Table 11.
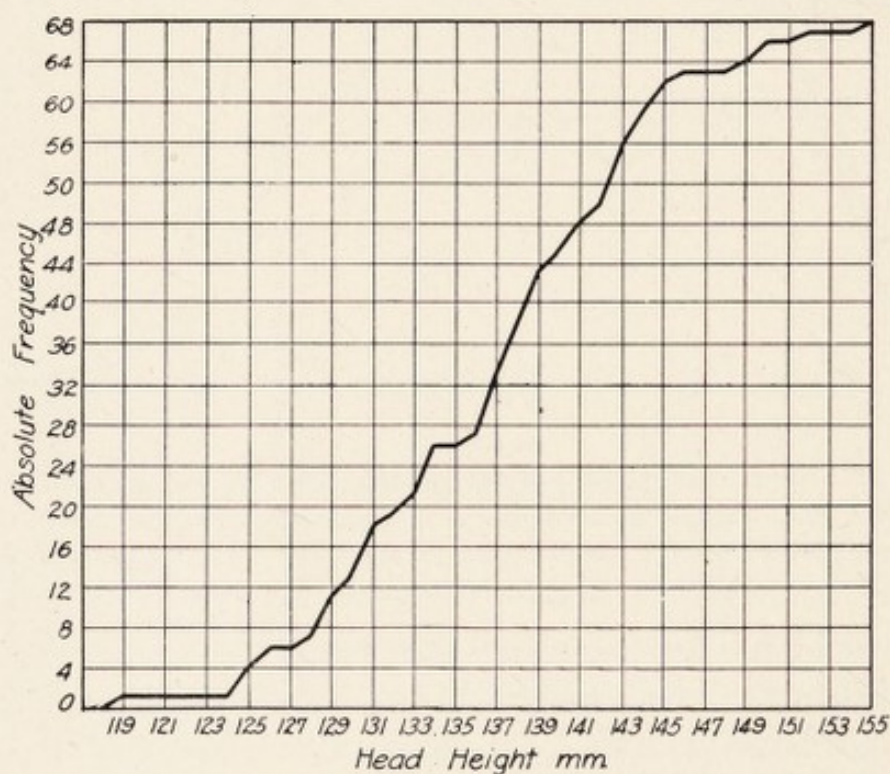
Fig. 43.—Integral curve of ungrouped frequencies of head height from Table 12.

This form of diagram shows the number of individuals having a head height greater or smaller than any assigned value. This

property is often useful. This integral form of diagram may, by a simple device discussed in detail by von Huhn[3] be made to show relative as well as, and along with, absolute accumulated frequencies. In Fig. 43, 68 individuals are 100 per cent. of this particular group or sample. Suppose, then, there is set up on the right-hand margin a division of the ordinal distance (= 68 individuals = 100 per cent.) into 10 equal parts. This scale will then be a percentage or relative scale, while that on the left-hand
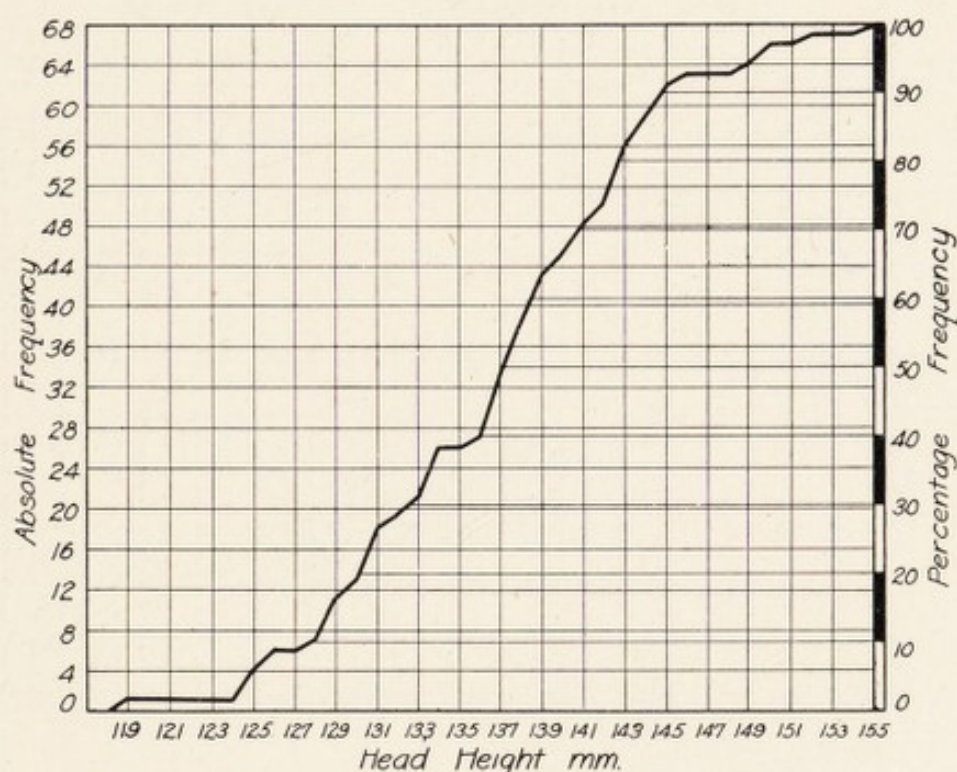
Fig. 44.—Like Fig. 43, but with added scale of relative or percentage frequencies.

margin still remains an absolute scale for frequencies in the same group. The resulting diagram is shown as Fig. 44.

The advantages of this form of diagram are at once apparent. It is seen, for example, that 90 per cent. of the group had head heights under 145 mm.; 10 per cent. were under 128 mm. in head height, etc. In a wide range of cases plotting in this manner will obviate all necessity of calculating percentages.

The student will note that the ogive and integral forms of plotting a frequency distribution are fundamentally the same. The only essential difference between Figs. 42 and 43 is that in the

case of the ogive (Fig. 42) frequencies are plotted along the abscissal axis, and in the integral (Fig. 43) along the y axis as usual. Also the scale of plotting is a little different in the two diagrams.

Non-cyclic Time Trend Diagrams

One of the commonest uses of the graphic method in statistics is to show the trend of events in time. The obviously simple way to do this is to make a *line diagram* with time as abscissa and the
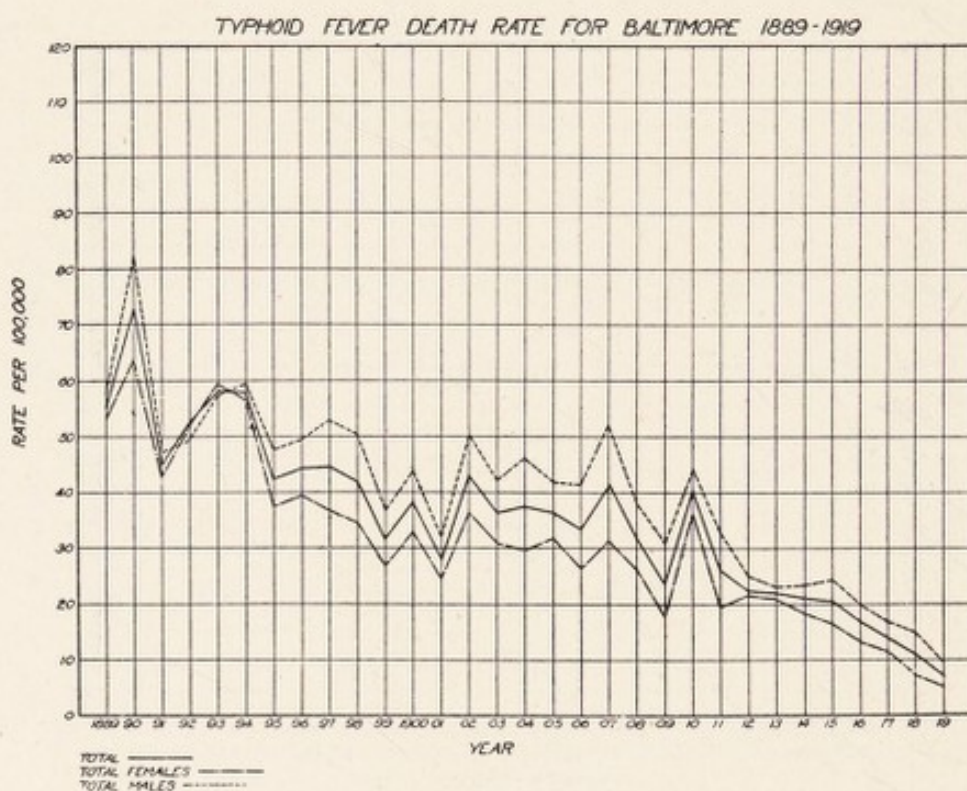
Fig. 45.—Death-rate from typhoid in Baltimore 1889–1919 inclusive for males, females, and total population. (From Howard, W. T., "The Natural History of Typhoid Fever in Baltimore, 1851–1919," Johns Hopkins Hospital Bulletin, vol. 31, pp. 276–286, 319–334, 1920.)

frequency of occurrence of the event in question as ordinate. Thus suppose it is desired to show the decline in the death-rate from typhoid fever in Baltimore from 1889 to 1919 inclusive. A diagram like that shown in Fig. 45 may be prepared.

Now it would appear at first glance that this diagram gave an adequate representation of the facts. We see the line indicating a decline in the rate from about 55 to under 10 in the period covered. But actually the diagram is visually misleading. Why and how

it is so will now be shown. Suppose we wish to *compare* the decline in the death-rate from tuberculosis of the lungs with that in the death-rate from typhoid fever. Let us transfer from Baltimore as a universe of discourse to the United States Registration Area. In Table 13 are given the death-rates per 100,000 in the original registration states (Connecticut, Indiana, Maine, Massachusetts, Michigan, New Hampshire, New Jersey, New York, Rhode Island, and Vermont, and the District of Columbia) for each year from

TABLE 13

DEATH-RATES PER 100,000 POPULATION IN THE ORIGINAL REGISTRATION STATES
1900 TO 1920 INCLUSIVE

Year.	a Tuberculosis (all forms).	b Typhoid fever.
1900	195.2	31.3
1901	189.8	27.5
1902	174.1	26.3
1903	177.1	24.6
1904	188.5	23.9
1905	180.9	22.4
1906	177.8	22.0
1907	175.6	20.5
1908	169.4	19.6
1909	163.3	17.2
1910	164.7	18.0
1911	159.0	15.3
1912	149.8	13.2
1913	148.7	12.6
1914	148.6	10.8
1915	146.7	9.2
1916	143.8	8.8
1917	147.1	8.1
1918	151.0	7.0
1919	124.9	4.8
1920	112.0	5.0

1900 to 1920 inclusive, for the causes of death (*a*) tuberculosis (all forms) and (*b*) typhoid fever. The data are taken from Mortality Statistics, 1916, p. 21 (rates for years 1900 to 1909 inclusive), and 1920, p. 19 (rates for years 1910 to 1920 inclusive). The reason for confining attention to the original registration states is that the area and population at risk may be comparable throughout.

Using the same graphic methods as in Fig. 45 and the data from Table 13 we get the result shown in Fig. 46.

From this diagram the conclusion which one's eye draws at once is that the decline in the tuberculosis rate has been much more rapid during this period than in the typhoid rate. The tuberculosis line seems to slope downward much more steeply.

But is the conclusion implied by this apparent difference in slope correct? The diagram presented in Fig. 46 does not enable an easy, direct answer to the question. Why it does not will be perceived if the following considerations are taken into account. Suppose that in each of a series of six places in a period
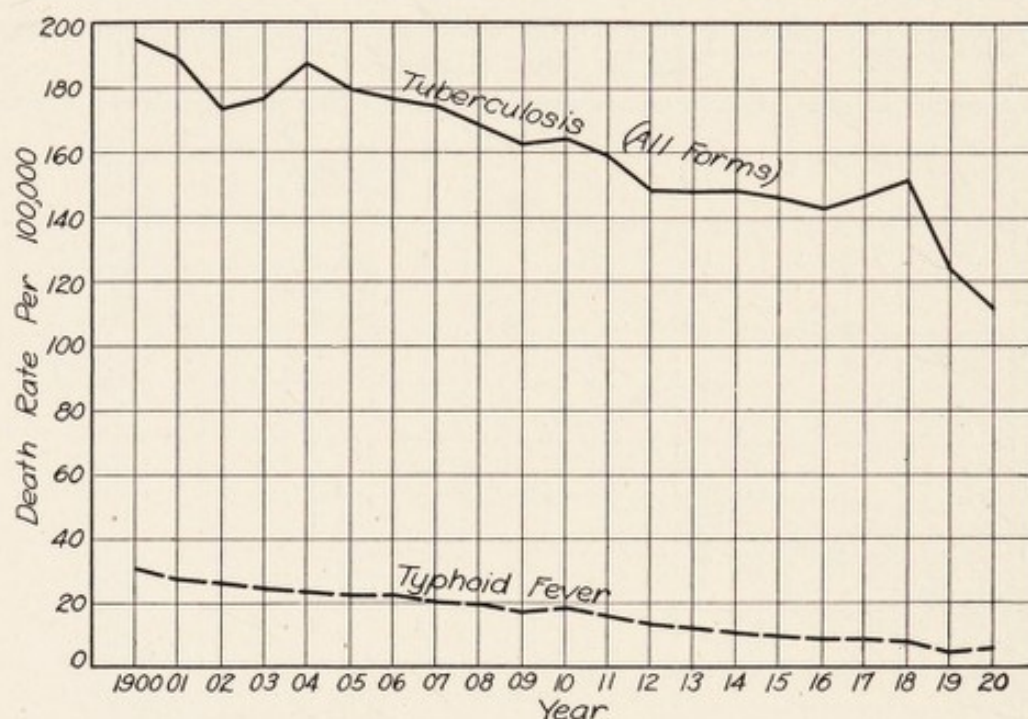
Fig. 46.—Death-rates from (a) tuberculosis (all forms) and (b) typhoid fever in the Registration Area, 1900–1920 inclusive. Arithmetic grid.

of time from a to b there occurred exactly 25 per cent. reduction in the number of deaths from a particular cause. But suppose further that, owing to the different absolute sizes of the places, the actual numbers of deaths which occurred in each of the six places, at the beginning of the period (time a) were respectively 5000, 4000, 3000, 2000, 1000, and 100. If then there was, as premised above, a reduction in mortality in the time period a to b of exactly 25 per cent., the numbers of deaths occurring at time b would be for the six places as follows: 3750, 3000, 2250, 1500, 750, 75. Now suppose this hypothetic case to be plotted

on an arithmetic grid as is Fig. 46.   The result will be as shown in
Fig. 47, A.

Anyone looking at this diagram would surely conclude that the
decline in mortality had been much more rapid in the first com-
munity than in the last.   Yet exactly the same rate of decline
(25 per cent.) was, by hypothesis, obtained in all the places.   To
produce a result *visually* correct all the lines ought to be parallel.
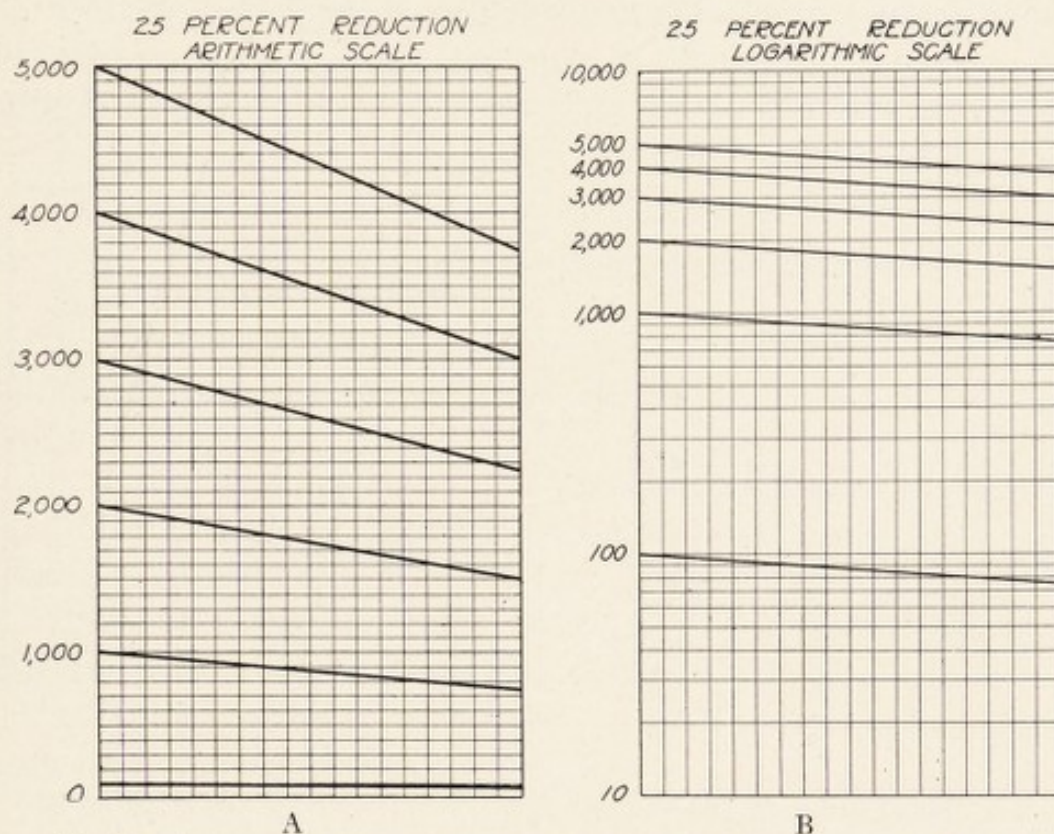
Fig. 47.—A, Diagram on arithmetic grid to show result of 25 per cent. reduction
in mortality in each of six places of different size.   Hypothetic case.   B, Showing the
result of plotting the same data as in A on an arithlog grid.

But plainly such a result cannot be attained by plotting these
data on an arithmetic grid.

Suppose now that the same data be plotted on a paper with a
grid ruling such that, while the abscissal scale is still graduated in
arithmetic progression (*i. e.*, with equally spaced steps), the
scale of the ordinates is divided not in arithmetic progression,
but in proportion to the logarithms of numbers in arithmetic pro-
gression.   Such a ruling is called an *arithlog* or *semi-logarithmic*
grid.   The result is shown in Fig. 47, B.

It is evident that there has been an almost magical transformation. The 25 per cent. reduction lines are now all parallel, as they ought to be if the diagram is to tell a visually correct story, and surely it is idle to plot diagrams if they are to tell a visually incorrect story when finished. For a diagram is plainly something to be looked at. It produces its results visually.

It will be well now to go back and replot the data of Fig. 46 on an arithlog grid. The result is that shown in Fig. 48.
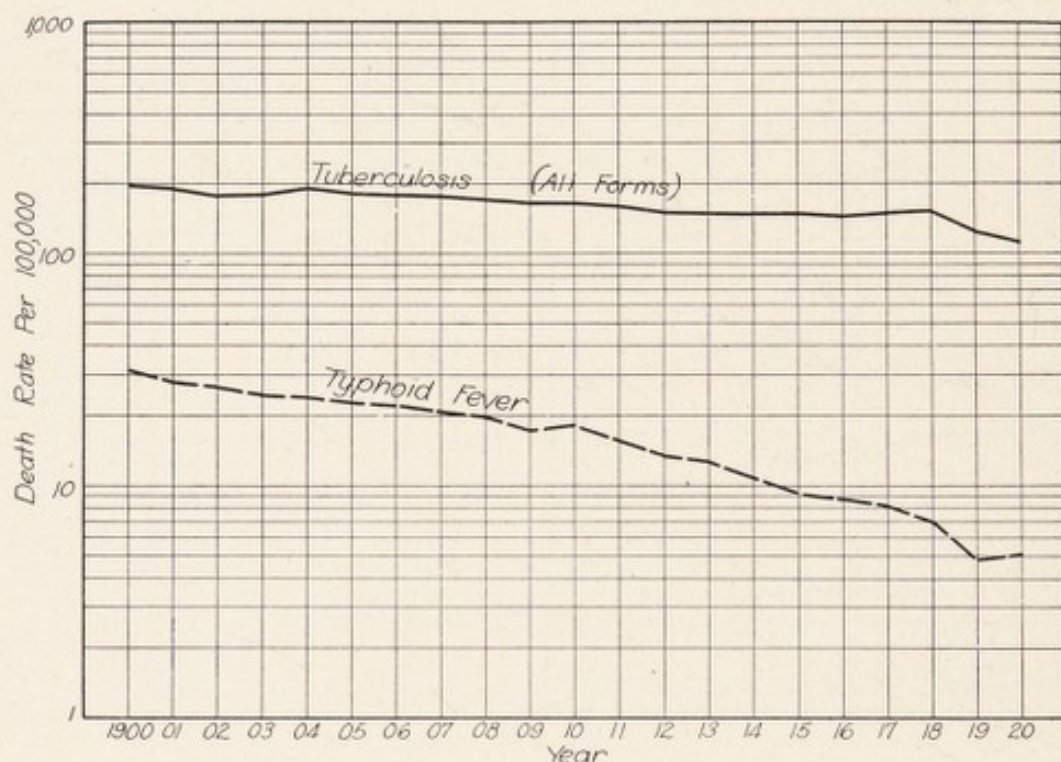
Fig. 48.—Death-rates from (a) tuberculosis (all forms) and (b) typhoid fever in the original registration states, 1900–1920 inclusive. Arithlog grid. Compare with Fig. 46.

The correct conclusion is now apparent. *Typhoid fever mortality has declined at a much more rapid rate in the period covered than has tuberculosis mortality.* And the fact is immediately apparent *visually*, as it ought to be if a diagram is used at all.

The advantages of the arithlog grid when trends are to be represented graphically has been emphasized by all recent American writers in this field, notably by Fisher,[4] Field,[5] and Whipple and Hamblen.[6] The papers of Fisher and Field especially should be

carefully read by the student for the full and scholarly discussion of this matter which they give.

Fisher sums up the advantages of this method of plotting trends (he calls a chart on an arithlog grid a "ratio chart") as follows:

"The eye reads a ratio chart more rapidly than a difference chart or a table of figures. We may recapitulate what most easily catches the eye as follows:

"1. If we see a curve ascending, and nearly straight, we know that the statistical magnitude it represents is increasing at a nearly uniform rate.

"2. If the curve is descending, and nearly straight, the statistical magnitude is decreasing at a nearly uniform rate.

"3. If the curve bends upward the rate of growth is increasing.

"4. If downward, decreasing.

"5. If the direction of the curve in one portion is the same as in some other portion it indicates the same percentage rate of change in both.

"6. If the curve is steeper in one portion than in another portion it indicates a more rapid rate of change in the former than in the latter.

"7. If two curves on the same ratio chart run parallel they represent equal percentage rates of change.

"8. If one is steeper than another the first is changing at a faster percentage rate than the second.

"9. The imaginary straight line most nearly representing, to the eye, the general trend of the curve, is its 'growth axis,' and represents the average rate of increase (or decrease); and the deviations of the curve from this growth axis are plainly evident without recharting.

"10. The slope of the imaginary line between any two points on a curve indicates the average rate of change between the two."

Whipple and Hamblen particularly discuss the use of this type of diagram in public health work.

Cyclic Time Trend Diagrams

A cyclic event is one whose frequency of occurrence varies in an orderly recurring manner. An example is found in the seasonal

AVERAGE WEEKLY CASE RATES FROM WHOOPING COUGH
NEW YORK CITY AND PHILADELPHIA. 1906–1912

Cases per 100,000 Population; Ages under Ten

Philadelphia

New York City

STATISTICIAN'S DEPT., THE PRUDENTIAL INSURANCE COMPANY OF AMERICA
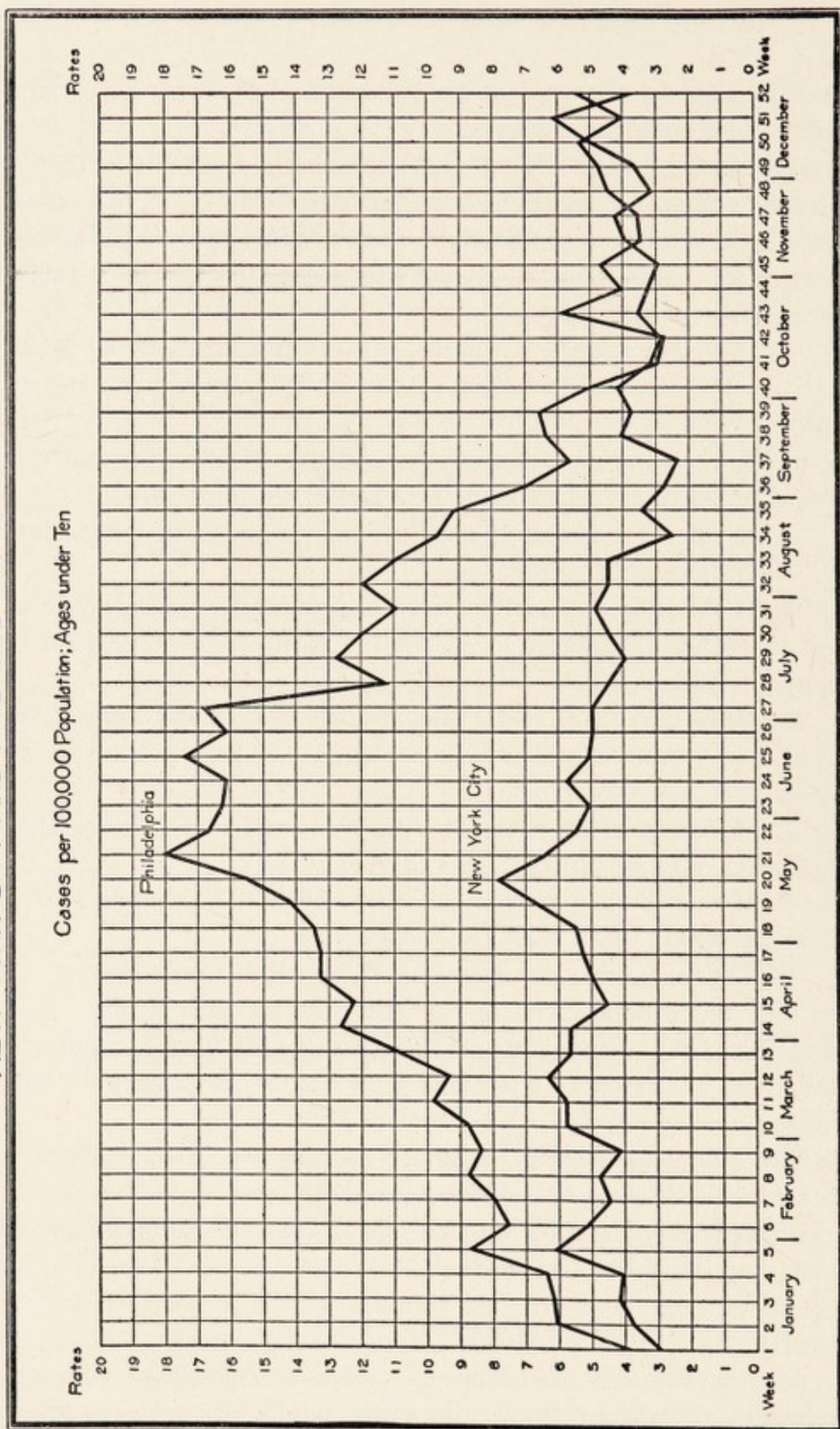
Fig. 49.—Average weekly case incidence rates from whooping-cough in two cities. (Reproduced by courtesy of the Statistician's Department of the Prudential Insurance Company.)

incidence of various diseases, as shown in Fig. 49 for whooping-cough in Philadelphia and New York City.

This diagram shows clearly that whooping-cough reaches its maximum incidence in the late spring months, and is less frequent at other periods of the year.

A method of plotting such cyclic events sometimes used is
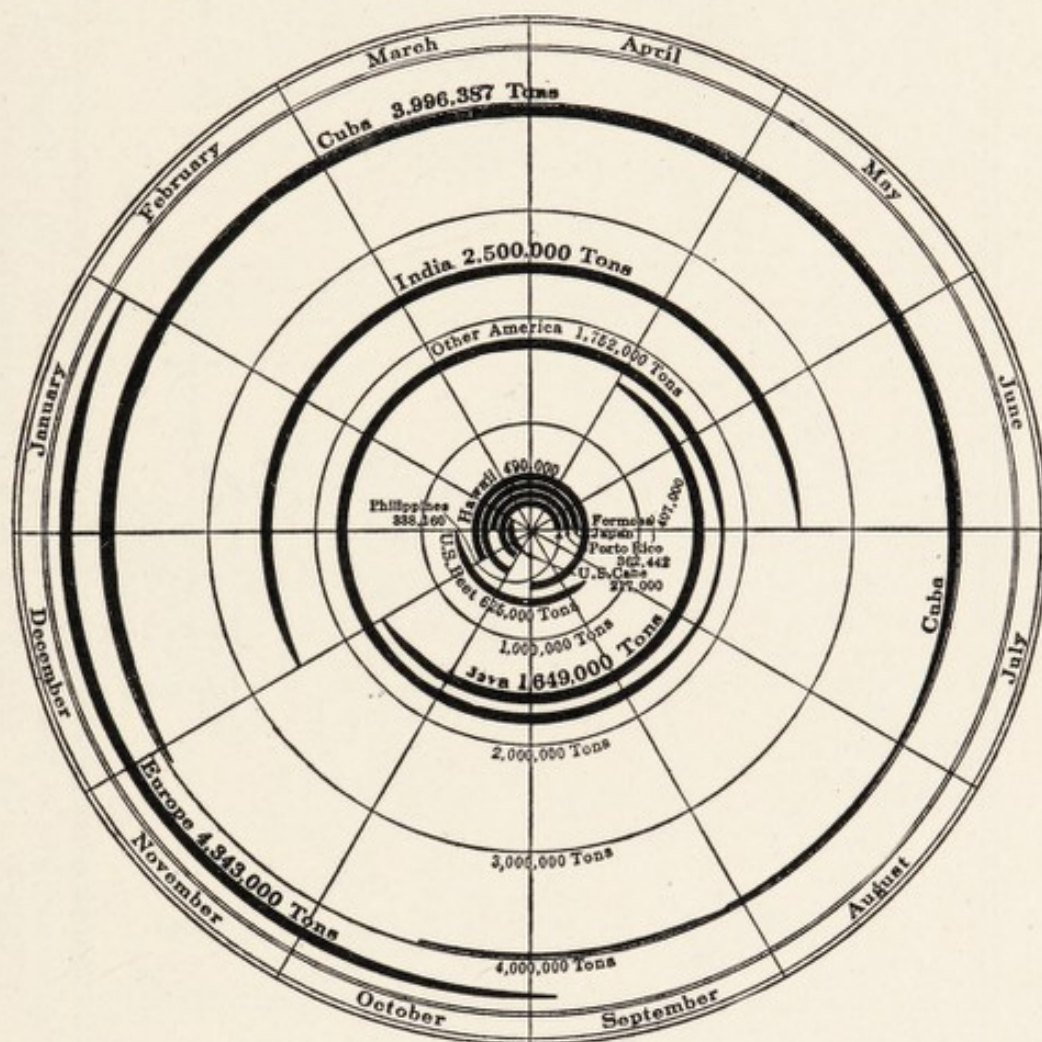
Fig. 50.—Diagram showing time of harvesting of principal sugar crops of the world. (Reproduced from source indicated in text, by permission of Mr. Earl D. Babst.)

through the employment of polar co-ordinates. In this type of diagram the frequencies corresponding to a given time are laid off as ordinates radiating from a central, polar point. On account of the greater familiarity which generally exists with regard to diagrams of the type of Fig. 49 these are perhaps to be preferred in ordinary statistical work to polar co-ordinate diagrams for cyclic events.
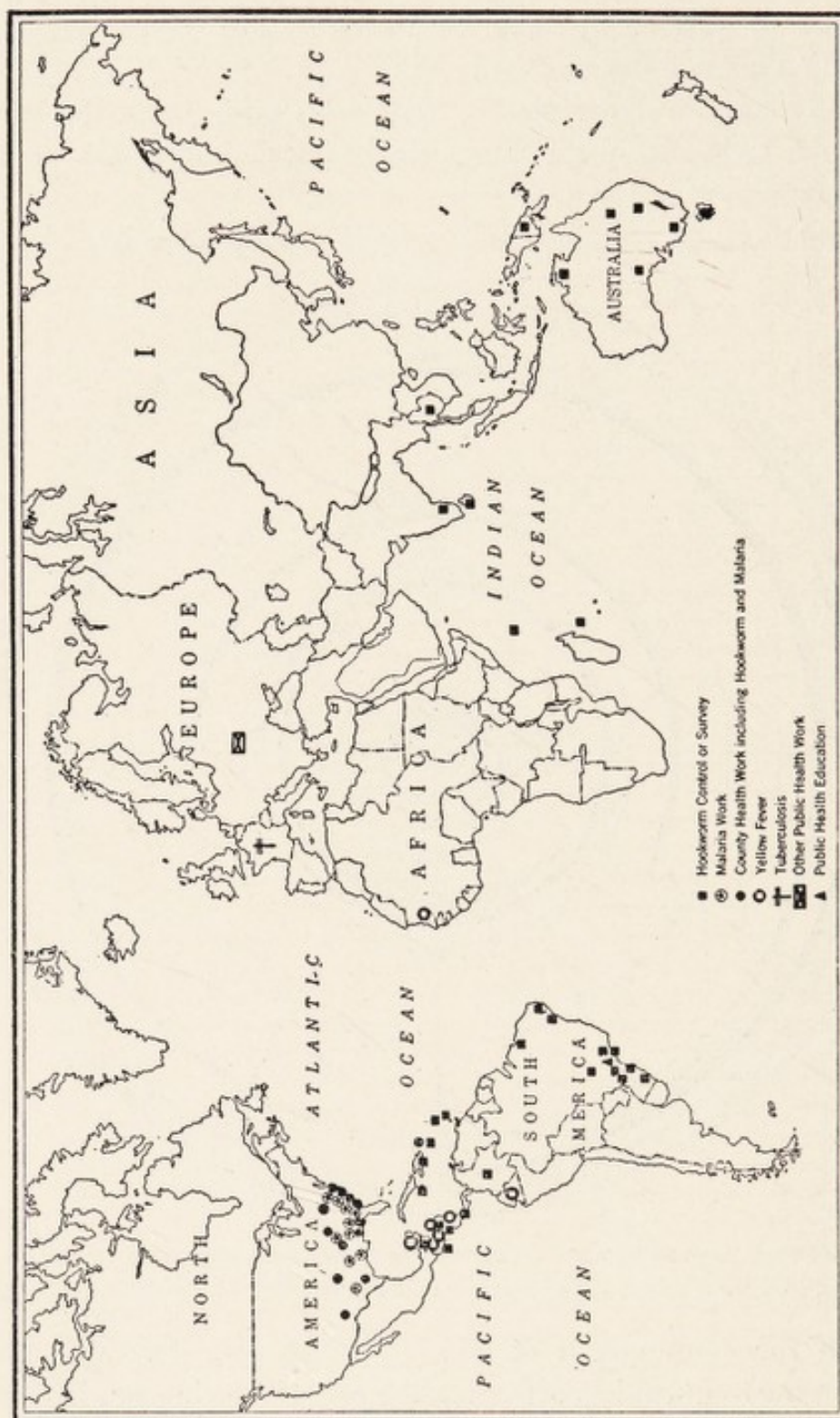
Fig. 51.—World map of activities of International Health Board during 1920. (Reproduced by permission of Mr. Wickliffe Rose from Seventh Ann. Rept. International Health Board, 1921.)

An interesting and useful method of showing graphically the time relations of certain kinds of cyclic phenomena is presented in Fig. 50. This diagram, taken from the Annual Report for 1922
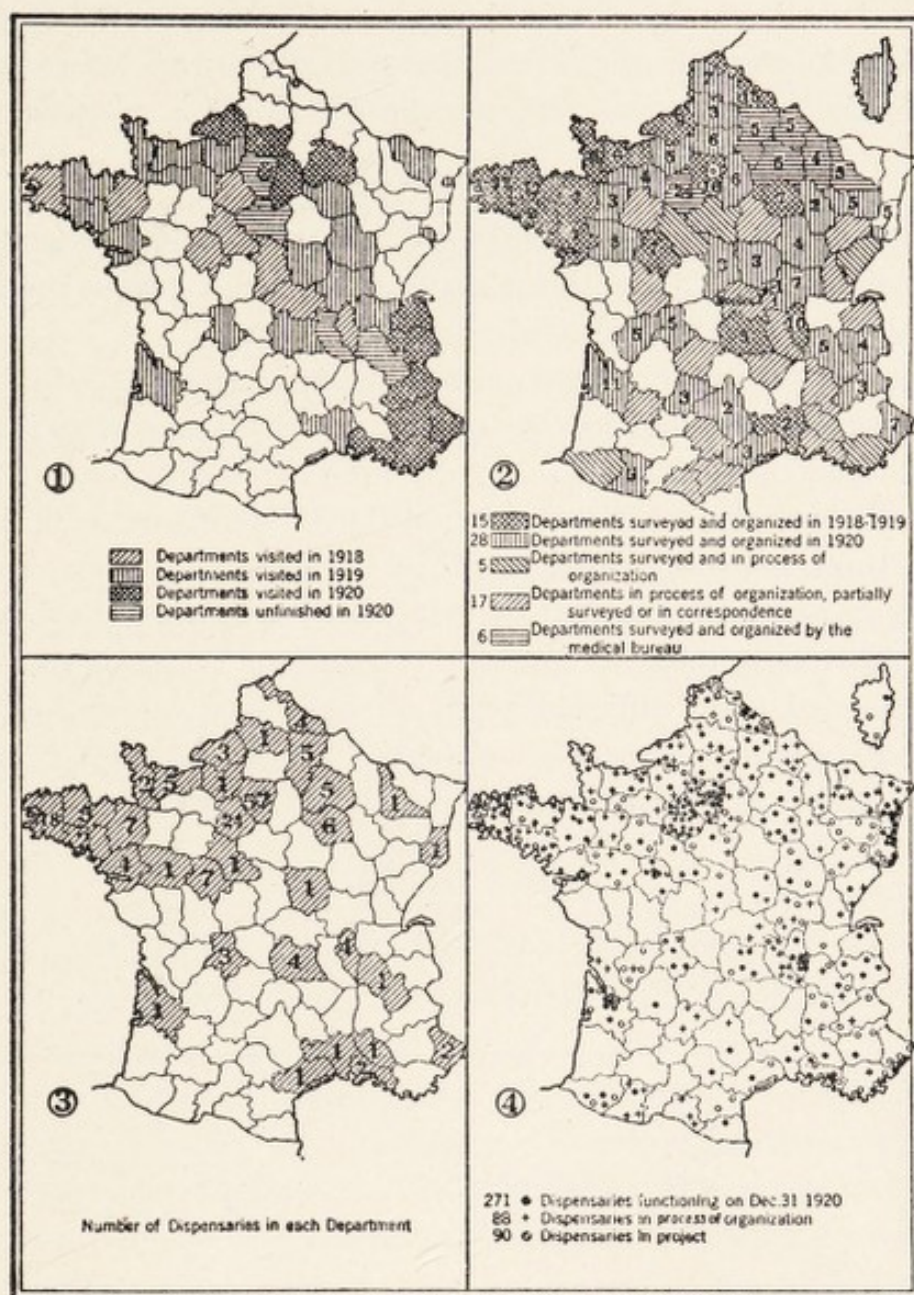
Fig. 52.—Organization and activities of Commission for the Prevention of Tuberculosis in France: 1. Work of educational division, showing departments visited by traveling exhibits during 1918, 1919, and 1920. 2. Work of division of departmental organization, showing departments in which antituberculosis organization has been effected or is in progress. 3. Number of tuberculosis dispensaries in each department co-operating with the Commission on December 31, 1920. 4. Total number of tuberculosis dispensaries functioning, in process of organization, or in project at the end of 1920. (Reproduced by permission of Mr. Wickliffe Rose from Seventh Ann. Rept. International Health Board, 1921.)

of the American Sugar Refining Company, shows the time relations of harvesting of the principal sugar crops of the world, the sizes of the respective crops being plotted to polar co-ordinates.

Statistical Maps

Maps may be usefully employed for the graphic presentation of certain types of data. Such maps are of two types in the main: (*a*) Spot maps and (*b*) shaded or colored maps.

In the spot map the *locality* of occurrence of an event is indicated by a properly located dot on the map. This type of map is much used in epidemiologic work. An example of such a map is given in Fig. 51, showing the distribution of the different sorts of activities of the International Health Board in 1920.

Figure 51 illustrates that by using different sorts of "spots" one can indicate a number of facts and relations on the same spot map.

In shaded maps different types of shading or coloring of areas are used to bring out statistical facts. Figure 52 gives examples of such maps, as well as another instance of a spot map.

Scatter Diagrams

For certain purposes it is useful to employ the device of placing dots instead of lines in a reference plane of rectangular co-ordinates
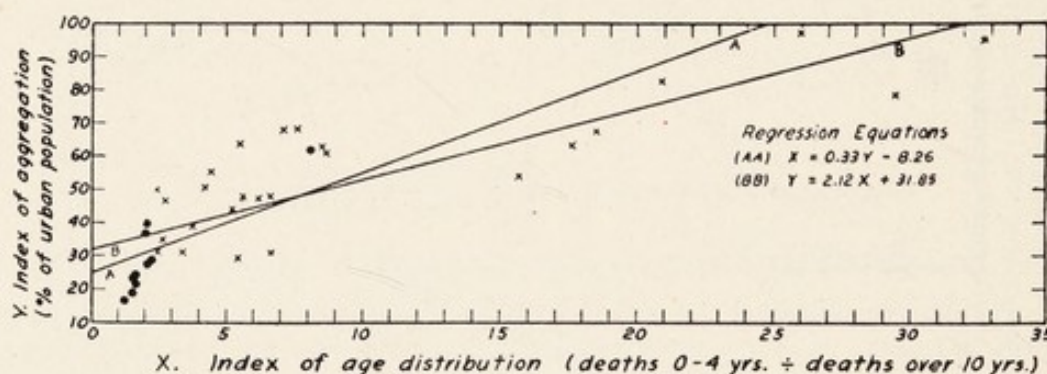
Fig. 53.—Scatter diagram showing the correlation between indices of aggregation of population (per cent. urban) and age distribution of deaths from measles, 1917–24, in 36 registration states, southern states (circled) included. (From Doull, J. A.: Amer. Jour. Hygiene, vol. 8, p. 635, 1928.)

to show the distribution of individual events or values. This scheme has particularly been used by biometricians to show graphically the distribution of individual variation of organisms relative to

two correlated variables. Such a diagram brings out clearly the "scatter" of the individual variates, whence the name of this type of diagram is derived. They are also sometimes called "spot" diagrams. An example is given in Fig. 53.

Scatter diagrams are to be regarded, in general, as preliminary, graphic aids, primarily useful in the working stage of a research, rather than as finished exhibits in the final presentation of results. However, in certain cases, such as the one illustrated in Fig. 53, they are valuable in emphasizing the distribution of the individual observations in the published presentation.

Nomograms

Up to this point in the discussion of graphic methods every case has dealt with the plotting of but *two* variables. Nomography is a development of graphic methods which permits the representation of theoretically *n* variables upon a plane surface. The invention of co-ordinate geometry was due to Descartes, who developed the idea of representing graphically two variables in a plane. Buache, in 1752, showed that a third variable could be added by the use of contour lines. D'Ocagne[9] hit upon the idea of collinear points as furnishing a method of dealing graphically with *n* variables in a plane. To him is due the name "nomography," which is given to this branch.

The outstanding usefulness of nomography is to facilitate the numerical solution of complex mathematical expressions and relations. An example of a nomogram for this purpose is to be found on page 34 of Pearson's "Tables for Statisticians and Biometricians."

Space is lacking here for any detailed development of this subject. The statistician and the medical man will, however, do well to master it, because it has many important applications in these fields. The best brief account in English is that of Hezlet.[7] Brodetsky's[8] book is a sound, if pedagogically somewhat inept introduction to the subject. An elementary treatise which is in some respects much better is that of Fréchet and Roullet.[11] D'Ocagne's[9] own writings are, of course, the final authority, but not particularly adapted to the medical man with a meager equip-

ment of mathematics. Also the two-volume treatise by Soreau[12] may be consulted.

A single example, of the simplest possible character, may be given here to indicate in some measure what a nomogram fundamentally is, and the logic underlying the construction of nomograms. Suppose we wish to set up a nomogram for the graphic solution of the expression

$$x = a + b$$

Lay off on two parallel lines scales with equally spaced divisions. The scales may be divided with any desired degree of fineness, may be of any length one pleases, and may be as far apart (or near together) as one pleases. One of these scales will be the a scale (*i. e.*, that upon which values of a are to be read) and the other the b scale. Now, plainly, it will be possible to draw somewhere between the a and b lines of Fig. 54 a third line parallel to the other two, and so graduated that if a straight-edge connects any value on a with any value on b the point where the straight-edge crosses x will give a reading on x which will satisfy the relation $x = a + b$. The problem is to find the location of the x line and its graduation. To do this is very simple, as shown in Fig. 54.

We know that

$$\text{when } a = -20 \text{ and } b = +15, x = -5$$
$$a = +15 \text{ and } b = -20, x = -5$$

If then we draw straight lines connecting these two particular values of a with the two connected values of b, the point where these two lines cross each other must, in the first place, lie on the x line, and in the second place must be the point on that line which is to be graduated -5. Again, we know that

$$\text{when } a = + 5 \text{ and } b = \quad 0, x = +5$$
$$a = -10 \text{ and } b = +15, x = +5$$

Draw these lines, and we shall have determined a second point on the x line. Two points being sufficient, we have now located the position in space and the direction of the x line. Its further graduation may be wrought out by continuation of the same

process, though to do it that way would be a highly unintelligent procedure in the case of so simple a relationship.

Two examples may be given of nomograms for dealing with medical problems. The first relates to the calculation of the surface area of the human body from known height and weight. Feldman and Umanski* have published a nomogram of the DuBois equation

$$S = 71.84\ W^{0.425}\ H^{0.725}$$
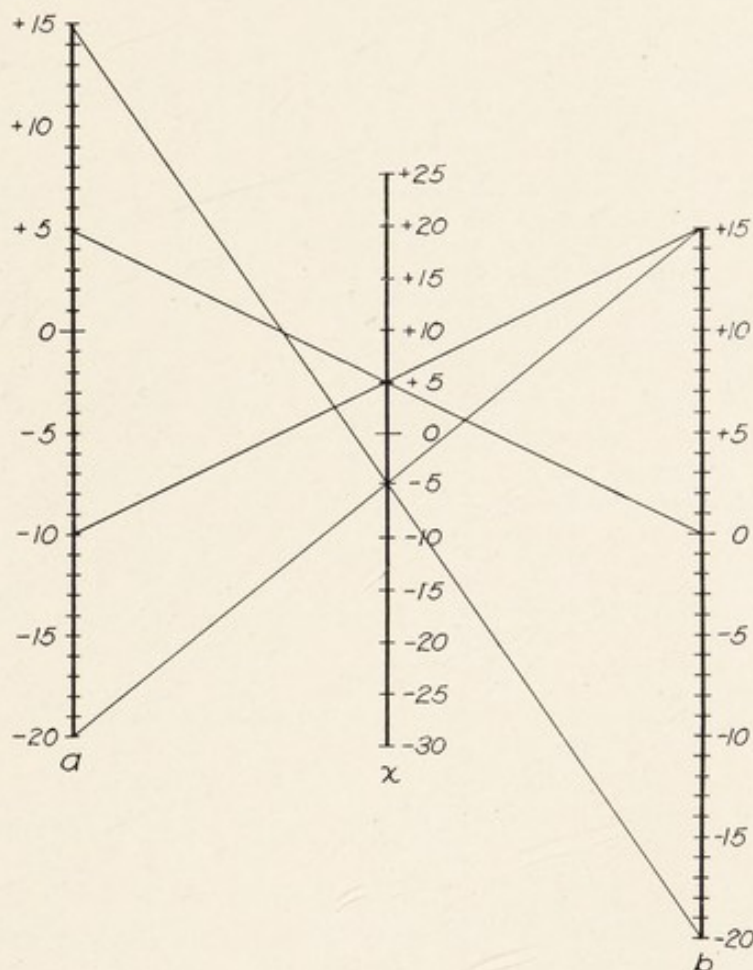
Fig. 54.—Construction of addition nomogram. See text.

This is reproduced as Fig. 55.

The second example is one of Lawrence J. Henderson's† nomo-

* Feldman, W. M., and Umanski, A. J. V.: The Nomogram as a Means of Calculating the Surface Area of the Living Human Body, Lancet, vol. 202, February 11, 1922, pp. 273, 274.

† Henderson, L. J.: Blood as a Physicochemical System, Jour. Biol. Chem., vol. 46, pp. 411–419, 1921.

grams relating six variables in the physiology of the blood. It is shown in Fig. 56.
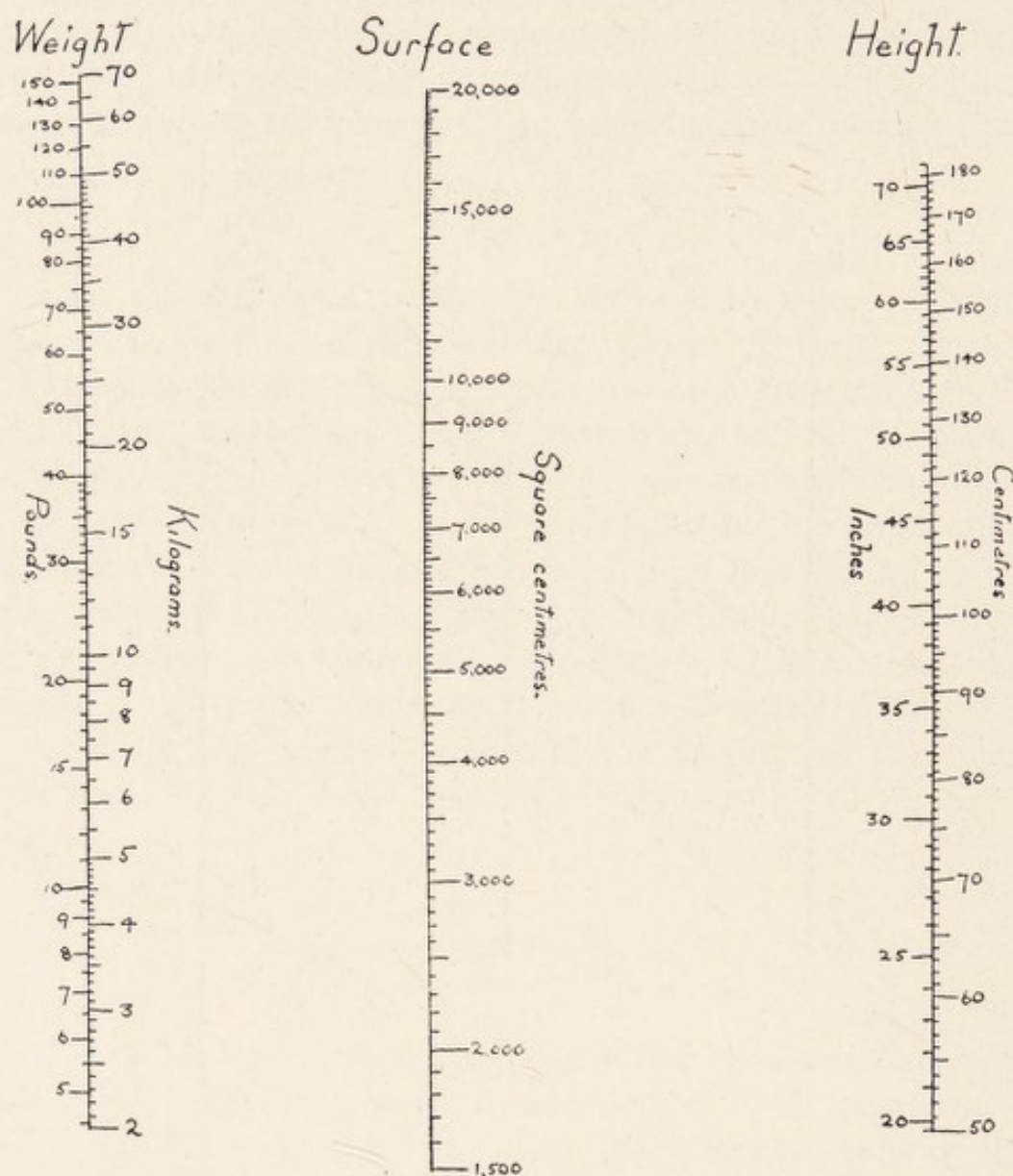
Fig. 55.—Nomogram for $S = 71.84 \ W^{0.425} \ H^{0.725}$, where S = surface in sq. cm., W = weight in kg., and H = height in cm. A straight line joining given values of W and H cuts the middle scale at the correct value of S. Thus a line joining the point 24 on the weight scale, with the point 110 on the height scale, will cut the surface scale at a point corresponding to 8375, which means that the surface area of a person 24 kilograms in weight and 110 cm. in height is 8375 sq. cm. (From Feldman and Umanski.)

The six variables involved in this nomogram are the free and combined oxygen of the whole blood, $[O_2]$ and $[HbO_2]$; the free

and combined carbonic acid of the serum, [H₂CO₃] and [BHCO₃];
the hydrogen-ion concentration of the serum, expressed as [pH];
and the chlorid concentration of the serum, [BCl].

This nomogram expresses at once the results of Barcroft
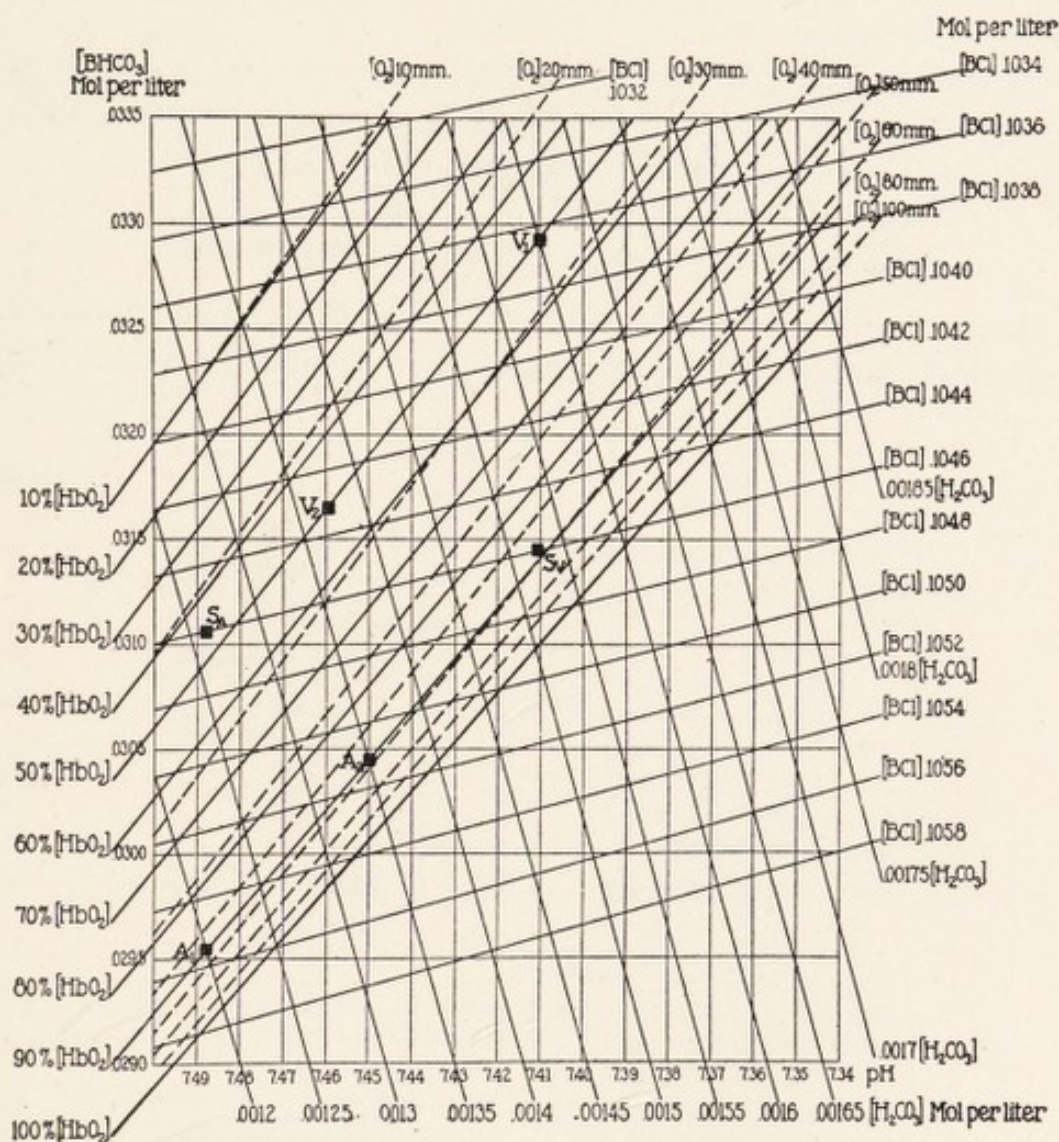
Fig. 56.—Nomogram for certain physicochemical relations of the blood. (From
L. J. Henderson.)

upon the oxygen dissociation curve of blood, and of Christiansen,
Douglas, and Haldane on the carbon dioxid dissociation curve,
as well as the peculiarities of the acid-base equilibrium, and of the
distribution of chlorids. Obviously it has the property that if values
are assigned to any two of the variables, all six are determined.

Henderson's original paper must be consulted for further discussion of this nomogram. Anyone interested in the application of nomography to medical problems should read the same author's Silliman Lectures.[13] They represent the most extensive, varied and penetrating application of the method to biological problems which has yet been made.

A further example illustrating the application of nomography to statistical material and problems is given in Chapter VIII, where a life table nomogram is presented and discussed.

ELEMENTARY STANDARDS IN GRAPHIC WORK

In 1915 a widely representative joint committee of engineering, statistical, economic, biologic, and other societies, interested in

1   The general arrangement of a diagram should proceed from left to right.

Fig. 1
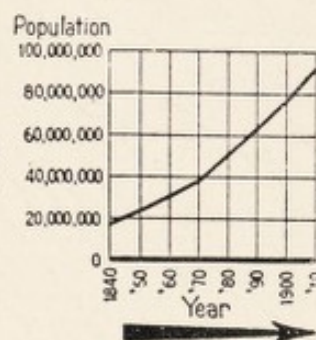
Year	Tons
1900.	270,588
1914.	555,031
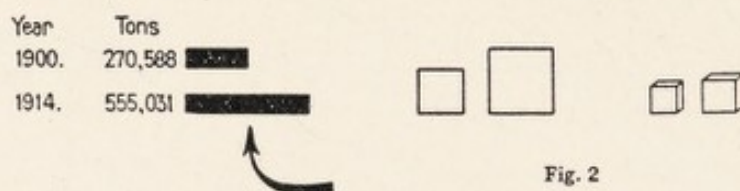
Fig. 2

2   Where possible represent quantities by linear magnitudes as areas or volumes are more likely to be misinterpreted.

3   For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.
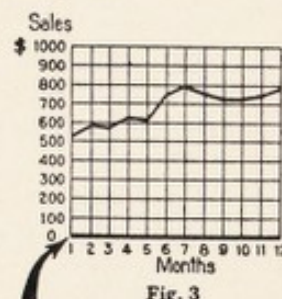
Fig. 3

the promotion of sound methods of graphic presentation of data, published[10] a preliminary report on standards. This re-

port is so valuable for the beginner in this type of work that, with the permission of the Chairman of the committee, Mr. Willard C. Brinton, its essential parts are here reproduced in full.

4   If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.
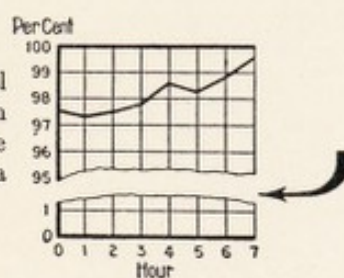
Fig. 4
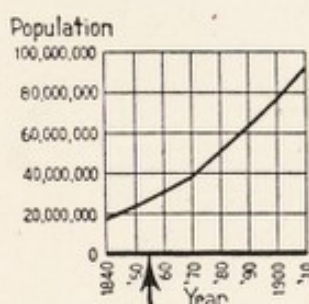
Fig. 5A
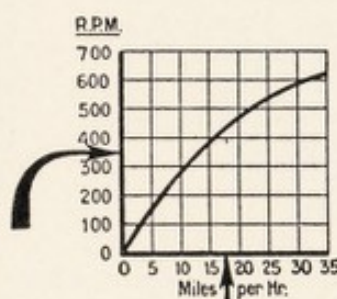
Fig. 5B

5   The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

Fig. 5C

Per Cent
Utilized

Fig. 6A

Relative
Cost
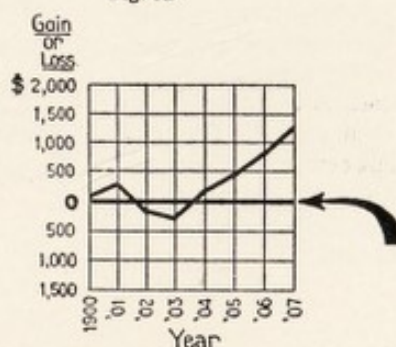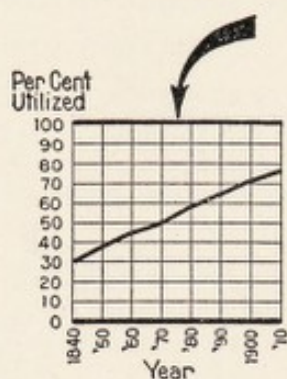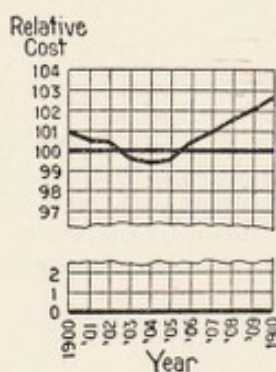
Fig. 6B

6 For curves having a
scale representing percentages,
it is usually desirable to em-
phasize in some distinctive
way the 100 per cent line or
other line used as a basis of
comparison.

Per Cent
of
People

Per Cent of Income
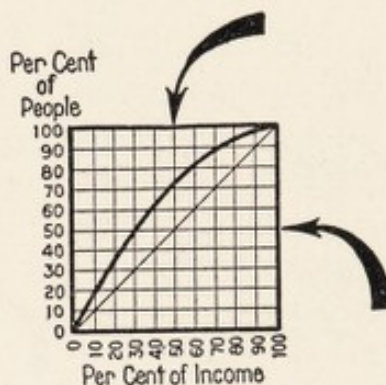
Fig. 6C

7 When the scale of
a diagram refers to dates,
and the period repre-
sented is not a complete
unit, it is better not to
emphasize the first and
last ordinates, since such
a diagram does not repre-
sent the beginning or end
of time.

Population
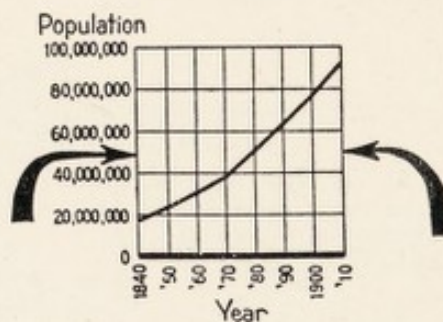
Fig. 7

8   When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.
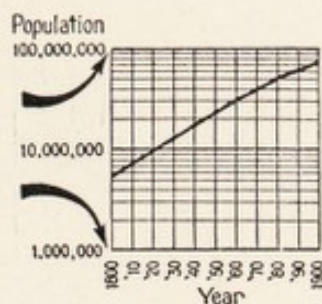
Fig. 8
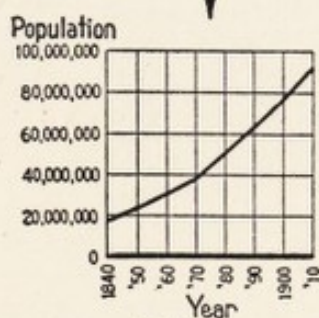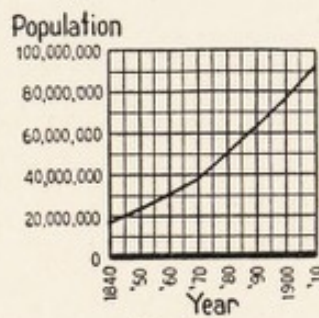
9   It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

Fig. 9A

Fig. 9B

10   The curve lines of a diagram should be sharply distinguished from the ruling.
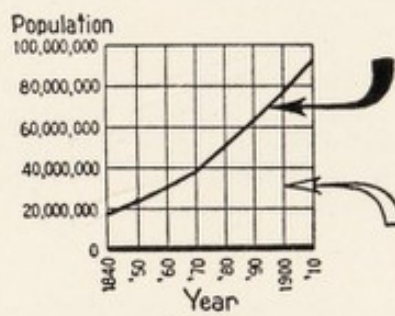
Fig. 10
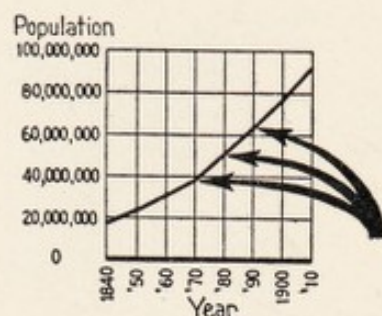
Fig. 11A
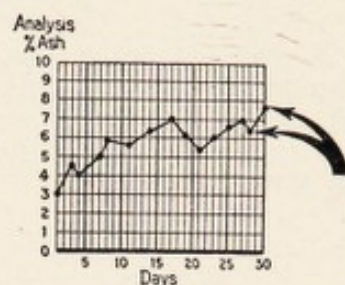
Fig. 11B

11 In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.
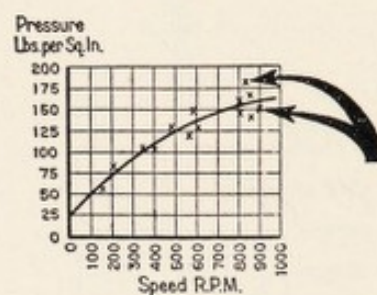
Fig. 11C

12 The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.
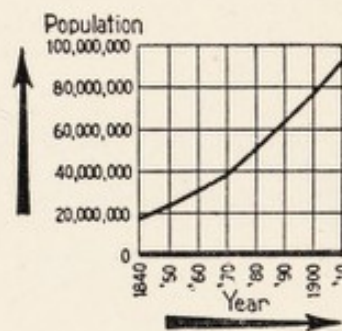
Fig. 12
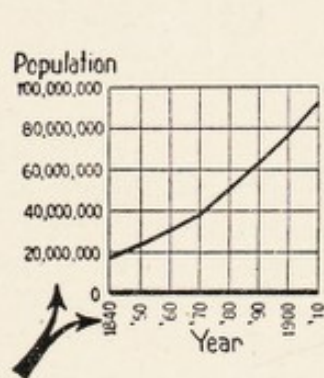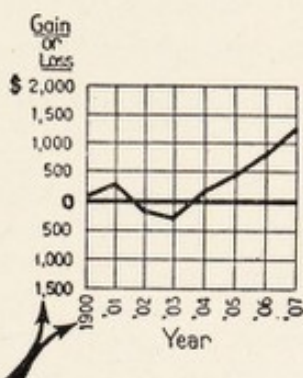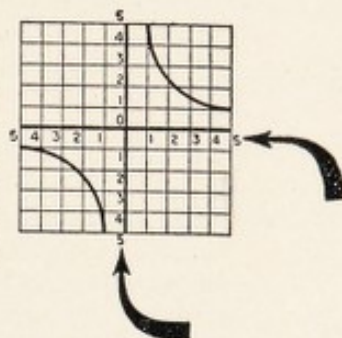
Fig. 13A   Fig. 13B   Fig. 13C

13 Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.
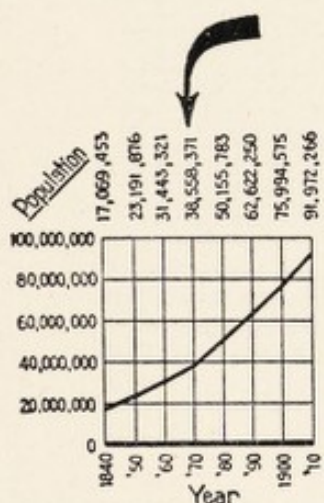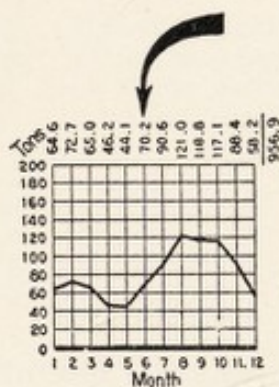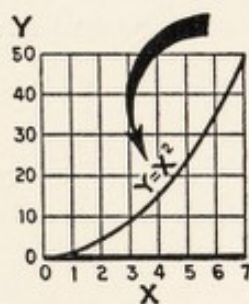
Fig. 14A   Fig. 14B   Fig. 14C

14 It is often desirable to include in the diagram the numerical data or formulae represented.

15   If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.

Year	Population
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,622,250
1900	75,994,575
1910	91,972,266
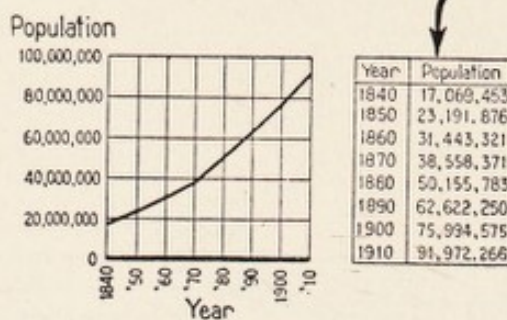
Fig.15

16   All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.
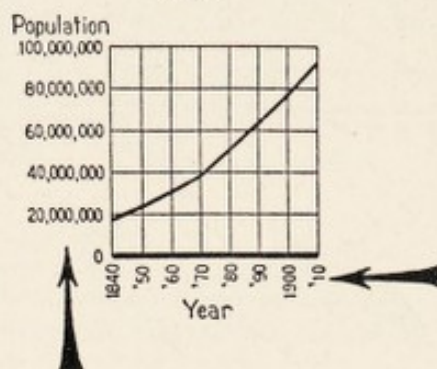
Fig 16

17   The title of a diagram should be made as clear and complete as possible. Sub-titles or descriptions should be added if necessary to insure clearness.
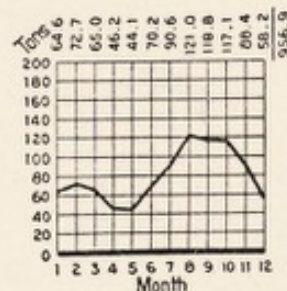
Aluminum Castings Output of Plant No. 2, by Months, 1914.

Output is given in short tons.

Sales of Scrap Aluminum are not included.

Fig. 17

SUGGESTED READING

1. Brinton, W. C.: Graphic Methods for Presenting Facts, New York, The Engineering Magazine Company, 1919.

2. Haskell, A. C.: How to Make and Use Graphic Charts, New York, Codex Book Company, 1919.

(References 1 and 2 are the best available general treatises on graphic methods. The student should go through them completely.)

3. Von Huhn, R.: A New Graphical Method for Comparing Performance with Program or Expectation, Science, N. S. vol. 47, pp. 642–645, 1918. (Deals with percentage accumulated frequency plotting.)

4. Fisher, I.: The "Ratio" Chart for Plotting Statistics, Quarterly Publ. Amer. Stat. Assoc., June, 1917, pp. 577–601. (Has bibliography on arithlog plotting.)

5. Field, J. A.: Some Advantages of the Logarithmic Scale in Statistical Diagrams, Jour. Pol. Econ., vol. 25, pp. 805–841, 1917.

6. Whipple, G. C., and Hamblen, A. D.: The Use of Semilogarithmic Paper in Plotting Death-rates, Public Health Reports, vol. 37, pp. 1981–1991, 1922.

7. Hezlet, R. K.: Article "Nomography" in Encyclopedia Britannica, 12th Edit., vol. 31, pp. 1139–1144, 1922. (Cf. also the same author's book, Nomography, 1913.)

8. Brodetsky, S.: A First Course in Nomography, London (G. Bell & Sons, Ltd.), 1920.

9. D'Ocagne: Traité de Nomographie, 1899; Calcul Graphique et Nomographique, 1908.

10. Joint Committee on Standards for Graphic Presentation. Preliminary report, Quart. Publ. Amer. Stat. Assoc., vol. 14, pp. 790–797, 1915.

11. Fréchet, M., et H. Roullet: Nomographie, Paris (Librairie Armand Colin), 208 pp., 1928.

12. Soreau: Nomographie ou Traité des abaques, 2 vols., Paris (Chiron), 1921.

13. Henderson, L. J.: Blood. A Study in General Physiology, New Haven (Yale University Press), 1928, pp. xix + 397.

CHAPTER VII

RATES AND RATIOS

In Chapter III the raw materials of statistics, the absolute frequencies of occurrence of events, were discussed. In many sorts of problems absolute frequencies will not alone suffice for the intelligent discussion of problems. The reason for this is simple. To say that in one city 2596 persons died of tuberculosis in a year, while in another city 1304 died in the same year of the same disease conveys no particularly useful information. It is essential to know, in addition, the *populations* of the two cities, at least. Otherwise it is impossible to form any conception of whether tuberculosis was more fatal in the one place than in the other. *In short, it is necessary to know the number exposed to the risk of the happening of a particular event, before the full significance of the statistics of that event can be appreciated.*

The calculation of *rates* in statistical work consists in arriving at frequencies of occurrence relative to the number exposed to risk of the occurrence. Properly calculated rates are said to measure:

In the case of deaths, the *force of mortality*.

In the case of births, the *force of natality*.

In the case of sickness, the *force of morbidity*.

The "force of mortality" is expressed as the proportion of those exposed to risk who die. Thus, if 100 persons are truly exposed to risk of dying within a given year, and 3 die, the force of mortality within the time limit of that year is 3 per cent.

It should be noted at the outstart of the discussion of rates that "number exposed to risk" does not always, or indeed usually, mean precisely the same thing as "number living." For example, suppose that in a particular community, say New York State in 1900, 452 persons died of puerperal septicemia, and in the same state

the same year there were living 7,284,461 persons. These facts do not imply that the true force of mortality of puerperal septicemia was $452 \div 7,284,461 = .00006$, or 6 per 100,000.

The true force of mortality must be quite different from this because:

(*a*) Males cannot have puerperal septicemia, and are, therefore, not at risk of dying from this disease.

(*b*) Females under ten or over sixty years of age are not exposed to risk of dying from this disease, because they are outside the reproductive period of life.

(*c*) Women not in the puerperium, *i. e.*, who have not recently been pregnant, are not exposed to risk of death from this disease.

So then it appears that from the figure of 7,284,461 living there must be subtracted at the start all the males, and then all the females except those in a certain physiologic state. The number of live births in New York State in 1900 was 143,156. Now, adding to this number 4 per cent. of itself, to correct roughly for stillbirths, multiple births, etc., the number 148,900 may be taken approximately to represent the number of women who during that year were in the puerperal state. So then the figure for force of mortality from this disease becomes roughly somewhere in the neighborhood of $452 \div 148,900 = .003$, or 300 per 100,000, a very different figure indeed from the 6 per 100,000 with which we started.

My colleague, Dr. W. T. Howard,[1] has discussed in detail the true risk of mortality in child-bearing, and his more precise and thorough treatment of the matter should be read in connection with the simple, rough example given above.

This same fallacy of using an incorrect figure for the exposed to risk often appears in medical statistics. An example may be cited. Litchfield and Hardman* reported excellent results in the treatment of laryngeal diphtheria by suction to remove the membrane. They presented a table, here reproduced as Table 14, to contrast their results before and after the use of this treatment.

* Litchfield, H. R., and Hardman, R. P.: Suction in the Treatment of Laryngeal Diphtheria, Jour. Amer. Med. Assoc., vol. 80, pp. 524–526, 1923.

TABLE 14

COMPARATIVE DATA ON TREATMENT OF LARYNGEAL DIPHTHERIA (LITCHFIELD AND
HARDMAN'S TABLE 1)

	May–December	
	1921.	1922.
Total cases of laryngeal diphtheria..................	158	106
No local treatment—mild cases.....................	43	21
Applicator treatment...............................	13	12
Applicator and intubation..........................	18	0
Intubation..	84	18
Suction...	0	46
Suction and intubation............................	0	9
Total deaths......................................	41	14
Mortality...	26− %	13+ %

Now, the mortality percentages given in the last line, 26− per cent. in 1921 (no suction treatment), and 13+ per cent. in 1922 (suction treatment in some cases), are reckoned on the basis 41/158 = .26, and 14/106 = .13. But it appears that in 1921 there were 43 cases so mild as to be given "no treatment" (text p. 526), and in 1922 there were 21 cases of the same sort. Clearly these 64 patients were not a proper part of the "universe of discourse," if that universe, as is the fact, concerns itself with discourse about different modes of treatment. They were *not treated*, therefore they cannot possibly have any bearing upon the relative merits of different kinds of local treatment, either one way or the other. Furthermore, none of them died, as, of course, was to be expected. Actually there were treated in 1921, 158 − 43 = 115 cases, and in 1922, 106 − 21 = 85 cases. Of these treated cases, 41 died in 1921, and 14 in 1922. Hence the true comparative mortality rates per cent. in the two years of this experience, are

$$\text{For 1921, } \frac{41 \times 100}{115} = 36 \text{ per cent.}$$

$$\text{For 1922, } \frac{14 \times 100}{85} = 16 \text{ per cent.}$$

Or, in other words, calculated on a proper basis the results in 1922 were even better relatively than those stated by the authors.

DEFINITION AND CLASSIFICATION OF RATES AND RATIOS

The basic *relative* figures of vital statistics may conveniently be divided into rates and ratios.

A *rate* has the following form:

$$R = \left(\frac{a}{a+b}\right),$$ (i)

which, expressed in words, means

Rate = $\left\{\begin{array}{l}\text{The number of times a specified kind of event actually occurs.}\\ \text{The whole number of exposures to risk of its occurrence, } i.\ e.,\\ \text{the number of times it actually occurs} + \text{the number of times it}\\ \text{might occur, but does not.}\end{array}\right\}$

The part of the right-hand member of the rate equation which is in brackets limits the universe of discourse to which the rate applies to a particular *kind* of event, as for example "death" or "birth."

A rate is also limited to a particular universe of discourse *in time*. This is done by preliminary definition. Thus a death-rate is "annual," referring to the deaths in a specified year, or "monthly" or "weekly," etc.

A rate as defined above states the result on an individual basis. Numerically it will, in this form, obviously be always a decimal fraction. In order to put rates into whole numbers rather than fractions, so that they may be more easily read and comprehended, it is the customary, and now generally conventionalized, practice to multiply rates on an individual basis, as above defined, by some multiple of 10. They thus become rates *per cent.* (when the rate on an individual base is multiplied by 100); or rates *per thousand* (when the rate on an individual base is multiplied by 1000), and so on.

The commonly employed rates in biostatistical work may be classified as follows:

A. *Death-rates* (Mortality rates).

    1. Observed actual death-rates, obtained by the direct application of equation (i), without assumptions:

       (*a*) Crude death-rates.

       (*b*) Specific death-rates.

       (*c*) Infant mortality rates.

       (*d*) Case fatality rates.

2. Theoretic death-rates based upon certain assumptions:
    (*a*) Standard (or standardized) death-rates.
    (*b*) Corrected death-rates.
(These theoretic death-rates will be considered in detail in Chapter IX, after certain requisite preliminaries have been explained in Chapter VIII.)

B. *Birth-rates* (Natality rates).
    1. Observed actual birth-rates obtained from equation (i):
        (*a*) Crude birth-rates.
        (*b*) Specific birth-rates.
    2. Theoretic birth-rates, based upon certain assumptions:
        (*a*) Standardized birth-rates.
        (*b*) Corrected birth-rates.

C. *Morbidity Rates.*
    1. Observed, actual:
        (*a*) Crude.
        (*b*) Specific.

D. Marriage Rates

E. Divorce Rates

As these two categories fall, in actual practice, rather in the field of demographic statistics than in that of medical statistics, they will not be further considered.

Each of the types above mentioned will be discussed in detail farther on.

Before doing so, however, it will be well to define and classify the *ratios* commonly used in biostatistics.

A *ratio* is a relative figure in fractional form, but distinguished from a rate by the fact that the denominator does not denote the number exposed to risk of occurrence of the event, whose frequency of occurrence is given by the numerator.

$$R_o = \left(\frac{a}{c + d} \right) \qquad \text{(ii)}$$

where

    R_o = a ratio,
    a = the number of times an event of some specified kind occurs,
    $c + d$ = the number of times some other kind of event, in general different from the a event, occurs, although in some cases $c = a$.

There are but two sorts of ratios at all commonly employed in biostatistical work, viz.:

(*a*) Death ratios.

(*b*) Birth-death ratio (or Vital Index).

Each of these different sorts of rates and ratios will now be discussed and illustrated in some detail. But before going on to this it is important to emphasize particularly one point. It is this: As defined above, each of the rates mentioned is mathematically an expression measuring a *probability*. When in a later chapter the discussion of the theory of probability is undertaken this fact about death-rates, birth-rates, etc., will be more easily and fully appreciated. But it is desired to bring it out here in anticipation of the more formal discussion of probability in order that the reader may fully realize from the start that what a death-rate or a birth-rate really measures, in a mathematical sense, is always a probability. The conventional use of the constant multiplier of 100 or 1000, etc., in stating rates tends somewhat to disguise (at least to the unwary) this fact, but in the detailed discussion of rates pains will be taken to state formally what probability it is that each particular rate measures.

CRUDE DEATH-RATES

Here the fundamental equation (i) becomes

$$R_c = \left(\frac{D}{P}\right)$$

where

R_c = crude death-rate,
D = deaths from all causes,
P = total population = $D + (P - D) = P$.

(Crude death-rates are usually stated "per 1000" or "per 100,000.")

Nothing could be less refined than this. The deaths are not separated as to cause, and the entire population is assumed to be at risk of death. The annual crude death-rate measures the probability of a person, regardless of age, sex, race, or occupation, dying within one year, from any cause whatever, in a population constituted in respect of its age, sex, racial and occupational dis-

14

tribution, as the population under discussion happens to be. A crude death-rate, in other words, is an absolutely accurate and precise measure of something which, because of its heterogeneous, composite, unanalyzed character, is not particularly worth measuring accurately. So many variables besides those essentially lethal can (and do) influence the stated values of crude death-rates as to make them rather untrustworthy for any but the broadest and roughest conclusions and estimates. Taken alone and by themselves, in the complete absence of any other knowledge than that furnished by the crude rates themselves, they must be employed with the utmost caution and reservation in comparisons of one locality or one time with another. The reasons for the great unreliability of crude rates for comparative purposes will more and more clearly appear as we proceed.

Another class of crude death-rates is given by the expression

$$R'_c = \left(\frac{D'}{P} \right)$$

where D' = deaths from a particular cause or group of causes only, and all the other letters have the same significance as before. Thus we might have the crude death-rate for tuberculosis of the lungs. This represents the first step in specification, but does not go far. Indeed R'_c may certainly be said in a good many cases to give a wholly *false* measure. It does not measure any rational probability, because P still is the total living population. But as we have seen earlier not all P is exposed to risk of dying, for example, of puerperal septicemia. Therefore the probability given R'_c is in that case a false one. R_c does measure a true probability, because all P *is* exposed always to the risk of dying of something or other, but it is not a very important or interesting probability. In short, R_c is rather a fool, while R'_c is a knave.

The crude rate from all causes R_c may be used with a fair degree of safety for comparing the *relative* mortality of the *same place* (city, state, etc.) at different *times*, provided the periods compared are not too far apart, and provided the place has not undergone rapid growth or decline in population during the period.

The reason for this is that in fairly stable, large communities
the age and sex constitution of the population changes only very
slowly. This fact is well illustrated by Fig. 57, which shows the
age distribution of the living population of Amsterdam, at nine
consecutive census periods (1829 to 1920 inclusive). It is at once
apparent that in this long period the age constitution of the popula-
tion of Amsterdam has not shown any very considerable changes.
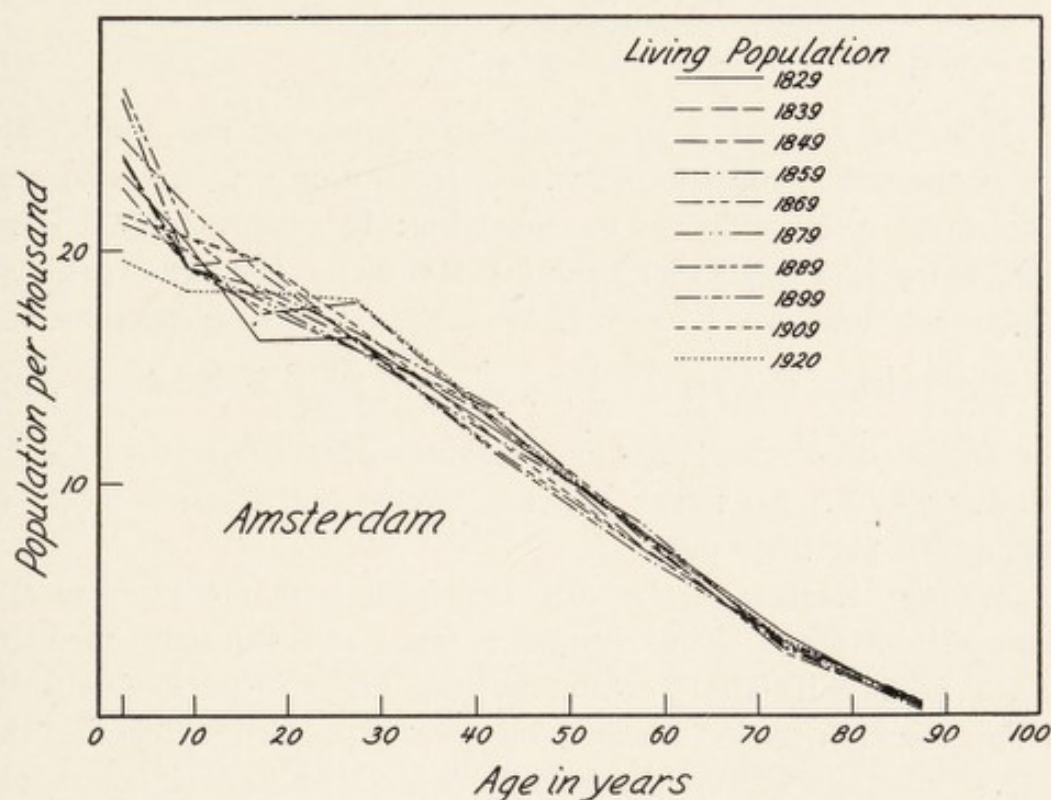
Fig. 57.—The proportion per thousand of the total population of Amsterdam falling in
different age classes, at each of nine census periods between 1829 and 1920.

It has been shown analytically by Lotka[2] that, under certain
conditions not widely different from those which prevail in large
human population aggregates, the age distribution tends to converge
toward a stable normal condition or state.

The crude rate from all causes R_c is wholly unreliable as an
index of the relative mortality in *different places*, unless it be first
shown by a preliminary investigation that the populations of the
places compared are substantially identical in age and sex dis-
tribution, a condition which is usually not carried out.

SPECIFIC DEATH-RATES

Here the fundamental equation becomes

$$R_s = \left(\frac{D_e}{E}\right),$$

where

R_s = specific death-rate,
D_e = deaths in a specified class of the population,
E = number exposed to risk of dying, in the same specified class of the population from which the deaths come.
(Specific death-rates are usually stated as "per 1000.")

In actual statistical practice at the present time death-rates are commonly made specific with reference only to age and sex. This means a situation like the following: In a community A there were living in a particular year say 100 *males*, the age of each of whom was between 12 and 12.99 years. Of these persons say 10 died within the year. Then $R_{as} = \left(\frac{10}{100}\right)$, which means that the annual death-rate, specific for age and sex (R_{as}), in this community was 0.1 for males between twelve and thirteen years of age, or 100 per thousand.

Specific death-rates are the true and best measures of the force of mortality. They furnish a real and meaningful measure of the probability that certain specified kinds of persons will die within the time period (usually one year) specified in forming the rate. From age specific death-rates (which the English commonly speak of as measures of "mortality at ages") is derived all the really fundamental knowledge which we have of the laws of mortality.

It will be well at this point to put before the reader a definite picture of the form of the specific death-rate curve from all causes. This is done in Table 15 and Fig. 58, in which the rates are specific for quinquennial age groups.

It will be noted that this specific death-rate curve has a characteristic form. Starting at a high point in earliest infancy the specific rate drops till it reaches a low point in the age group 10–14. From that point on it rises steadily, though not entirely evenly, till the end of the life span. The specific death-rates are lower

TABLE 15

AGE AND SEX SPECIFIC DEATH-RATES, PER 1000 LIVING, FROM ALL CAUSES FOR THE
U. S. REGISTRATION AREA (EXCLUSIVE OF NORTH CAROLINA) IN 1910.
(Author's Computation from Census Bureau Data.)

Ages.	Males.	Females.
Under 1	124.4	143.4
1– 4	15.1	13.8
5– 9	3.7	3.5
10–14	2.5	2.4
15–19	4.1	3.7
20–24	6.0	5.2
25–29	6.8	6.1
30–34	8.0	6.8
35–39	9.8	7.8
40–44	11.6	8.9
45–49	14.5	11.0
50–54	18.5	14.6
55–59	25.7	20.6
60–64	36.1	29.4
65–69	51.4	44.3
70–74	75.1	66.8
75–79	112.2	100.9
80–84	168.1	155.9
85–89	237.9	222.7
90–94	313.0	309.7
95–99	410.2	368.9
100 and over	494.2	471.7

in females than in males at every age period in life except the first (under 1).

Specific death-rates can obviously be calculated for each separate cause of death, and will furnish exact and useful information about comparative forces of mortality.

The most extensive compilation of age specific death-rates for the United States is contained in *Mortality Rates 1910–1920*[10] published by the Census Bureau. The text and tables of this report should be studied carefully, in order to get a general understanding of human mortality statistics. They will be found useful for reference in many connections.

It is apparent that the specificity of death-rates may be extended to any degree, provided the necessary data relative to population and to deaths are available. For a really penetrating insight into the forces of mortality, both for purposes of research

and the administration of public health, death-rates ought to be made specific for the following factors:

1. Age.
2. Sex.
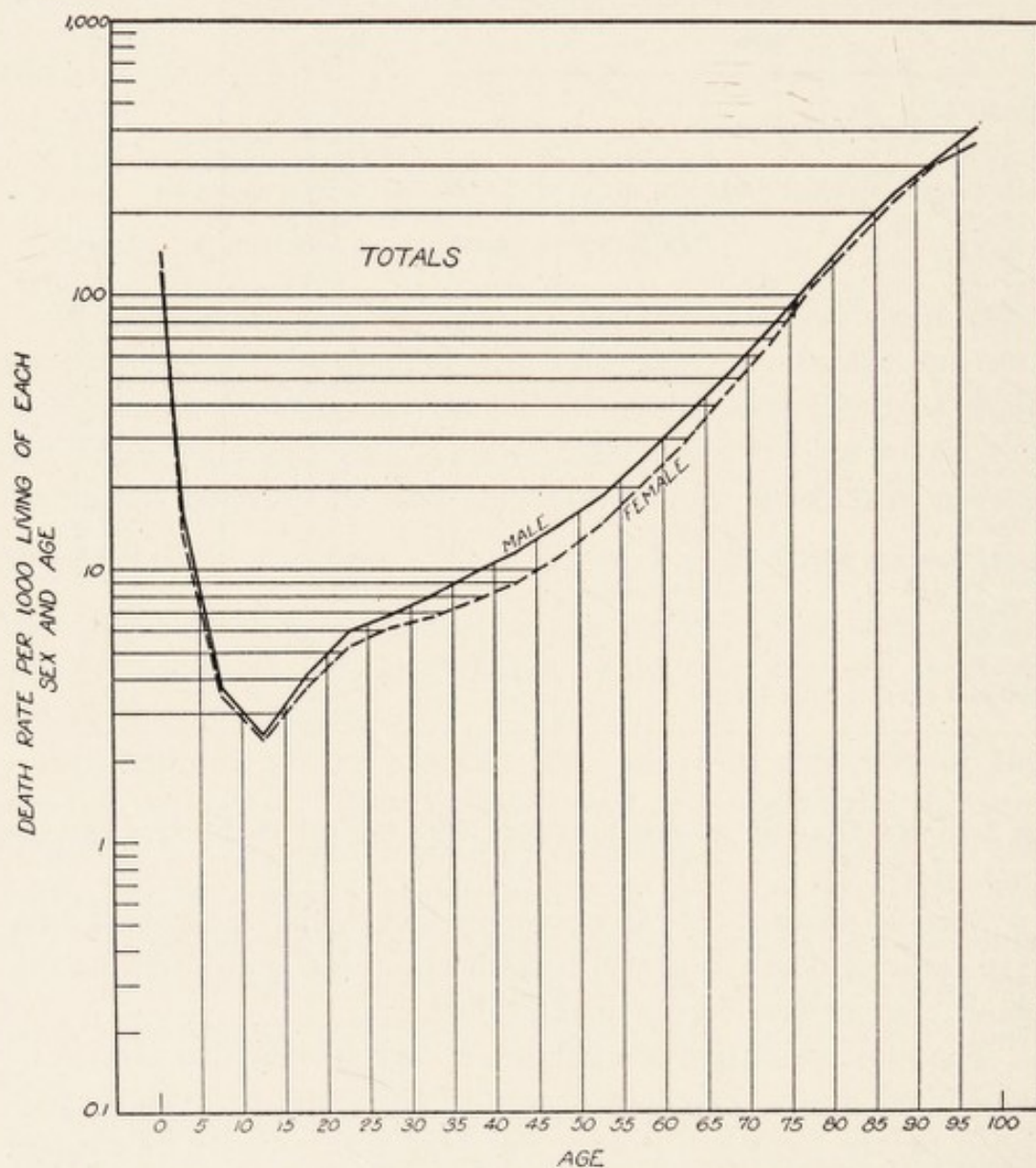
Fig. 58.—Age and sex specific death-rates from all causes for the U. S. Registration Area (exclusive of North Carolina) in 1910. Plotted from data of Table 15, on an arithlog grid.

3. Race (or country of birth of person and parents at least). Race will include color.
4. Occupation.
5. Locality of dwelling (urban or rural).

Each of these factors more or less profoundly influences the force of mortality. Death certificates carry the necessary data (at least theoretically, and actually if properly filled out) regarding deaths. Every ten years the census collects the necessary data regarding the population. If only these data could be properly tabulated and published it would be possible to calculate in census years the death-rates specific for the above five factors. Eventually this will surely be done. The sciences of medicine and hygiene will imperiously demand it. In the meantime we make shift to get along by groping in the dark in respect of all factors except age, sex, and urban or rural dwelling.

The sort of probability which a death-rate specific for the above five factors would measure is, for example, the probability that a male person, aged twenty, native born of native white parents, living in the country and by occupation a farmer, would die within one year.

INFANT MORTALITY RATES

Here the fundamental equation (i) becomes

$$R_i = \left(\frac{D_i}{B}\right),$$

where

R_i = infant mortality rate,
D_i = deaths of infants under one year of age,
B = births.
(Infant mortality rates are usually stated as "per 1000.")

The question which will inevitably occur to the reader's mind at this point is: Why not use the age specific death-rate for age under one as the measure of infant mortality? To which the answer is, Such would be the practice if it were not for the difficulty of getting accurately (or annually) a count of the population under one year of age. But because this is difficult and the results are known to contain large errors, whereas the registration of births is or can be made accurate, the form of death-rate given above is generally used as the measure of infant mortality rather than the simple age specific death-rate under one.

The theory on which the formula for R_i, given above, is based, is obvious. The number of babies *born* in a given year is held to

be at least a fair index of the number of babies exposed to risk of dying within the year under one year of age. Actually, of course, it does not measure the exposed to risk of dying under one year. Because, consider a given calendar year; the baby born on December 1st of that year is only exposed for one month to risk of dying under one year of age *within that calendar year*. But, on the other hand, given a fairly stable population, and accurate birth registration, the error in the absolute value of the infant mortality rate introduced by the relations just mentioned, will be a *constant* one over fairly long periods of time, and, because constant, negligible when the rates are used for comparative purposes.

In the present state of knowledge upon the subject it is impossible to state *exactly* what the probability is that is measured by R_i.

The infant mortality rates, as defined by R_i, for American cities of 100,000 or more population in 1920 are given in Table 16.

It will be noted from this table that there is great variation among the different cities in the rate of infant mortality. This variation has been discussed biometrically elsewhere.[3] Its significance, from the standpoint of public health and preventive medicine, is very great. In the paper referred to it was pointed out that the facts of variation make it clearer where the fundamental administrative problems of control of infant mortality lie than perhaps could be done in any other way. The first step in the solution of any problem is obviously a clear definition of the problem itself. We see, as we pass from city to city, town to town, or rural county to rural county, that the rate of infant mortality varies greatly. In a hypothetic commonwealth where the most perfect administrative control over infant mortality possible or conceivable had been attained this variation would to a considerable extent disappear, the only residue of diversity between communities in respect of infant mortality being such as arose either (1) purely by the operation of chance, that is, from random sampling, or (2) from the racial composition of the several populations, or (3) from fundamentally uncontrollable environmental differences, such as climate, soil, etc., or (4) from some combination of these factors (1) to (3). Now with the actually

TABLE 16

INFANT MORTALITY RATES (DEATHS UNDER ONE YEAR OF AGE PER 1000 LIVE BIRTHS) IN REGISTRATION CITIES OF 100,000 POPULATION OR MORE IN 1920. (Rearrangement of Data from Birth Statistics, 1920, p. 26.)

Cities.	1920 rate.
Lowell, Mass.	135
Fall River, Mass.	129
New Bedford, Mass.	122
Scranton, Pa.	119
Richmond, Va.	114
Pittsburgh, Pa.	111
Kansas City, Kans.	108
Baltimore, Md.	106
Syracuse, N. Y.	105
Detroit, Mich.	104
Buffalo, N. Y.	103
Boston, Mass.	101
Norfolk, Va.	100
Hartford, Conn.	99
Grand Rapids, Mich.	99
Reading, Pa.	99
Cambridge, Mass.	96
Columbus, Ohio.	96
Youngstown, Ohio.	95
Milwaukee, Wis.	94
Bridgeport, Conn.	92
Omaha, Neb.	92
Washington, D. C.	91
Indianapolis, Ind.	91
Philadelphia, Pa.	91
Yonkers, N. Y.	89
Toledo, Ohio.	89
New Haven, Conn.	87
Cleveland, Ohio.	87
Louisville, Ky.	86
Springfield, Mass.	85
Worcester, Mass.	85
New York, N. Y.	85
Dayton, Ohio.	85
Rochester, N. Y.	84
Akron, Ohio.	84
Cincinnati, Ohio.	82
Albany, N. Y.	77
St. Paul, Minn.	73
Salt Lake City, Utah.	72
Los Angeles, Calif.	71
Oakland, Calif.	71
Spokane, Wash.	71
Minneapolis, Minn.	65
San Francisco, Calif.	62
Portland, Ore.	60
Seattle, Wash.	57

existing condition of variation between different communities in respect of infant mortality, it is obvious that there probably are definite and presumably in large degree determinable reasons for

each large particular difference which exists. Just as obviously, before administrative control can effectively wipe out these mortality differences and get all communities at or near the level of the lowest, we must know something about the determining causes upon which they depend. Efforts to reduce infant mortality have in the past been attended with considerable success. With the advance of general sanitation the death-rate under one year of age has fallen enormously. Greenwood quotes some interesting figures on the point from Farr, which we may well reproduce here to show how enormous has been the improvement:

TABLE 17

SHOWING THE REDUCTION IN THE MORTALITY OF INFANCY AND EARLY CHILDHOOD.
(After Greenwood.)

Period.	1730–49.	1750–69.	1770–89.	1790–1809.	1810–29.
Percentage deaths under five years...	74.5	63.0	51.5	41.3	31.8

But after such a decline as these figures indicate, to continue the reduction presents a difficult problem to the administrative official. The easy part of the conflict has happened and is in the past. To continue the good fight with the same relative measure of success, one presumably needs to know more precisely than is now known the pattern of the causal nexus which controls and determines the rate of infant mortality. The problem confronts the administrative official or the altruistic organization in a specific rather than a general manner. City A has a death-rate under one year of age so low that even the most sanguine of hygienic optimists would hardly undertake seriously to reduce it further by any significant amount. In City B, on the other hand, babies die like flies, only somewhat more rapidly. City B differs in many respects from A. Some of these respects are such as to be easily within the power of control of a health official. Others, such as climate or the racial composition of the population, for example, are obviously beyond the possibility of any control or modification. Others lie between the two extremes, and offer practical diffi-

culties of varying degrees. What one needs to know is which particular line of effort will in practice yield the largest return. And it is *real* knowledge, not *a priori* logic, that is wanted. Let a

TABLE 18

Frequency Distribution Showing Variation in Infant Mortality in Birth Registration Area of United States

Deaths per 1000 births in specified years.	Total population cities of 25,000 and over.*				Total population cities of under 25,000.*				Total population rural counties.				White population cities of 25,000 and over.*		White population cities under 25,000.*		White population rural counties.		Colored population cities of 25,000 and over.*		Colored population cities under 25,000.*		Colored population rural counties.	
	1915	1916	1917	1918	1915	1916	1917	1918	1915	1916	1917	1918	1917	1918	1917	1918	1917	1918	1917	1918	1917	1918	1917	1918
0–19					1	1	1		1		3	1					1	9						1
20–39		1				2			2	9	33	32					30	33					2	1
40–59	2	1	5	1	11	4	12	6	49	45	152	174	1			1	71	63				1	8	5
60–79	16	18	22	17	25	27	49	37	130	125	396	342	4	1	6	3	66	64			1		26	22
80–99	27	24	50	43	44	42	76	65	99	107	316	298	14	13	8	5	46	38		1	1	1	29	19
100–119	29	34	45	40	35	39	61	48	52	57	140	165	7	8	8	6	13	16	3	1	3		36	42
120–139	13	14	13	27	20	23	24	38	17	21	59	64	1	5	3	4	3	8	1	1		4	40	37
140–159	9	5	7	13	11	7	13	15	6	15	18	31			1	5	1	2	6	1	3	1	28	30
160–179	1	2	1	1	3	5	5	15	1	2	4	11				2	1		5	2		2	21	19
180–199	1		1	2	3	6	4	4			4	8							6	3	1	4	18	17
200–219								6	1		1	1							3	2	2	1	5	10
220–239							1				1								2	5	5	4	8	12
240–259								1												6	3	2	4	8
260–279																				2	2	1	1	1
280–299																					2	4	2	4
300–319																					1	2	1	2
320–339								1												1	1		1	1
340–359																				1	1			
360–379																							1	1
380–399																			1					
400–419																								1
420–439																								
440–459																								1
460–479																				1				
480–499																								
500–519																								
520–539																								
540–559																								
560–579																								
580–599																								
600–619																							1	
	98	99	144	144	153	156	236	236	358	381	1127	1127	27	27	26	26	232	234	27	27	26	26	232	234

* In 1910.

single example illustrate. It has been maintained that excessive infant mortality is primarily the resultant of the so-called "degrading influence" of poverty, and such a contention stirs a warmly

sentimental feeling of agreement in the minds of a well-meaning public, zealous to do good. This relationship obviously *ought* to be true, therefore to a too-common type of mind it must be and is

TABLE 19

INFANT MORTALITY RATES (DEATHS UNDER ONE YEAR PER 1000 BIRTHS) FOR VARIOUS COUNTRIES

(Rearrangement of Data from Birth Statistics, 1920, p. 40, and Birth Statistics, 1925, Part II, pp. 67–69. Numbers in parentheses denote the year to which the rate applies.)

Country.	Male.		Female.	
Hungary	281.9 (1915)	181.0 (1925)	244.6 (1915)	153.8 (1925)
Russia	264.9 (1909)	....	236.9 (1909)	....
Chile	260.9 (1918)	262.2 (1925)	248.2 (1918)	253.2 (1925)
Ceylon	227.8 (1919)	181.3 (1925)	217.3 (1919)	162.1 (1925)
Austria	204.2 (1913)	170.7 (1920)	174.6 (1913)	141.8 (1920)
Japan	181.8 (1917)	151.1 (1925)	164.2 (1917)	133.3 (1925)
German Empire	177.1 (1914)	115.8 (1925)	149.2 (1914)	93.9 (1925)
Prussia	177.1 (1914)	....	150.2 (1914)	....
Italy	174.5 (1916)	125.6 (1925)	157.7 (1916)	113.0 (1925)
Jamaica	167.7 (1919)	184.7 (1925)	155.4 (1919)	162.6 (1925)
Bulgaria	166.1 (1911)	....	145.7 (1911)	....
Spain	163.5 (1917)	....	146.1 (1917)	....
Serbia	144.7 (1910)	....	132.4 (1910)	....
Belgium	132.1 (1912)	104.4 (1925)	107.2 (1912)	82.4 (1925)
Uruguay	124.7 (1920)	121.1 (1925)	109.5 (1920)	108.7 (1925)
France	122.7 (1913)	98.9 (1925)	101.7 (1913)	78.6 (1925)
Finland	122.6 (1918)	92.9 (1925)	107.5 (1918)	76.6 (1925)
Scotland	112.9 (1919)	103.5 (1925)	89.6 (1919)	77.0 (1925)
Denmark	101.3 (1919)	90.5 (1925)	81.2 (1919)	68.6 (1925)
United Kingdom	101.3 (1919)	89.7 (1922)	79.0 (1919)	68.8 (1922)
England and Wales	100.0 (1919)	84.0 (1925)	77.6 (1919)	65.7 (1925)
Ireland	97.3 (1919)	75.0* 97.7† (1925)	77.5 (1919)	60.5* 74.4† (1925)
Switzerland	96.9 (1918)	63.6 (1925)	79.1 (1918)	52.9 (1925)
United States (birth registration area)	95.1 (1920)	79.5 (1925)	76.1 (1920)	63.3 (1925)
Australian Commonwealth	76.7 (1920)	58.8 (1925)	61.1 (1920)	47.7 (1925)
Sweden	76.6 (1916)	....	62.5 (1916)	....
Norway	70.6 (1917)	54.3 (1924)	57.0 (1917)	45.9 (1924)
The Netherlands	55.2 (1919)	66.0 (1925)	43.9 (1919)	50.3 (1925)
New Zealand	53.6 (1918)	44.0 (1925)	43.0 (1918)	35.6 (1925)

* Irish Free State.
† Northern Ireland.

true. But Greenwood and Brown,[4] in what may fairly be regarded as one of the most thoroughly sound, critical, and penetrating contribution which has yet been made to the problem of infant mortality,

are unable "to demonstrate any unambiguous association between poverty . . . and the death-rate of infants."

The plain fact is that before control or ameliorative measures can be applied with the maximum of efficient economy to the general public health problem of infant mortality there is need to know a great deal more than we now do about the quantitative influence of the general factors which induce spatial and temporal differences in the rate of that mortality. But first it will be helpful to get an adequate conception of the magnitude and character of the differences themselves.

The distribution of variation in infant mortality in cities and rural areas in the United States is shown in Table 18, taken from the paper cited.

The infant mortality rates of various countries are given in Table 19.

It is evident from Table 19 that in the period covered by the data the infant mortality rate declined notably in most of the countries. The outstanding exceptions to this rule are Chile, Jamaica, Uruguay, and The Netherlands. It will also be noted that in every country listed the infant mortality rates are lower for females than for males.

CASE FATALITY RATES

Here the fundamental equation becomes

$$R_F = \left(\frac{D_c}{C}\right),$$

where

R_F = case fatality rate,
D_c = deaths amongst recognized cases of the disease for which the rate is calculated,
C = cases of the disease.
(Case fatality rates are usually expressed as "per 100," occasionally as "per 1000.")

This is, provided age, sex, race, occupation, and locality of dwelling are taken into account, the most refined form of specific death-rate. Because, in the most exclusive sense, those who *have* a given disease are the most truly exposed to risk of dying of that disease at that time. The case fatality rate for typhoid, for ex-

ample, measures the probability that a person who has typhoid will die at that time (*i. e.*, within the course of the attack) of that disease.

Unfortunately, our knowledge of true case fatality rates, even for the commonest diseases, is very meager, because of the inadequacy of the reporting of morbidity. The case fatality rate is, of all the data of biostatistics, the most interesting to the clinician, because of its obvious bearing upon prognosis. The most reliable data in existence on case fatality rates are those derived from the experience of great hospitals. But these do not give a true scientific picture of the situation for two reasons: First, a hospital population is an adversely selected population. In the main, the cases which get into a hospital are those in which the prognosis at a fairly early stage of the disease is thought, often on the best of grounds, to be in some degree unfavorable. Consequently, hospital case fatality rates tend to be unduly high. This state of affairs becomes grossly exaggerated when it is the practice for the hospitals of a city to send to one particular hospital, usually that one supported by the municipality, the greater part of their cases which upon entrance are seen to be either moribund or of very bad prognosis.

In the second place, the treatment of a disease in a hospital may significantly influence, in a differential manner, the course of the disease, as compared statistically with the treatment given on the average outside.

There is a wonderful field open to the quantitatively inclined student of medicine, in the procuring and biometric analysis of accurate case fatality rates.

BIRTH-RATES

The *crude* birth-rate is given by

$$R_B = \left(\frac{B}{P}\right),$$

where

     R_B = crude birth-rate,
     B = number of births (but exclusive of still-births) in a given time, as a year
     P = total living population.
(Crude birth-rates are usually stated as "per 1000" or "per 10,000.")

This rate is obviously a most crude measure of the reproductive capacity of a population. To begin with, not all living persons are exposed to the risk of having a baby. Only females, and those between certain ages (roughly from ten to sixty as outside limits) are liable to this occurrence. Furthermore, under existing conditions of law and public sentiment, in the main the giving of birth to babies is confined to *married* women within the age limits stated. So then to arrive at anything like a true general measure of the force of natality it will be essential first to differentiate between legitimate and illegitimate births, and between living and still-births, and in the second place, to use as the denominator of the rate fraction for legitimate babies the number of married women between the age limits of say ten and sixty years.* For the illegitimate rate the denominator must be, of course, the unmarried women within the same age limits.

As to the reliability and significance of crude birth-rates, as commonly calculated with the total population for denominator, much the same considerations apply as have already been set forth for crude death-rates. They can be used for comparison of different places only with the utmost caution, because differences in the age and sex constitution of the populations compared, quite regardless of their true forces of natality, may have most profound effects upon the rates. So long as the population of a given place is changing only slowly in its composition, its crude birth-rates are fairly comparable *inter se* at different times, as, for example, in successive years. In the routine official birth statistics of the United States it is the crude birth-rate which is tabulated.

For a considerable number of years the crude birth-rate has been falling in most civilized countries. A general conspectus of birth-rate statistics for different countries is shown in Table 20, taken from Knibbs[5] for the years down to and including 1912, and from the Registrar-General's Statistical Review of England and Wales for 1927 (Text, p. 117) for 1913 through 1927.

* The limits usually taken are 15 and 45, 50 or 55. Actually, however, there are occasionally recorded births from mothers under fifteen and over fifty-five years of age. There are not many such, of course, but still it is a physiologic fact that there is a minute risk that some women may become pregnant and bear a child at or very near the extreme ages of ten and sixty that have been stated above.

TABLE 20

CRUDE BIRTH-RATES FOR VARIOUS COUNTRIES—1860–1927—PER 10,000 OF THE POPULATION

Year.	Australia.	England and Wales.	Scotland.	Ireland.	France.	Prussia.	Italy.	Switzerland.	Norway.	Sweden.	Denmark.	Netherlands.	Belgium.	Austria.	Hungary.	Mean.
1860	426	343	356	..	262	386	..	..	..	348	..	319	306	379	..	381
1861	423	346	349	..	269	377	..	..	..	326	318	354	308	372	..	344
1862	433	350	346	..	265	372	..	..	..	334	310	332	301	379	..	342
1863	417	353	350	..	269	395	..	..	..	336	311	364	318	403	..	352
1864	429	354	356	240	266	397	379	..	..	336	303	357	315	403	..	345
1865	421	354	355	257	265	393	385	..	..	328	314	361	314	378	..	344
1866	398	352	354	262	264	393	390	..	..	331	322	354	327	379	421	350
1867	404	354	351	260	264	371	367	..	..	308	305	354	321	366	388	340
1868	405	358	353	268	257	369	354	..	..	275	312	349	325	379	424	341
1869	387	348	343	267	257	379	372	..	..	282	295	343	316	393	426	339
1870	387	352	346	277	255	383	369	298	..	288	305	361	323	396	417	339
1871	380	350	345	281	229	383	370	291	292	304	302	354	310	389	430	331
1872	371	356	349	278	267	397	379	300	297	300	303	360	323	391	410	339
1873	374	354	348	271	260	396	363	299	299	308	308	362	325	399	422	339
1874	368	360	356	266	262	401	349	305	307	309	309	364	326	397	427	334
1875	359	354	352	261	259	407	377	320	312	312	319	366	325	399	450	345
1876	360	363	356	264	262	407	392	330	318	308	326	371	332	400	463	350
1877	350	360	353	262	255	399	370	323	318	311	324	366	323	387	436	343
1878	354	356	349	251	252	387	362	316	311	298	317	361	315	386	431	337
1879	358	347	343	252	251	390	378	308	320	305	320	367	315	392	458	340
1880	352	342	336	247	246	378	339	298	307	294	318	355	311	380	428	323
1881	353	339	337	245	249	370	380	300	300	291	323	350	314	377	429	351
1882	345	338	335	240	248	367	371	291	309	294	324	353	312	391	438	331
1883	348	335	328	235	248	371	372	288	309	289	318	343	305	382	448	328
1884	356	336	337	239	247	376	390	285	310	300	334	349	305	387	456	334
1885	357	329	327	235	243	377	386	280	313	294	326	344	299	376	448	328
1886	354	328	329	232	239	377	370	280	309	298	325	346	296	380	456	328
1887	356	319	317	231	235	377	389	280	308	297	320	337	294	382	442	326
1888	355	312	313	228	231	374	375	278	308	288	317	337	291	379	438	322
1889	346	311	309	227	230	371	383	276	297	277	313	332	295	379	437	313
1890	350	302	304	223	218	366	358	264	303	280	306	329	287	367	403	311
1891	345	314	312	231	226	377	372	278	309	283	309	337	296	370	423	319
1892	337	304	307	225	223	363	362	274	296	270	295	320	289	362	404	309
1893	328	307	308	230	228	375	365	277	307	274	305	338	295	379	426	316
1894	308	296	299	230	223	366	355	273	298	271	301	327	290	367	415	307
1895	304	303	300	233	217	369	349	273	306	275	300	328	285	381	418	310
1896	284	296	304	237	225	369	348	281	304	272	305	327	290	380	405	309
1897	282	296	300	235	222	365	347	283	300	267	298	325	290	375	403	306
1898	271	293	301	233	218	367	335	285	303	271	302	319	286	363	377	302
1899	273	291	298	231	219	363	339	290	309	264	297	321	288	373	393	303
1900	273	287	296	227	214	361	330	286	301	270	297	316	289	373	393	301
1901	272	285	295	227	220	362	326	290	296	270	297	323	294	366	378	300
1902	267	285	293	230	217	355	334	285	289	265	292	318	284	371	389	298
1903	253	285	294	231	211	344	317	274	288	257	287	316	275	353	369	290
1904	264	280	291	236	209	347	329	273	281	258	289	314	271	356	374	290
1905	262	273	286	234	206	335	327	269	274	257	284	308	261	339	363	285
1906	266	272	286	235	206	337	321	269	267	257	285	304	257	350	365	285
1907	268	265	277	232	197	330	317	262	264	255	282	300	253	340	367	281
1908	266	267	281	233	201	327	337	264	263	257	285	297	249	337	369	282
1909	267	258	273	234	195	317	327	255	263	256	282	291	237	334	377	278
1910	268	251	262	233	196	305	333	250	261	247	275	286	237	325	357	273
1911	272	244	256	232	187	294	315	242	259	240	267	278	229	314	350	265
1912	286	238	259	230	190	289	324	241	256	237	267	281	226	313	363	267

TABLE 20—*Continued*

Year.	Australia.	England and Wales.	Scotland.	Irish Free State.	Northern Ireland.	France.	Germany.	Italy.	Switzerland.	Norway.	Sweden.	Denmark.	Netherlands.	Belgium.	Austria.	Hungary.	Mean.
1913	282	241	255	228		182	275	317	232	251	232	256	283	224	297	343	260
1914	279	238	261	226		179	268	311	224	251	229	256	283	204	233	345	252
1915	271	218	239	220		116	204	305	195	236	216	242	263	161	184	236	220
1916	266	210	229	210		95	152	240	189	242	212	244	266	129	147	168	200
1917	263	178	203	198		105	139	195	185	251	209	237	262	113	139	160	189
1918	250	177	205	200		122	143	181	187	246	203	241	250	113	141	153	187
1919	235	185	217	200		126	200	214	186	227	198	226	244	163	180	289	206
1920	255	255	281	222		213	259	318	209	261	236	254	283	221	224	324	254
1921	250	224	252	202		207	253	303	208	240	215	240	274	218	229	316	242
1922	247	204	235	195	233	193	229	302	196	231	196	222	259	204	232	306	230
1923	238	197	228	205	239	192	210	294	194	225	188	223	260	204	225	292	226
1924	232	188	219	211	227	187	205	284	188	211	181	218	251	199	217	268	218
1925	229	183	213	208	220	189	207	278	184	200	175	210	242	198	206	283	214
1926	220	178	209	206	225	188	195	272	182	197	169	205	238	190	192	273	209
1927	217	166	198	203	213	181	183	269	174	182	161	..	231	182	178	252	199
Mean	318	292	300	238*		220	333	337	261	279	268	290	318	270	335	376	

* This mean is for Ireland as a whole for the years 1864–1921.

SPECIFIC BIRTH-RATES

Age specific birth-rates may be formed if the necessary statistical data are available in accordance with exactly the same principle as was used in forming age specific death-rates. The number of women of a given age, or within a given small age group, is used as the denominator, and the number of babies born in a year to women

TABLE 21

AGE SPECIFIC BIRTH-RATES COMPUTED FROM AUSTRALIAN (1911) DATA. (Data from Knibbs,[5] p. 325.)

Age of mothers.	Total married women.	Number who bore a child during the year.	Specific birth- (or fertility) rate.†
19 and under	8,716	4,146	476
20–24	65,959	25,957	394
25–29	110,591	33,817	306
30–34	113,310	25,682	227
35–39	105,550	16,839	160
40–44	95,573	6,763	71
45 and over	82,933	713	9
Totals	582,632	113,917	196

† Births per 1000 married women of indicated age.

in this age group as the numerator of the rate fraction. Such figures measure the *fertility* of women of the specified class. Matthews Duncan[11] long ago showed that the fertility rate varied in a definite and lawful manner with age. Some recent statistics to the same purpose are presented in Table 21, adapted from Knibbs.[5]

It is to be understood that the figures in Table 21 do not refer to first births only, but to all births regardless of their order. It is seen that the age specific birth-rates are highest in the earlier years, and decrease in value with advancing age. It will be remembered that all Australian birth-rates are relatively high as compared with various other countries.

There is a good deal of confusion in the use of the terms "fertility" and "fecundity." The writer some years ago discussed* this terminology in the following words:

"We would suggest that the term 'fecundity' be used only to designate the innate potential reproductive capacity of the individual organism, as denoted by its ability to form and separate from the body mature germ cells. Fecundity in the female will depend upon the production of ova and in the male upon the production of spermatozoa. In mammals it will obviously be very difficult, if not impossible, to get reliable quantitative data regarding pure fecundity. On the other hand, we would suggest that the term 'fertility' be used to designate the total actual reproductive capacity of *pairs* of organisms, male and female, as expressed by their ability when mated together to produce (*i. e.*, bring to birth) individual offspring. Fertility, according to this view, depends upon and includes fecundity, but also a great number of other factors in addition. Clearly it is fertility rather than fecundity which is measured in statistics of birth of mammals."

Standardized and corrected birth-rates of populations may be calculated on principles discussed in Chapter IX for death-rates.

* Pearl, R., and Surface, F. M.: Data on the Inheritance of Fecundity Obtained from the Records of Egg Production of the Daughters of "200-egg" Hens, Maine Agr. Exp. Sta. Annual Report, 1909, pp. 49–84.

MORBIDITY RATES

The fundamental equation for a crude morbidity rate is as follows:

$$R_M = \left(\frac{M}{P}\right)$$

where

R_M = crude morbidity rate,
M = number of persons sick, either from all causes together or from some one particular cause (in the latter case the rate, of course, is the crude morbidity rate for that disease) in a given stated time,
P = the total population.
(Morbidity rates are stated sometimes as "per 1000," sometimes as "per 10,000," and sometimes as "per 100,000.")

Such a figure measures the *incidence rate* of sickness in the population, either in general or for particular diseases. It is subject to many, if not all, of the same difficulties that crude death- and birth-rates are. Unfortunately, however, there exist so few statistics relatively regarding morbidity that it is somewhat academic to be too critical regarding any morbidity rates. Anything in the nature of age and sex specific morbidity rates is practically non-existent at the present time.

But there is no doubt that morbidity statistics are, by and large, of all statistics the most potentially valuable to the administrative public health official. The United States Public Health Service is taking a leading position in the development of morbidity statistics. The student who is particularly interested in the subject should apply to that Service for publications.

It is not fair to measure the effectiveness of public health work entirely in terms of mortality, because much of its effectiveness in actual fact has nothing to do with mortality, but with morbidity. This fact shows itself in every-day language. We have boards of *health*, not boards of mortality, and quite rightly so. Some of the human ailments against which public health work directs its most effective work are diseases which at the worst are not particularly fatal. An example is uncinariasis—hookworm disease. It would be folly to attempt to measure the social worth of the campaign against this distressing ailment in terms of mortality. What this work accomplishes is not primarily a reduction in mortality, but a positive increase in the sum total of human happiness and well-

being, individual, social, and economic. The same considerations apply to many other lines of public health work, indeed, to most of them. The most important causes of *death*, taken by and large, are not the ones against which hygiene and sanitation are, in the present state of knowledge and of the organization of society, particularly effective. But this fact should in nowise be taken to mean that public health efforts have no great value.

DEATH RATIOS

A death ratio measures the probability that in a given total number of deaths from all causes a particular one will be from one particular cause, say tuberculosis of the lungs. The fundamental equation is

$$Rt_D = \left(\frac{D'}{D}\right),$$

where

Rt_D = the death ratio,
D' = deaths from a particular cause (or group of causes) in a specified time interval,
D = total deaths from all causes in the same time interval.
(Death ratios are usually stated as "per 100," or "per 1000.")

This statistical constant has been much criticized, and has in consequence largely fallen out of general use, on the ground that both D' and D are variable quantities affected by the same biologic forces, and that in consequence it is never possible to tell with any degree of accuracy what portion of the derived value of Rt_D is due specifically to D' and what to D. Undue weight has undoubtedly been given to this criticism. In principle the same criticism applies to any rate, for P in a crude death- or birthrate, or any more precisely defined part of P, is not an invariable quantity. As a matter of fact Rt_D may be a very valuable statistical datum if used intelligently, and there is no statistical datum whatever that can be relied upon to give correct results if unintelligently employed. The criterion as to the usefulness of Rt_D is simply and solely whether the probability which it measures is, in the particular premises set by the study in hand, an intelligible probability. If it is, Rt_D has validity and usefulness.

The death ratio has in recent years been most effectively em-

ployed in researches on tuberculosis by Greenwood and Tebb. It has been employed by Arne Fisher[12] as a basis for computing life tables from a knowledge of deaths alone.

THE BIRTH-DEATH RATIO OR VITAL INDEX

The writer[7] has elsewhere suggested that the term "vital index" be used to designate that measure of a population's condition which is given by the ratio of births to deaths within a given time. It may fairly be said that there is no other statistical constant which furnishes so adequate a picture as this of the net biologic status of a population as a whole at any given moment. If the ratio 100 Births/Deaths is greater than 100, the population is in a growing and in so far healthy condition. If it is less than 100, the population is *biologically* not holding its own. Depopulation may not be actually occurring if there is a sufficient amount of immigration to make up the deficiency in births. But fundamentally and innately the condition is not a sound one from a biologic standpoint, though under certain circumstances it may conceivably be from a social standpoint. It is curious, in view of the obvious significance of this statistic, the vital index of a population, that so little attention is paid to it by demographers. It is a highly sensitive measure of the immediate biologic status, in the evolutionary sense, of a nation or any subgroup of people. Wernicke* discussed it in 1889, but did not use it in the most effective manner or form. Sundbärg† proposed its use as a "measure of civilization" of different peoples. Rubin‡ criticized Sundbärg, but only in respect of technic, proposing as a measure of civilization D^2/B in place of D/B, where D = deaths and B = births. Pell§ has dealt with the idea implicit in the birth/death ratio, but in an inadequate manner. The most extensive and comprehensive discussion of the vital

* Wernicke, J.: Das Verhältniss zwischen Geborenen und Gestorbenen in historischer Entwicklung und für die Gegenwart in Stadt und Land, Jena, 1889, vi and 91 pp. 8vo.

† Sundbärg, G.: Dodstalen sassom Kulturmätare, Nationalökonomiska Föreningens Forhandlingar, i Aaret, 1895, Stockholm, 1896.

‡ Rubin, M.: A Measure of Civilization, Jour. Roy. Stat. Soc., vol. 60, pp. 148–161, 1897.

§ Pell, C. E.: The Law of Births and Deaths, London (Unwin), 1921, 192 pp.

index in the literature is that by Sweeney.[18] He computed the vital index for all countries of the world for which adequate sta-

TABLE 22

VITAL INDICES OF VARIOUS ELEMENTS IN THE POPULATION OF REGISTRATION STATES, CITIES IN REGISTRATION STATES, AND RURAL PORTIONS OF THE REGISTRATION STATES IN THE BIRTH REGISTRATION AREA (1915–18 INCLUSIVE)

State and group		1915—Vital Index				1916—Vital Index				1917—Vital Index				1918—Vital Index			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
Connecticut	Cities	82.9	355.8	94.8	195.4	81.8	340.6	85.1	189.3	90.4	331.0	82.0	196.1	72.9	219.2	75.1	143.7
	Rural	75.7	292.0	60.4	149.0	73.5	282.9	81.3	146.4	77.8	293.2	82.7	149.2	65.7	212.2	65.0	120.2
	Total	80.5	339.7	80.0	180.9	79.2	326.6	84.3	176.7	86.4	322.6	82.1	182.8	70.8	217.7	73.4	137.3
District of Columbia	Total	117.0	97.5	85.2	123.0	119.7	93.5	87.3	125.9	126.6	103.7	85.5	131.7	102.1	78.5	64.7	104.2
Indiana	Cities	•	•	•	•	•	•	•	•	144.0	174.5	71.5	158.2	124.1	142.5	66.3	134.9
	Rural	•	•	•	•	•	•	•	•	172.5	54.3	59.1	166.3	153.2	53.1	53.8	149.0
	Total	•	•	•	•	•	•	•	•	162.7	121.8	68.2	163.3	142.8	108.5	63.3	143.6
Kansas	Cities	•	•	•	•	•	•	•	•	149.8	58.1	70.6	150.0	116.2	72.3	65.6	114.7
	Rural	•	•	•	•	•	•	•	•	223.7	68.6	74.5	208.2	190.4	50.5	67.4	177.9
	Total	•	•	•	•	•	•	•	•	207.3	26.6	72.2	195.3	171.9	50.1	66.4	162.2
Kentucky	Cities	•	•	•	•	•	•	•	•	135.8	35.9	47.4	123.1	105.6	25.1	34.7	98.5
	Rural	•	•	•	•	•	•	•	•	241.4	29.7	91.0	236.4	203.1	38.6	74.5	199.6
	Total	•	•	•	•	•	•	•	•	221.9	29.7	76.0	202.9	183.1	29.3	60.3	177.0
Maine	Cities	75.6	188.6	71.4	131.5	73.8	163.8	25.0	122.4	79.4	173.8	75.0	128.4	70.7	124.9	53.8	106.2
	Rural	105.6	156.0	87.5	136.4	105.7	146.2	18.7	135.9	115.0	161.6	8.3	145.5	96.5	112.5	21.1	119.9
	Total	98.8	169.0	80.0	135.1	98.1	151.3	21.9	132.3	106.3	166.7	85.0	143.0	90.2	117.8	34.4	116.2
Maryland	Cities	•	•	•	•	138.3	166.4	82.9	136.8	137.7	152.1	80.8	152.7	96.2	107.7	63.7	105.9
	Rural	•	•	•	•	177.0	90.1	128.6	173.5	173.3	82.6	125.1	168.7	129.8	67.6	92.8	126.3
	Total	•	•	•	•	157.5	144.9	106.8	164.6	155.1	132.2	103.1	160.1	111.9	96.8	78.8	114.9
Massachusetts	Cities	86.7	276.2	113.2	186.4	87.0	251.2	101.1	176.3	92.1	246.5	111.3	179.7	70.1	171.5	97.0	129.6
	Rural	80.5	226.4	70.6	145.1	79.0	202.1	73.5	135.3	77.4	207.4	128.4	135.6	63.0	147.8	86.2	104.8
	Total	85.1	267.1	105.8	177.0	85.1	242.3	96.0	167.2	88.5	239.6	113.3	169.9	68.5	167.4	95.4	124.5
Michigan	Cities	104.5	234.5	86.3	205.4	138.0	227.2	66.7	195.7	143.0	226.4	79.1	198.5	132.3	192.6	86.8	179.2
	Rural	182.2	143.5	83.1	197.9	172.2	140.5	68.1	186.7	171.7	139.4	64.1	185.3	155.9	123.8	77.0	167.4
	Total	165.0	187.8	85.2	201.2	157.2	184.4	67.1	190.9	158.8	184.2	74.5	191.5	145.2	159.7	84.1	173.0
Minnesota	Cities	173.0	166.3	59.9	211.5	163.2	148.2	51.4	194.5	170.8	136.3	67.7	194.5	137.0	105.1	61.1	156.4
	Rural	282.0	118.1	18.1	264.2	277.3	104.6	88.2	252.8	289.9	96.7	50.0	254.9	208.0	79.3	42.9	195.1
	Total	240.7	134.9	51.8	244.9	232.1	120.2	55.3	230.5	242.1	111.2	66.1	231.5	181.3	88.7	59.2	180.7
New Hampshire	Cities	70.3	293.7	160.0	163.0	72.4	248.4	200.0	153.9	69.2	239.5	114.3	147.1	60.9	164.1	100.0	113.7
	Rural	90.9	172.0	50.0	124.7	90.0	150.3	30.0	121.3	87.4	133.3	60.0	113.3	69.7	101.0	90.0	89.3
	Total	82.8	240.9	110.0	140.9	83.1	206.7	69.2	135.4	79.8	195.2	91.7	128.4	65.9	139.3	71.4	100.7
New York	Cities	88.4	273.5	94.9	179.5	88.5	255.4	101.6	172.5	95.6	246.0	96.3	175.2	79.4	187.1	84.5	137.1
	Rural	109.6	140.7	85.3	128.1	107.4	138.5	79.7	125.6	105.1	128.4	69.9	121.4	88.5	106.6	54.8	101.2
	Total	95.8	253.6	93.6	166.5	94.2	238.2	98.5	160.8	98.5	228.7	92.6	161.8	82.1	176.3	80.8	128.4

tistics were available, and over as long a period of time in each case as he could. The student should read his book.

In Table 22 are shown four vital indices for urban, rural, and

TABLE 22—*Continued*

State and group		1915—Vital Index				1916—Vital Index				1917—Vital Index				1918—Vital Index			
		A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
North Carolina	Cities									148.0	60.9	78.1	147.5	99.1	32.9	62.2	98.7
	Rural									266.8	160.3	190.3	270.4	223.7	47.0	145.0	224.4
	Total									255.4	106.3	173.1	258.2	209.2	41.6	33.8	209.3
Ohio	Cities									136.1	210.5	64.6	167.1	117.2	66.1	66.1	39.3
	Rural									157.5	113.3	73.8	156.3	138.3	98.5	70.1	137.5
	Total									147.0	182.2	67.2	162.0	127.7	143.3	67.2	138.5
Pennsylvania	Cities	117.2	273.3	95.2	179.2	110.4	253.8	87.6	166.5	114.6	243.1	74.4	166.3	81.1	153.7	59.3	112.5
	Rural	152.5	385.7	87.4	207.6	141.8	353.0	80.7	191.0	146.7	332.5	76.0	192.9	104.0	174.2	56.4	128.3
	Total	135.7	314.9	93.3	193.1	126.7	290.4	86.0	178.4	130.9	276.3	74.7	179.0	92.6	162.2	58.6	120.0
Rhode Island	Cities	74.3	219.5	83.2	158.3	72.5	216.4	76.1	152.7	79.2	217.8	89.5	159.2	64.0	172.5	91.6	126.0
	Rural	70.5	319.4	65.8	157.4	79.7	331.9	113.6	175.2	84.0	358.7	123.8	182.9	63.8	207.5	29.8	127.7
	Total	73.5	232.2	80.6	158.2	73.9	231.2	78.9	156.5	80.1	234.8	91.9	163.1	64.0	117.8	83.7	126.3
Utah	Cities									244.3	91.2	140.0	227.3	186.0	79.4	75.0	181.0
	Rural									390.7	97.1	33.3	338.9	312.7	71.6	66.7	274.2
	Total									339.4	94.7	81.8	297.8	265.4	74.8	72.4	238.5
Vermont	Cities	109.3	158.4	100.1	147.9	109.4	153.0	100.0	147.5	114.7	147.8	100.0	147.3	92.7	72.1	—	104.7
	Rural	121.3	138.4	133.3	147.2	110.7	135.5	75.0	134.8	114.3	134.7	125.0	136.6	96.5	91.3	100.0	111.5
	Total	119.9	142.5	120.0	147.3	110.5	139.0	77.7	136.7	114.4	137.3	116.7	138.2	95.9	86.5	66.7	110.4
Virginia	Cities									163.1	163.6	91.6	170.7	115.6	101.5	71.6	117.4
	Rural									255.3	125.4	159.2	252.6	200.7	103.6	137.0	200.1
	Total									233.4	144.7	139.2	232.6	177.7	102.4	117.1	176.8
Washington	Cities									169.1	123.2	63.0	184.8	132.2	84.3	57.6	140.3
	Rural									201.5	116.2	42.3	203.8	168.4	91.7	56.8	168.0
	Total									286.6	119.9	58.5	194.8	150.1	87.5	57.4	153.6
Wisconsin	Cities									178.4	142.2	75.0	194.8	143.2	115.9	68.2	156.9
	Rural									266.0	57.6	37.1	209.1	217.9	57.5	65.2	185.9
	Total									231.5	89.8	60.4	203.6	187.9	81.6	67.2	174.8
Totals	Cities	100.5	267.5	93.1	181.7	100.5	247.8	89.2	172.5	117.0	228.3	79.6	173.0	93.2	166.9	66.8	132.0
	Rural	141.1	215.4	82.5	179.0	137.7	199.9	109.0	170.1	177.7	156.5	146.2	187.4	144.8	118.8	118.4	150.8
	Total	117.8	252.4	91.4	180.7	116.3	234.1	94.2	171.6	148.1	205.2	114.3	179.8	118.8	151.8	93.7	140.6

* Not in the Birth Registration Area in designated year.

total births and deaths of each state in the Birth Registration Area for the years 1915 to 1918 inclusive.

The significance of the several indices is as follows:

$$\text{Vital index } A = \frac{100 \text{ (births of whites of native parents)}}{\text{Deaths of all native whites}}$$

In this index the births and deaths come from an identical group of the population. The children born were, of course, native, and their parents were also native born. The deaths were of native born, *i. e.*, the same group as the parents of the births. All racial elements (white) are included in births and deaths, but all are Americans in the sense of nativity.

$$\text{Vital index } B = \frac{100 \text{ (births of whites, both parents foreign)}}{\text{Deaths of foreign-born whites}}$$

Here again both births and deaths come from an identical group. The births are children of foreigners in this country. The deaths are of foreigners in this country.

$$\text{Vital index } C = \frac{100 \text{ (births of negroes)}}{\text{Deaths of negroes}}$$

This needs no discussion.

$$\text{Vital index } D = \frac{100 \text{ (births of whites)}}{\text{Deaths of whites}}$$

This is for comparison with C. Both C and D are true vital indices, in the sense that the parents of the births in the numerator are drawn from the same population group as the deaths in the denominator.

Unfortunately, on the basis of present published official compilations of statistics, these four are the only significant vital indices which can be drawn up. For any really deep understanding of what the biologic effect is of racial fusion, and of a new environment, on the net vitality of populations we ought to have a whole series of racially specific vital indices. Here again there is little practical hope of getting these from purely official sources. Some one must come forward and finance a comprehensive and thorough investigation along these lines from outside.

The facts about Indices A, B, C, and D are set forth in Table 22. In this table a figure in *italics* indicates that the absolute number of births and deaths on which the index is based is in each case less than 100. It will be noted that there are few such cases, and that they are practically all among the negroes of the northern states.

This table presents many points of interest. We may compare vital indices *A* and *B*, which indicate the relative biologic vigor of the native-born and the foreign-born populations in this country. Taking *totals* (the last line of the table) we note that for each year Index *B* is much larger than Index *A*. Generally speaking the foreign population produced in this country approximately two or more babies for every death, on the average during the years here studied with the exception of the last. The native population (as defined in Vital Index *A*) produced only a small fraction over one baby for each death. In other words, the portion of the native population dealt with in Table 22, even when so broadly defined as by Index *A*, was, in the period 1915–18, in about the same state as France before the war, and not in as vigorous a state as the French population was in 1920 and 1921.

The vital indices of Table 22 are crude indices. We need age-specific vital indices for native- and foreign-born populations.

Let us put the matter in this way: Suppose that a gigantic corral were constructed with two compartments. Suppose that, further, there were put into one of these compartments, on a given date, all the native-born women aged twenty to twenty-four inclusive say, while into the other compartment were put all the foreign-born women in the country of the same ages. Suppose them all to be told that they were to stay there for one year, but that men could have free access to the corrals for purposes of repro-duction. Finally, suppose that similar corrals were constructed, and the women impounded in them, for each age group, from say ten to fourteen at one extreme to fifty-five and over at the other extreme.

In any one compartment of any one corral during the year (*a*) some of the women would have babies, and (*b*) some of the women would die. If we kept statistical record of these events we could, at the end of the year, calculate the age specific vital index for each group of women. It would not be the general population vital index because no male deaths were included. But it would be an age-specific vital index for the females as reproductive units.

The results of exactly such an experiment for the women of the Birth Registration Area in the year 1919 are shown in Table 23.

<div align="center">

TABLE 23

AGE-SPECIFIC VITAL INDICES FOR NATIVE-BORN AND FOREIGN-BORN WOMEN IN
B. R. A. 1919

</div>

Ages.	Births from mothers born in U. S.	Deaths of native-born females.	Vital indices for native women.	Births from foreign-born mothers.	Deaths of foreign-born females.	Vital indices for foreign women.
10–14	391	5,002	7.82	15	268	5.60
15–19	77,048	7,763	992.50	10,768	759	1418.71
20–24	258,876	11,854	2183.87	74,247	2,120	3502.22
25–29	250,548	13,189	1899.67	102,429	3,317	3088.00
30–34	166,777	11,813	1411.81	83,326	3,583	2325.59
35–39	101,638	10,603	958.58	56,414	3,723	1515.28
40–44	33,832	9,511	355.71	18,878	3,566	529.39
45–49	3,202	10,092	31.73	1,866	4,120	45.29
50–54	68	10,926	.62	54	4,968	1.09
55 and over	26	96,919	.03	13	47,478	.02
Totals	892,406	187,672	....	348,010	73,902	

The figures in Table 23 show plainly enough that at every age between fifteen and fifty-four inclusive the foreign-born women have higher *specific* vital indices than native-born women. How much so is shown graphically in Fig. 59.

As a reproductive machine the foreign-born woman far excels the native born. For each native-born woman dying between twenty and twenty-four years of age, the native-born women as a group produce approximately 22 babies. But for each foreign-born woman dying between twenty and twenty-four, the foreign-born women as a whole produce 35 babies. It is in these five years that women, under conditions of life as now socially organized in the United States, do their best work biologically for the race, "best" being taken here in the sense of biologic efficiency and economy.

If we had specific vital indices for populations of lower animals in different environmental situations we should be in a position to know a great deal more than we now do as to the method of evolution. For it is the net balance between births and deaths which is the most significant information that can be had about the progress of the struggle for existence.

It may be objected in Table 23 that we have put all births

(both male and female) against only female deaths. The thought
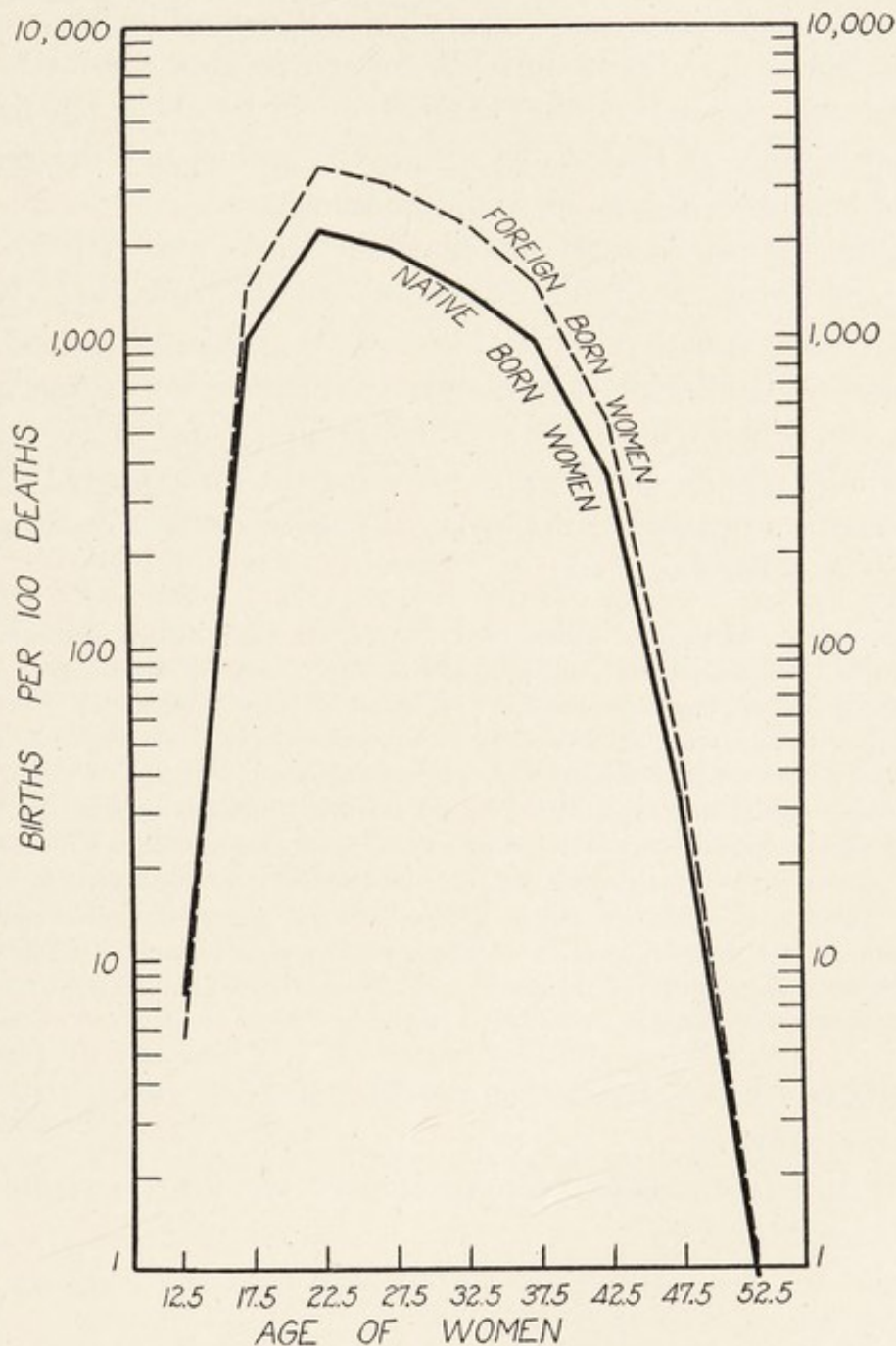in doing this was that, after all, females have to produce *all* the

Fig. 59.—Showing the differences in specific vital indices for native-born and
foreign-born women in 1919. Solid line, native-born women; dash line, foreign-
born women.

babies, whether the latter are boys or girls. If one wishes to
postulate the problem in this way: how many new reproductive

machines (females) do women of a specified age produce as a class for each similar reproductive machine lost by death?·then, of course, one should take only female births in computing the specific vital indices. The result would be, of course, that the births and consequently the indices in Table 23 would be about one-half as large absolutely as they really are in that table, but the general *form* of the curve of Fig. 59 would be unchanged.

In the Eighth Annual Report of the Census on "Birth, Still-birth, and Infant Mortality Statistics" for the year 1922 (Washington, Government Printing Office, 1924) Table M (pp. 17, 18) and Table N (p. 19) give age-specific vital indices for the age group 15–44, and separately in five-year age groupings for native whites, foreign-born whites, and negroes, covering the three years 1920–22. These figures will repay careful study. The Report (p. 17) comments on them as follows:

"For native women aged fifteen to forty-four, the three highest indices are for Utah (3100), Nebraska (3014), and Virginia (2898.7), and the three lowest are for California (1378.7), Massachusetts (1424.7), and New York (1592.9).

"Native white women aged twenty to twenty-four have a vital index of 3630.6, while foreign-born white women of the same age have a vital index of 4795.3. For native white women of this age the lowest vital index (2592.3) appears for Massachusetts and the highest (5808.6) for Nebraska. For foreign-born white women of this age comparatively high indices appear, for example, for Pennsylvania (6697.4) and for Connecticut (5822.2). Similar differences throughout the table emphasize once more the fact that foreign-born white women as a class have more children than native white women. The vital indices available for Negroes are much lower than those for native white women."

LeBlanc[14] has discussed age specific vital indices for the Japanese population.

For further discussion of vital indices see Pearl and Burger,[8] Pearl[*] and Miner.[†]

[*] Pearl, R.: Seasonal Fluctuations in the Vital Index of a Population, Proc. Nat. Acad. Sci., vol. 8, pp. 76–78, 1922.

[†] Miner, J. R.: The Probable Error of the Vital Index of a Population, Ibid., vol. 8, pp. 106–108, 1922.

SUGGESTED READING

1. Howard, W. T.: The Real Risk-rate of Death to Mothers from Causes Connected with Childbirth, Amer. Jour. Hyg., vol. 1, pp. 197–233, 1921.
2. Lotka, A. J.: The Stability of the Normal Age Distribution, Proc. Nat. Acad. Sci., vol. 8, pp. 339–345, 1922.

3. Pearl, R.: Biometric Data on Infant Mortality in the United States Birth Regis-
tration Area, 1915–1918, Amer. Jour. Hyg., vol. 1, pp. 419–439, 1921.

4. Greenwood, M., and Brown, J. W.: An Examination of Some Factors Influencing
the Rate of Infant Mortality, Jour. Hyg., vol. xii, pp. 5–45, 1912.

5. Knibbs, G. H.: The Mathematical Theory of Population, of its Character and
Fluctuations, and of the Factors Which Influence Them, Appendix A, vol. i,
Census of the Commonwealth of Australia, 1917.

   (The student will find this a useful reference work, containing many sug-
   gestive ideas and results. The present writer disagrees fundamentally with
   some of the underlying philosophy and technic of the mathematical treatment
   developed by Knibbs, and believes that the beginner will do well to leave that
   part of the work strictly alone, as being a somewhat unsound guide.)

6. Greenwood, M., and Tebb, A. E.: An Inquiry Into the Prevalence and Etiology
of Tuberculosis Among Industrial Workers, with Special Reference to Female
Munition Workers, Med. Res. Comm., Spec. Rept. Ser. No. 22, London, 1919.
   (Excellent critical discussion of death ratios.)

7. Pearl, R.: The Vitality of the Peoples of America, Amer. Jour. Hyg., vol. i, pp.
592–674, 1921.

8. Pearl, R., and Burger, M. H.: The Vital Index of the Population of England and
Wales, 1838–1920, Proc. Nat. Acad. Sci., vol. 8, pp. 71–76, 1922.

9. Farr's Vital Statistics. (For complete reference see list at end of Chapter II,
Item 12.)

   (To get a real grasp of the meaning and use of death- and birth-rates every
   student should read and read again the writings of the great master, Farr. There
   one will see how, by the use of such rates, most of what can now be regarded
   as the laws of mortality and natality were worked out.)

10. Bureau of the Census: Mortality Rates 1910–1920, with Population of the Federal
Censuses of 1910 and 1920 and Intercensal Estimates of Population, Washington
(Government Printing Office), 1928, 681 pp.

11. Duncan, J. Matthews: Fecundity, Fertility, Sterility, and Allied Topics, Edin-
burgh (A. and C. Black), 1866, pp. xvi + 378.

12. Fisher, A.: An Elementary Treatise on Frequency Curves and Their Application
in the Analysis of Death Curves and Life Tables. Translated from the Danish
by E. A. Vigfusson, New York (Macmillan), 1922, pp. xv + 240.

13. Sweeney, J. S.: The Natural Increase of Mankind, Baltimore (Williams and
Wilkins Co.), 1926, 185 pp.

14. LeBlanc, T. J.: Specific Vital Indices for Japan, 1925, Human Biology, vol. 1,
pp. 198–213, 1929.

CHAPTER VIII

LIFE TABLES

A LIFE table is a particular conventional method of presenting the most fundamental and essential facts about the age distribution of mortality. It has many points of usefulness. The chief one, and the one which is mainly responsible for having secured for life tables the position of respectability and importance that they now enjoy, is that on them depends the successful operation of the great commercial enterprise which is somewhat naïvely called "life insurance." But beyond all this commercial application life tables have, in respect of their fundamental structure, an essential place in vital statistics. It is impossible for the student fully to grasp the significance of certain matters which will be discussed as we proceed unless he knows beforehand the main features, at least, of the anatomy of a life table. It is to furnish this background that the present chapter finds a place in this book. It is not the intention to go at all into the details as to how life tables are constructed, for two reasons: In the first place, there is an extensive and easily available literature on the subject. In the second place, the details of actuarial science are not likely to be of immediate interest or use to the medical man.

THE ANATOMY OF A LIFE TABLE

Suppose one could so arrange affairs that 100,000 babies would be born all at the same identical instant of time, and in such circumstances that each one could be observed then and subsequently without break of continuity in the observations until the very last one had died as a centenarian. If a record were kept of the course of events, something like this would be bound to emerge. Some of the 100,000 babies would die in the first day after birth. Let us say there were observed to be d_1 of these. Then on the morning of the second day there would be surviving out of the original 100,000 who started life together the day before only

$$l_1 = 100,000 - d_1.$$

It is perceived that when this experiment started there were exposed to risk of dying within the first day, or, in other words, within the first twenty-four hours after birth, 100,000 individuals. Within this time period there actually died d_1 individuals. Therefore it follows from the principles laid down in the last chapter that the specific death-rate in this first day, provided we consider a day as a not further divisible unit or instant of time, which is to say that we consider the whole 100,000 to be exposed to risk over the whole day,*

$$q_1 = \frac{d_1}{100,000}.$$

But both our observations and the babies are continuing. In the second day d_2 individuals were observed to die. Hence on the morning of the third day there were surviving

$$l_2 = (100,000 - d_1) - d_2$$

and the death-rate during the second day was, on the same assumptions as before,

$$q_2 = \frac{d_2}{(100,000 - d_1)}$$

We have postulated that these observations are to be carried on without break until the last one of the original group has passed away. If so, the bookkeeping at the end of the process will at least contain columns as follows:

x	d_x	l_x	q_x
(Age, in days, months, years, or whatever units one pleases, but best stated as an interval.)	(The number dying within the age interval stated in the x column.)	(The number surviving at the beginning of the age interval stated in the x column.)	(The rate d_x/l_x, i. e., the number dying in the age interval given in the x column divided by the number of survivors at the beginning of that interval.)
0-1 1-2 etc.		100,000	

---

* This assumption is, of course, of an arbitrary character. Actually the exposed to risk over the whole day is the integration of the number exposed to risk at each infinitesimal instant of time in the whole day. But what is here attempted is only to give the medical reader an understanding of the *gross* anatomy of a life table. If he wants a knowledge of the *microscopic* anatomy he must get a text which treats of that subject. References to such are given at the end of the chapter.

This is the skeleton of a life table. To this skeleton there are sometimes added certain other functions derived from these three, d_x, l_x, and q_x. For the vital statistician two of these functions only are of particular interest and importance. The first of these is what is called the "expectation of life," but in the interest of accuracy should always be called the "mean after lifetime." It is designated as $\overset{\circ}{e}$ symbolically. It gives the number of years which will, on the average, be subsequently lived by each person who has attained any stated age. The expectation of life *at birth* is the average age at death of all the 100,000 who start life together. But it should always be kept in mind that the average age at death of persons in the general population does not usually give the expectation of life at birth of the same people. This would only be true if the age distribution of the living population were identical with that of the stable life table population L_x. Furthermore, the mean age at death of one population is not comparable with the same constant from another population, unless the two populations have identical age distributions of the living. This fact was first pointed out by Farr many years ago.

The second important derived constant of a life table is L_x, which gives, by age groups, the stationary living population, unaffected by emigration and immigration, which, assuming the mortality rates given by q_x, would result if 100,000 persons were born alive uniformly throughout each year. One important use of this figure will appear in a later chapter.

HUMAN LIFE TABLES

In order that the reader may have a still more concrete realization of what a life table looks like, Table 24 and Figs. 60, 61, and 62 are inserted. The table is that portion of Glover's[1] life table for both sexes in the original registration states in 1910, which carries the constants in which we are here interested.

TABLE 24
LIFE TABLE FOR BOTH SEXES IN THE ORIGINAL REGISTRATION STATES, 1910.
(Glover's Table 2.)

Age interval.	Of 100,000 persons born alive:		Rate of mortality per thousand.	Complete expectation of life.	Stationary population.* Population in current age interval.
Period of lifetime between two exact ages.	Number alive at beginning of age interval.	Number dying in age interval.	Number dying in age interval among 1000 alive at beginning of age interval.	Average length of life remaining to each one alive at beginning of age interval.	Including only those in current month or year of age.
x to $x+1$	l_x	d_x	$1000q_x$	$\overset{\circ}{e}_x$	L_x
1	2	3	4	5	6

INFANT MORTALITY—FIRST YEAR OF LIFE BY AGE INTERVALS OF ONE MONTH

Months.			Monthly rate.	In years.	
0–1	100,000	4377	43.77	51.49	8,060
1–2	95,623	1131	11.83	53.76	7,921
2–3	94,492	943	9.98	54.32	7,835
3–4	93,549	801	8.57	54.78	7,762
4–5	92,748	705	7.60	55.17	7,700
5–6	92,043	635	6.90	55.51	7,644
6–7	91,408	579	6.33	55.81	7,593
7–8	90,829	533	5.87	56.08	7,547
8–9	90,296	492	5.45	56.33	7,504
9–10	89,804	456	5.08	56.56	7,465
10–11	89,348	421	4.72	56.76	7,428
11–12	88,927	389	4.38	56.95	7,394

LIFE TABLE FOR WHOLE RANGE OF LIFE BY AGE INTERVALS OF ONE YEAR

Years.			Annual rate.	In years.	
0–1	100,000	11,462	114.62	51.49	91,853
1–2	88,538	2,446	27.62	57.11	87,095
2–3	86,092	1,062	12.34	57.72	85,529
3–4	85,030	666	7.83	57.44	84,683
4–5	84,364	477	5.65	56.89	84,116
5–6	83,887	390	4.66	56.21	83,692
6–7	83,497	327	3.91	55.47	83,333
7–8	83,170	274	3.30	54.69	83,033
8–9	82,896	234	2.82	53.87	82,779
9–10	82,662	204	2.47	53.02	82,560
10–11	82,458	187	2.27	52.15	82,365
11–12	82,271	180	2.19	51.26	82,181
12–13	82,091	182	2.22	50.37	82,000
13–14	81,909	193	2.36	49.49	81,812
14–15	81,716	210	2.57	48.60	81,611
15–16	81,506	232	2.84	47.73	81,390
16–17	81,274	256	3.16	46.86	81,146
17–18	81,018	285	3.52	46.01	80,875
18–19	80,733	315	3.89	45.17	80,576
19–20	80,418	344	4.28	44.34	80,246
20–21	80,074	375	4.68	43.53	79,887
21–22	79,699	398	5.00	42.73	79,500
22–23	79,301	412	5.19	41.94	79,095
23–24	78,889	418	5.29	41.16	78,680
24–25	78,471	425	5.42	40.38	78,259

*Unaffected by emigration and immigration, which, assuming the mortality rates in column 4, would result if 100,000 persons were born alive uniformly throughout each year.

TABLE 24—*Continued*

Age interval.	Of 100,000 persons born alive:		Rate of mortality per thousand.	Complete expectation of life.	Stationary population.* Population in current age interval.
Period of lifetime between two exact ages.	Number alive at beginning of age interval.	Number dying in age interval.	Number dying in age interval among 1000 alive at beginning of age interval.	Average length of life remaining to each one alive at beginning of age interval.	Including only those in current month or year of age.
x to $x+1$	l_x	d_x	$1000q_x$	$\overset{\circ}{e}_x$	L_x
1	2	3	4	5	6

LIFE TABLE FOR WHOLE RANGE OF LIFE BY AGE INTERVALS OF ONE YEAR

Years.			Annual rate.	In years.	
25–26....	78,046	432	5.54	39.60	77,830
26–27....	77,614	440	5.67	38.81	77,394
27–28....	77,174	451	5.85	38.03	76,949
28–29....	76,723	465	6.06	37.25	76,491
29–30....	76,258	479	6.28	36.48	76,019
30–31....	75,779	493	6.51	35.70	75,532
31–32....	75,286	511	6.78	34.93	75,030
32–33....	74,775	530	7.09	34.17	74,510
33–34....	74,245	550	7.40	33.41	73,970
34–35....	73,695	568	7.72	32.66	73,411
35–36....	73,127	588	8.04	31.90	72,833
36–37....	72,539	605	8.33	31.16	72,237
37–38....	71,934	617	8.59	30.42	71,626
38–39....	71,317	631	8.84	29.68	71,001
39–40....	70,686	644	9.11	28.94	70,364
40–41....	70,042	658	9.39	28.20	69,713
41–42....	69,384	674	9.72	27.46	69,047
42–43....	68,710	693	10.09	26.73	68,364
43–44....	68,017	716	10.52	25.99	67,659
44–45....	67,301	740	10.99	25.26	66,931
45–46....	66,561	766	11.52	24.54	66,178
46–47....	65,795	795	12.08	23.82	65,397
47–48....	65,000	821	12.63	23.10	64,589
48–49....	64,179	846	13.18	22.39	63,756
49–50....	63,333	873	13.77	21.69	62,897
50–51....	62,460	897	14.37	20.98	62,012
51–52....	61,563	929	15.08	20.28	61,098
52–53....	60,634	970	16.01	19.58	60,149
53–54....	59,664	1025	17.17	18.89	59,151
54–55....	58,639	1084	18.49	18.21	58,097
55–56....	57,555	1153	20.03	17.55	56,978
56–57....	56,402	1225	21.72	16.90	55,790
57–58....	55,177	1289	23.37	16.26	54,532
58–59....	53,888	1346	24.97	15.64	53,215
59–60....	52,542	1404	26.73	15.03	51,840
60–61....	51,138	1462	28.58	14.42	50,407
61–62....	49,676	1521	30.62	13.83	48,915
62–63....	48,155	1587	32.96	13.26	47,361
63–64....	46,568	1656	35.55	12.69	45,740
64–65....	44,912	1718	38.25	12.14	44,053
65–66....	43,194	1773	41.06	11.60	42,308
66–67....	41,421	1826	44.08	11.08	40,508
67–68....	39,595	1877	47.41	10.57	38,657
68–69....	37,718	1928	51.12	10.07	36,754
69–70....	35,790	1974	55.14	9.58	34,803

*Unaffected by emigration and immigration, which, assuming the mortality rates in column 4, would result if 100,000 persons were born alive uniformly throughout each year.

TABLE 24—*Concluded*

Age interval.	Of 100,000 persons born alive:		Rate of mortality per thousand.	Complete expectation of life.	Stationary population.*
					Population in current age interval.
Period of lifetime between two exact ages.	Number alive at beginning of age interval.	Number dying in age interval.	Number dying in age interval among 1000 alive at beginning of age interval.	Average length of life remaining to each one alive at beginning of age interval.	Including only those in current month or year of age.
x to $x+1$	l_x	d_x	$1000q_x$	$\overset{\circ}{e}_x$	L_x
1	2	3	4	5	6

LIFE TABLE FOR WHOLE RANGE OF LIFE BY AGE INTERVALS OF ONE YEAR

Years.			Annual rate.	In years.	
70–71....	33,816	2013	59.52	9.11	32,810
71–72....	31,803	2044	64.29	8.66	30,781
72–73....	29,759	2065	69.38	8.22	28,726
73–74....	27,694	2072	74.82	7.79	26,658
74–75....	25,622	2070	80.78	7.38	24,587
75–76....	23,552	2057	87.37	6.99	22,523
76–77....	21,495	2028	94.35	6.61	20,481
77–78....	19,467	1981	101.74	6.25	18,476
78–79....	17,486	1920	109.78	5.90	16,526
79–80....	15,566	1854	119.10	5.56	14,639
80–81....	13,712	1786	130.28	5.25	12,819
81–82....	11,926	1696	142.17	4.96	11,078
82–83....	10,230	1565	153.06	4.70	9,448
83–84....	8,665	1409	162.58	4.45	7,960
84–85....	7,256	1255	172.97	4.22	6,628
85–86....	6,001	1103	183.80	4.00	5,449
86–87....	4,898	954	194.85	3.79	4,421
87–88....	3,944	816	206.84	3.58	3,536
88–89....	3,128	689	220.13	3.39	2,784
89–90....	2,439	571	234.31	3.20	2,154
90–91....	1,868	466	249.62	3.03	1,635
91–92....	1,402	371	264.66	2.87	1.216
92–93....	1,031	289	279.90	2.73	886
93–94....	742	219	295.12	2.59	633
94–95....	523	162	310.17	2.47	442
95–96....	361	117	325.02	2.35	302
96–97....	244	83	339.74	2.24	202
97–98....	161	57	354.55	2.14	132
98–99....	104	39	369.73	2.04	85
99–100...	65	25	385.46	1.95	53
100–101...	40	16	401.91	1.85	32
101–102...	24	10	419.14	1.76	19
102–103...	14	6	437.37	1.67	11
103–104...	8	4	456.77	1.59	6
104–105...	4	2	477.48	1.50	3
105–106...	2	1	500.22	1.41	2
106–107...	1	1	524.82	1.33	1

* Unaffected by emigration and immigration, which, assuming the mortality rates in column 4, would result if 100,000 persons were born alive uniformly throughout each year.

The following diagrams illustrate the important functions of a life table. The first (Fig. 60) shows the form of the life table
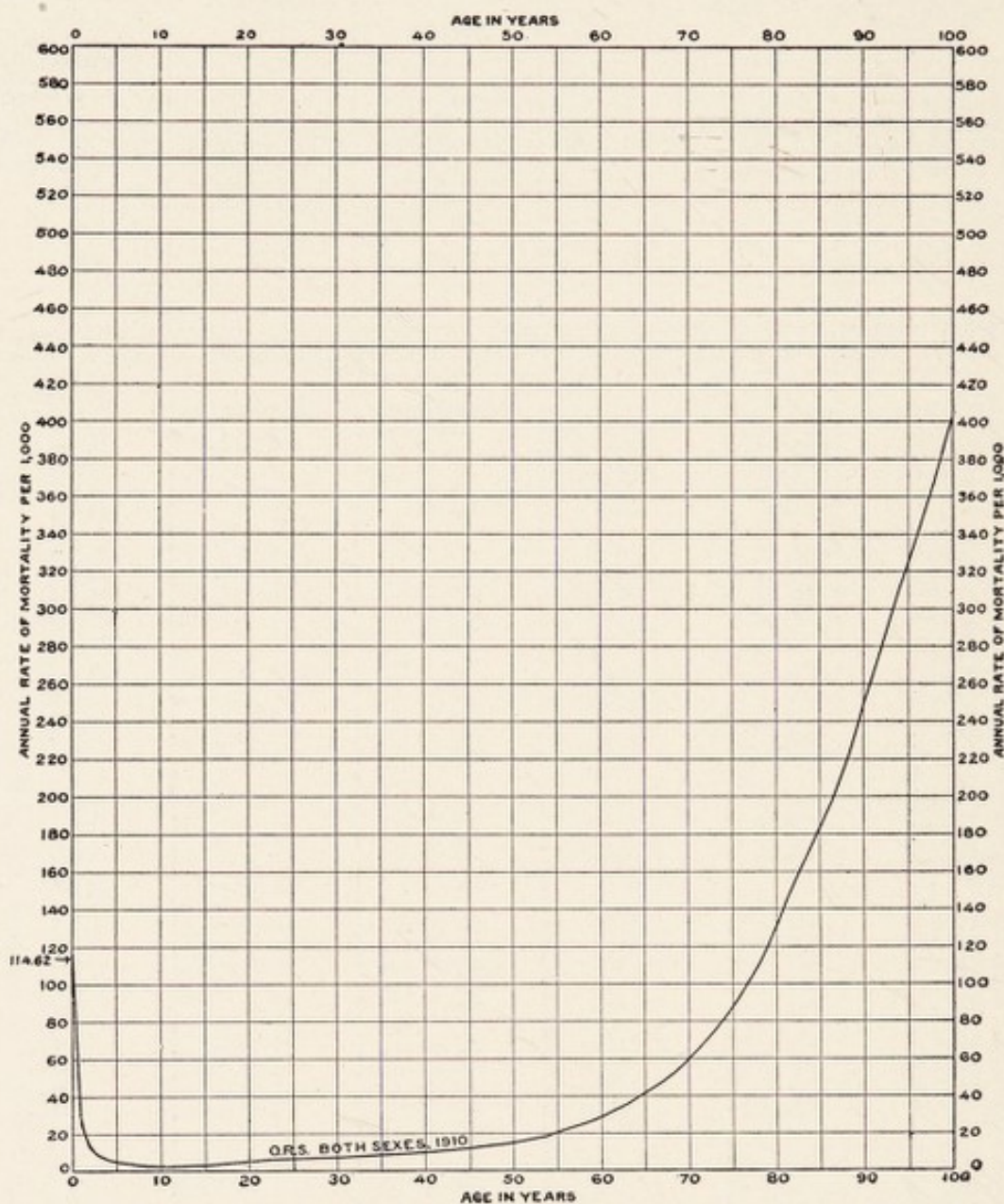

Fig. 60.—Annual mortality rate per thousand. The original registration states, both sexes, 1910 (from Glover,[1] p. 243).

specific death-rate curve (q_x), being the plot of this column of Table 24 above.

The next diagram (Fig. 61) shows the form of the l_x curve. Here the data for a number of different countries are included.

The picture shows in a striking way the usefulness of the life table method in the comparative study of mortality.
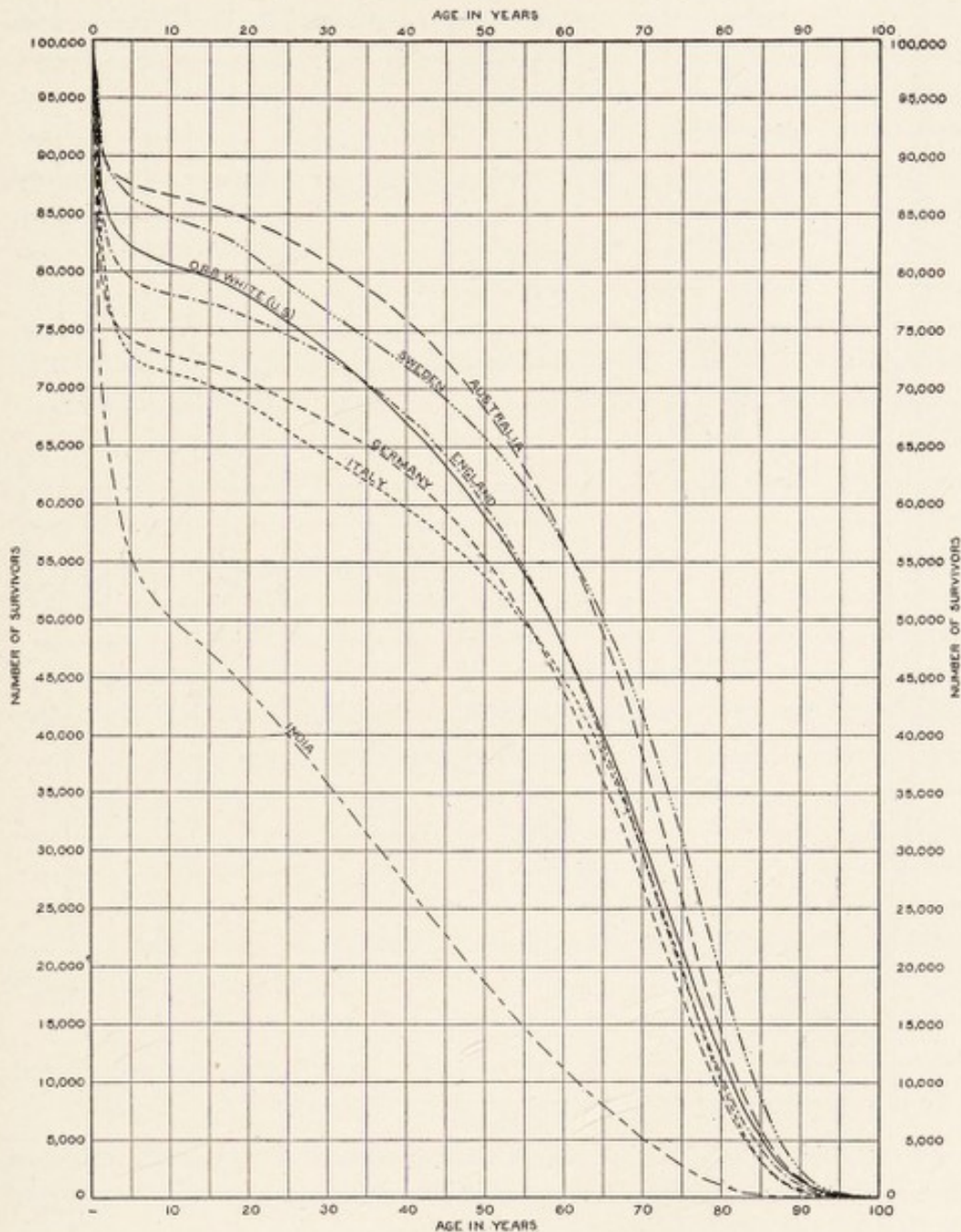

Fig. 61.—Number of survivors out of 100,000 born alive. Australia, England, Germany, India, Italy, Sweden, and whites in the original registration states. Males, 1901–10 (from Glover,[1] p. 260).

The next diagram (Fig. 62) shows the form of the d_x curve, and again the life tables of several countries are drawn upon for comparison.

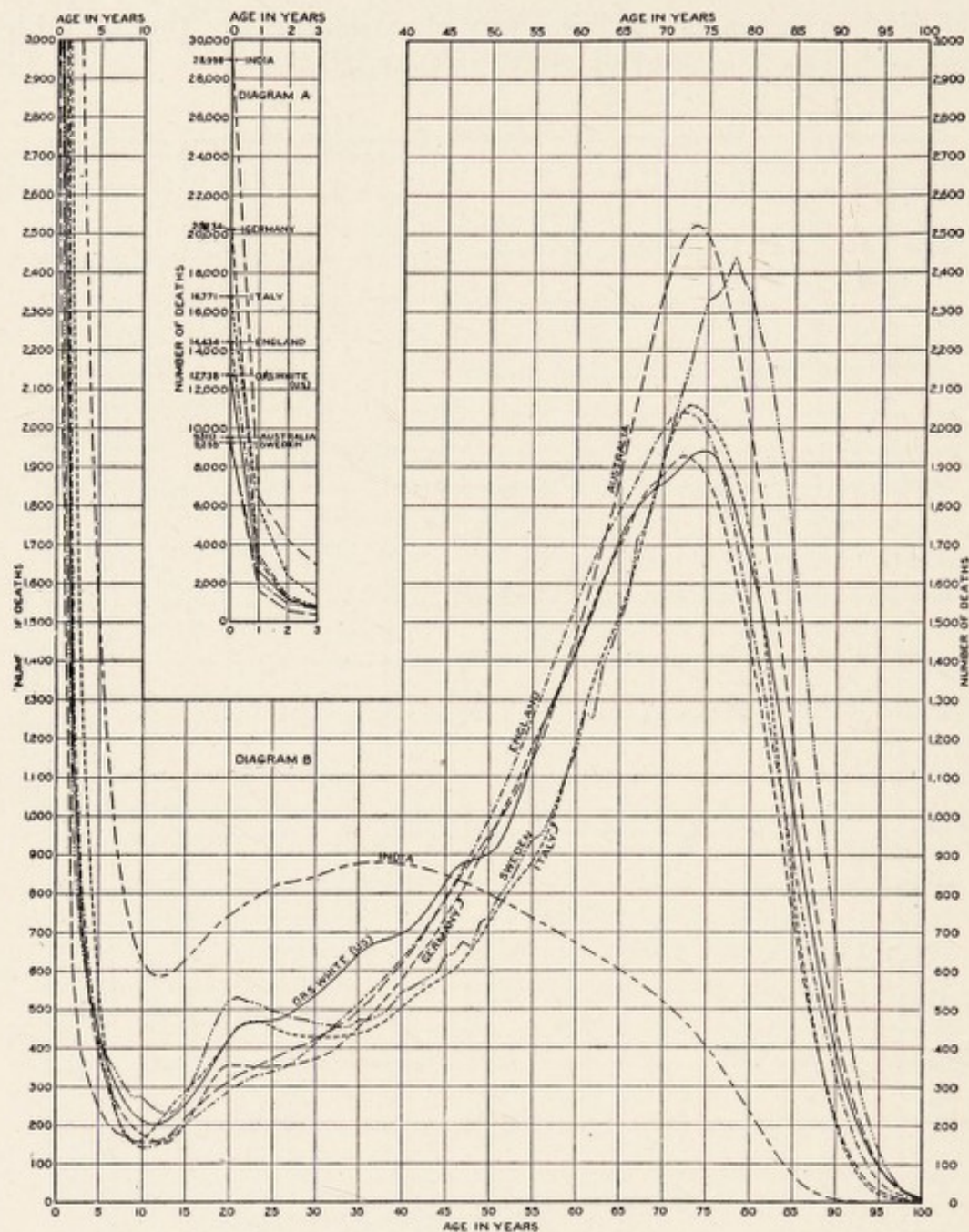Fig. 62.—Number of deaths out of 100,000 born alive.   Australia, England, Germany, India, Italy, Sweden, and whites in the original registration states. Males 1901–10 (from Glover,[1] p. 270).

A LIFE TABLE NOMOGRAM

As a further help toward understanding the structure and meaning of life tables a nomogram devised by Pearl and Reed[4] may be presented here.

If d_x be used, as in what has preceded, to indicate the deaths

at age x, or, more correctly, the number of persons dying in the interval from x to $x + 1$, where 1 denotes one unit of age, which theoretically can be made as small as one likes, then

$$l_x = d_x + d_{x+1} + d_{x+2} + \cdots + d_\omega = \overset{\omega}{\underset{x}{\Sigma}} d_x,$$

denotes the number of survivors at age x, and we may define a quantity

$$L_x = \frac{l_x + l_{x+1}}{2},$$

which will be the number living between any two ages x and $x + 1$.

The instantaneous death rate, or the probability of dying in the age interval x to $x + 1$ is

$$q_x = \frac{d_x}{l_x}.$$

Let us define another quantity as

$$T_x = L_x + L_{x+1} + L_{x+2} + \cdots + L_\omega = \overset{\omega}{\underset{x}{\Sigma}} L_x,$$

which gives the population in current and all older age intervals.

Then the expectation of life, or mean after-life time at age x will be

$$e_x = \frac{T_x}{l_x}.$$

The reciprocal of this last quantity, or l_x/T_x will give the death-rate at age x and over.

Finally the number living at age x per death at that age will be the reciprocal of q_x or l_x/d_x.

Consider now Fig. 63. This is a diagram* in two parts, plotted on an arithlog grid, namely, one in which the abscissal divisions are arithmetically equal, and represent ages, and the ordinate divisions are proportional to the logarithms of the numbers set down by their side. On the left hand, or larger one of the two diagrams, in which the logarithmic grid extends to 5 decks, there are plotted 3

* On account of the size of the page the original diagram is much reduced here. This, however, is unimportant because the only purpose of Fig. 63 is to show how the nomogram is constructed. The student should draw this nomogram for himself on a large sheet of 5-deck arithlog paper.

lines, namely d_x, l_x, and T_x as defined above, for Glover's[1] life table for males in the Original Registration States in 1910. Let us for convenience call this larger diagram the *nomogram base*. The d_x, l_x and T_x data were as a matter of fact in this case taken directly from Glover's table. But the data for plotting T_x could equally well have been got by accumulating by successive additions the l_x values of Glover's table, beginning with l_ω and adding back-
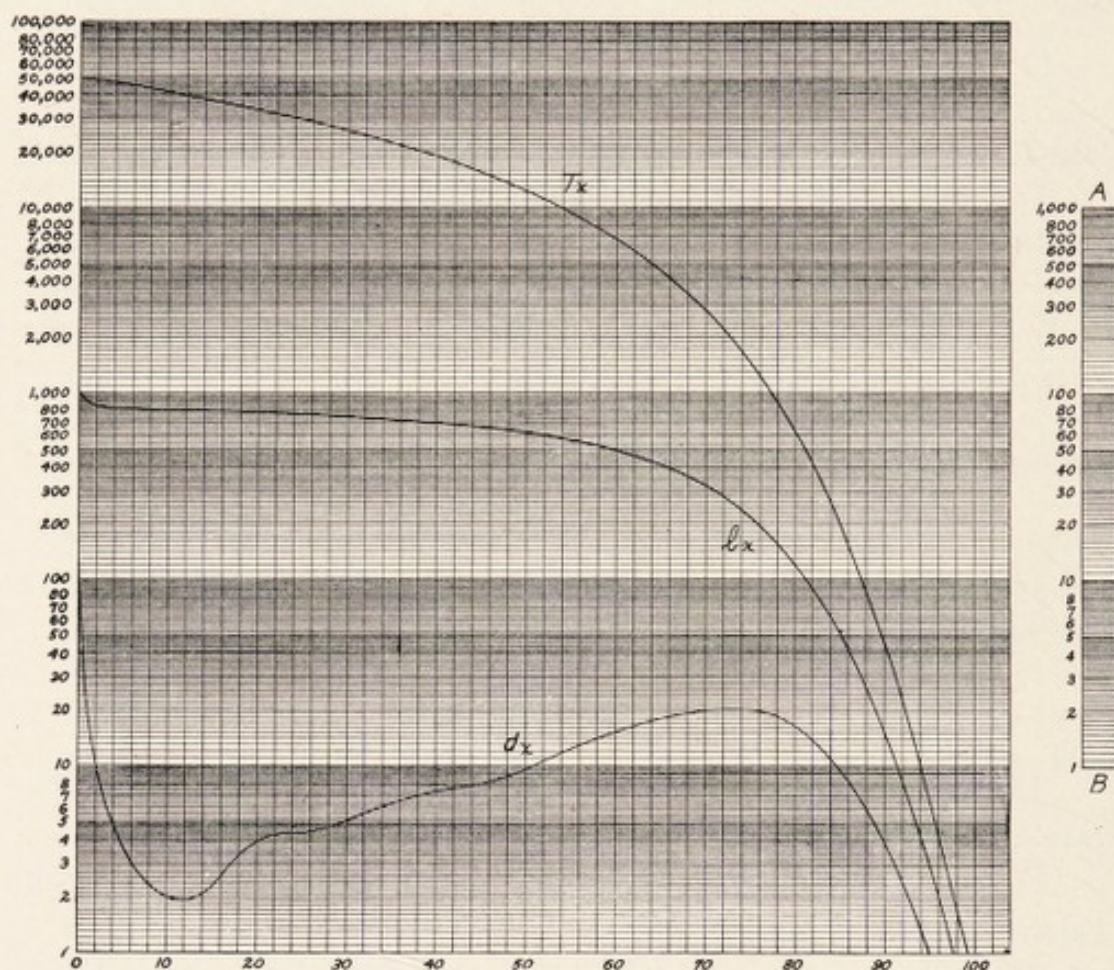
Fig. 63.—A life table nomogram.   For explanation see text.

ward, *i. e.*, toward the beginning of life.   It is important to note this point because many life tables do not table T_x.   The right hand, or smaller one of the diagrams, which we may for convenience call the *nomogram scale*, is simply the same logarithmic scale as that of the larger of the two diagrams, but extending only from 1 to 1000.

Now consider the nomogram *scale* of Fig. 63 to be cut free from the rest of the sheet and therefore freely movable.

We then have the following rules:

Rule I. To find the instantaneous death-rate q_x. Place the *nomogram scale* on the nomogram *base* in such a manner that A is on the l_x line at the age x. Then read the nomogram *scale* at the point where, at age x, it is cut by the d_x line of the nomogram *base*. The value so read will be q_x, the death-rate per 1000 living.

Rule II. To find the death-rate at age x and over. Place the nomogram *scale* on the nomogram *base* in such a manner that A coincides with the T_x line at age x. Then read the nomogram *scale* at the point where, at age x, it is cut by the l_x line of the nomogram *base*. The value so read will be the death rate per 1000 living at age x and over.

Rule III. To find the expectation of life e_x. Place the nomogram *scale* on the nomogram *base* in such a manner that B coincides with the l_x line, at age x. Then read the nomogram *scale* at the point where, at age x, it is cut by the T_x line of the nomogram *base*. The value so read will be the expectation of life at age x, in years.

Rule IV. To find the number of living persons of age x per death at that age. Place the nomogram *scale* on the nomogram *base* in such a manner that B coincides with the d_x line, at age x. Then read the nomogram *scale* at the point where, at age x, it is cut by the l_x line of the nomogram base.

It will have been noted that the *nomogram scale* is identical with the scale of the *nomogram base* throughout the range of the former, namely, from 1 to 1000. This fact indicates at once that the use of the nomogram *scale* may be replaced by an ordinary pair of dividers. The rules then take the following form.

Label one point or leg of the dividers A and the other B.

Rule I bis. Place leg A on the l_x line at age x. Bring leg B to coincide with the d_x line at age x. Lift the dividers and place leg A at 1000 on the nomogram *base* scale. Then read q_x at the point below where leg B touches the same scale.

Rule II bis. Place leg A on the T_x line at age x and bring leg B to coincide with the l_x line at age x. Lift the dividers and place leg A at 1000 on the nomogram *base* scale. Then read the death rate at age x and over on the point below 1000 where leg B touches the same scale.

Rule III bis. Proceed as in Rule II *bis*, but after lifting the dividers place leg *A* at 1 on the nomogram *base* scale and read expectation of life at the point above 1 where leg *B* touches the same scale.

Rule IV bis. Proceed as in Rule I *bis*, but after lifting the dividers place leg *A* at 1 on the nomogram *base* scale and read number of living persons of age x per death at that age at the point above 1 where leg *B* touches the same scale.

The proof of the above rules is evident from the equations. All that this nomographic treatment essentially does is to take advantage of the property of logarithms which enables division to be accomplished by a process of subtraction. The subtracting of the logarithms is done geometrically. Simple as the idea involved is its very useful application to the functions of a life table has apparently not hitherto been systematically made. J. A. Field (*loc. cit.*, Chap. VI) pointed out that when l_x is plotted on arithlog paper the slope of the tangent at any point of the curve is q_x. This principle has been made use of in certain cases in the graphic presentation of life curves. But a trial convinces one at once that only a very rough approximation to q_x may be obtained in this way.

There are, of course, a number of useful corollaries of the four rules given above, which will occur to the student. We shall mention only one here, by way of numerical illustration of the kind of service which this life table nomogram may render. As the world approaches more and more closely to a condition of population saturation, the populations of the various demographic units will obviously come nearer and nearer to the conditions of life table stability. This is a condition in which births bear a fixed and constant relation of equality to deaths and there is no alteration of the situation by migration. That we are now somewhat definitely approaching such a condition in most highly industrialized and civilized countries is evidenced by the concomitant decline of birth- and death-rates, and the closer and closer approach of both these rates to each other. The location of the exact levels at which these rates will stabilize presents an interesting problem. It has been suggested that a stable death-rate of the order of 7 or even 5

per thousand is quite within the range of possibility; is in fact almost certain to be attained in comparatively few years.

Now, by Rule II, if (*a*) 1000 on the nomogram *scale* is placed, at age 0 (birth), on the nomogram *base* so to coincide with the T_x line at that age, and (*b*) the point on the age 0 ordinate of the nomogram *base* corresponding to the nomogram scale graduation 7 is marked, and finally (*c*) the nomogram *scale* is then moved so that B coincides with the point just marked on the nomogram *base*, it is then found that the line T_x cuts the nomogram *scale* at approximately 143. This means that in order to have a stabilized death-rate of 7 per thousand in a population unaffected by migration, the mean or average duration of life (expectation of life at birth) would have to be approximately 143 years! Under the same conditions a death-rate stabilized at 5 would mean an expectation of life at birth of exactly 200 years! A death-rate stabilized at 10, under the above restriction, means an expectation of life at birth (mean after life time) of exactly 100 years! Of course such death-rates as these of which we have been speaking are only attained under one or the other or a combination of three conditions: (*a*) a constantly increasing rate of growth of the population by an ever-increasing birth rate, or (*b*) by immigration into the population of persons of those ages where the age specific death-rates are low and a concomitant or subsequent migration out of the population of persons at advanced ages, where specific death rates are high, to die elsewhere, or (*c*) a combination of an immigration of persons of favorable ages and an increasing birth-rate sufficient always to offset the necessity of old persons emigrating to die elsewhere. But no one of these conditions is compatible, by definition, with a stable population or a stable death-rate in the sense of a life table.

The usefulness of this nomogram in experimental work on duration of life, as for example the investigations on *Drosophila* discussed in the next section, is great. In such experimental work d_x (in the true actuarial sense) is directly observed, and can be put upon a per thousand base and directly plotted. From this l_x, and in turn T_x, can be plotted. Then by the aid of this nomographic method all the important functions of the life table can be read directly, without the necessity of any computation whatsoever.

The case can never be quite so simple for human life tables, because there we cannot observe the true life table d_x line directly. It must be computed from the statistical data.

LIFE TABLES FOR LOWER ORGANISMS

Life tables can and should be computed for other forms of life besides man. Their importance for the study of organic evolution can scarcely be overestimated. Owing to the general lack in biologic literature, however, of the basic observational data necessary for the construction of a life table, only the merest beginning has been made in this direction.

An example of a complete life table for another organism, the fruit-fly, *Drosophila melanogaster*, is given in Tables 25 and 26, and Fig. 64. These life tables were worked out in the author's laboratory.[5, 6] Only two *Drosophila* life tables are given here. Similar tables for females will be found in the original publication. The l_x curves in the diagram show the similarity of the findings to those in man, remembering that the fly curves are plotted on an arithlog grid and that they have no infant mortality component.

The "vestigial" (Table 26) is a mutant form of *Drosophila*, characterized by minute, functionless wings, and a shorter life span than the normal form (Table 25).

An interesting problem now presents itself. How shall one compare the mortality of two organisms whose total life spans are so widely different in extent of time that it is in practice quite impossible to measure or express them in the same unit?

Various methods have been used for making this comparison. The one originally used by the author[5] has been criticised by Greenwood,[8] whose valuable discussion of the whole problem should be read. The method which appears to be the most valid and least open to statistical objections is one which is a particular application of a general method devised[7] for the purpose of comparing the relative variability of different organisms.* In the present case the mode of application of this principle is to regard the observed mean age at death (duration of life) of the individuals

* See the section on Graphic Representation of Relative Variability in Chapter XIII of this book.

TABLE 25

Life Table for Drosophila—Wild Type. Line 107—Males

Age in days.	l_x	q_x	e_x	Age in days.	l_x	q_x	e_x
1	1000	0.2	45.8	46	551	43.5	12.3
2	1000	0.6	44.8	47	527	46.6	11.8
3	999	1.0	43.8	48	502	49.8	11.3
4	998	1.3	42.9	49	477	53.2	10.9
5	997	1.7	41.9	50	452	56.9	10.4
6	995	2.0	41.0	51	426	60.8	10.0
7	993	2.4	40.1	52	400	65.0	9.6
8	991	2.8	39.2	53	374	69.5	9.2
9	988	3.2	38.3	54	348	74.2	8.8
10	985	3.5	37.4	55	322	79.2	8.4
11	981	3.9	36.5	56	297	84.5	8.0
12	978	4.3	35.7	57	272	90.2	7.7
13	973	4.7	34.8	58	247	96.2	7.4
14	969	5.1	34.0	59	223	102.5	7.0
15	964	5.5	33.2	60	200	109.2	6.7
16	958	6.0	32.3	61	179	116.3	6.4
17	953	6.4	31.5	62	158	123.8	6.1
18	947	6.9	30.7	63	138	131.7	5.9
19	940	7.4	29.9	64	120	139.9	5.6
20	933	7.9	29.2	65	103	148.8	5.3
21	926	8.5	28.4	66	88	157.9	5.1
22	918	9.0	27.6	67	74	167.6	4.9
23	910	9.6	26.9	68	62	177.7	4.7
24	901	10.3	26.1	69	51	188.3	4.4
25	892	10.9	25.4	70	41	199.4	4.2
26	882	11.7	24.6	71	33	211.0	4.1
27	872	12.4	23.9	72	26	223.1	3.9
28	861	13.3	23.2	73	20	235.8	3.7
29	849	14.1	22.5	74	15	248.9	3.5
30	837	15.1	21.8	75	12	262.6	3.4
31	825	16.1	21.1	76	9	276.8	3.2
32	811	17.2	20.5	77	6	291.5	3.1
33	798	18.3	19.8	78	4	306.7	2.9
34	783	19.5	19.1	79	3	322.5	2.8
35	768	20.8	18.5	80	2	338.7	2.6
36	752	22.3	17.9	81	1	355.5	2.4
37	735	23.8	17.3	82	1	372.7	2.2
38	717	25.4	16.7				
39	699	27.2	16.1				
40	680	29.1	15.5				
41	660	31.1	14.9				
42	640	33.3	14.4				
43	619	35.5	13.8				
44	597	38.0	13.3				
45	574	40.7	12.8				

TABLE 26

LIFE TABLE FOR DROSOPHILA—VESTIGIAL—MALES

Age in days.	l_x	q_x	e_x	Age in days.	l_x	q_x	e_x
1	1000	0.0	14.1	26	72	162.7	5.8
2	1000	9.0	13.1	27	61	162.5	5.7
3	991	18.1	12.2	28	51	162.0	5.6
4	973	27.4	11.7	29	43	161.1	5.5
5	946	36.7	10.7	30	36	160.8	5.4
6	912	45.8	10.1	31	30	160.7	5.3
7	870	55.4	9.6	32	25	161.5	5.1
8	821	64.7	9.1	33	21	163.6	4.9
9	768	73.8	8.6	34	18	167.7	4.6
10	712	82.8	8.2	35	15	174.5	4.4
11	653	91.5	7.9	36	12	184.8	4.1
12	593	100.0	7.6	37	10	198.5	3.8
13	534	108.1	7.3	38	8	219.6	3.4
14	476	115.9	7.0	39	6	246.0	3.1
15	421	123.1	6.8	40	5	279.6	2.8
16	369	129.9	6.7	41	3	320.9	2.5
17	321	136.2	6.5	42	2	370.4	2.3
18	277	141.8	6.4	43	1	427.7	2.0
19	238	146.8	6.3	44	1	491.9	1.7
20	203	151.2	6.2				
21	172	154.8	6.1				
22	146	157.8	6.0				
23	123	160.0	6.0				
24	103	161.5	5.9				
25	86	162.4	5.9				

in each different species to be compared as representing a *biologically* equivalent point in each of their several life cycles, and then to represent every other absolute age as a percentage deviation from the mean duration of life of each particular species. Thus, for example, if the mean absolute duration of life for a particular species is fifty days (= 100 per cent. on this plan), then an individual organism of that species dying at age seventy-five days will be recorded as having a relative duration of life of 150 per cent., because it lived half again as long as the average of the individuals of the species to which it belongs. Furthermore, as is usual in life table work, frequencies are put upon a relative rather than an absolute basis.

In Table 27 and Fig. 65 are shown the results of certain comparisons made in this way between a few widely separated organisms, in respect of the order in time of their dying.

The mean absolute durations of life for the forms shown in Table 27 are given in Table 28.

From Table 27 and Fig. 65 certain results are at once apparent:

1. The life curves fall clearly into three groups. The first of these, *Group A*, approximates to the rectangular type of survivor-
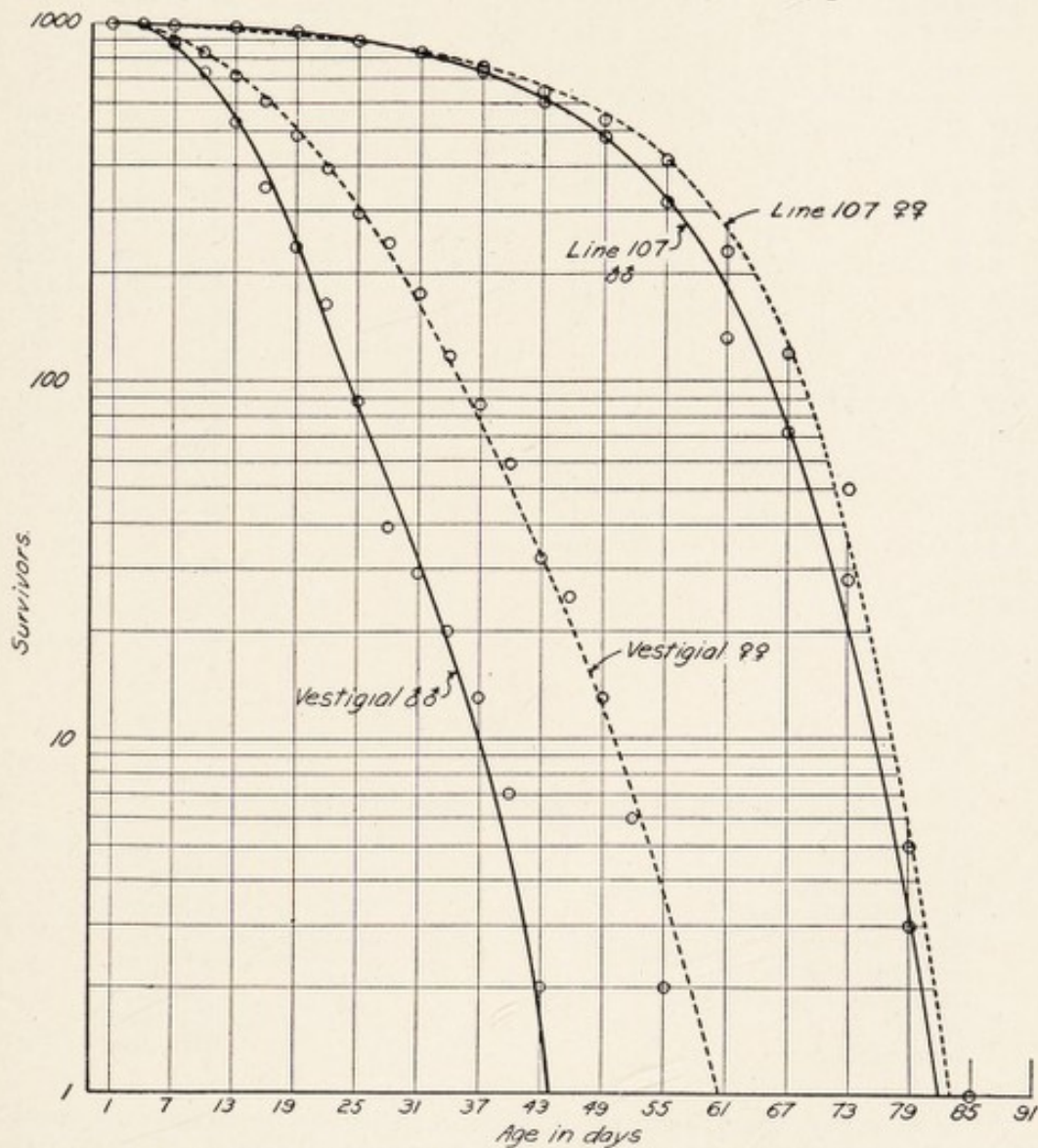

Fig. 64.—Diagram showing the observed and graduated l_x points for (*a*) line 107 wild type, and (*b*) vestigial flies. The small circles are the observations, and the smooth lines the fitted curves from the equations.

ship curve, in which (in the limiting case) all the individuals live to a certain age and then die together at the same instant. This limiting type is approached, though of course not *precisely* realized, in the present material by (*a*) the rotifer *Proales* and (*b*) the fly

TABLE 27

SURVIVORSHIP DISTRIBUTION (l_x) FOR AGES EXPRESSED AS PERCENTAGE DEVIATIONS FROM MEAN DURATION OF LIFE

Percentage deviation from mean duration of life.	Drosophila wild (107) ♂♂	Drosophila vestigial. ♂♂	Drosophila starved. ♂♂	Proales decipiens.	Hydra fusca.	Blatta orientalis.	Agriolimax.	Mice (Hill's data).	Automobiles.
−100	1000	1000	1000	1000	1000	1000	1000	1000	1000
− 80	988	993	991	1000	875	994	787	988	979
− 60	945	924	981	996	747	977	664	961	912
− 40	867	795	967	967	629	908	583	901	801
− 20	741	635	914	849	526	752	519	775	654
0	556	469	537	543	440	513	454	548	488
+ 20	322	323	71	147	367	258	398	247	325
+ 40	118	211	5	1	302	83	330	40	191
+ 60	19	132	....	....	242	15	259	1	94
+ 80	1	80	....	....	182	1	190	....	38
+100	....	49	....	....	119	....	129	....	12
+120	....	30	....	....	59	....	81	....	3
+140	....	18	....	....	18	....	46		
+160	....	11	....	....	3	....	24		
+180	....	5	....	....	....	....	11		
+200	....	2	....	....	....	....	5		
+220	....	....	....	....	....	....	2		
+240	....	....	....	....	....	....	1		

TABLE 28

MEAN DURATION OF LIFE IN ABSOLUTE TIME UNITS

Hydra fusca[a]	54.89 days
Proales decipiens[b]	5.95 days
Agriolimax agrestis[c]	4.12 months
Blatta orientalis[d]	40.89 days
Drosophila, Wild (107) ♂♂	45.81 days
Drosophila, Starved ♂♂	44.09 hours
Mouse[e]	636.50 days
Automobiles[f]	7.04 years

[a] Life table calculated from data in Hase, A., Über die deutschen Süss-wasser-Polypen *Hydra fusca*, etc. Arch. f. Rassen-u. Gesellschaft-Biologie, Bd. 6, pp. 721–753, 1909.

[b] Pearl, R., and Doering, C. R.: A Comparison of the Mortality of Certain Lower Organisms with that of Man, Science, N. S., vol. 57, pp. 209–212, 1923.

[c] Life table calculated from data in Szabó, I., and Szabó, M.: Lebensdauer, Wachstum und Altern, studiert bei der Nacktschneckenart *Agriolimax agrestis* L. Biologia Generalis, Bd. 5, pp. 95–118, 1929.

[d] Life table calculated from data in Rau, P.: The Biology of the Roach, *Blatta orientalis* Linn. Trans. Acad. Sci., St. Louis, vol. 25, pp. 57–79, 1924.

[e] Life table calculated from Hill data given by Greenwood.[g]

[f] Data from Griffin, C. E.: The Life History of Automobiles, Michigan Business Studies, vol. i (University of Michigan), 1928, p. v + 42.

Drosophila under conditions of complete starvation. In this group
the characteristic feature of the mortality is that there is no death-
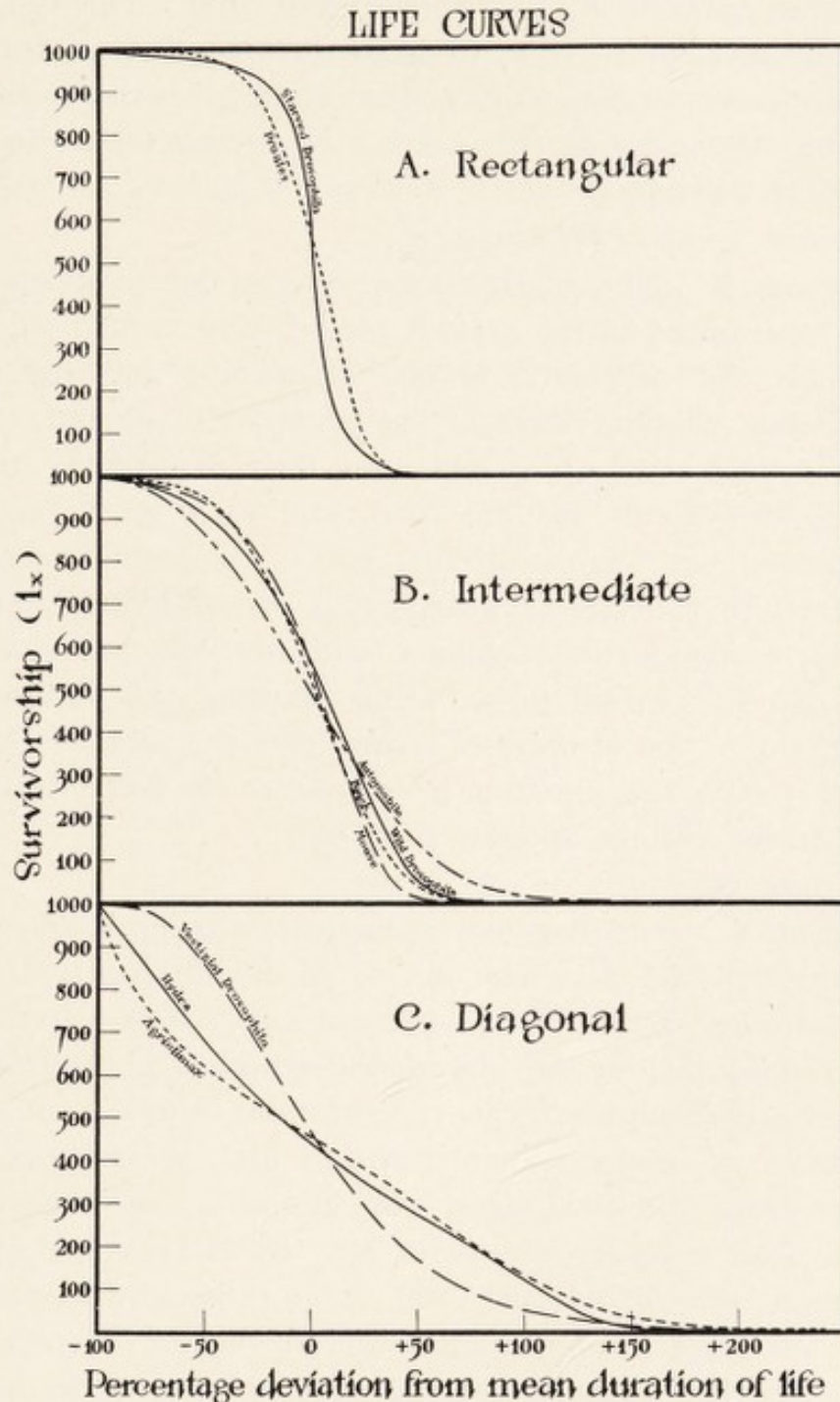rate at all (or a negligible one) until the upper end of the life span

Fig. 65.—Survivorship lines for various species of animals and the automobile,
on a relative time base. For each form represented the mean duration of life is taken
as 100 per cent. on the abscissal side, and all other ages (time durations) are expressed
as percentage deviations (plus or minus) from this mean. Further explanation in text.

17

is nearly reached. Then there is an explosive outbreak of mortality which kills all the individuals within a short time interval. In these cases the upper end of the life span stands, in terms of relative age, to the mean duration of life as roughly 140 : 100. This type of life curve may most properly be designated as the limit of negatively skew life curves, because the d_x curve which gives rise to this type of rectangular l_x line is characterized by negative skewness (cf. Chap. XIII, Skewness).

The next group of life curves is

2. *Group B.* The intermediate type of survivorship curve. This is represented in the present material by normal wild (107) *Drosophila*, the cockroach (*Blatta orientalis*), and the mouse. The common characteristics of the order of dying out of these forms are, first, that the death-rates tend to increase smoothly with age throughout the life span, and at a more rapid rate in the second half than in its first half. This would seem to be the characteristic mode of wearing out of non-living, man-made machines, exemplified by Griffin's automobile life curve here depicted. In the second place, forms showing this intermediate, "wearing-out" type of order of dying have the total life span, in terms of relative age, standing in relation to the mean duration of life, as, on the average, roughly 185 : 100.

We come next to

3. *Group C.* The diagonal type of survivorship curves, with a constant death-rate until near the end of the life span (in the theoretical limiting case). In the present material the fresh water polyp *Hydra fusca*, the slug *Agriolimax agrestis*, and the *Drosophila* mutant vestigial approach this type of order of dying out. These forms have as common characteristics, first, an approach to a constant death-rate at all ages, and, second, a very wide ratio of total life span to mean duration of life. On the average, for these three forms, this ratio is 300 : 100, or, in other words, some individuals live three times as long as the average of the population. If man had this characteristic the upper limit of his life span would be around 175 to 180 years instead of around 100 years as it is.

Space cannot be given here for further discussion of these

matters. The student who is interested in them will find further
details in references 5 and 6 at the end of this chapter.

STATIONARY POPULATIONS

The stationary population of a life table serves a useful purpose
as a standard in the computation of certain derived rates to be
discussed in the next chapter. For this purpose it is desirable to
have this function on the basis of a total population of 1,000,000

TABLE 29

STATIONARY LIFE TABLE POPULATION OF 1,000,000 PERSONS. NUMBER LIVING IN
EACH YEARLY INTERVAL OF AGE

Age interval.	Persons per million in current age interval.	Age interval.	Persons per million in current age interval.	Age interval.	Persons per million in current age interval.
0– 1	17,841	35–36	14,146	70– 71	6373
1– 2	16,916	36–37	14,031	71– 72	5979
2– 3	16,612	37–38	13,912	72– 73	5579
3– 4	16,448	38–39	13,791	73– 74	5178
4– 5	16,338	39–40	13,667	74– 75	4776
5– 6	16,255	40–41	13,540	75– 76	4375
6– 7	16,186	41–42	13,411	76– 77	3978
7– 8	16,127	42–43	13,278	77– 78	3589
8– 9	16,078	43–44	13,141	78– 79	3210
9–10	16,036	44–45	13,000	79– 80	2843
10–11	15,998	45–46	12,854	80– 81	2490
11–12	15,962	46–47	12,702	81– 82	2152
12–13	15,927	47–48	12,545	82– 83	1835
13–14	15,890	48–49	12,383	83– 84	1546
14–15	15,851	49–50	12,216	84– 85	1287
15–16	15,808	50–51	12,045	85– 86	1058
16–17	15,761	51–52	11,867	86– 87	859
17–18	15,708	52–53	11,683	87– 88	687
18–19	15,650	53–54	11,489	88– 89	541
19–20	15,586	54–55	11,284	89– 90	418
20–21	15,516	55–56	11,067	90– 91	318
21–22	15,441	56–57	10,836	91– 92	236
22–23	15,363	57–58	10,592	92– 93	172
23–24	15,282	58–59	10,336	93– 94	123
24–25	15,200	59–60	10,069	94– 95	86
25–26	15,117	60–61	9,791	95– 96	59
26–27	15,032	61–62	9,501	96– 97	39
27–28	14,946	62–63	9,199	97– 98	26
28–29	14,857	63–64	8,884	98– 99	17
29–30	14,765	64–65	8,556	99–100	10
30–31	14,671	65–66	8,217	100–101	6
31–32	14,573	66–67	7,868	101–102	4
32–33	14,472	67–68	7,508	102–103	2
3–34	14,367	68–69	7,139	103–104	1
34–35	14,259	69–70	6,760	104–105	1

persons living. The necessary computations have been done for three age class ranges and the results are presented in Tables 29, 30, and 31, on the basis of the L_x data of Table 24 above. This

TABLE 30

STATIONARY LIFE TABLE POPULATION OF 1,000,000 PERSONS. NUMBER LIVING IN EACH FIVE-YEARLY INTERVAL OF AGE

Age interval.	Persons per million in current age interval.
0– 4	84,155
5– 9	80,682
10– 14	79,628
15– 19	78,513
20– 24	76,802
25– 29	74,717
30– 34	72,342
35– 39	69,547
40– 44	66,370
45– 49	62,700
50– 54	58,368
55– 59	52,900
60– 64	45,931
65– 69	37,492
70– 74	27,885
75– 79	17,995
80– 84	9,310
85– 89	3,563
90– 94	935
95– 99	151
100–104	14

TABLE 31

STATIONARY LIFE TABLE POPULATION OF 1,000,000 PERSONS. NUMBER LIVING IN EACH TEN-YEARLY INTERVAL OF AGE

Age interval.	Persons per million in current age interval.
0– 9	164,837
10–19	158,141
20–29	151,519
30–39	141,889
40–49	129,070
50–59	111,268
60–69	83,423
70–79	45,880
80–89	12,873
90–99	1,086
100 and over	14

then is the population derived from the life table for the original registration states in 1910, both sexes together.

It is important that the student should have a clear mental

picture of the age distribution of a stationary life table population, and of the manner in which it differs from the actually existing general population upon which the life table is computed. Accordingly there is inserted here Table 32 (p. 262). Table 32 exactly
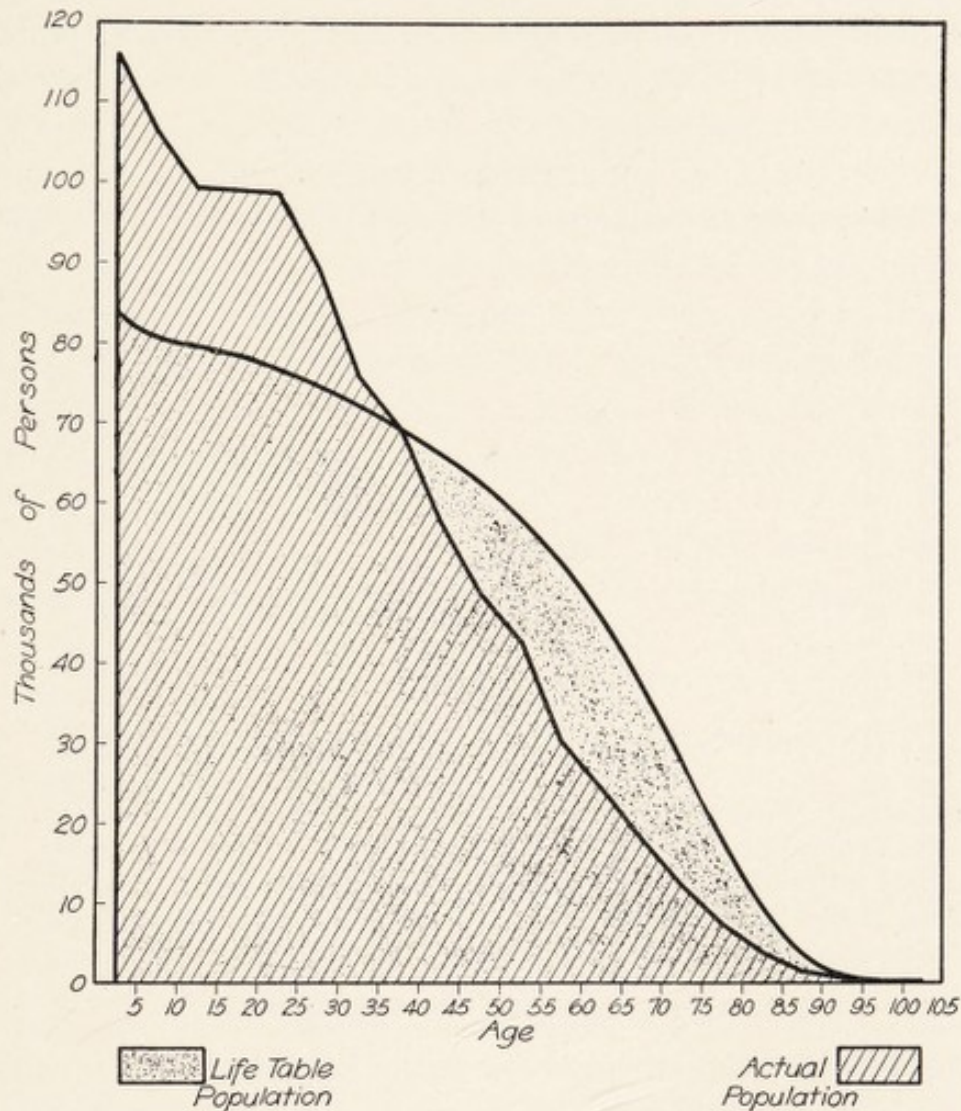

Fig. 66.—Diagram comparing the standard million of (a) the life table stationary population (stippled area), and (b) the actual population (cross-hatched area), both for the year 1910, and for both sexes together. (Data of Tables 30 and 32.)

corresponds to Table 30 in arrangement, but gives the age distribution per million of the population of the United States of both sexes actually living in 1910 by quinquennial age groups.

Figure 66 compares the life table standard million (from Table 30) with the standard million of the actual population.

From this diagram it is apparent that the essential difference between actual and life table populations in this country consists in the former having an excess of persons in early life (up to age thirty-eight years roughly) and a defect of persons of all ages beyond that. This difference arises mainly from two causes: excess of births over deaths and of immigration over emigration in the actual population.

TABLE 32

STANDARD MILLION OF ACTUAL LIVING POPULATION (BOTH SEXES) IN THE UNITED STATES, 1910

Age interval.	Persons per million.
0– 4	115,806
5– 9	106,321
10– 14	99,203
15– 19	98,728
20– 24	98,656
25– 29	89,104
30– 34	75,947
35– 39	69,672
40– 44	57,314
45– 49	48,682
50– 54	42,491
55– 59	30,358
60– 64	24,696
65– 69	18,294
70– 74	12,132
75– 79	7,269
80– 84	3,505
85– 89	1,338
90– 94	365
95– 99	80
100–104	39

THE CONSTRUCTION OF LIFE TABLES

The statement has already been made that it is not the intention to go here into the methods actually employed in the construction of a life table. It, however, seems only fair to outline the procedure in general terms. The starting-point is the determination, from recorded statistics of living *population* at ages, and *deaths* at ages (and for the early part of life *births*, because of the inadequacy at those ages of census counts of population, and because of the rapidity of the flow of vital events in the first year of life) of the *specific death-rates* at ages. From these specific death-rates (in the sense of the vital statistician), which are symbolically designated

as m_x values, the q_x's of the life table are derived. The q_x values are then subjected to a more or less elaborate process of *graduation* or *smoothing*, the purpose of which is to eliminate such portion of the minor fluctuations in their values as may reasonably be supposed due to chance. This smoothing process is where the heavy mathematics of actuarial work comes in. Around this phase of the subject a highly esoteric cult has grown up. In its fundamental and essential principles the smoothing process is simple enough to be grasped by any intelligent person, but, like many other things, when finally dressed out in all its symbolic panoply it is forbidding.

After the q_x values have been graduated the rest of the work of constructing a life table is simple, even if tiresome in its extent. The q_x's are successively applied to an l_x group starting with 100,000 at age zero (birth) to determine the d_x's. When this is done one has l_x, d_x, and q_x for each age interval. From the l_x's and d_x's the $\overset{\circ}{e}_x$'s are easily calculated.

Short methods for the construction of life tables in public health work have been discussed by Hayward[3] and Snow.[9]

SUGGESTED READING

1. Glover, J. W.: United States Life Tables, 1890, 1901, 1910, and 1901–1910, Washington (Bureau of the Census), 1921.

    (This book is, at the present time, perhaps the most complete treatise in existence on the construction of life tables. It gives the methods in detail, as well as a large number of life tables. It should form a part of the library of every medical man, health officer, and vital statistician. It may be obtained from the Superintendent of Documents, Washington, D. C., at a price of $1.25 per copy, cloth bound.)

2. Henderson, R.: Mortality Laws and Statistics, New York, 1915, pp. v and 111.
    (This is an excellent brief elementary treatise on life table construction.)

3. Hayward, T. E.: On Life Tables: Their Construction and Practical Application, Jour. Roy. Stat. Soc., vol. 62, pp. 443–483, 1899, and pp. 683–702, 1899, vol. 63, pp. 625–636, 1900; Notes on Life Tables, Ibid., vol. 65, pp. 354–358, 1902; pp. 680–684, 1902.

4. Pearl, R., and Reed, L. J.: A Life Table Nomogram, Amer. Jour. Hygiene, vol. 5, pp. 330–334, 1925.

5. Pearl, R., and Parker, Sylvia L.: Experimental Studies on the Duration of Life. IX. New Life Tables for *Drosophila*, Amer. Nat., No. 58, pp. 71–82, 1924.

6. Pearl, R.: The Rate of Living. Being an Account of Some Experimental Studies of the Biology of Life Duration, New York (Alfred A. Knopf), 1928, pp. 185. (In this book the author's earlier experimental work in problems of life duration is summarized.)

7. Pearl, R.: The Graphic Representation of Relative Variability, Science, vol. 65, pp. 237–241, 1927.

8. Greenwood, M.: "Laws" of Mortality from the Biological Point of View, Jour. Hyg., vol. 28, pp. 267–294, 1928.

9. Snow, E. C.: An Elementary Rapid Method of Constructing an Abridged Life Table. *In* Supplement to the Seventy-fifth Annual Report of the Registrar General...England and Wales. Part II. Abridged Life Tables. London (H. M. Stationery Office), 1920, pp. xlv + 65 (Cmd. 1010); price, 1s. 6d.

CHAPTER IX

STANDARDIZED AND CORRECTED DEATH-RATES

It has been seen in Chapter VII (Table 15 and Fig. 58) that the specific death-rates are characteristically different at different ages. The fact is also brought out strikingly by the q_x curve of the life table. Now this circumstance must obviously have important consequences in regard to the use of general death-rates at all ages to measure the comparative mortality in different communities. For suppose two communities to have absolutely *identical* specific death-rates at different ages. But suppose, further, that one of the communities is primarily a manufacturing place, and in consequence has a large excess of young adults in its population, whereas the other is primarily a residence city for elderly, retired persons. The former will have relatively few persons of advanced age where the specific death-rates are high. The latter will have relatively many of such persons. In consequence of this difference in the age distribution of the *living* the two places are bound to have quite different general death-rates at all ages, even though, as postulated, all the specific death-rates are identical in the two places.

It therefore follows that crude death-rates at all ages should be *corrected* to allow for differences in the age distribution of the general population. This may be done by the use of what are called *standardized* and *corrected death-rates.*

STANDARDIZED DEATH-RATES

A *standardized death-rate* is an *abstract* or *theoretic* figure derived by applying the specific death-rates of the general population, or of some standard imaginary population, to the actually existing age and sex distribution of the living population of a particular locality to determine what would be the number of deaths in that locality if the specific death-rates of the standard population prevailed there, and then dividing the number of deaths so obtained by the actual total living population of the locality.

In the calculation of the standardized death-rate the actual deaths in the locality do not enter at all. Expressed in a formula the case is like this:

$$R_{St} = \frac{\Sigma\,(P_x \times q_x)}{\Sigma\,P_x}$$

where

R_{St} = a standardized death rate,

P_x = actual living population of age x in the community for which the rate is calculated,

q_x = the specific death-rate at age x in the general population, or in the life table population, or in some other arbitrarily chosen standard population,

Σ denotes summation over all values of x.

(Standardized death-rates are usually expressed as per 1000 of population.)

An example will make the case clear.

Suppose we take the life table population of the original Registration states in 1910, as determined by Glover, as a standard of reference, and confine attention, for the sake of simplicity, to age alone, dealing with both sexes together, we find the following specific death-rates at ages in that population to be as given in Table 33.

TABLE 33

LIFE TABLE DEATH-RATES, FROM TABLE 24 SUPRA

Age interval.	Rate of mortality per thousand living in current age interval.
Under 5	37.19
5– 9.9	3.44
10–19.9	2.93
20–39.9	6.64
40–59.9	15.28
60–79.9	56.22
80 and over	190.61
All ages together	19.42

Now an examination of the Mortality Statistics reveals that in the year 1910 the *crude death-rate* was,

In Providence, R. I. .... 17.66 per thousand
In Seattle, Wash. .... 10.05 " "

But the census of 1910 revealed further that the living populations of these two cities were constituted in respect of age as shown in Table 34.

TABLE 34

ACTUAL LIVING POPULATION IN 1910 OF PROVIDENCE AND SEATTLE

Age interval.	Population in thousands of Providence, R. I.	Population in thousands of Seattle, Wash.
Under 5...........................	21.814	17.043
5– 9.9...........................	18.707	15.123
10–19.9...........................	38.315	32.666
20–39.9...........................	83.563	109.340
40–59.9...........................	46.482	49.817
60–79.9...........................	14.111	10.140
80 and over......................	1.058	.590
Totals...........................	224.050	234.719

It is at once apparent that while these two cities were of about the same total size in 1910, the age distributions of the two populations were widely different. Providence had a great many more young people under twenty than had Seattle. Seattle, on the contrary, had many more young adults (twenty to thirty-nine) than had Providence. Plainly, Seattle would be bound to have a lower crude death-rate than Providence, because there were in the population *fewer* persons to whom high specific death-rates apply, and *more* persons to whom low specific rates apply.

Now, according to the rule set forth above, to get the standardized death-rate it is merely necessary to perform the operations shown in Table 35.

TABLE 35

EXPECTED DEATHS IN PROVIDENCE AND SEATTLE IN 1910 IF THE LIFE TABLE DEATH-RATES PREVAILED

Age interval.	Providence population × Life table specific death-rates (=deaths which would have occurred in Providence if life table rate of mortality had existed there).	Seattle population × Life table specific death-rates (=deaths which would have occurred in Seattle if life table rate of mortality had existed there).
Under 5............	811.26	633.83
5– 9.............	64.35	52.02
10–19.............	112.26	95.71
20–39.............	554.86	726.02
40–59.............	710.24	761.20
60–79.............	793.32	570.07
80 and over........	201.67	112.46
Totals...........	$3247.96 = \Sigma (P_x \times q_x)$	$2951.31 = \Sigma (P_x \times q_x)$

Hence

$$\text{For Providence } R_{St} = 1000 \left(\frac{3247.96}{224050}\right) = 14.50$$

$$\text{For Seattle} \quad R_{St} = 1000 \left(\frac{2951.31}{234719}\right) = 12.57$$

These figures tell us that if identical forces of mortality had operated in Providence and Seattle, the crude rates of the two places would have been different in the ratio indicated, solely because of differences in the age constitution of the living population. But it cannot have failed to impress one that it is a curious use of words to call this standardized rate a death-rate *of Providence*, for example, because in its calculation no account whatever is taken of the *deaths* which occurred in Providence. Providence's statistics only enter into the situation at all in respect of the living, not the dead. But surely a death-rate may not unreasonably be required to have in it something about the deaths which really occurred.

Can this be done on the basis of only such data as are now in hand? It can, and in this way. It has already been seen from Table 33 that, in the life table population which we are taking as a standard, the death-rate for all ages together is 19.42 per thousand. Now then it is obvious that the standardized rates which have been obtained above for Providence and Seattle *differ* from the death-rate for all ages in the standard population, *only* because of the differences in the age distribution of the living in the actual populations of Providence and Seattle respectively, and of the living in the standard population. Therefore it follows that the ratio

$$\frac{\text{Death-rate in standard population}}{\text{Standardized death-rate of local population}}$$

will give a *correction factor* which will measure the amount by which the *crude* death-rate of the local population is altered from the death-rate at all ages of the standard population, *as a result solely of the difference between the two populations in respect of the age distribution of the living.*

We then have

$$\text{Correction factor for Providence} = \frac{19.42}{14.50} = 1.339$$

$$\text{Correction factor for Seattle} = \frac{19.42}{12.57} = 1.545$$

These figures indicate that the crude death-rates of both cities are *lower* than they would be if their living populations had the same age distribution as the standard population, even though both cities had the same specific forces of mortality *that they actually do.* If the correction factor were less than 1 it would mean that the crude death-rates were *higher* than they would be in a population of the same age distribution as the standard.

Now, as has been seen, the *crude death-rate* of Providence was 17.66, and of Seattle 10.05.   So then,

17.66 × 1.339 = 23.65 = a death-rate for Providence in which is included (*a*) the specific forces of mortality peculiar to Providence (introduced implicitly in the crude rate 17.66); and (*b*) an allowance for the peculiar age distribution of the living population of Providence, which brings it to identity with the age distribution of the standard population.

Similarly for Seattle, we have

10.05 × 1.545 = 15.53 = a death-rate for Seattle which has the same properties as those described above for Providence.

CORRECTED DEATH-RATES

A *corrected death-rate* is an *abstract* or *theoretic* figure got by applying the specific death-rates observed in a local population to the age and sex distribution of some arbitrarily chosen standard population.   A corrected death-rate is, in short, just the reverse of a standardized death-rate.   It answers questions like the following: What would be the death-rate of city *A* if instead of having the actual age distribution of the population which it has, it had an age distribution identical with that of the standard population? How much of the difference in the crude death-rates of cities *A* and *B* is to be attributed to the fact that the age distributions of the populations are different in the two places?

The formula for a corrected death-rate is,

$$R_{Co} = \frac{\Sigma \, (L_x \times R_{sx})}{\Sigma \, (L_x)}$$

where

R_{Co} = a corrected death-rate,
L_x = the number of persons of age x in the standard population,
R_{sx} = the specific death-rate at age x observed in the particular locality for which the corrected rate is being calculated,
Σ denotes summation over all values of x.
(Corrected death-rates are usually expressed as per 1000 of population.)

Coming back to the Providence-Seattle example we have already had given in Table 34 the populations of these two cities at ages. Table 36 gives the deaths at ages in columns (1) and (2). By dividing each figure in column (1) of Table 36 by the corresponding population figure of Table 34, we shall get the specific death-rates of Providence set down in column (3), and similarly for Seattle in column (4).

TABLE 36

SPECIFIC DEATH-RATES PER THOUSAND OF PROVIDENCE AND SEATTLE

Age interval.	Deaths in Providence. (1)	Deaths in Seattle. (2)	Specific death-rate in Providence (per 1000). (3)	Specific death-rate in Seattle (per 1000). (4)
Under 5.................	1175	453	$\dfrac{1175}{21.814} = 53.86$	$\dfrac{453}{17.043} = 26.58$
5– 9....................	74	50	$\dfrac{74}{18.707} = 3.96$	$\dfrac{50}{15.123} = 3.31$
10–19...................	144	107	$\dfrac{144}{38.315} = 3.76$	$\dfrac{107}{32.666} = 3.28$
20–39...................	596	623	$\dfrac{596}{83.563} = 7.13$	$\dfrac{623}{109.340} = 5.70$
40–59...................	854	625	$\dfrac{854}{46.482} = 18.37$	$\dfrac{625}{49.817} = 12.55$
60–79...................	954	447	$\dfrac{954}{14.111} = 67.61$	$\dfrac{447}{10.140} = 44.08$
80 and over...........	182	103	$\dfrac{182}{1.058} = 172.02$	$\dfrac{103}{.590} = 174.58$
Totals.................	3979	2408		

The next step is to multiply the appropriate standard population figures derived from Tables 29, 30, and 31 of the preceding chapter by the specific death-rates of Table 36 above, to get the number of deaths which would have occurred in Providence and Seattle had their living population been that of our standard million, and their specific forces of mortality as they were. This is done in Table 37.

TABLE 37

DEATHS EXPECTED IN 1910 IN PROVIDENCE AND SEATTLE IF THEIR POPULATIONS HAD HAD THE SAME AGE DISTRIBUTION AS THE STATIONARY LIFE TABLE POPULATION

Age interval.	Persons in standard population in thousands. (1)	(1) × Providence specific death-rates per 1000. (2)	(1) × Seattle specific death-rates per 1000. (3)
Under 5.......	84.155	84.155 × 53.86 = 4,532.6	84.155 × 26.58 = 2,236.8
5- 9.........	80.682	80.682 × 3.96 = 319.5	80.682 × 3.31 = 267.1
10–19.........	158.141	158.141 × 3.76 = 594.6	158.141 × 3.28 = 518.7
20–39.........	293.408	293.408 × 7.13 = 2,092.0	293.408 × 5.70 = 1,672.4
40–59.........	240.338	240.338 × 18.37 = 4,415.0	240.338 × 12.55 = 3,016.2
60–79.........	129.303	129.303 × 67.61 = 8,742.2	129.303 × 44.08 = 5,699.7
80 and over....	13.973	13.973 × 172.02 = 2,403.6	13.973 × 174.58 = 2,439.4
Totals......	1,000.000	23,099.5	15,850.3

Whence we have: .

$$\text{For Providence: } R_{Co} = 1000 \, \frac{23,100}{1,000,000} = 23.10$$

$$\text{For Seattle: } \quad R_{Co} = 1000 \, \frac{15,850}{1,000,000} = 15.85$$

It will be noted at once that these corrected death-rates are nearly the same as those got by the correction factor from the standardized rates above. There are thus available two different methods of computation for getting corrected death-rates. The method given in this section is the more refined and exact.

The same principle as that which has been illustrated in Table 37 can be successively applied, provided the necessary data are at hand, to correct death-rates for a whole series of variables. Actually, the necessary data are usually not available, so that when a corrected death-rate is spoken of, all that is commonly meant is a death-rate corrected for the age and sex distribution of the population.

THE SIGNIFICANCE OF STANDARD POPULATIONS IN CALCULATING CORRECTED DEATH-RATES

It will have been perceived by the thoughtful that all that a corrected death-rate is *is a weighted average of the local specific death-rates*, the weighting being in proportion to the portions in each age group of the population chosen as the standard. Looking at a

corrected death-rate in this way one is led to ask the question: What is the best system of weights to choose, or, in other words, What shall be taken as the standard million of population?

The answer to this question depends in part, as do all similar questions of weighting, upon what answer is given to the further question: What do you want to do with the corrected death-rate after you get it? If one's point of view is to seek what would be the value of a local death-rate if the locality had the average population distribution of the whole country of which it is a part, the standard population will be so chosen as to be nearly or quite identical with the actually existing population of the whole country. This is the usual procedure. The Registrar-General of England and Wales uses as a standard of reference the age and sex distribution of the actual population of England and Wales over a period of years.

If, on the other hand, one is interested in getting as stable a standard, both in time and space, as is possible, the L_x population of a life table will be better than any actually existing population. This will, however, just because it is not a growing population, be quite different from most existing populations in respect of age distribution, as has already been seen in the preceding chapter. Table 38 shows a standard million of the population of the United States in 1910 distributed to the same age classes as used in the Providence-Seattle example. It is obviously quite different from the life table standard population given in Table 37 on page 271.

TABLE 38

A STANDARD MILLION FROM THE ACTUAL LIVING POPULATION OF THE UNITED STATES
IN 1910

Age interval.	Population both sexes U. S., 1910.	Population basis, 1,000,000.
0– 4	10,631,364	115,806
5– 9	9,760,632	106,321
10–19	18,170,743	197,931
20–39	30,605,272	333,379
40–59	16,418,526	178,845
60–79	5,727,683	62,391
80 and over	488,991	5,327
Total	91,803,211*	1,000,000

* This total does not include "ages unknown."

Suppose we calculate the corrected death-rates of Providence and Seattle, weighting the specific death-rates with the million of Table 38 as a standard. The result is that shown in Table 39.

TABLE 39

EXPECTED DEATHS IN PROVIDENCE AND SEATTLE IN 1910, ON BASIS OF ACTUAL UNITED STATES POPULATION AS STANDARD

Age interval.	Persons in actual population, both sexes, in thousands. (1)	(1) × Providence specific death-rates per 1000. (2)	(1) × Seattle specific death-rates per 1000. (3)
0– 5.........	115.806	115.806 × 53.86 = 6,237.3	115.806 × 26.58 = 3,078.1
5– 9.........	106.321	106.321 × 3.96 = 421.0	106.321 × 3.31 = 351.9
10–19........	197.931	197.931 × 3.76 = 744.2	197.931 × 3.28 = 649.2
20–39........	333.379	333.379 × 7.13 = 2,377.0	333.379 × 5.70 = 1,900.3
40–59........	178.845	178.845 × 18.37 = 3,285.4	178.845 × 12.55 = 2,244.5
60–79........	62.391	62.391 × 67.61 = 4,218.3	62.391 × 44.08 = 2,750.2
80 and over....	5.327	5.327 × 172.02 = 916.4	5.327 × 174.58 = 930.0
Total........	1,000.000	18,199.6	11,904.2

Whence the

Corrected death-rate for Providence = 18.20
Corrected death-rate for Seattle   = 11.90

These values, for perfectly obvious reasons, are smaller than those got above on the basis of the L_x population and are much nearer absolutely to the crude rates. The *ratios* of the death-rates for the two cities are as follows:

$$\text{Crude} = \frac{17.66}{10.05} = 1.76$$

$$\text{Corrected } (L_x \text{ pop. standard}) = \frac{23.10}{15.85} = 1.46$$

$$\text{Corrected (actual pop. standard)} = \frac{18.20}{11.90} = 1.53$$

It is seen that the judgment of the *relative* mortality rates of Providence and Seattle is not sensibly altered if use is made of the life table population or of the actually existing population of the whole country as standard. The *ratios* are only .07 apart. But both ratios are far from that derived from the *crude* rates.

One can obviously build up standard populations in various ways. One which has been used is to take a million persons so

18

distributed as to age (and sex if one wishes) as to yield 1000 deaths per year on the basis either of (*a*) the specific death-rates of the actual population of the whole country, or of (*b*) the specific death-rates of the life table.

On the whole, the matter is really one of arbitrary choice, governed essentially by taste and viewpoint as to purpose, rather than strict logic. My own preference is for the L_x population of the life table as a standard, because of its inherent stability. If one recognizes that any corrected death-rate is *at best* a purely artificial figure, there will be no need to worry over the artificiality of a life table population as a standard.

From a purely biologic viewpoint probably the most significant system would be one which weighted equally each specific death-rate and averaged. This is the same as assuming an equal number of persons in each age group of the standard population. This idea is not likely to appeal to public health officials or to professional official vital statisticians. It is based upon these considerations. Provided the subsamples at ages are sufficiently large each to give a reliable rate, having regard to the probable errors, any age and sex specific death-rate is a definite quantitative biologic attribute of the group to which it applies. It differs between group A_x and group B_x because of one or the other *or both* of the following factors, and for no other reason:

1. The organisms composing A_x are inherently different from those composing B_x.

2. The environment of A_x is different from that of B_x.

The simple, unweighted average of age specific death-rates gives in a single numeric value not any measure of the public health, but an excellent measure of a highly significant biologic situation. It offers a method of getting a little nearer to an adequate appreciation of the relative influence of constitution and environment in determining mortality rates.

OTHER APPLICATIONS OF THE CORRECTED RATE PRINCIPLE

While we have considered so far in this chapter only examples of forming standardized and corrected death-rates, the student should understand that the method which has been used is of much wider

and more general application. In fact, the method is theoretically perfectly general. It can be employed to correct any crude *rate*, as defined at the beginning of Chapter VII, for the influence of any number of variables for which the requisite data are available.

In further illustration of the method it is proposed now to give another example in a different field than death or death-rates. The material for the example is given in a paper by Dr. Robert H. Riley,* Director of the Maryland State Department of Health, dealing with the disease infantile paralysis (poliomyelitis). His Table 1 includes the case incidence of the disease during the 1928 epidemic outbreak in five states. Here two only of these states, California and Minnesota, will be taken for purposes of illustration. In California 289 cases occurred, and in Minnesota 221. We have then the following *crude incidence rates*, the population being estimated population as of 1928.

For California: $100,000 \dfrac{289}{4,556,000} = 6.3 =$ crude incidence rate per 100,000.

For Minnesota: $100,000 \dfrac{221}{2,722,000} = 8.1 =$ crude incidence rate per 100,000.

On the basis of the crude rates alone, Minnesota appears to have had about a third heavier incidence of the disease than California.

Table 40 gives (*a*) the age incidence of the cases, as reported by Riley, (*b*) the estimated populations (in thousands) for the same

TABLE 40

CASES OF POLIOMYELITIS, ESTIMATED POPULATIONS, AND AGE SPECIFIC INCIDENCE RATES PER 100,000 OF POLIOMYELITIS IN 1928 IN CALIFORNIA AND MINNESOTA

Age in years.	California.			Minnesota.		
	Cases of poliomyelitis. (*a*)	Population in thousands. (*b*)	Incidence rate. (*c*)	Cases of poliomyelitis. (*a*)	Population in thousands. (*b*)	Incidence rate. (*c*)
Under 1.....	16	73	0.000219	0	57	0
1 to 4.....	73	292	.000250	72	242	.000298
5 to 9.....	89	374	.000238	68	283	.000240
10 to 14.....	35	346	.000101	36	267	.000135
15 to 19.....	35	323	.000108	23	250	.000092
20 and over..	41	3148	.000013	22	1623	.000014
Totals....	289	4556	....	221	2722	

* Riley, R. H.: Poliomyelitis, Jour. Amer. Med. Assoc., vol. 94, pp. 550–557, 1930.

ages in 1928, and (c) the age specific incidence rates, got in each case by dividing, line by line, (a) by (b).

Using the standard million of the stationary life table population for reference, we have in Table 41 the number of cases of poliomyelitis expected to occur in that population under the age specific incidence rates of California and Minnesota respectively.

TABLE 41

EXPECTED INCIDENCE OF POLIOMYELITIS IN CALIFORNIA AND MINNESOTA IF BOTH HAD THE SAME STANDARD POPULATION (THE STATIONARY LIFE TABLE POPULATION)

Age in years.	Stationary life table population. (a)	California.		Minnesota.	
		Age specific incidence rates from Table 40. (b)	Expected cases in specified standard million of population. (a) × (b)	Age specific incidence rates from Table 40. (b)	Expected cases in specified standard million of population. (a) × (b)
Under 1.....	17.841	0.000219	3.9	0	0
1 to 4.....	66.314	.000250	16.6	.000298	19.8
5 to 9.....	80.682	.000238	19.2	.000240	19.4
10 to 14.....	79.628	.000101	8.0	.000135	10.7
15 to 19.....	78.513	.000108	8.5	.000092	7.2
20 and over..	677.022	.000013	8.8	.000014	9.5
Totals....	1,000,000	....	65.0	....	66.6

It thus appears that the incidence rates of poliomyelitis in California and Minnesota in 1928, when corrected to the same age distribution of the population (that of the stationary life table population), have the following values:

$$\text{California: } 100,000 \, \frac{65.0}{1,000,000} = 6.5 \text{ cases per } 100,000 \text{ population.}$$

$$\text{Minnesota: } 100,000 \, \frac{66.6}{1,000,000} = 6.7 \text{ cases per } 100,000 \text{ population.}$$

The difference in the crude rates of the two states thus disappears upon correction for age differences.

In the illustrations given in this chapter the correction has been for differences in age distribution of different populations. The same method can be used to correct for differences in population distributions relative to sex, color, race, occupation, and, indeed, any other factor for which the necessary data are available.

SUGGESTED READING

1. Brownlee, J.: The Use of Death-rates as a Measure of Hygienic Conditions, Medical Research Council, Spec. Rept. Series, No. 60, pp. 80, 1922.
2. Greenwood, M., et al.: Value of Life-tables in Statistical Research, Jour. Roy. Stat. Soc., vol. 85, pp. 537–560, 1922.
      (These two papers should always be read together.   By so doing the reader will preserve his mental balance.)
3. Collis, E. L., and Greenwood, M.: The Health of the Industrial Worker, London (J. and A. Churchill), 1921.
      (Especially Chapter III on Statistical Methods.)
4. Greenwood, M.: Is the Statistical Method of Any Value in Medical Research? The Lancet, July 26, 1924, p. 153.

CHAPTER X

THE PROBABLE ERROR CONCEPT

PERHAPS the simplest and most direct way in which statistical methods can be of practical use to the medical man in his everyday problems is by giving him a means of measuring and stating precisely the degree of reliability which attaches to any particular set of results or conclusions he may reach. Only a little consideration of the matter will be necessary to convince anyone that the reliability or trustworthiness of any conclusion is in some way a function of the number of cases upon which it is based. If the number of cases determined forms but a small sample of all the cases it would be possible to collect, it is probable that there will be considerable fluctuation among the results given by such small sampling.

As an illustration of the effect of random sampling, let us consider the following case: In any large city, or a state, or indeed, any large population aggregate, the *average age at death* of persons dying at the same calendar date should be identical for all dates, except for the influence of two factors, viz., (*a*) chance, or random sampling, and (*b*) long seasonal waves arising from such considerations as that relatively more infants die in hot summer weather than in the colder seasons of the year. In any short period, say ten consecutive days, the second factor (*b*) would not operate in any sensible degree, and we should expect the persons dying on each of these consecutive ten days to show the same average age, except for the fluctuations due to chance alone. How considerable these fluctuations may be is shown in Table 42, which gives the number of deaths and the age at death of those dying during ten consecutive days in 1916 in Baltimore.

Here we have a fluctuation in the average, based on samples of from 30 to 50 individuals, amounting to more than twenty-two years, arising from random sampling alone. Such an illustration

TABLE 42

MEAN AGE AT DEATH OF THOSE DYING IN THE STATED DAYS IN BALTIMORE

Date.	Number of deaths.	Mean age at death in years.
January 13, 1916.	31	30.16
January 14, 1916.	40	43.80
January 15, 1916.	27	40.59
January 16, 1916.	48	48.21
January 17, 1916.	32	48.34
January 18, 1916.	41	51.90
January 19, 1916.	39	46.82
January 20, 1916.	31	52.39
January 21, 1916.	39	51.62
January 22, 1916.	57	39.40

emphasizes the fact that before conclusions can safely be drawn from differences between numbers it is necessary to know something about the "probable errors" of those numbers.

Another example of random fluctuations may be given: In "Who's Who" the names are entered in alphabetic order. If one takes five names in order as they are given and determines the average age at which these five persons married, and then takes the next five names in order and does the same thing, and so on, there is no reason why the average ages at marriage should not be identical for all such groups of five, *except for the operation of chance.* Five is a small sample, and we know from practical experience of life that probably the first set of five ages at marriage so chosen will not give quite the same average as the second set, and so on.

Table 43 gives the result of such an experiment with ' Who's Who." I opened Vol. X (1918–19) at random and the page chanced to be 680. This is in the letter D and the first name on that page is William Franklin Dana. I then calculated the age at marriage for each person in order, without any omissions whatever, except such as were occasioned by (*a*) failure of the person to have married, or (*b*) absence of birth date or marriage date, or both. The figures obtained are given in the upper half of Table 43. As soon as the fifth age of each set of five was set down the average for that group was calculated before going on to the next name. This was kept up till ten groups or fifty names had been taken out.

When this first series was done and the means plotted, it was decided to take a second fifty names from another part of the alphabet. So the book was opened again at random and the page chanced to be 2486, with the first name Frederic Singer. The same procedure as before for fifty consecutive names gave the bottom half of Table 43.

TABLE 43

SHOWING THE AVERAGE AGE AT MARRIAGE OF TEN CONSECUTIVE GROUPS OF FIVE PERSONS EACH, TAKEN IN ORDER FROM "WHO'S WHO" IN LETTER D Beginning at p. 680.

Age at marriage.	Age at marriage.	Age at marriage.	Age at marriage.	Age at marriage.
I { 22 34 35 34 25	III { 30 30 26 26 31	V { 30 39 28 30 35	VII { 28 38 41 46 38	IX { 31 28 33 30 28
Average 30.0	Average 28.6	Average 32.4	Average 38.2	Average 30.0
II { 23 30 33 36 33	IV { 29 26 21 26 26	VI { 33 45 23 32 36	VIII { 32 27 24 32 28	X { 28 25 33 50 28
Average 31.0	Average 25.6	Average 33.8	Average 28.6	Average 32.8

A SECOND GROUP LIKE ABOVE, BUT FROM LETTER S

Beginning at p. 2486.

Age at marriage.	Age at marriage.	Age at marriage.	Age at marriage.	Age at marriage.
I { 33 25 28 31 28	III { 32 30 28 27 36	V { 28 35 35 29 22	VII { 25 37 31 32 32	IX { 32 28 28 27 32
Average 29.0	Average 30.6	Average 29.8	Average 31.4	Average 29.4
II { 29 31 23 30 27	IV { 31 24 26 30 25	VI { 28 29 45 25 35	VIII { 23 27 30 27 31	X { 24 24 33 30 29
Average 28.0	Average 27.2	Average 32.4	Average 27.6	Average 28.0

The means of the two series are shown graphically in Fig. 67, the solid line showing the group means for the 50 persons whose names began with D, and the broken line the group means for the persons having names beginning with S.

Table 43 and Fig. 67 show a number of interesting things about random sampling and the phenomenon we call chance. In the first place, the fluctuations of the group averages are large, considering the inherent stability of the phenomenon with which we
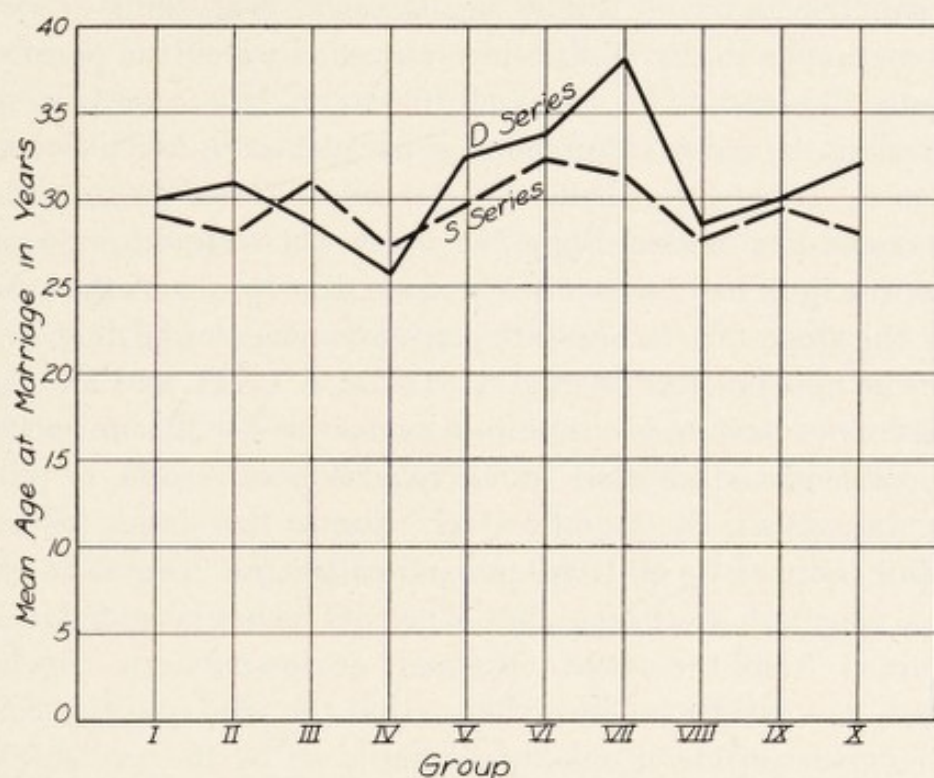
Fig. 67.—Group averages of age at marriage of persons taken at random. (Data from Table 43 above.) The Roman numerals indicate the order of the groups from the starting-points indicated in the text. Solid line = data from upper half of table. Broken line = data from lower half of table.

are dealing. In the D series Group IV has a mean age at marriage of 25.6 years, while Group VII has a mean of 38.2, almost thirteen years higher. In the second place the means of the D series do not fluctuate about a straight horizontal line. Instead there are three more or less well-defined trends, downward from Group I to IV, upward from Group IV to VII, and generally downward from Group VII to the end.

In the third place, the S series does not show such extreme

fluctuations of the group means, nor generally such high absolute values of these means, as does the D group. In the fourth place, there is *apparently* a curious suggestion of a rough parallelism in the courses of the lines of means for the D and S series. Probably not a few non-statistically trained experimental investigators would be apt to say, if they performed a series of 10 experiments and got results like those shown in the D series, and then repeated the series and got results like those shown in the S series, that the second series *confirmed* the first. So it does in respect of everything *except* the apparent trends in the D series, in respect of which the parallelism is wholly illusory. The case well illustrates how easy it is to be deceived by the *general* impression of parallelism of two lines known each to be subject to chance fluctuations. As a matter of fact if one counts the cases in Fig. 67 in which, between two consecutive points, the lines have slopes in the same direction, and the cases in which the slopes are in opposite directions, it is found that in four out of the nine possible cases (I–II, II–III, VI–VII, and IX–X) the D and S lines have opposite slopes, against five with similar slopes.

A conventional measure of the reliability of results, or put the other way about, of their "scatter" due to the chance effects of sampling, is used by statisticians and called the "probable error." It is a constant so chosen that when its value is added to and subtracted from the result obtained, or the numeric conclusion reached, it is exactly an even chance that the true result or conclusion lies either inside or outside the limits set by the probable error in the plus and minus direction. For example, if it is stated that the mean age at death of persons dying in Baltimore is 39.83 ± 2.60 years, it means that the mathematical probability that the *true* average age falls between 37.23 years $(39.83 - 2.60)$ and 42.43 years $(39.83 + 2.60)$ is exactly equal to the mathematical probability that the true age falls outside those limits.

The significance of any result is to be judged by its relation to its probable error. A simple theorem in probability tells us that the probable error of the difference between any two independent quantities (*i. e.*, quantities such that there is no correlation between their errors) is equal to the square root of the sum of the squares of the probable errors of the quantities entering into the difference.

It will be perceived then that the probable error of a difference will necessarily be larger than either of the two probable errors entering into its determination. Every student of elementary geometry knows that the hypothenuse of a right triangle is longer than either of the other sides. The square of the hypothenuse is equal to the sum of the squares of the other two sides, just as the square of the probable error of a difference is equal to the sum of the squares of the probable errors of the two quantities entering into the difference. It should be particularly noted by the student that this expression for the probable error of a difference is true only under the particular condition stated above, as to absence of correlation of errors. The general formula, true in all cases, is given in the third line of Table 61, p. 361.

As an example of the use of the probable error of a difference, suppose that a physician found, after administering a standard dose of a drug to a considerable number, say 150 people, that the pulse rate was $81.12 \pm .20$ beats per minute, while the normal condition in the same group was $79.68 \pm .15$ beats per minute. Would he be justified in concluding that the drug significantly increased the heart rate, or is the apparent increase simply a result of chance, arising from sampling? We have the following very simple calculation:

$$\text{Difference} = 81.12 - 79.68 = 1.44,$$
$$(.20)^2 + (.15)^2 = .0400 + .0225 = .0625,$$
$$\sqrt{.0625} = .25$$

Or we see that the difference in the two cases is $1.44 \pm .25$. The difference, small as it is absolutely, is approximately six times its probable error. Is a difference six times its probable error likely to arise from chance alone, or does it represent a really significant difference?

There has grown up a certain conventional way of interpreting probable errors, which is accepted by many workers. It has been practically a universal custom among biometric workers to say that a difference (or a constant) which is smaller than twice its probable error is probably not significant, whereas a difference (or constant) which is three or more times its probable error is either "certainly," or at least "almost certainly," significant.

Now such statements as these derive whatever meaning they may possibly have from the following simple mathematical considerations. Assuming that the errors of random sampling are distributed strictly in accordance with the normal or Gaussian curve, which will be discussed in some detail in the next chapter, it is a simple matter to determine from any table of the probability integral the precise portion of the area of a normal curve lying outside any original abscissal limits, or, in other words, the probability of the occurrence of a deviation as great as or greater than the assigned deviation. To say that a deviation as great or greater than three times the probable error is "certainly significant" means, strictly speaking, that the area of the normal curve beyond 3 P. E. on either side of the central ordinate is negligibly small. As a matter of fact this is not true, unless one chooses to regard 4.3 per cent. as a negligible fraction of a quantity. There are certainly many common affairs of life in which it would mean disaster to "neglect" a deviation of 4 per cent. of the total quantity involved.

In order that a more adequate conception may be had of just what the probable error, and various multiples of it, mean, Figs. 68 to 71 are inserted here. They show the areas of the normal curve inside and outside certain limits.

From these diagrams one may perceive exactly what is meant when he says, for example, that a difference which is three times its probable error is *certainly* significant. He means that the sum of the two cross-hatched areas in Fig. 70 is a wholly negligible quantity in comparison with the blank area under the curve in the same figure. Everyone will agree, after looking at Fig. 71, that a conclusion based upon a difference four or more times its probable error is practically safe, so far as concerns purely statistical considerations.

Table A of Appendix III (p. 438) sets forth, for a series of ratios between a statistical deviation and the "probable error" of the distribution, first, the probability that a deviation as great as or greater than the given one will occur, and second, the odds against the occurrence of such a deviation. The probabilities are expressed on a percentage basis, on the ground that they will probably in this way make a more direct appeal to the average mind, since we are

more accustomed to thinking in terms of parts per 100 than per any other number. A single example will indicate how the table is to be used. Suppose one has determined the mean of each of two

-1 -.5 0 +.5 +1
Lower   Upper
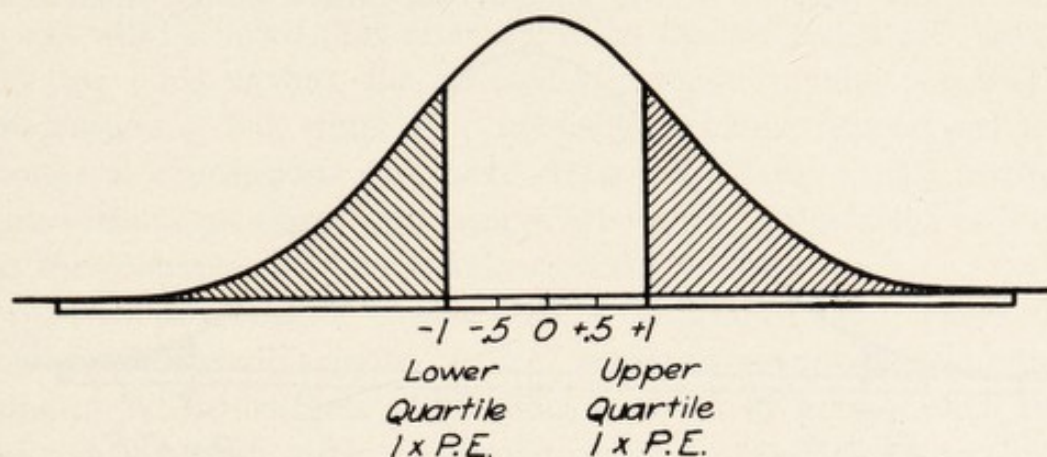Quartile   Quartile
1 x P.E.   1 x P.E.

Fig. 68.—The area of a normal curve inside (blank) and the area outside (cross-hatched) the lower and upper quartiles. The quartiles are the points on the abscissa where perpendiculars to the base cut off just one-quarter of the total area of the curve at each end. By definition of the probable error given above, it is seen that the quartile distance on the x axis is $1 \times$ P. E. The sum of the two cross-hatched areas is exactly equal to the blank area in the center.

comparable series of measurements. These means, which may be called A and B, differ by a certain amount. The difference is found
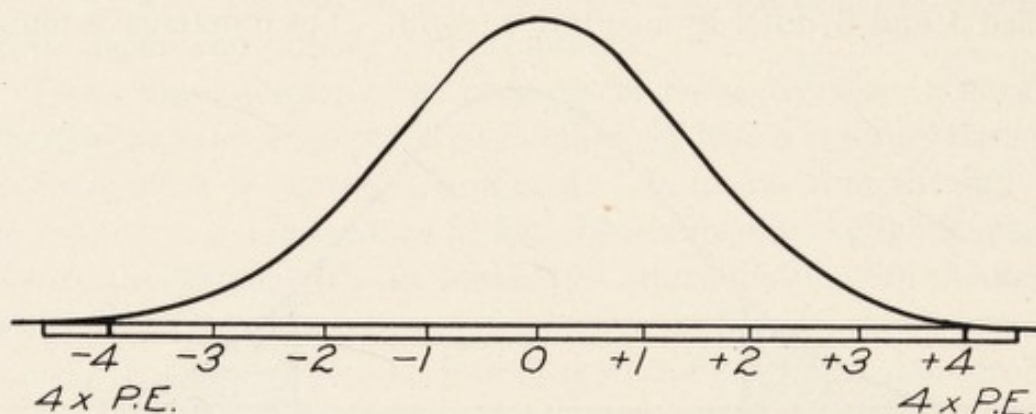

-2   -1   0   +1   +2
2x P.E.   2x P. E.
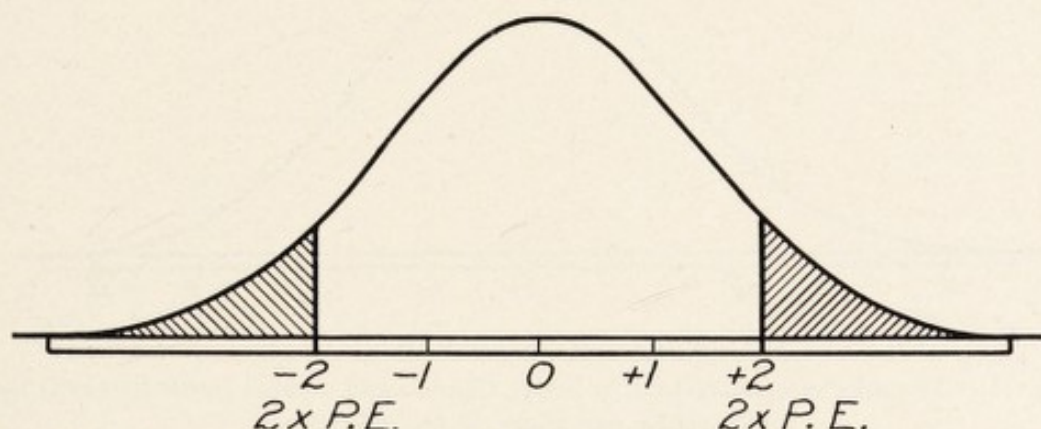
Fig. 69.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *twice the probable error*.

to be, let us say, 3.2 times as large as the probable error of the difference. Is one mean *significantly* larger than the other? Or, put in another way, what is the probability that the difference arose

purely as a result of random sampling (as a result solely of chance)? Under the argument 3.2 in the table we find the probability of the occurrence of a deviation as great or greater than this to be 3.09. This means that if, in the general population from which our
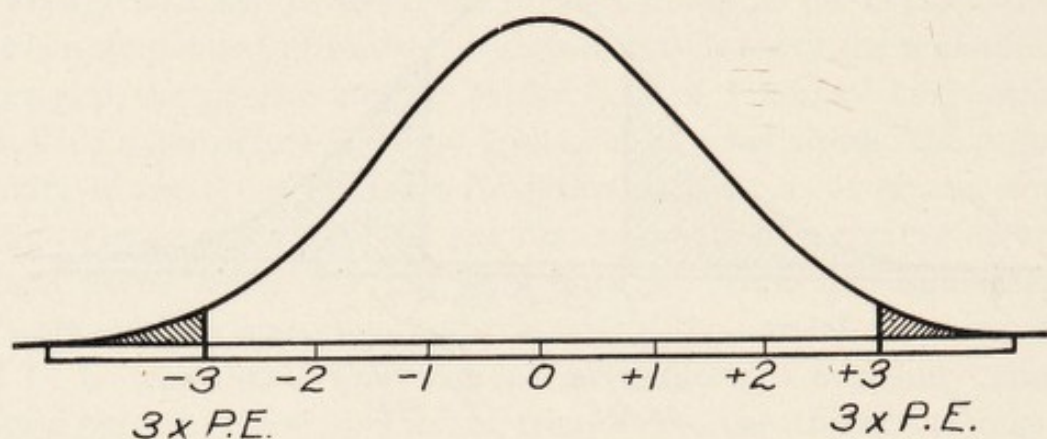
Fig. 70.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *three times the probable error.*

samples are drawn, the means A′ and B′ were truly and absolutely *identical*, and we drew successively 100 pairs of samples of the size which led to the two observed means, and took the difference between the averages in the case of each of the 100 pairs, there would

Fig. 71.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *four times the probable error.*

be about 3 cases in the 100 trials in which the difference would be as great as or greater than that actually found between the two observed means A and B with which we started this discussion. Or, from the next column, the odds against the occurrence of a difference as great or greater than this in proportion to its probable

error, are 31.36 to 1, if chance alone were operative in the determination of the event. If one wants to call this "certainty" he has a perfect right to do so. The table merely defines quantitatively his particular conception of certainty.

It will be noted that after the ratio, deviation ÷ P. E., passes 3.0 the odds against the deviation increase rapidly, reaching a magnitude at 8.0, which is, practically speaking, beyond any real power of conception. We have started the table at 1.0 because this is the point where the chances are even. A deviation as large as the probable error is as likely to occur as not.

From this table it is seen that a deviation of four times the probable error will arise by chance less often than once in a hundred trials. When one gets a difference as great or greater than this he may conclude with reasonable certainty that it did not arise by chance alone, but may have significant meaning.

SUGGESTED READING

1. Brownlee, J.: The Theory of Probable Error and Its Application to Vital Statistics, Transactions of the Royal Sanitary Institute, London, vol. 34, pp. 87–106, 1914.
    (This reference may most profitably be read after the next chapter has been studied.)
2. Yule, G. U.: Introduction to the Theory of Statistics, Chapters XIII and XVII particularly.

CHAPTER XI

ELEMENTARY THEORY OF PROBABILITY

THE TOSS OF A PENNY

THE tossing of a coin is a classical event in the discussion of probability. Let us examine somewhat carefully what this event consists of and involves. Consider first the penny. It is a simple mechanism, but possesses two important structural characteristics. These are:

1. It is *thin*. By this we mean, more precisely, that it is a right cylinder, having its height very small as compared with its diameter.

2. The two ends of the cylinder which we call a penny are so marked as to be distinguishable from one another. One of these ends is called the head, the other the tail.

Now the general experience of mankind with structures like a penny, that is, with exceedingly short cylinders, is that only in one or the other of two positions are they in *stable* equilibrium. These positions are respectively, standing on the head end or standing on the tail end. Everyone knows that a penny on its edge (which is of course the side of the cylinder) is in a highly unstable position, so much so in point of fact that, except by an excess of precaution which would physically be exceedingly difficult and expensive of attainment, a penny will not stand free of support on its edge for more than an extremely short time. *Why* everyone knows this is simply and solely because he has tried it. That is, his personal and racial experience with machines or structures like pennies, *and this experience alone*, has taught him that they will not stand on edge. No amount of *a priori* reasoning, in the complete absence of experience, could safely lead to this conclusion.

Since pennies then always do come to rest with either head or tail uppermost following any disturbance of their previous state of rest, we are led to a further question. Is there anything in the *structure* of the penny which makes it any more easy for it to come to rest after a disturbance of its prior state of equilibrium on its

head end than on its tail end, or *vice versa?* Again we call upon our general experience of machines and structures. Plainly that experience gives us no warrant for believing that the slight differences in the pattern of the two ends of a penny do, in fact, sensibly favor either the head or the tail position of equilibrium in any particular case.

We have now gained two important results, both based upon general experience, personal and racial. They are that when a structure like a penny comes to rest after a disturbance, *the structure itself determines* that there are only two possible positions of stable equilibrium, and that there is nothing in the structure itself which makes one of these any easier of attainment than the other.

So much for the structure of the penny. Now for its tossing. Tossing can be interpreted as any disturbance of a prior state of equilibrium. Is there anything in the tossing which makes it easier for the penny to come to rest, when it does so come, with one end rather than the other uppermost? Plainly this depends upon how the tossing is done. Suppose a penny to be sitting on its tail end (that is, head up) on the desk before me. If I carefully grasp two opposite points of its periphery between my thumb and forefinger and raise it just one millimeter from the table, and then let go, it will again come to rest with head up. I can repeat this performance industriously forever, and it will always come to rest head up. The same result will happen if I raise it just two millimeters before I let go. How do I know this? From past experience of falling bodies in air, and in particular from experience of excessively short cylinders falling distances less than their diameter in air. So then we see that it is possible to disturb the stable equilibrium of a penny at rest, and have it always return to the same position of rest. Equally it is possible so to disturb the penny that it will always return to the *opposite* position of equilibrium to that which it had before. I have only to give it a sufficiently strong flip at the start of a fall through a distance a little more than its own diameter to turn it over just once in the course of that fall.

But now suppose I drop the penny from a much greater height than those we have spoken about; or literally *toss* it, that is, pick it up from the table and throw it into the air; or set it spinning like a top on its edge; or roll it across the table or floor on its edge.

19

Then I have fundamentally altered the situation. No longer have I disturbed the equilibrium in such a way as to make it easier for the penny to come to rest on one of its ends rather than the other, as was the case in the examples discussed in the previous paragraph. On the contrary, by these operations of tossing described in this present paragraph, I have in each case *lost control* of the future movements of the penny as soon as it leaves my hand. An indefinitely large number of circumstances can influence its course before it comes to rest. But since I cannot control these circumstances, I call them *random*. So long as I could control the circumstances I could predict with positiveness and certainty the final position of rest of the penny, knowing what I did about its structure. Still knowing just as much as before about the structure of the penny, and it being just as fixed and determinate as before, I have lost my power of prediction because I have introduced, in the tossing, *and only in the tossing*, an element of *randomness*.

What do we mean by randomness? Only this, that a penny tossed at random is one tossed in such a way that the attainment of one of the possible states of equilibrium is not more favored than the other *in or by the act of tossing*. Therefore, since, as we have seen, the structure of the penny does not favor one position of rest more than the other, and the method of tossing does not favor one more than the other, there is nothing so far to enable us to assert, on the basis of what is known by experience, that the penny will more often come to rest on one end than on the other end.

Can we then assert the opposite, namely, that the penny *will*, under the conditions of structure and tossing named, come to rest with the head end uppermost as often as with the tail end uppermost? Here we come to a sharp division of opinion among students of the foundation of the theory of probability. There are those who maintain that solely on the basis of experience with structures like pennies and random tossing, or even without experience by pure induction from the structure of a penny and from the abstract idea of randomness, we are able by *a priori* reasoning to assert that the penny tossed at random will come to rest as often with head uppermost as with tail. These persons, in short, assert that fundamentally our notions of probability are purely *a priori*.

But this view overlooks, as it seems to me, a most important consideration. How can one know that the *only* things concerned in determining which of the alternative positions of equilibrium of a penny shall eventuate, are things related solely to the structure of the penny and the randomness of the tossing? Plainly he cannot know *a priori*. In fact this is one of the most important things he wants to find out in a research on penny tossing. *A priori* one could not possibly assert that there might not be some wholly unknown and unperceived cosmic principle influencing the coming to rest of pennies. At not so remote times in the history of human thought it might easily have been solemnly asserted that a demon, or some other supernatural agent, interested himself in penny tossing.

And today the only way to prove that a demon is *not* involved in the affair is *to try the case*. Now what is found when one tries it, by tossing a normal penny a great many times in a random way, is that in fact the penny comes to rest in the long run just about as many times, and no more, with head uppermost as with tail uppermost. But this is just what would be expected if *the only things concerned* were the structure of the penny and the randomness of the tossing. Hence it may reasonably be concluded, *on the basis of this experience*, and on this basis alone, that there are no supernatural agencies involved, and that in these two factors of structure and randomness we have the sole essential elements.

By this long argument I hope it has been made clear that the only basis we have for saying that when a penny is tossed at random it is as likely, or probable, that it will come to rest with the head up as with the tail up, *is the basis of experience*.* This experience, summarized, is of three sorts:

A. Experience of machines or structures like pennies, namely, cylinders excessively short in proportion to their diameter. This experience teaches that such structures can attain a steady state of rest only when lying on one end or the other, namely, with either head up or tail up.

* The student will find the same point of view which has been developed here as to the experiential basis of our knowledge of probability expressed in more general terms in the opening chapter of Professor Julian L. Coolidge's text-book.[8]

B. Experience of random tossing; namely, of uncontrolled phenomena, in which because of the lack of control one outcome is not more favored than another. This experience teaches that after a penny is randomly tossed the tosser has lost all control of which end shall be uppermost, head or tail, when it comes to rest.

C. Experience of tossing pennies many times. This experience teaches that if a true penny is tossed many times it will come to rest about one-half the times with the head up, and about one-half of the times with the tail up.

THE MATHEMATICS OF SIMPLE PROBABILITY

A penny can by virtue of its structure come to rest either head up or tail up. Suppose we call the times it happens the first of these ways a, and the times it happens the second b. Therefore the total possible times it can come to rest will be $a + b$. If the penny is tossed at random it is as likely to fall the a way (*i. e.*, H) as the b way (*i. e.*, T). In any one toss but one *actual occurrence* can happen (namely, the penny must come to rest on an end, not on the edge), though there are *two possible* ways in which the occurrence can happen (namely, it may come to rest on either the H or the T end). The mathematical measure of simple probability is taken *as the ratio in which (1) the number of times a particular specified event occurring at random in a class of events either has happened, or by inference from actual experience of similar events could have happened, is to (2) the whole number of times all kinds of events possible in the class either have happened, or, by inference from experience of similar events, could have happened.*

The numerical appreciation or determination of actual occurrences and of possible ways is, and must always be, based upon experience; but this experience may be of either of two sorts, namely, general experience of particular structures (as in the case of the penny), or particular statistical experience of events. But, however the numerical determination is derived, the form of the probability statement remains the same, a ratio or fraction; and no greater validity necessarily or absolutely inheres in the one

method of arriving at the numerical determination than in the other, so far as the resulting probability is concerned.

To return now to the penny:

The probability that after any one particular random toss a penny will come to rest with the head end up is, upon the reasoning given above,

$$p = \frac{a}{a+b}$$

In any one particular toss of one penny clearly either

$$a = 1,$$
$$\text{or } b = 1$$

and the whole number of possible ways in which the event can happen is $1 + 1$, whence

$$p = \frac{1}{1+1} = \tfrac{1}{2}$$

Similarly, the probability that after any one particular toss it will come to rest with the tail end up is

$$q = \frac{b}{a+b} = \frac{1}{1+1} = \tfrac{1}{2}$$
$$p + q = 1.$$

These results tell us that on any given single random toss of one penny it is an even or equal chance (or probability) that the penny will come to rest with head up. It is a certainty $(p + q = 1)$ that it will come to rest with either head or tail up.

Thus in the numerical expression of the probability of resting with head up after *one* random toss, the numerator of the fraction must be 1 because the specifications are that it shall be head up, and not otherwise. The denominator must be 2 because the whole number of possible ways is either head or tail $(= 2)$.

Suppose the penny to be tossed at random n times. How many times out of the n will it probably come to rest head up (H)?

Plainly pn, because one toss does not influence the next, nor the next, nor any other toss whatever. Therefore the number of H's in n trials must be n times the probability of H on one trial, which is $\tfrac{1}{2}$, as we have seen.

Now suppose we are dealing not with a particular structure

like a penny, but a series of events and wish to know the probability of occurrence of a particular kind of event in this series. Following the rule that the probability is the ratio of the frequency of actual occurrences of the specified sort to the total number of possible ways, we count in the statistical experience the occurrences of the specified kind and make the result the numerator of the probability fraction, and count the total number of all occurrences in the universe under discussion and put this result as the denominator.

Example: On the basis of the experience of the U. S. Birth Registration Area in 1919, what is the probability that any individual baby born in that area will be a male?

In 1919 male births = 705,593 = a
In 1919 total births = 1,373,438 = $a + b$

Therefore the probability that a given birth would be male is

$$p = \frac{705,593}{1,373,438} = .5137$$

The chance that a given birth would be a female is

$$q = 1 - p = 1 - .5137 = .4863$$

Or there were about fifty-one chances in a hundred that a given birth would be of a male.

The principles stated above regarding the fraction which measures probability may be extended to any number of mutually exclusive events equally capable of happening. Thus

$$p = \frac{a}{a + b + c + \ldots}$$

$$q = \frac{b}{a + b + c + \ldots}$$

$$r = \frac{c}{a + b + c + \ldots}$$

etc.

$$p + q + r + \ldots = 1$$

Example: What is the probability of drawing any number of just three figures from the entire list of numbers which can be formed from the first seven digits, it being specified that any digit can be used but once in forming any number?

The number of different three figure numbers which can be formed from the first seven digits is

$$210 = a$$

The whole number of different numbers (of 1 digit, 2 digits, etc.) which can be listed from the first seven digits is

$$13,699 = a + b + c + d + e + f + g$$

Therefore

$$p = \frac{210}{13,699} = \frac{1}{65}$$

The probability of drawing any one *particular* three figure number, say 123, is

$$p = \frac{1}{13,699}$$

But at this point some one will say: How do you know that just 210 different three figure numbers can be made up from the first seven digits? Or that the total of different numbers of all sizes from these seven digits is just 13,699?

To answer these pertinent questions it will be necessary to ask the reader to review briefly, as a digression from the main probability argument, which under all the circumstances will perhaps be pardoned, a small portion of his elementary college algebra, which the medical man has perhaps forgotten.

PERMUTATIONS

The number of different ways in which the three letters *a*, *b*, and *c* can be arranged (or permuted) in groups of three is plainly

$$a\ b\ c$$
$$a\ c\ b$$
$$b\ a\ c$$
$$b\ c\ a$$
$$c\ a\ b$$
$$c\ b\ a$$

These six different arrangements are the *permutations* of three things taken three at a time.

Generally we may write

$$_nP_r = n\,(n-1)\,(n-2)\,(n-3)\,\ldots\ldots(n-r+1) = \frac{\underline{|n}}{\underline{|(n-r)}}$$

which means that the number of permutations of n things taken r at a time $(_nP_r)$ is equal to factorial n, $\left(\underline{|n}\right)$, divided by factorial n minus r, $\left(\underline{|(n-r)}\right)$.

From this it will be perceived that

$$_nP_n = \underline{|n}$$

which in the case of our three letter example becomes

$$_3P_3 = 3 \times 2 \times 1 = 6,$$

just precisely the result we got experimentally.

The total number of permutations of n things taken singly, by twos, by threes, etc., is found by summing $_nP_r$ for all values of r from 1 to n.

Call this sum $\Sigma\,_nP_r$.

Then it can be proved that

$$\Sigma\,_nP_r = \underline{|n} + \frac{\underline{|n}}{1} + \frac{\underline{|n}}{1.2} + \frac{\underline{|n}}{1.2.3} + \ldots\ldots + \frac{\underline{|n}}{\underline{|(n-1)}}$$

$$= \underline{|n}\left(1 + \frac{1}{1} + \frac{1}{1.2} + \frac{1}{1.2.3} + \ldots\ldots + \frac{1}{\underline{|(n-1)}}\right)$$

It can further be shown that the series in the parenthesis approximates more and more closely in value the longer it is, to a number conventionally called e, which is the base of the Napierian system of logarithms, and has the value

$$e = 2.7182818\ldots\ldots$$

Hence it follows that for large values of n

$$\Sigma\,_nP_r = e\,\underline{|n}\ \text{approximately.}$$

The question at once arises: How large does n have to be to

make this approximation close enough for practical statistical purposes? The answer can be given by an example.

When $n = 9$, obviously not an excessively large number,
$\Sigma \, _nP_r = 986,410$, by the $e \lfloor n$ approximation,
$\Sigma \, _nP_r = 986,409$, exactly.

For the convenience of the reader a brief table of permutations and their sums is given as Table 44.

TABLE 44

VALUES OF PERMUTATIONS

Permutations of

r \ n	10	9	8	7	6	5	4	3	2	1
1..	10	9	8	7	6	5	4	3	2	1
2..	90	72	56	42	30	20	12	6	2	
3..	720	504	336	210	120	60	24	6		
4..	5,040	3,024	1,680	840	360	120	24			
5..	30,240	15,120	6,720	2520	720	120				
6..	151,200	60,480	20,160	5040	720					
7..	604,800	181,440	40,320	5040						
8..	1,814,400	362,880	40,320							
9..	3,628,800	362,880								
10..	3,628,800									
$\Sigma \, _nP_r$	9,864,100	986,409	109,600	13,699	1956	325	64	15	4	1

At a time.

COMBINATIONS

How many different *combinations* of three letters each can be made from the four letters a, b, c, and d? This is not the same problem as before. Now each combination of three letters must be different, not in respect of the order of the letters, but merely of the letters themselves. Thus only one of the combinations *abc* and *cab* can be used, because each contains the same letters, a, b, and c.

Writing down the possibilities we get

$$a\ b\ c$$
$$a\ b\ d$$
$$a\ c\ d$$
$$b\ c\ d$$

No other combination can be written which will not contain, in some arrangement, the same three letters that are in one or another of the four groups above.

Using a similar notation to that of permutations we have

$$_nC_r = \frac{\lfloor n}{\lfloor r \ \lfloor (n-r)}$$

which tells us how to find the number of different combinations of n things taken r at a time.   The example of the letters becomes,

$$_4C_3 = \frac{4 \times 3 \times 2 \times 1}{(3 \times 2 \times 1) \times (1)} = \frac{24}{6} = 4,$$

which again coincides with the experimental result.   In passing it may be noted that if r be put equal to n we have

$$_nC_n = \frac{\lfloor n}{\lfloor n} = 1$$

which again is reasonable, since obviously only one combination of n things taken all together can possibly be made.

For the sum of combinations, that is, the total combinations of n things taken singly, by twos, etc., we have

$$\Sigma \ _nC_r = n + \frac{n.(n-1)}{1.2} + \frac{n.(n-1)(n-2)}{1.2.3} + \ldots + n + 1$$

But the right-hand side of the equation, as will appear from the discussion of the binomial theorem in a later section, is

$$(1+1)^n - 1.$$

Hence

$$\Sigma \ _nC_r = 2^n - 1.$$

Again, for the sake of convenience, a brief table of combinations is inserted as Table 45.

TABLE 45

VALUES OF COMBINATIONS

Combinations of

n / r	10	9	8	7	6	5	4	3	2	1
1......	10	9	8	7	6	5	4	3	2	1
2......	45	36	28	21	15	10	6	3	1	
3......	120	84	56	35	20	10	4	1		
4......	210	126	70	35	15	5	1			
5......	252	126	56	21	6	1				
6......	210	84	28	7	1					
7......	120	36	8	1						
8......	45	9	1							
9......	10	1								
10......	1									
$\Sigma \ _nC_r$	1023	511	255	127	63	31	15	7	3	1

At a time.

It will be noted from this table that in each column the values rise to a maximum and then decline.

$$\text{The maximum } {}_nC_{\frac{1}{2}} = \frac{\lfloor\underline{n}}{\left(\lfloor\underline{\tfrac{n}{2}}\right)^2} \quad \text{when } n \text{ is even.}$$

$$\text{The maximum } {}_nC_r = \frac{\lfloor\underline{n}}{\lfloor\underline{\tfrac{n+1}{2}}\ \lfloor\underline{\tfrac{n-1}{2}}} \quad \text{when } n \text{ is odd.}$$

Approximations to $\lfloor\underline{n}$

In all practical work with probability it is useful to have an easily computed approximation to the value of $\lfloor\underline{n}$ in cases when n is large. In Pearson's "Tables for Statisticians and Biometricians" and also in Glover's "Tables of Applied Mathematics" a table of log $\lfloor\underline{n}$ for $n = 1$ to 1000 is given. But for still higher values an approximation is needed. A number of such formulæ are available.

Stirling's:

$$\lfloor\underline{n} = \sqrt{2\pi n} \times n^n\, e^{-n} \times \left\{ 1 + \frac{1}{12n} + \frac{1}{288n^2} + \cdots \right\}$$

Forsyth's:

$$\lfloor\underline{n} = \sqrt{2\pi}\ \left\{ \frac{\sqrt{n^2 + n + \frac{1}{6}}}{e} \right\}^{n+\frac{1}{2}}$$

This is accurate to $\dfrac{1}{240n^3}$.

To indicate the closeness of such approximations we may calculate $\lfloor\underline{n}$ for $n = 2$.
The result is

$$\lfloor\underline{n} = 1.999479 \text{ (Forsyth)}$$
$$\text{Actual error} = .000521$$
$$\frac{1}{240n^3} = .00052083$$

Hence it may be concluded that for all practical purposes Forsyth's approximation is sufficiently accurate. It is, in the opinion probably of most computers, somewhat easier and quicker of calculation than Stirling's approximation.

THE PROBABILITY OF CONCURRENT EVENTS

Suppose this question is put: If two pennies are tossed at random together, what is the probability that both will show heads when they come to rest?

What are the possibilities? Let us call one of the pennies A to distinguish it from the other B. Then we have, as possibilities,

$$AH, BH$$
$$AH, BT$$
$$AT, BH$$
$$AT, BT.$$

From this it appears that the favorable event AH, BH, can occur in but one way, out of a total of four ways in which any event may happen under the specifications (namely, of two pennies tossed together). Hence

$$p = \tfrac{1}{4}$$

The probability that the pennies will fall one head and one tail is evidently,

$$p = \tfrac{2}{4}.$$

Now let us consider these results analytically. Any one throw of the two pennies must necessarily result in a *combination* of the character A −, B −, where the dashes may be either H or T. But considering the A penny *alone*, the probability that it will be AH after any particular toss is, as we have already seen, $\tfrac{1}{2}$. This means that in *n* successive tosses of the A penny alone it will come AH approximately one-half of the times and AT one-half of the times. This fact will not be altered by virtue of the fact that B is tossed with A, because if the tossing is random neither penny affects the other. Consequently it must happen that in about one-half of *n* tosses of the two together the constitution of the result must be of the form AH, B −, or numerically the result will be $\tfrac{1}{2}$ *n* AH, B −. But now the B penny, which is associated with each of these AH pennies in the $\tfrac{1}{2}$ *n* throws, will be subject to the same influences as though it were tossed alone. Consequently we shall have in these $\tfrac{1}{2}$ *n* tosses these results:

$$\tfrac{1}{2} \, (\tfrac{1}{2} \, n) \text{ AH, BH, and}$$
$$\tfrac{1}{2} \, (\tfrac{1}{2} \, n) \text{ AH, BT.}$$
$$\text{But } \tfrac{1}{2} \, (\tfrac{1}{2} \, n) \text{ AH, BH} = \tfrac{1}{4} \, n \text{ AH, BH.}$$

Continuing, let us consider next the one-half of the n tosses in which the A penny falls T.   By the same reasoning as before, we shall get

$$\tfrac{1}{2} (\tfrac{1}{2} n) \text{ AT, BH, and}$$
$$\tfrac{1}{2} (\tfrac{1}{2} n) \text{ AT, BT.}$$

But the $\tfrac{1}{4} n$ AH, BT, and $\tfrac{1}{4} n$ AT, BH clearly must be added together, since they are the cases in which head and tail occur together, and it makes no difference which penny is head or which tail, so that we have for the probability of the two pennies falling one head and one tail,

$$\tfrac{1}{2} n \begin{cases} \text{AH, BT or} \\ \text{AT, BH.} \end{cases}$$

So, then, the complete result is,

$$\tfrac{1}{4} n \text{ AH, BH} = 2 \text{ heads,}$$
$$\tfrac{1}{2} n \begin{cases} \text{AH, BT} \\ \text{or} \\ \text{AT, BH} \end{cases} = 1 \text{ head, 1 tail,}$$
$$\tfrac{1}{4} n \text{ AT, BT} = 2 \text{ tails.}$$

Whence we arrive at the rule:

If the separate probabilities of each of several independent events are respectively $p_1, p_2, p_3 \ldots \ldots$, the probability of their all occurring together is

$$P = p_1 \times p_2 \times p_3 \ldots \ldots$$

The concurrence of events implied in this rule and the discussion which has led up to it may be either in time, or in space but not in time, or in both space and time.   Thus in the case of tossing two pennies together, the probability of $\tfrac{1}{4}$ that they will fall **HH** would plainly not be affected in any way if one of the pennies were tossed say a fraction of a second later than the other, nor, indeed, if it were tossed several seconds, or minutes, or days, or any other time unit, later, provided, as always, that all the tossing was random in character.   Hence it is seen that the probability of **HH** with two pennies is the same, $\tfrac{1}{4}$, whether they are tossed together or successively.

The simple theorems in probability so far developed have many practical applications in medical work.   An example from actual experience may be given in illustration:

A physician has seen in the whole of his lifetime's practice 23,464 patients. Of these patients, 1474 had some disease of the gall-bladder or ducts. Also of the same 23,464 patients 454 had glycosuria from some cause or other. Of the 454 patients exhibiting glycosuria, 372 were cases of diabetes mellitus. Now in the whole experience 24 patients exhibited *both* disease of the gall-bladder and glycosuria, and 13 had *both* gall-bladder disease and diabetes mellitus.

The question now is: Were gall-bladder disease *and* glycosuria more or less often associated together in this series than would be expected if chance or random association were the only influence bringing them together?

In the experience of this physician the probability that a patient had disease of the gall-bladder and ducts was

$$p_1 = \frac{1474}{23,464} = .0628$$

The probability that a patient had glycosuria was

$$p_2 = \frac{454}{23,464} = .0193$$

The probability of a patient having both gall-bladder disease and glycosuria was

$$P = p_1 \times p_2 = .0628 \times .0193 = .001212$$

There would then be expected, from random assortment of diseases alone, in this series a total of

$$23,464 \times .001212 = 28.4$$

patients showing both these morbid conditions. Actually there were 24 such patients. Whence we may at once conclude that the association of the gall-bladder disease and glycosuria observed in this series of 23,464 patients was approximately what might have been expected from the operation of chance alone.

The case for diabetes mellitus and gall-bladder disease is somewhat different.

Here

$$p_1 = .0628 \text{ as before}$$
$$p'_2 = \frac{372}{23,464} = 0159$$
$$P = p_1 \times p'_2 = .000999$$

and the number of cases expected is

$$23,464 \times 000999 = 23\ 4$$

while actually only 13 occurred with the combination. Hence it may be concluded that in this series diabetes mellitus and diseases of the gall-bladder and ducts actually occurred together in the same patients only slightly more than half as often as they would be expected to from chance alone.

THE POINT BINOMIAL

Let us now consider what will happen in n trials regarding an event for which the probability of occurrence is p, and the probability of failure is $q = 1 - p$.

1. The probability that the event will occur at every trial is evidently

$$p \times p \times p \times p \ldots = p^n$$

Thus if we toss together at random four pennies the probability that they will fall all heads, **HHHH**, is

$$\tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = (\tfrac{1}{2})^4 = \tfrac{1}{16}$$

2. The probability that in any one throw $n - 1$ particular pennies will give successes (say heads) and one particular penny a failure (tail) is

$$p \times p \times p \times \ldots \ldots \times q = p^{n-1} \cdot q$$

But this result can occur n different ways, as is plain from the four pennies, which may give three heads and one tail, as follows:

$$
\begin{array}{cccc}
H & H & H & T \\
H & H & T & H \\
H & T & H & H \\
T & H & H & H \\
\end{array}
$$

Hence the complete probability that the event will occur $n - 1$ times and fail once is

$$n \, p^{n-1}.q$$

or in the penny case

$$4 \, (\tfrac{1}{2})^3 \, (\tfrac{1}{2}) = \tfrac{4}{16}$$

3. The probability that in any one throw $n - 2$ particular pennies will give successes and 2 particular pennies failures is

$$p \times p \times \,\dots\dots\, \times q \times q = p^{n-2}.q^2$$

But again this may happen in

$$\frac{n \, (n-1)}{1.2} = {}_nC_r$$ (remembering that in the formula given above for ${}_nC_r$ some factors cancel in numerator and denominator).

different ways, as can be seen from the example of tossing four pennies, where the combination of two heads and two tails may occur as follows:

$$\begin{matrix} \text{H H T T} \\ \text{H T H T} \\ \text{H T T H} \\ \text{T H H T} \\ \text{T H T H} \\ \text{T T H H} \end{matrix}$$

Hence the complete probability of the event occurring $n - 2$ times and failing twice is

$$\frac{n \, (n-1)}{1.2} \, p^{n-2}.q^2,$$

which in the penny example is

$$\frac{4.3}{2} \, (\tfrac{1}{2})^2 \, (\tfrac{1}{2})^2 = \tfrac{6}{16}$$

4. And so the same process may be continued. But enough detail has been presented to make it evident that:

If n trials be made of an event for which the probability of occurrence is p and the probability of failure is q, the probability of each of the several possible occurrences is given by the appropriate term in the expansion of the binomial

$$(p + q)^n.$$

5. If $p = q = \frac{1}{2}$, as in the case of the penny, the point binomial will be symmetric, as shown in Fig. 72, which gives the results for the four-penny example.

But within fairly wide limits p and q may have any values. Thus consider the results of throwing four dice together. In the
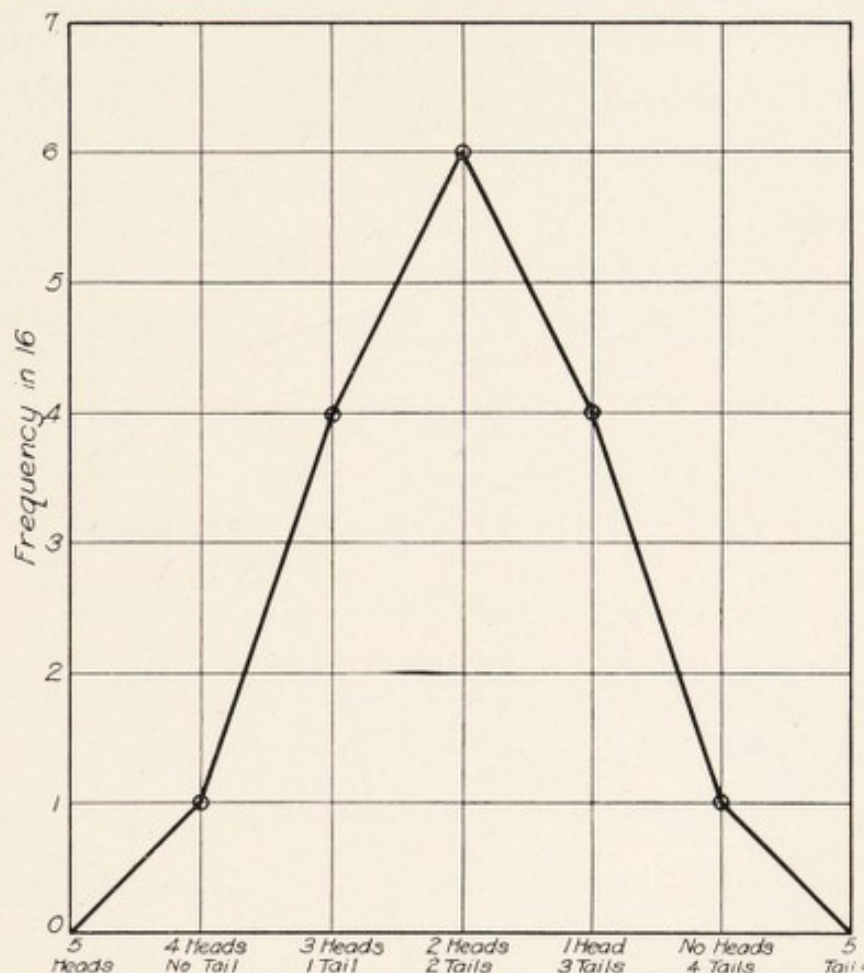
Fig. 72.—The results of tossing four pennies together at random, as given by the binomial $(\frac{1}{2} + \frac{1}{2})^4$.

case of dice the probability of any particular face of the die coming up after one random throw of one die is

$$p = \tfrac{1}{6}$$

whence

$q = \tfrac{5}{6}$ = the probability that this particular face will not come up.

Hence for the probabilities of getting different numbers of 6's with 4 dice thrown together at random we require the successive terms of

$$(\tfrac{1}{6} + \tfrac{5}{6})^4$$

20

These are:

$$p^n = \frac{1}{1296} = \text{probability that all 4 dice will fall with the 6 face up.}$$

$$np^{n-1}q = \frac{20}{1296} = \text{probability that 3 dice will fall 6's and 1 die something other than 6}$$

$$\frac{n\,(n-1)}{1.2}\,p^{n-2}q^2 = \frac{150}{1296} = \begin{array}{l}\text{probability that 2 dice will fall 6's, and the other 2 some-}\\ \text{thing other than 6.}\end{array}$$

$$\frac{n\,(n-1)\,(n-2)}{1.2.3}\,p^{n-3}q^3 = \frac{500}{1296} = \begin{array}{l}\text{probability that 1 die will fall 6 and the other}\\ \text{three something else.}\end{array}$$

$$\frac{n\,(n-1)\,(n-2)\,(n-3)}{1.2.3.4}\,p^{n-4}q^4 = \frac{625}{1296} = \text{probability that no die will fall 6.}$$
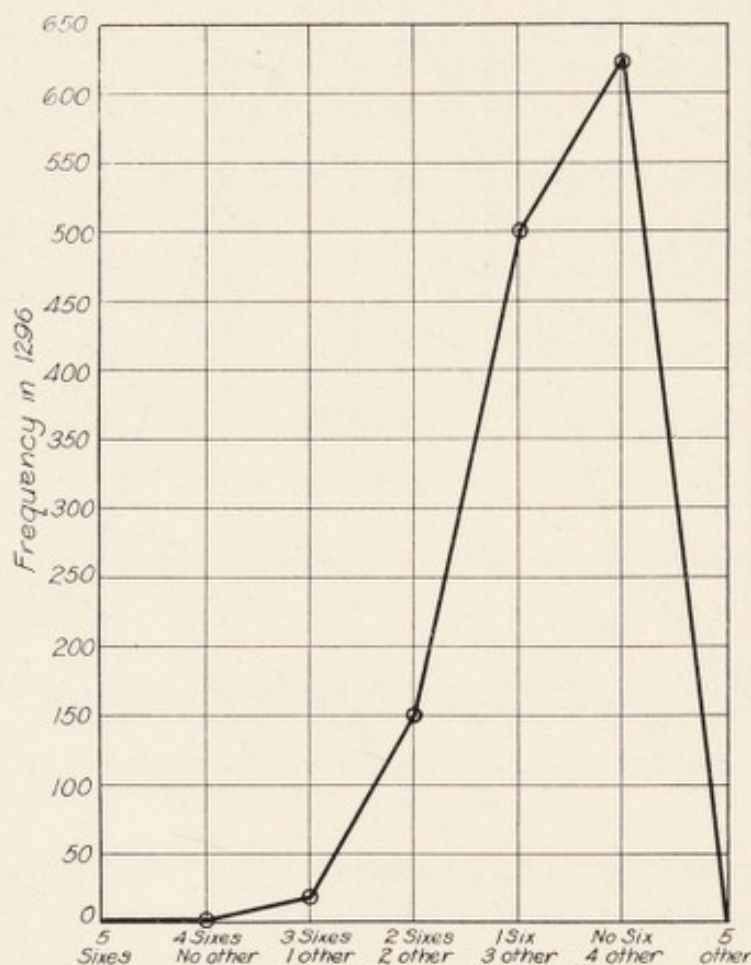
Fig. 73.—The probability of getting different numbers of 6's in the throws of 4 dice together, as given by $(\frac{1}{6} + \frac{5}{6})^4$.

This distribution is shown graphically in Fig. 73, and its asymmetry or skewness is apparent.

The student must bear always in mind in connection with the graphical representations of the point binomial in this section and

elsewhere, that the terms of the binomial are *true ordinates*, and not frequency areas. Consequently the lines connecting the circles to form a polygon are not a correct representation of actuality. Theoretically the circles in such a diagram as Fig. 73 stand alone by themselves. The lines are put in simply as a convenience, to enable the eye to get the sweep of the ordinates as a whole.

6. The probability of an event occurring *t or more times* in n trials is the sum of the terms of $(p + q)^n$ from p^n up to the term in $p^t . q^{n-t}$.

The consequences and usefulness of this proposition are far reaching and will bear careful examination.

Let us start with an example. Suppose ten pennies to be tossed together at random. For the results we have

$$(\tfrac{1}{2} + \tfrac{1}{2})^{10} = \frac{1 + 10 + 45 + 120 + 210 + 252 + 210 + 120 + 45 + 10 + 1}{1024}$$

These fractions are reduced to decimals in Table 46.

TABLE 46

SUCCESSIVE TERMS OF $(\tfrac{1}{2} + \tfrac{1}{2})^{10}$

Ordinal number of term.	Value of term.	Term measures the probability that there will be, in any one throw
1.	.000977	10 heads, 0 tail
2.	.009766	9 heads, 1 tail
3.	.043945	8 heads, 2 tails
4.	.117187	7 heads, 3 tails
5.	.205078	6 heads, 4 tails
6.	.246094	5 heads, 5 tails
7.	.205078	4 heads, 6 tails
8.	.117187	3 heads, 7 tails
9.	.043945	2 heads, 8 tails
10.	.009766	1 head, 9 tails
11.	.000977	0 head, 10 tails
Total.	1.000000	

There is then about one chance in a thousand that on any one throw the 10 pennies will all fall head. There is approximately one chance in four that there will be 5 heads and 5 tails on any one throw, and so on.

The ordinates of Table 46 are plotted in Fig. 74.

What, now, is the probability that on any one throw there will fall *six or more heads?* By the rule given above, and obviously from general principles discussed earlier, this probability is:

The probability for 6 heads　　　　　　　　= 　.205078
+ The probability for 7 heads　　　　　　　= +.117187
+ The probability for 8 heads　　　　　　　= +.043945
+ The probability for 9 heads　　　　　　　= +.009766
+ The probability for 10 heads　　　　　　　= +.000977
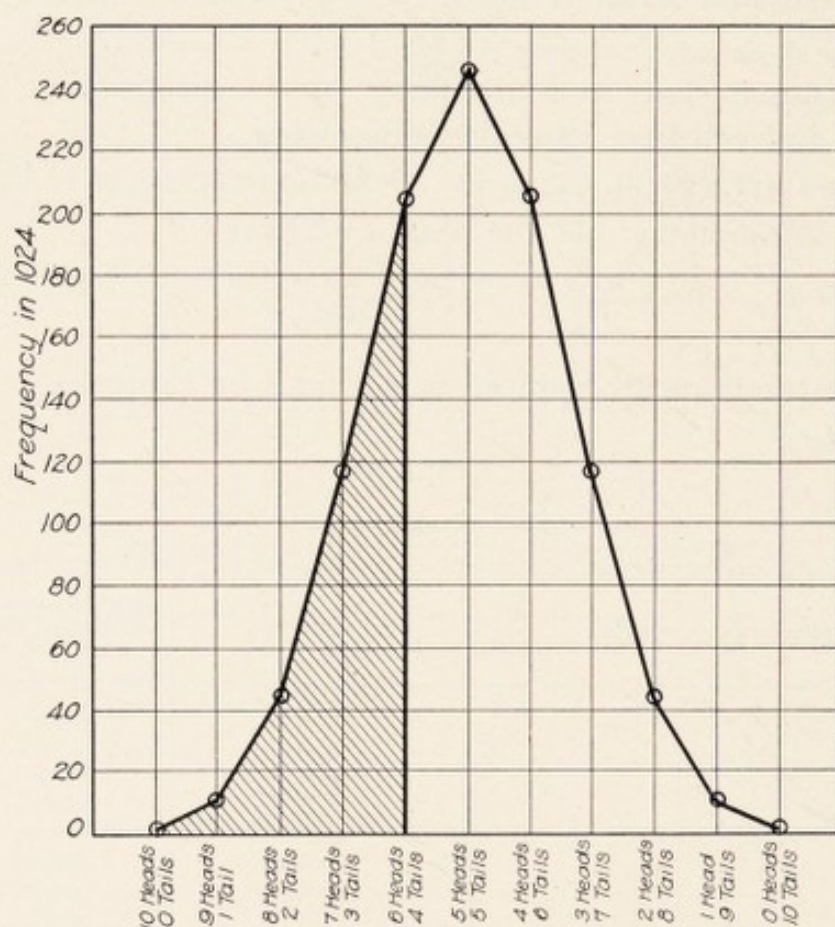Complete probability of 6 or more heads on one throw　=　.376953

Fig. 74.—The binomial $(\frac{1}{2} + \frac{1}{2})^{10}$. The meaning of the cross-hatched area is explained in the text.

Or, it appears that there are approximately thirty-eight chances in one hundred, or a little more than one in three of throwing 6 or more heads at one toss of the 10 pennies. In the diagram the cross-hatched portion shows the ordinates summed. The ratio of the area of the cross-hatched portion to the total area is, for reasons which will appear in the next section, approximately that of the total probability of .38 given above.

7. In all of the discussion of the point binomial so far nothing has been said specifically about abscissas. The discussion has been wholly about ordinates, and in the tables and diagrams we have simply named in words the situation relative to the pennies at each point at which an ordinate was erected. But this is plainly not a neat or complete procedure. It is time now to see if something different cannot be done relative to abscissas.

Consider the symmetric binomial, where $p = q = \frac{1}{2}$. The structure resulting from its expansion is a series of points, which if connected by lines as we have done, form a polygon, shaped like a Napoleonic cocked hat, the line rising from each end to a peak in the middle. Now suppose instead of designating each abscissal point at which an ordinate is erected by a descriptive term, such as "6 heads, 4 tails," we measure the distance of each such point from the center of the polygon where the highest ordinate is (or when n is odd, from a point half-way between the two equal central ordinates), using as the yardstick for the measurement some function of the shape of the curve, or of the spread of its two limbs. Every one is bound to agree that such a procedure would be fair enough, provided the yardstick were at hand.

Now several such yardsticks are available, and have, indeed, been used at different times in the history of the subject. The one which has at the present time come to be almost universally used, because of its significance in the higher mathematical development of the subject, is

$$\sigma = \sqrt{n\,p\,q}$$

This quantity, which is perceived to be easily calculated, and which for the present we shall call simply by its symbol *sigma*, will be more fully discussed in a later chapter, and its mechanical and geometric meaning explained. Here it need only be pointed out that every point on the abscissal axis can be numerically defined as some multiple of σ since it itself is a distance along that axis.

So then we may set up Table 46 in another form, as shown in Table 47 on page 310.

Normally, of course, one would never carry so many places of decimals in x/σ. But this example will indicate that the position

TABLE 47

ABSCISSÆ IN TERMS OF x/σ, AND ORDINATES OF $(\frac{1}{2} + \frac{1}{2})^{10}$

x/σ	y
− 3.162278	.000977
− 2.529822	.009766
− 1.897366	.043945
− 1.264911	.117187
− .632456	.205078
0	.246094
+ .632456	.205078
+ 1.264911	.117187
+ 1.897366	.043945
+ 2.529822	.009766
+ 3.162278	.000977

of any abscissal point can be expressed in terms of σ with any desired degree of accuracy.

THE NORMAL CURVE

It has been pointed out that to get the probability that an event will occur *t or more* times in *n* trials it is necessary only to sum the terms of the binomial up to the one in $p^t.q^{n-t}$. This is a simple enough matter when *n* is small or, at any rate, not very large. But how if one is confronted with this problem? Suppose a city to have 10,000 births per annum, and further suppose that long experience of that city has demonstrated, on the average, that the probability of any given birth being of a male is $p = .52$. What is the probability that in a given year, say next year, there will be born 5300 *or more* male babies? To answer this by the point binomial route requires the calculation and summing of the successive terms in the binomial $(.52 + .48)^{10,000}$ from the end of the curve to the term in which p has the exponent 5300. Plainly the labor involved in this procedure would far outweigh any possible significance which could attach to the result.

Let us examine what happens as the exponent *n* of the binomial increases in value. Figure 75 shows this graphically for a small range of values of *n*, but a sufficient number to bring out the point. In plotting this diagram all the deviations are taken in the form $4(x/\sigma)$, and the sums of the ordinates of all the polygons are made the same.

Now what this diagram shows is that, as *n* increases, the polygon
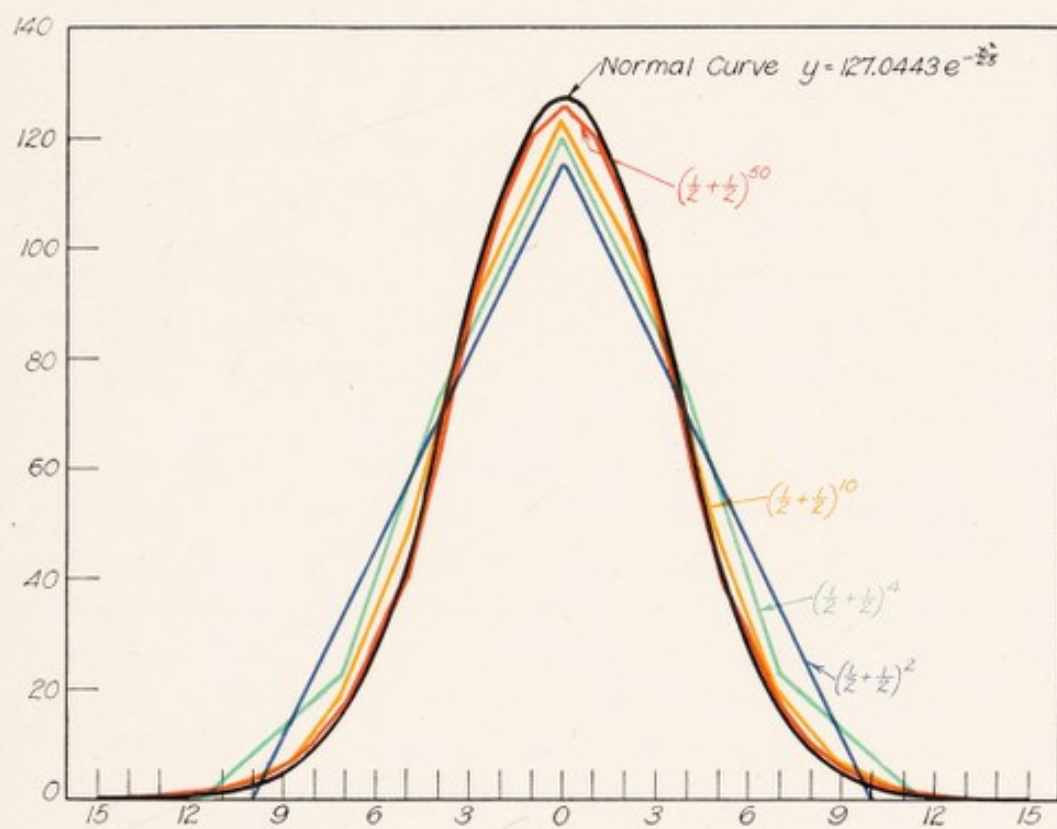
Fig. 75.—Point binomials for several values of *n*, and a superimposed normal curve.

got by connecting the tops of the ordinates of the binomial increases its number of sides as would be expected. Furthermore, the binomial approaches in its form closer and closer to the smooth curve as n increases. Now suppose n to increase indefinitely in value. The resulting polygon would come closer and closer to the smooth curve, but would never quite reach it because, after all, however large n might be, if it were still finite, the resulting figure would still be a polygon, that is, made up of many short but still straight sides, whereas the curve is everywhere curving.

But suppose we went on to the binomial

$$(\tfrac{1}{2} + \tfrac{1}{2})^{\infty}$$

Then each side of the "polygon" would be infinitely short, corresponding to a point in a smooth curve, and each such point may be thought of as a straight line of infinite shortness. Furthermore, each ordinate of this "polygon" would be infinitely close to the next one. This "polygon" would then have come to coincide exactly with the smooth curve, and, in short, have become identical with it.

In other words, the smooth curve is what is known mathematically as the *limit* of the point binomial, as n of the binomial increases. But this result opens out wonderful possibilities. For, plainly, if the equation to the smooth curve is known it can be integrated over any portion of its range. These integrations may be performed once and for all, for this curve reduced to standard area of say 1, and tabled. Then, *in so far as the curve is a good approximation to the binomial*, these integrations can be used in place of the tedious finite summation of the terms of the binomial, and the derived probabilities read off from the table of these integrations, without any work at all. Now it is apparent from Fig. 75 that with n no larger than 50 the smooth curve is a quite sufficiently close approximation to the binomial for all practical statistical purposes, and we shall be quite justified in so using it in practical work.

All this has been done. The integrals of the smooth curve, which has the equation

$$y = \frac{n}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

have been calculated and tabled. Such tables are known as tables of the probability integral. A short table of this kind, but quite extensive enough for most practical statistical work, is given in Appendix IV of this book. It carries the argument—the deviation from the center measured on the x/σ yardstick—to two places of figures, and the function to four places. Besides the area the individual ordinate corresponding to the same argument is given in each case.

The curve itself is known as the *normal curve*, or from its discoverers, the De Moivre-Gauss-Laplace curve of error. It has many and varied properties and uses in statistics, space for the discussion of which is lacking in this book. It may truly be said to be the very corner-stone of the foundation of the statistical treat-
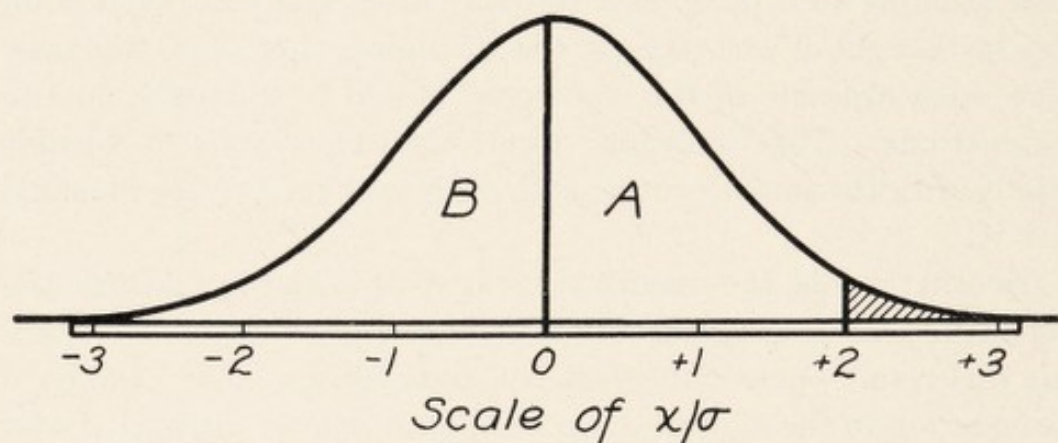
Fig. 76.—Diagram of probability example given in the text.

ment of observational data, whether quantitative or qualitative in character.

As an example of the use of the probability integral to replace finite summation of the terms of the point binomial we may take the case propounded above regarding the sex ratio of births.

Here

$$n = 10,000, p = .52, q = .48$$

Hence

$$\sigma = \sqrt{n\,p\,q} = \sqrt{10,000 \times .52 \times .48} = 49.96$$
$$x = 5300 - 5200 = 100$$
$$x/\sigma = \frac{100}{49.96} = 2.00$$

Thus we have the situation depicted graphically in Fig. 76.

Now in the table in Appendix IV there is found against the argument 2.00 the figure .4772. This means that, taking the total area of the curve as 1, the area of that part of the curve (A) between the mid-ordinate ($x/\sigma = 0$) and the ordinate where $x/\sigma = 2$, is .4772. Therefore the fraction of the area of the whole curve up to the ordinate where $x/\sigma = 2$ will be B + A = .5 + .4772 = .9772. Hence the area of the *rest of the curve*, which measures the probability of positive deviations of 2 x/σ *and greater*, will be 1 − .9772 = .0228. Or we say that the chances are about $2\frac{1}{4}$ in a hundred that in any given year in our hypothetic city there will be 5300 or more male babies born. Or, put in another way, we should not expect, on the premises stated in the example, 5300 male births in a year to be equalled or exceeded oftener than between two or three times in a century.

THE RELATION BETWEEN σ AND THE PROBABLE ERROR

We have used in the discussion in this chapter σ as the yardstick to measure deviations. In an earlier chapter the probable error has been used for the same purpose, though that phase of the matter was not then emphasized. What is the relation between the two? It is a simple one, that given by the following equation:

$$\text{P. E.} = .6744898 \ldots \sigma.$$

SUGGESTED READING

1. Peirce, C. S.: A Theory of Probable Inference. *In* Studies in Logic by Members of the Johns Hopkins University, Boston, 1883, pp. 126–181.
    (This is a classic. No student of probability or statistics can be properly said to have laid his basic foundations until he has mastered this essay.)
2. Venn, J.: The Logic of Chance, 3d ed., London, 1888 (Macmillan).
    (Suffers from a curiously diffuse and wandering style, but sound as to doctrine.)
3. Laplace: A Philosophical Essay on Probabilities, New York (Wiley), 1902. (Translated by Truscott and Emory.)
4. Edgeworth, F. Y.: Methods of Statistics, Jour. Roy. Stat. Soc., Jubilee vol. 1885, pp. 181–217.
5. Yule, G. U.: An Introduction to the Theory of Statistics, 6th ed., Chapter XV.
6. Galton, F.: Natural Inheritance, London, 1889 (Macmillan), Chapters IV and V.
    (These two chapters contain perhaps the clearest and simplest account of the structure and significance of the normal curve anywhere to be found.)

7. Fisher, A.: The Mathematical Theory of Probabilities, vol. i, 2d ed., New York
    (Macmillan), 1922, Chapters I–VI, inclusive.

        (This book brings a fresh and original viewpoint and modes of expression
    to the old problems.   It will probably be found rather difficult by most medical
    readers.)

8. Coolidge, J. L.: An Introduction to Mathematical Probability, Oxford (Clarendon
    Press), 1925, pp. xii + 216.   (This is one of the best of the modern text-books
    on the theory of probability.)

9. Pearson, K.: Historical Note on the Origin of the Normal Curve of Errors, Bio-
    metrika, vol. 16, pp. 402–404, 1924.

CHAPTER XII

SOME SPECIAL THEOREMS IN PROBABILITY

In this chapter will be discussed some special developments and applications of the theory of probability likely to be of particular use to the medical worker.

THE CHI-SQUARE TEST

A. The Goodness of Fit of Theory to Observation

In 1900 Professor Karl Pearson published a paper[1] which opened with the following sentence in italics: "*The subject of this paper is to investigate a criterion of the probability on any theory of an observed system of errors, and to apply it to the determination of goodness of fit in the case of frequency curves.*" This paper has now become a classic, and from it, and later papers elaborating and extending the theory which it embodies, have come some of the most important developments in modern statistical work.

Pearson showed that the probability sought in his opening sentence is given by the expression

$$P = \frac{\int_{X}^{\infty} e^{-\frac{1}{2}\chi^2} \chi^{n-1} \, d\chi}{\int_{0}^{\infty} e^{-\frac{1}{2}\chi^2} \chi^{n-1} \, d\chi}$$

The expression $\chi^2 = $ constant is the equation of a generalized "ellipsoid," all over the surface of which the frequency of the system of errors or deviations is constant.

Tables showing the value of P for different values of χ^2 are now available in Pearson's "Tables for Statisticians and Biometricians." The consequence is that the application of the chi-square test of goodness of fit is an extremely simple matter.

Translating the matter abruptly from abstract mathematical notations to concrete statistical data, the value of χ^2, for a system

of observed and theoretical frequencies, is the sum of the ratios of the squared differences between theoretical and observed frequencies, to the theoretical frequencies, for all such pairs of observed and theoretical frequencies.

An example will make the meaning clear. Table 48 gives the observed frequencies (m'_r) of the weight of the brain in 416 adult Swedish males,* and the theoretical frequencies (m_r) given by the normal curve,

$$y = 78.0401\, e - .01106\, x^2$$

fitted to the observed frequencies. These theoretical frequencies in Table 48 are *areas* of the normal curve standing over the abscissal intervals indicated in the first column.

TABLE 48

OBSERVED AND THEORETICAL FREQUENCIES OF SWEDISH MALE BRAIN WEIGHT, TO SHOW THE METHOD OF CALCULATING χ^2

Grams of brain weight.	Observed (m'_r)	Calculated (m_r)	$\dfrac{(m_r - m'_r)^2}{m_r}$
Under 1100	0	.981	.981
1100–1149	1	2.9	1.24
1150–1199	10	8.5	.26
1200–1249	21	20.3	.02
1250–1299	44	39.0	.64
1300–1349	53	60.4	.91
1350–1399	86	75.2	1.55
1400–1449	72	75.3	.14
1450–1499	60	60.8	.01
1500–1549	28	39.4	3.29
1550–1599	25	20.6	.94
1600–1649	12	8.7	1.25
1650–1699	3	2.9	.003
1700–1749	1	.8	.05
1750 and over	0	.036	.036
Totals	416	415.817	11.320 $= \chi^2$

The number of frequency groups here is 15,† and χ^2 (the sum of the quantities in the last column of the table) = 11.32. From

* From p. 41 of Pearl, R.: Biometrical Studies on Man. I. Variation and Correlation in Brain-weight, Biometrika, vol. 4, pp. 13–104, 1905.

† This is plainly arbitrary since the comparison could be set up with more or fewer classes (as, for example, by lumping the tail classes at each end). In cases where the tail classes are found to contribute heavily to the value of χ^2 they should be so lumped together.

Pearson's Tables (p. 27), with $n' = 15$ and $\chi^2 = 11$, we read $P = .686036$. For $n' = 15$ and $\chi^2 = 12$ we read $P = .606303$. This result means that if the brain weight of Swedish males in general varied absolutely and strictly in accordance with the normal law of error, it would be expected that in about 65 out of every 100 trials of the matter, each trial being made with a random sample of 416 individuals, the divergence between observation and theory would be greater than it is in this case. Therefore, it may be concluded that the normal curve gives an excellent fit to these observations.

In the use of the chi-square test of the goodness of fit of theory to observation the following points must always be kept in mind:

1. That the test in this form is valid only for *frequencies*, not for ratios, rates, or time ordinates.

2. That the theoretical frequencies must be *areas* above the abscissal class ranges, and not mid-ordinates.

3. That if the frequencies are very small and scattering toward the tails of the curve, as is often the case, a more reliable estimation of P will be obtained if the tail frequencies are lumped together in two single classes, one at each tail end of the curve.

THE FOUR-FOLD TABLE

One of the most common applications of the χ^2 test arises in cases where we have knowledge of the frequencies of each of the four possible combinations of two attributes in respect of presence or absence. Such knowledge is conveniently presented in a table of the following type:

1st Attribute

2nd Attribute		+	−	Totals
	+	a	b	$(a + b)$
	−	c	d	$(c + d)$
	Totals	$(a + c)$	$(b + d)$	$(a + b + c + d) = N$

In this table a, b, c, and d are the frequencies with which two attributes have been observed according to the indicated presence $(+)$ or absence $(-)$ of each.

The question to be answered by such a table is as to whether or not there is any significant association between the two attributes. This question may be answered by considering the divergence of the observed table from a theoretical table constructed on the assumption that the two attributes are completely independent. If the attributes *are* completely independent, the *theoretical* frequency to be expected in the upper left hand corner of the table (for which the observed frequency is a) would obviously be $\frac{(a+b)}{N} \cdot \frac{(a+c)}{N} \cdot N$, in accordance with the theorem regarding the probability of concurrent events, set forth in Chapter XI, *supra*, because the probability of the presence of the first attribute alone is $\frac{(a+c)}{N}$; the probability of the presence of the second attribute alone is $\frac{(a+b)}{N}$; and the probability of both the first and the second attribute being present together will be $\frac{(a+c)}{N} \cdot \frac{(a+b)}{N}$, and in turn the expected number of such combined occurrences will be this last probability multiplied by N. Similar expressions can be written for each of the other cells. Thus the expression for the theoretical expected frequency for the case of presence of the first attribute and absence of the second (for which the *observed* frequency is c) will be $\frac{(a+c)}{N} \cdot \frac{(c+d)}{N} \cdot N$.

The χ^2 test of the preceding section can therefore be applied to determine whether or not there is any significant association between the two attributes under consideration. The only additional point to be noted is that since the theoretical frequencies are determined from the marginal totals of the observed frequencies, there is but one independent comparison to be made, although four cells are under consideration. The fact that there is but one degree of freedom may be shown by considering the difference between theoretical and observed for each cell in turn, when it will be found that all four differences are the same. For example:

$$\frac{(a+b)}{N} \cdot \frac{(a+c)}{N} \cdot N - a = \frac{a^2 + ab + ac + bc - a^2 - ab - ac - ad}{a+b+c+d}$$

$$= \frac{bc - ad}{a+b+c+d}.$$

and

$$\frac{(a+b)}{N} \cdot \frac{(b+d)}{N} \cdot N - b = \frac{ab + b^2 + ad + bd - ab - b^2 - bc - bd}{a+b+c+d}$$

$$= \frac{ad - bc}{a+b+c+d}.$$

These two differences are alike except for sign and it will be found on trial that the other two differences are arithmetically equal to $\frac{ad - bc}{a+b+c+d}$.

So then we shall have the following theoretical four-fold table for the assumed case of complete independence of the two attributes:

1st Attribute

	+	−	Totals
+	A	B	$(A+B) = (a+b)$
−	C	D	$(C+D) = (c+d)$
Totals	$(A+C) = (a+c)$	$(B+D) = (b+d)$	N

2nd Attribute

Here the capital letters indicate the theoretical frequencies and the small letters the observed frequencies as before.

Then, in accordance with the preceding section,

$$\chi^2 = \frac{(A-a)^2}{A} + \frac{(B-b)^2}{B} + \frac{(C-c)^2}{C} + \frac{(D-d)^2}{D}.$$

In making a comparison between two or more series of observations the number of frequency classes involved in the comparison must be taken into consideration. R. A. Fisher has shown that

$$n = (r-1)(c-1),$$

where n is the number of "degrees of freedom" in making such comparisons, r is the number of rows, and c the number of columns in the table.

In Pearson's Tables (pp. 26–28) the tabular argument is in terms of $n' = (r-1)(c-1) + 1$.

Thus we need a table to show the values of P associated with χ^2 values when we have but one degree of freedom, that is when n'

of the notation of the previous section $= 2$ (that is, $(r - 1)(c - 1)$ $+ 1 = 2$). But since, as is shown later, $x^2 = \left(\dfrac{x}{\sigma}\right)^2$ Table B of Appendix III may be used. It will only be necessary to take the square root of the observed x^2 and enter Table B with that value as argument.

As an illustration of the application of x^2 to a four-fold table, consider the following table* which shows the individuals under consideration, tested as to, first, whether or not they had enlarged spleens, and, second, as to whether or not their blood films showed presence of malaria parasites.

Enlarged Spleen

		+	−	Totals
Parasites	+	740	743	1483
	−	1287	2731	4018
	Totals	2027	3474	5501

We may compute χ^2 directly by inserting in each cell the theoretical frequency, and then using the formula given above. Thus we have

Enlarged Spleen

		+	−	Total
Parasites	+	*546.45* 740	*936.55* 743	1483
	−	*1480.55* 1287	*2537.45* 2731	4018
	Total	2027	3474	5501

(Theoretical frequencies in *italics*.)

$$x^2 = \frac{(546.5 - 740)^2}{546.5} + \frac{(936.5 - 743)^2}{936.5} + \frac{(1480.5 - 1287)^2}{1480.5} + \frac{(2537.5 - 2731)^2}{2537.5}$$

$$= 148.6$$

* Data taken from paper by H. C. Clark, M. D., entitled "A Comparison of the Spleen and Parasite Rates as Measure of Malaria Incidence in the Races of the Mainland of Central America." Seventeenth Annual Report of Medical Department of United Fruit Co., 1928.

If we take the algebraic form of this four-fold table, and operate with it as with the table above, we may arrange χ^2 in this form

$$\chi^2 = \frac{(ad - bc)^2 \, N}{(a + b) \, (c + d) \, (a + c) \, (b + d)}.$$

For the above example this gives

$$\chi^2 = \frac{[(740 \times 2731) - (743 \times 1287)]^2 \, 5501}{1483 \times 4018 \times 2027 \times 3474}$$

$$= \frac{(1064699)^2 \, 5501}{(5958694) \, (7041798)} = \frac{623585 \times 10^{10}}{419599 \, (10)^8}$$

$$= 148.6,$$

which is the same as the previous result.

If we look up in Table B of Appendix III in which $n = 1$, P for $\chi^2 = 148$ ($\chi = 12.1655$) we find that this value of $\chi = \frac{x}{\sigma}$ is outside the range of the table. Thus P must be less than 0.00000000026. This indicates that the actually observed concurrent frequency of enlarged spleen and parasites in the blood is significantly different from the theoretical concurrent frequency when there is no association between the two attributes. Hence we may conclude that these two attributes are definitely and significantly associated.

We may examine this association from still another point of view by considering the two groups of persons having positive and negative blood films and forming the proportions of individuals in each group that have enlarged spleens.

Thus

p_1 = per cent. of enlarged spleens in blood + group = 49.899
p_2 = per cent. of enlarged spleens in blood − group = 32.031

The difference between these two percentages is 17.868, and this difference may be considered in terms of the standard error of the difference in order to determine its significance. Since it is a question as to whether or not these two percentages could have arisen from the same universe, we may use the p of the marginal total for the determination of the standard error of p_1 and p_2.

Thus, by principles explained in Chapter XI,

$$p = 100 \cdot \frac{2027}{5501} = 36.848$$

Then, again, by the theory of simple probability developed in Chapter XI,

$$\sigma_{p_1} = \sqrt{\frac{(36.848)\ (63.152)}{1483}} = \sqrt{1.56915} = 1.253$$

$$\sigma_{p_2} = \sqrt{\frac{(36.848)\ (63.152)}{4018}} = \sqrt{.57915} = .761$$

$$\sigma_{p_1 - p_2} = \sqrt{\sigma^2_{p_1} + \sigma^2_{p_2}} = \sqrt{2.14830} = 1.4657$$

$$\frac{\text{Difference}}{\text{Standard error of difference}} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{17.868}{1.4657} = 12.19$$

Now $\sqrt{\chi^2} = \sqrt{148.6} = 12.19$, and the fact that $\chi^2 = \left(\frac{x}{\sigma}\right)^2$ can be shown in general.

When we test a four-fold table by χ^2, we are, therefore, testing as to whether or not there is any significant difference between any two of the properly contrasted proportions of the table.

THE CHI-SQUARE COMPARISON OF TWO OBSERVED SAMPLES

A third application of the χ^2 test which we owe to Pearson[2] should be widely useful to medical men.   Problems of the following sort arise constantly: Given two frequency distributions of phenomena, what is the probability, on the one hand, that the two can be regarded as random samples from the same population, whose characteristics are known only from the samples; or, put the other way about, what is the probability that the one distribution is really *different* from the other to a greater degree than could reasonably be supposed to have arisen by the operation of chance alone?

Pearson shows that if we let the population from which the two samples, if undifferentiated, are supposed to be drawn be given by the class frequencies

$$m_1,\ m_2,\ m_3,\ m_4 \ldots\ldots\ldots\ m_p,\ m_q \ldots\ldots\ldots\ m_s$$

the total population being M, and let the samples be given by the frequencies in the same classes:

									Total
First sample	f_1	f_2	f_3	$\cdots$	f_p	f_q	$\cdots$	f_s	N
Second sample........	f'_1	f'_2	f'_3	$\cdots$	f'_p	f'_q	$\cdots$	f'_s	N'

where the totals N and N' differ widely or little, and then form a quantity

$$\chi^2 = S_1^s \left\{ \frac{N N' \left(\frac{f_p}{N} - \frac{f'_p}{N'} \right)^2}{f_p + f'_p} \right\}$$

where S_1^s denotes summation of like quantities from 1 to s, that then the required probability that the two samples are undifferentiated, i. e., did come as random samples from the same population, may be found by looking out the value of P corresponding to the ascertained χ^2 and n' (the number of classes) from the tables given on pp. 26–29 of Pearson's "Tables for Statisticians and Biometricians."

Let an example make the theorem plain. MacDonald* gave the following distributions of hair color of children attacked (a) with scarlet fever and (b) with measles, from data collected in the Glasgow Corporation Fever Hospitals.

The question is: Do scarlet fever and measles attack individuals indifferently and at random so far as concerns hair pigmentation? Or, in other words, are the scarlet fever and measles distributions, in respect of hair color, different from each other only by so much as might arise by chance in samples of the size of these?

TABLE 49

DATA ON THE INCIDENCE OF SCARLET FEVER AND MEASLES IN RELATION TO HAIR PIGMENTATION

(MACDONALD'S DATA)

Hair color.	Number of cases of	
	Scarlet fever.	Measles.
Black	12	0
Dark	289	85
Medium	1109	367
Fair	360	184
Red	94	25
Totals	1864	661

* MacDonald, David: Pigmentation of the Hair and Eyes of Children Suffering from the Acute Fevers; Its Effect on Susceptibility, Recuperative Power, and Race Selection, Biometrika, vol. 8, pp. 13–39, 1911.

The distributions are shown graphically in Fig. 77. The numerical work is set forth in Table 50.
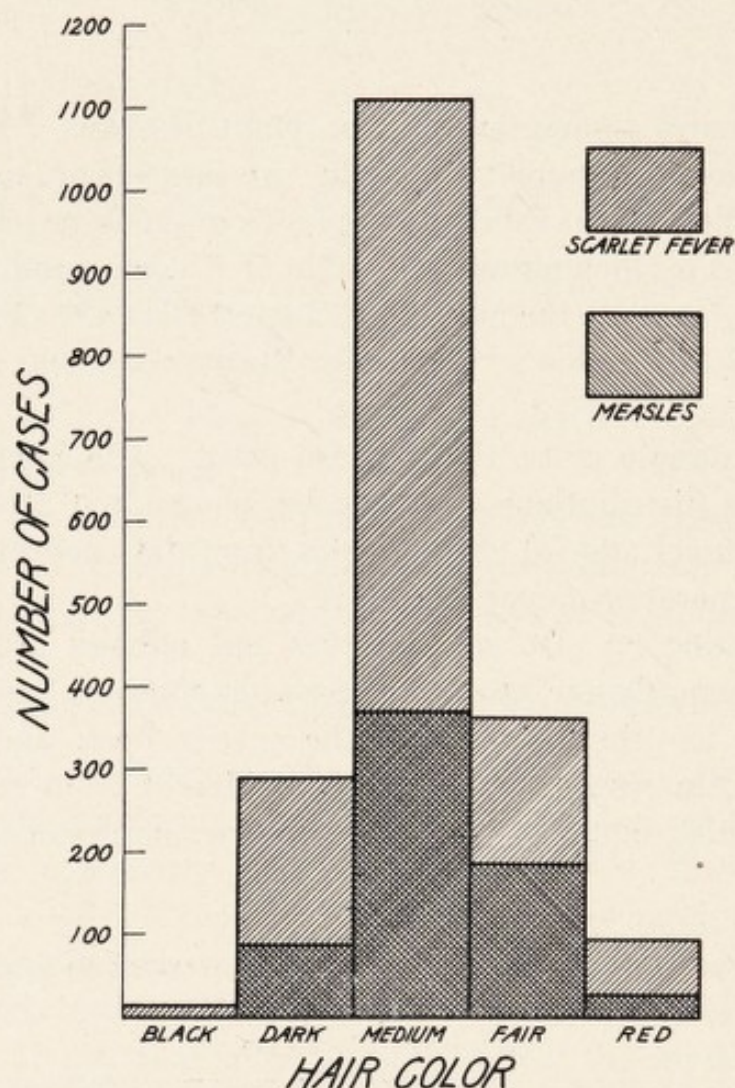
Fig. 77.—Distribution of scarlet fever and measles in respect of hair color of those attacked.

From Table 50.

$$\chi^2 = NN' \times .000,0211 = 1864 \times 661 \times .000,0211 = 26.00$$

P from the tables is about .000,03. In other words, the odds are more than 33,000 to 1 against the occurrence of two such divergent samples of hair color if they were *random* samples from the same population. We can conclude that they are really differentiated samples, or that scarlet fever and measles do not attack

TABLE 50

NUMERICAL WORK TO CALCULATE PROBABILITY THAT THE MEASLES AND SCARLET FEVER DISTRIBUTIONS OF TABLE 49 ARE RANDOM SAMPLES OF THE SAME POPULATION

		Black.	Dark.	Medium.	Fair.	Red.	Totals.
Scarlet fever..............	(i)	12	289	1109	360	94	1864
Measles...................	(ii)	0	85	367	184	25	661
(i) + (ii).................	(iii)	12	374	1476	544	119	2525
(i)/1864..................	(iv)	.0064	.1551	.5950	.1931	.0504	1.0000
(ii)/661..................	(v)	.0000	.1286	.5552	.2784	.0378	1.0000
(iv) − (v)................	(vi)	+.0064	+.0265	+.0398	−.0853	+.0126	
Square of (vi)............	(vii)	+.000,041	+.000,702	+.001,584	+.007,276	+.000,159	
(viii)	(viii)	.000,0034	.000,0019	.000,0011	.000,0134	.000,0013	.000,0211
(vii) + (iii).............							

Totals column formulas:

$$f,$$
$$f+f'$$
$$f/N$$
$$f'/N'$$
$$(f/N - f'/N')^2$$
$$\frac{(f/N - f'/N')^2}{f+f'}$$

indifferently all individuals whatever their hair pigmentation; or, that scarlet fever and measles are differential in their selection.

It will be seen that the arithmetical work is not difficult, and the usefulness of the method in drawing correct conclusions from many classes of medical data is great. One caution must always be kept in mind. The validity of the method depends upon the data tested being *frequencies*. It is not directly applicable to rates, indices, or true ordinates.

PRACTICAL PROBLEMS OF SAMPLING

In the practical affairs of life perhaps the most frequent use of the statistical method which is made, either consciously or unconsciously, is to form a judgment of the probable constitution of an unknown universe, on the basis of the constitution of a sample of known constitution drawn at random from it.

For example, suppose it to be assumed that, in order to justify mass treatment for hookworm infestation in a population, 70 to 80 per cent. of the people must harbor the worms. How, by a process of sampling in making examinations, shall it be ascertained that this proportion of the people does, in fact, probably harbor the worms?

This is not an easy or simple problem. Much research still needs to be done on the general problem of which the one cited is a particular case, before we shall be able to proceed with entire precision, and certainty of the validity of all the methods employed in its solution. But in the meantime the problem is of such great practical importance to every scientific worker that it seems desirable to discuss it in some detail here.

In the first place it can be seen at once that an adequate judgment of the constitution can only be arrived at if:

(*a*) The sample is a *good* one.

(*b*) The sample is an *adequate* one.

By a "good" sample is meant one which is fairly *representative* qualitatively of the universe from which it is drawn. By an "adequate" sample is meant one which is *large enough* in point of numbers to satisfy the requirements of the theory of probability.

To get a good sample, if we are working in the realm of living things, is a biologic problem primarily and fundamentally. How

shall it be gone about? Evidently the general criterion is that the
sample should contain at least one individual from each of the
classes of the universe known from prior experience to be differen-
tiated in any important particular from all other classes in the
universe. Thus, to consider the hookworm case. We know,
quite apart from hookworm problems at all, that mankind is
differentiated everywhere into classes in respect of

(*a*) Age.
(*b*) Sex.
(*c*) Race (or color).
(*d*) Geographic location.

That is to say, at any given instant of time it is known that a
human population contains a number of people forming a class
ranging in age from birth to nine years, another class aged ten to
nineteen years, etc. It contains a class of persons like each other,
but different from all the rest, in respect of being males. It con-
tains perhaps a class of persons who are white, and another class
who are colored. It contains a class of persons who all live in
town A, another class of persons who live on farms in county B, etc.

These are all perfectly well-known and certain differentiations
of the population. Whatever else may be peculiarly distributed
among the individuals of our universe, it is *certain* that any universe
of human beings from which it is proposed to draw a sample will
contain some or all of these four differentiations which have been
mentioned. Plainly, then, any sample, to be qualitatively repre-
sentative of the universe, must contain some individuals from each
of the differentiated classes. Thus, to have a representative sample
from the population of a given locality relative to our hookworm
problem, it would be necessary to take as a minimum one person
in each decade of age, or say 10 in all.   But there should also within
each decade be at least one male and one female, and one white
and one colored person, making 4×10 or 40 in all. Of course
practically there may be no negroes at all in the locality, or there may
be no persons ninety to ninety-nine years of age, and so on, in any of
which events the necessary sample will be, by so much, reduced.

As regards geographic location the procedure must be in
principle the same.   The whole universe dealt with covers a certain

area. To get a representative sample it will therefore be necessary to lay down over the whole area an imaginary network, in which all the meshes are of equal and not too large area, and then draw a sample relative to the other differentiations from within each mesh.

The meaning of all this discussion is that it is both practically and theoretically wise to make all probabilities *specific* relative to already known differentiations of the universe from which the sample is drawn. Crude probabilities for whole universes in which differentiation is known to exist, rarely have any particular practical significance. Thus one might ask what is the probability that a warm-blooded animal will shave tomorrow morning, and put into the denominator of the fraction all the elephants, tigers, other mammals, and birds; but supposing there were accurate data to do all this, the resulting probability would have only a very academic interest, because everyone already knows beforehand from direct observational experience that elephants and eagles, for example, do not shave.

This reasoning applies to the hookworm problem in this way. In a county the situation actually may be this: On four or five plantations in one corner of the county 90 per cent. of the negro laborers are infested. Nowhere else in the county nor among the whites is there more than 1 per cent. of infestation. This is the *real* situation, but is unknown to the workers who come into that county to clean up hookworm by an efficient campaign. By what general procedure shall the real fact become most speedily known? Now, plainly, a completely random sample of the county taken as a whole, and the probability deduced therefrom would be quite misleading, and of no practical use in bringing about the prompt treatment of the negroes on the heavily infested plantations. But suppose the imaginary network to have been laid down and each mesh sampled, with due regard to the other differentiations of color, age, and sex. Then it would at once appear that virtually all the efforts should be directed to one mesh. Furthermore, if the individuals to form the sample in each mesh were chosen relative to the other differentials, color, sex, and age, so that the sample should contain the two races, the two sexes, and the different ages, in roughly the proportion that they existed *in the population of the mesh*, then it would at once appear that it was the *negroes* only who needed mass treatment.

We now come to the question of how many individuals should be included in the sample taken in the way indicated from each mesh, or, in short, how large must a sample be to be adequate? This is a *mathematical* problem, and, as will appear as we go on, a problem to which no fixed or unique general answer can be given. What size of sample is adequate depends in part upon the constitution of the population. How this works out we may now consider.

Suppose a population of any absolute size whatever, say N, except for the restriction that it shall be at least ten times as large as any sample m drawn from it.

Further, suppose that the proportion of hookworm infestation in N is actually (though unknown to us):

(*a*) 10 per cent.
(*b*) 20 per cent.
(*c*) 30 per cent.
(*d*) 40 per cent.
(*e*) 50 per cent.
(*f*) 60 per cent.
(*g*) 70 per cent.
(*h*) 80 per cent.
(*i*) 90 per cent.

Suppose now we take samples from N, of m individuals in each sample, and examine certain consequences which flow from different values of m.

We may then set up the following table (Table 51), which shows in each cell two figures. These figures are the *lower* (light) and *upper* (heavy) limiting *whole* numbers of individuals who will be found to have hookworm infestation, on the average, in only one sample of the size named out of every 200 such samples tried of the same size, if the general population from which the sample is drawn is actually infested in the degree indicated by the percentage figure at the top of the column. That is to say, to take a concrete example, if 90 per cent. of the population are really infested, in a random sample of 100 from that population there will not be found fewer than 82 persons showing infestation as often as once in 200 trials. Odds of 199 to 1 are sufficiently wide to constitute certainty

in most practical statistical matters. These odds indicate a far smaller fluctuation or error than inheres in the original observational or experimental data of biology generally.

TABLE 51
Sampling Limits

Size of sample.	Actual percentage of occurrence in population N.																	
m	10%.		20%.		30%.		40%.		50%.		60%.		70%.		80%.		90%.	
10	0	4	0	6	0	7	0	8	0	10	2	10	3	10	4	10	6	10
15	0	5	0	7	0	10	1	11	2	13	4	14	5	15	8	15	10	15
20	0	6	0	9	0	12	2	14	4	16	6	18	8	20	11	20	14	20
25	0	7	0	11	1	14	3	17	6	19	8	22	11	24	14	25	18	25
30	0	8	0	12	2	16	5	19	7	23	11	25	14	28	18	30	22	30
35	0	9	0	14	3	18	6	22	9	26	13	29	17	32	21	35	26	35
40	0	9	1	15	4	20	8	24	11	29	16	32	20	36	25	39	31	40
45	0	10	2	16	5	22	9	27	13	32	18	36	23	40	29	43	35	45
50	0	11	2	18	6	24	11	29	15	35	21	39	26	44	32	48	39	50
60	0	12	4	20	8	28	14	34	20	40	26	46	32	52	40	56	48	60
70	0	14	5	23	11	31	17	39	24	46	31	53	39	59	47	65	56	70
80	1	15	6	26	13	35	20	44	28	52	36	60	45	67	54	74	65	79
90	1	17	8	28	15	39	24	48	32	58	42	66	51	75	62	82	73	89
100	2	18	9	31	18	42	27	53	37	63	47	73	58	82	69	91	82	98
110	2	20	11	33	20	46	30	58	41	69	52	80	64	90	77	99	90	108
120	3	21	12	36	23	49	34	62	45	75	58	86	71	97	84	108	99	117
130	4	22	14	38	25	53	37	67	50	80	63	93	77	105	92	116	108	126
140	4	24	15	41	28	56	41	71	54	86	69	99	84	112	99	125	116	136
150	5	25	17	43	30	60	44	76	59	91	74	106	90	120	107	133	125	145
160	6	26	18	46	33	63	48	80	63	97	80	112	97	127	114	142	134	154
170	6	28	20	48	35	67	51	85	68	102	85	119	103	135	122	150	142	164
180	7	29	22	50	38	70	55	89	72	108	91	125	110	142	130	158	151	173
190	8	30	23	53	40	74	58	94	77	113	96	132	116	150	137	167	160	182
200	9	31	25	55	43	77	62	98	81	119	102	138	123	157	145	175	169	191
300	16	44	42	78	69	111	98	142	127	173	158	202	189	231	222	258	256	284
400	24	56	59	101	96	144	134	186	174	226	214	266	256	304	299	341	344	376
500	32	68	76	124	123	177	171	229	221	279	271	329	323	377	376	424	432	468
600	41	79	94	146	151	209	209	271	268	332	329	391	391	449	454	506	521	559
700	49	91	112	168	178	242	246	314	315	385	386	454	458	522	532	588	609	651
800	58	102	130	190	206	274	284	353	363	437	444	516	526	594	610	670	698	742
900	66	114	149	211	234	306	322	398	411	489	502	578	594	666	689	751	786	834
1000	75	125	167	233	262	338	360	440	459	541	560	640	662	738	767	833	875	925

The manner in which Table 51 was calculated needs some discussion. First, for each value of m and of the percentages of infes-

tation the sigma (σ) of the point binomial was calculated. Thus for 60 per cent. of infestation and $m = 100$

$$\sigma = \sqrt{100 \times .6 \times .4}$$

The values so obtained were multiplied by 2.58, which is the x/σ value which cuts off just a little more than .005 of the tail area of the normal curve. The value so obtained was then subtracted from the mean number expected on each set of m, p, and q values, to obtain the lower (light) entries in the table, and added to it to obtain the upper (heavy) entries. The tabled values were adjusted to whole numbers from the values computed to three places of decimals by taking for each light entry the next *lower* whole number, and for each heavy entry the next *higher* whole number, regardless of the value of the decimal portion. This was, of course, to create a margin of safety, beyond the strictly accurate decimal values.

There may be some inclined to object to the procedure outlined above, on the ground that in the case of the extremely skew binomials, say where $p = .9$ and $q = .1$, there will be scant justification for replacing the areas of the binomial with those of the normal curve, as has been done in the formation of Table 51. Wishing to see just how much there was in this objection, and also desiring

TABLE 52

Ordinates of Point Binomial, When $n = 10$. Sum of All Ordinates $= 1.00$

Favorable occurrences.	$p = .5$ $q = .5$	$p = .6$ $q = .4$	$p = .7$ $q = .3$	$p = .8$ $q = .2$	$p = .9$ $q = .1$
10	.00	.01	.03	.11	.35
9	.01	.04	.12	.27	.39
8	.04	.12	.23	.30	.19
7	.12	.21	.27	.20	.06
6	.21	.25	.20	.09	.01
5	.25	.20	.10	.03	.00
4	.21	.11	.04	.01	.00
3	.12	.04	.01	.00	.00
2	.04	.01	.00	.00	.00
1	.01	.00	.00	.00	.00
0	.00	.00	.00	.00	.00
Sum	1.01	.99	1.00	1.01	1.00

TABLE 53

ORDINATES OF THE POINT BINOMIAL WHEN n = 50.   SUM OF ALL ORDINATES = 1.000000

Favorable occurrences.	$p = .5$ $q = .5$	$p = .6$ $q = .4$	$p = .7$ $q = .3$	$p = .8$ $q = .2$	$p = .9$ $q = .1$
50	.000000	.000000	.000000	.000014	.005154
49	.000000	.000000	.000000	.000178	.028632
48	.000000	.000000	.000004	.001093	.077943
47	.000000	.000000	.000028	.004371	.138565
46	.000000	.000000	.000140	.012840	.180904
45	.000000	.000002	.000551	.029531	.184925
44	.000000	.000011	.001771	.055371	.154104
43	.000000	.000047	.004770	.087012	.107628
42	.000000	.000169	.010989	.116922	.064278
41	.000002	.000527	.021978	.136409	.033329
40	.000009	.001440	.038619	.139819	.015183
39	.000033	.003491	.060185	.127108	.006135
38	.000108	.007563	.083830	.103275	.002215
37	.000315	.014738	.105017	.075470	.000719
36	.000833	.025967	.118948	.049864	.000211
35	.001999	.041547	.122347	.029919	.000056
34	.004373	.060589	.114700	.016362	.000014
33	.008746	.080785	.098314	.008181	.000003
32	.016035	.098737	.077247	.003750	.000001
31	.027006	.110863	.055757	.001579	.000000
30	.041859	.114559	.037039	.000612	
29	.059799	.109103	.022677	.000218	
28	.078826	.095879	.012811	.000072	
27	.095962	.077815	.006684	.000022	
26	.107957	.058361	.003223	.000006	
25	.112275	.040464	.001436	.000001	
24	.107957	.025938	.000592	.000000	
23	etc.	.015371	.000225		
22	symmetrical	.008417	.000079		
21	to first	.004257	.000026		
20	half.	.001987	.000008		
19	....	.000854	.000002		
18	....	.000338	.000001		
17	....	.000123			
16	....	.000041			
15	....	.000012			
14	....	.000003			
13	....	.000001			
12	....	.000000			
Sum	.9999997*	.999999	.999998	.999999	.999999

* Of all 51 terms.

to give the reader of this book a concrete idea of the behavior of binomials with different values of p and q, I asked my assistant, Dr. Flora Sutton, to calculate the ordinates of a series of binomials. The results are given in Tables 52 and 53.

Consider the most unfavorable case in Table 51 where $n = 10$, and the percentage of occurrence is 90. The table says, on the basis of normal curve areas, that if 90 is the true unknown percentage, we shall not get, with samples of 10, fewer than 6 favorable occurrences. Summing the ordinates of the binomial in the last column of Table 52, we have $\overset{5}{\underset{0}{\Sigma}} = 0.00$. To more than the degree of refinement that anyone ought to work with on the basis of samples of 10, the normal curve area adequately approximates the sum of the terms of the binomial, in the case which is of all in Table 51 most unfavorable to the normal curve.

We might let the case rest here, but it seems desirable to present another table for the binomial having $n = 50$. This is done in Table 53.

Again, let us test the worst case. Table 51 states that if the true but unknown composition of the population is 90 per cent. events of the favorable sort one will not expect to get in samples of 50 fewer than 39 favorable cases, oftener than five times in a thousand. From the last column of Table 53 the sum of the terms of the binomial up to 39 is .003220, or about 3 cases in 1000 trials. Up to 40 the sum is .009355 or 9 cases in 1000 roughly. For all practical statistical purposes it is apparent that Table 51 is a safe guide.

The practical uses of Table 51 are obviously manifold. It enables one, either from direct reading or interpolation between tabled values, to answer many questions which arise in experimental work, in field work, in epidemiologic enquiries, and, indeed, wherever in the whole range of scientific investigation a problem of sampling confronts one.

SUGGESTED READING

1. Pearson, K.: On the Criterion That a Given System of Deviation from the Probable in the Case of a Correlated System of Variables is Such That it Can Reasonably be Supposed to Have Arisen from Random Sampling, Phil. Mag., 1900, pp. 157–175.
2. Pearson, K.: On the Probability That Two Independent Distributions of Frequency Are Really Samples from the Same Population, Biometrika, vol. 8, pp. 250–254, 1911.

3. Pearson, K.: On a Brief Proof of the Fundamental Formula for Testing the Goodness of Fit of Frequency Distributions, and On the Probable Error of "P," Phil. Mag., vol. 31, pp. 369–378, 1916.

> (At this stage the really earnest student will find it helpful to do some reading in logic. The following two modern books in this field are suggested as a start. The second one was written by a distinguished medical man.)

4. Schiller, F. C. S.: Formal Logic, a Scientific and Social Problem, New York (Macmillan), 1912.
5. Mercier, C.: A New Logic, Chicago (Open Court Publishing Co.).

CHAPTER XIII

THE MEASUREMENT OF VARIATION
THE FREQUENCY DISTRIBUTION

WHEN one measures with a sufficient degree of precision a number of occurrences of any natural event whatever, he encounters the phenomenon of variation. No two occurrences are exactly alike, whether we are concerned with a physiologic event, such as pulse-rate or body temperature, or a morphologic matter, such as brain weight or cephalic index, or what not. If one measures exactly many events of the same kind and arranges the results in progressive order he will form a *frequency distribution* of variation (cf. Chapters IV and VI *supra*). An example of such a distribution is given in Table 54 and is exhibited graphically as a histogram in Fig. 78.

TABLE 54

FREQUENCY DISTRIBUTION OF VARIATION IN PULSE BEATS PER MINUTE IN ENGLISH CONVICTS*

Pulse beats per minute.	Frequency of occurrence.
44.5– 48.4	2
48.5– 52.4	5
52.5– 56.4	17
56.5– 60.4	57
60.5– 64.4	90
64.5– 68.4	150
68.5– 72.4	120
72.5– 76.4	131
76.5– 80.4	109
80.5– 84.4	86
84.5– 88.4	62
88.5– 92.4	42
92.5– 96.4	15
96.5–100.4	18
100.5–104.4	9
104.5–108.4	5
108.5–112.4	3
112.5–116.4	3
Total	924

A word should be said about the designation of the class limits in the first column of Table 54. The pulse rates, as actually

* Whiting, M. H.: A Study of Criminal Anthropometry, Biometrika, vol. 11, pp. 1–37, 1915.

335

recorded by the physicians who took the data originally, which went into the first class were rates of 45, 46, 47, and 48 beats per minute. But looking at the matter from the viewpoint of exact measurement a physician's record of 45 beats per minute really includes on the average all those rates which, with precise physical instruments for timing and recording beats, would fall between 44.500 . . . beats and 45.499 . . . beats per minute. Consequently the class limits are set down in the way shown in Table 54.
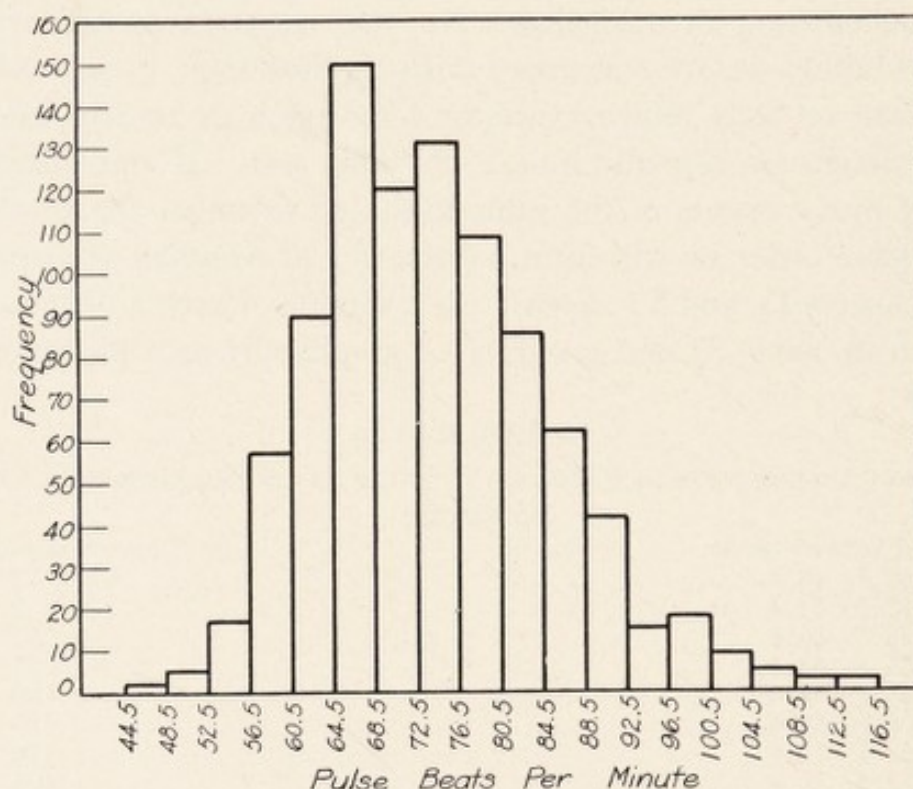
Fig. 78.—Histogram showing frequency distribution of variation in pulse beats per minute in English convicts. (Data of Table 54.)

This distribution shows in a rather typical manner the general characteristics of frequency distributions of variation, or variation curves, as they may briefly, if less precisely, be called. We see the "cocked hat" shape, with which we became familiar in Chapter XI, indicating that the most frequent occurrence of variates is, in general, near the middle of the distribution. Toward the ends the frequency becomes smaller and smaller till it disappears. The distribution has but a single peak. It might be thought, at first inspection, that there were two peaks, one on the class 64.5–68.4,

and the other on the class 72.5–76.4 beats per minute. But the depression on class 68.5–72.4, which gives rise to the impression of two peaks, is not significantly different from the frequency on the classes to either side of it, having regard to probable errors, and consequently means nothing. It is, in fact, merely a result of random sampling. How do we know this?

If, of N values, N_1 lie below X and N_2 above it, the probable error of N_1 or N_2 is

$$\pm .67449 \sqrt{\frac{N_1 N_2}{N}}$$

It is an even chance that N times the true proportion of values below X lies between $N_1 + .67449 \sqrt{\frac{N_1 N_2}{N}}$ and $N_1 - .67449 \sqrt{\frac{N_1 N_2}{N}}$. (Cf. Sheppard, Biometrika, Vol. II, p. 178.) So then we have for the data of Table 54 the results shown in Table 55.

TABLE 55

PROBABLE ERRORS OF FREQUENCIES

X	N_1	N_2	$P.E.$	X	N_1	N_2	$P.E.$
44.4	0	924	....	84.4	767	157	$\pm$ 7.7
48.4	2	922	$\pm$ 0.95	88.4	829	95	$\pm$ 6.2
52.4	7	917	$\pm$ 1.8	92.4	871	53	$\pm$ 4.8
56.4	24	900	$\pm$ 3.3	96.4	886	38	$\pm$ 4.1
60.4	81	843	$\pm$ 5.8	100.4	904	20	$\pm$ 3.0
64.4	171	753	$\pm$ 8.0	104.4	913	11	$\pm$ 2.2
68.4	321	603	$\pm$ 9.8	108.4	918	6	$\pm$ 1.6
72.4	441	483	$\pm$ 10.2	112.4	921	3	$\pm$ 1.2
76.4	572	352	$\pm$ 10.0	116.4	924	0	....
80.4	681	243	$\pm$ 9.0				

We thus see that in the region from 64.5 to 76.4 pulse beats per minute the probable error of the frequencies is about 10. None of the differences between neighboring frequencies is of the order of $4 \times 10 = 40$, which would have to be the case to make any deflection in this region of the curve significant.

CALCULATION OF MOMENTS

Having in this way satisfied ourselves that we are dealing with an essentially unimodal curve, we may proceed to its analysis, to the end that we may have quantitative expressions of the characteristic features of variation in pulse-rate. The first step in the mathematical analysis of any frequency distribution is to calculate

certain quantities known in theoretic mechanics as "moments of inertia." The arithmetic of this process for our pulse rate example is set forth in Table 56. We shall first calculate the moments about an arbitrary origin, at the lower range end, and then later transfer to the mean or center of gravity of the distribution. The first steps in the calculation are shown in Table 56.

TABLE 56
CALCULATION OF MOMENTS

Midpoint of pulse-rate class.	Frequency Z.	x Deviation from origin in class units.	Zx	Zx^2	Zx^3	Zx^4
46.5........	2	0	0	0	0	0
50.5........	5	1	5	5	5	5
54.5........	17	2	34	68	136	272
58.5........	57	3	171	513	1,539	4,617
62.5........	90	4	360	1,440	5,760	23,040
66.5........	150	5	750	3,750	18,750	93,750
70.5........	120	6	720	4,320	25,920	155,520
74.5........	131	7	917	6,419	44,933	314,531
78.5........	109	8	872	6,976	55,808	446,464
82.5........	86	9	774	6,966	62,694	564,246
86.5........	62	10	620	6,200	62,000	620,000
90.5........	42	11	462	5,082	55,902	614,922
94.5........	15	12	180	2,160	25,920	311,040
98.5........	18	13	234	3,042	39,546	514,098
102.5.......	9	14	126	1,764	24,696	345,744
106.5.......	5	15	75	1,125	16,875	253,125
110.5.......	3	16	48	768	12,288	196,608
114.5.......	3	17	51	867	14,739	250,563
Totals.....	924	..	6399	51,465	467,511	4,708,545

For the moments about the arbitrary origin at a pulse-rate of 46.5, we have, S denoting summation.

$$\nu_1 = \frac{S(Zx)}{S(Z)} = \frac{6399}{924} = 6.925325$$

$$\nu_2 = \frac{S(Zx^2)}{S(Z)} = \frac{51,465}{924} = 55.698052$$

$$\nu_3 = \frac{S(Zx^3)}{S(Z)} = \frac{467,511}{924} = 505.964286$$

$$\nu_4 = \frac{S(Zx^4)}{S(Z)} = \frac{4,708,545}{924} = 5095.827922$$

Since we shall have to use powers of these quantities in the subsequent calculations, it will be well to keep six places of decimals

for the present, in order to ensure the degree of arithmetical accuracy we shall want at the end. Keeping the decimals at this stage has nothing whatever to do with the accuracy or reliability of the original data. It is a purely arithmetical matter.

The next step is to determine, from these moments about the lower range end as origin, the values of the moments about the mean. Letting π denote a moment about the mean, and observing that ν_1, ν_2, ν_3, and ν_4 denote moments about any arbitrary origin, we have

$$\pi_1 = 0 \text{ (by definition of the mean)}$$
$$\pi_2 = \nu_2 - \nu_1{}^2$$
$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1{}^3$$
$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1{}^2\nu_2 - 3\nu_1{}^4$$

It should be understood that the above equations for the π's are valid regardless of the origin about which the ν's are taken. In the particular example shown in Table 56 the origin was taken at the center of the class marking the lower observed range end. But the equations for the π's given above would be equally valid if the origin had been taken at the upper range end, or somewhere in the middle.

For the pulse-rate example we have:

$$\pi_2 = 55.698052 - 47.960126 = 7.737926$$
$$\pi_3 = 505.964286 - 1157.181336 + 664.278920 = 13.061870$$
$$\pi_4 = 5095.827922 - 14015.868476 + 16027.713552 - 6900.521118 = 207.151880$$

To the values of the moments given above it is necessary to make certain corrections, to allow for the fact that individual observations have been grouped in forming the frequency distribution. The corrections generally used, called after their discoverer, Sheppard's corrections, are applicable when, as in our present example, the curve has reasonably high contact at both ends of the range. For corrections of the moments of entirely general applicability see Biometrika, Vol. 12, pp. 231–258. Using μ to designate a corrected moment about the mean as origin, Sheppard's corrections are:

$$\mu_1 = 0$$
$$\mu_2 = \pi_2 - \tfrac{1}{12}. \quad (\tfrac{1}{12} = .083333)$$
$$\mu_3 = \pi_3$$
$$\mu_4 = \pi_4 - \tfrac{1}{2}\pi_2 + \tfrac{7}{240}. \quad (\tfrac{7}{240} = .029167)$$

We then have, from the pulse-rate example,

$$\mu_2 = 7.654593$$
$$\mu_3 = 13.061870$$
$$\mu_4 = 203.312084$$

Besides the moments themselves, we shall need two simple functions of them, viz.:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3},$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2}.$$

For the pulse-rate example these have values as follows:

$$\beta_1 = \frac{170.612448}{448.503991} = .380403$$

$$\beta_2 = \frac{203.312084}{58.592794} = 3.469916$$

With the moments of the distribution in hand, the foundation is laid for the determination of the various physical constants which define and describe the several aspects of the phenomenon of variation. These constants may conveniently be divided into three groups as follows:

(1) Constants defining the type or center of variation.
(2) Constants measuring dispersion or degree of variation.
(3) Constants measuring the shape of the variation curve.

CONSTANTS DEFINING THE TYPE OR CENTER OF VARIATION

The first thing one wishes to know, when considering variation philosophically, is something about the central or typical condition, about which the variation groups itself. There are three constants commonly used to define different aspects of type, and together they give a sufficient picture of the central or typical condition. They are the mean, the median, and the mode.

The Mean

The arithmetic mean or average is mechanically the center of gravity of the frequency distribution. If the histogram of Fig. 78 were cut out of sheet metal of uniform thickness, and then exactly

balanced on a knife edge set at right angles to the base line or x axis, the point where the knife edge intersected the base would be the average or mean number of pulse beats per minute of the group of 924 observations included in the distribution. This being so, it will be readily perceived from the most elementary mechanical principles, the frequencies being regarded as masses concentrated at the midpoints of the class sub-ranges on the x axis, that the mean must be distant from the arbitrary origin, about which the first raw moments are taken, by the amount of ν_1.

Thus in the pulse-rate example we have:

Pulse beats at point of arbitrary origin	=	46.5
Number of class units, from origin to mean (ν_1)	= 6.925	
Number of pulse beats per class unit	= 4	
Number of pulse beats from origin to mean	=	27.700
Mean number of pulse beats		74.200

The probable error of the mean, when n the number of observations (S (Z) in the notation used in our example) is 15 or more, is

$$\text{P. E. Mean} = \pm \chi_1 \sigma.$$

where $\chi_1 = .6744898/\sqrt{N}$, and is tabled in Pearson's "Tables for Statisticians and Biometricians."[6] σ is the standard deviation, a constant already encountered in Chapter XII and further discussed below. When a mean or average is based upon less than 15 observations, the paper of "Student"[3] should be consulted for the method of procedure to determine the reliability of the mean.

In our present case we have

$$\text{Mean pulse-rate} = 74.200 \pm .246 \text{ beats per minute.}$$

The Median

The median is the value of the varying character (*i. e.*, the point on the x axis) above and below which exactly 50 per cent. of the variates fall. In our present example 462 (*i. e.*, $\frac{1}{2}$ of 924) pulse-rate observations fall below the median value, and 462 above it.

The arithmetic of determining the median is most simple. It can best be illustrated by example. We have seen in Table 55 that 441 observations show pulse beats of 72.4 per minute or less. One-half of all observations is 462. Therefore it is clear that the median

value must fall somewhere in the 72.5 − 76.4 class, and the distance into that class where it falls is evidently in the proportion which 462 − 441 = 21 is to the whole frequency in that class, which is 131. So then what is needed is to determine what 21/131 of 4 pulse beats is, 4 beats being the class unit. This equals 0.641 pulse beat. Consequently, the median is 72.5 + .641 = 73.141 beats per minute.

We should, of course, get the same result if we count $\frac{1}{2} N$ down from the upper range end to determine the median, as we get if we count $\frac{1}{2} N$ up from the lower range end, as was done in the preceding paragraph. This is in fact so. On our example the frequency from the upper end down to 76.4 (Table 55) is 352. That is, there are 352 individuals with pulse rates of 76.5, or above. $\frac{1}{2} N - 352 = 110$. 110/131 of 4 pulse beats equals 3.359. 76.5 − 3.359 = 73.141, which is the same value that was obtained in working from the other range end.

It is to be noted that the median is smaller than the mean, *i. e.*, lies to the left of it in the distribution. This means that the curve as a whole is asymmetric or skew toward the right end or large values of the pulse-rate. We shall return to this point later.

The probable error of the median is:

$$\text{P. E. Median} = 1.25332 \times \text{P. E. mean.}$$

So we have for a final result

$$\text{Median pulse-rate} = 73.141 \pm .308 \text{ beats per minute.}$$

The Mode

The mode is the value of the varying character which, in the theoretic, true variation curve, exhibits the maximum frequency of occurrence. Owing to the probable errors of individual frequencies arising from random sampling, to which attention has already been called, the true mode may not coincide exactly with the most frequent class in the observed distribution. This means merely that the particular observed sample with which we are dealing has, by chance, a particular class near the center of the distribution occurring more frequently than it should, in relation to all the other frequencies in the distribution. Mathematically,

the mode is the point on the theoretic curve which graduates the observations, where $\frac{dy}{dx} = 0$.

The mode is distant from the mean by a quantity

$$d = \chi \times \sigma$$

χ is a constant called the *skewness*, and obviously is the fraction which the modal distance d is of the standard deviation σ, since

$$\chi = \frac{d}{\sigma}.$$

The equation for the skewness χ in terms of the moment coefficients will be given in a later section (p. 357). It should be expressly noted that this χ (skewness) is *not* the same thing as the χ_1 discussed above. Then

$$Mode = Mean - d$$

The probable error of the modal distance d, in the general case, may be found from Table 40 in Pearson's "Tables for Statisticians and Biometricians." For most practical statistical purposes what one wishes to know is whether d is significantly different from zero, *i. e.*, whether the mode is separated from the mean by an amount greater than might probably have arisen by chance. In the normal or Gaussian curve, which, as we have seen, is a symmetric unimodal, "cocked hat" curve having the equation

$$y = \frac{N}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

the mean and the mode coincide, or $d = 0$, with a probable error of

$$\text{P. E.}_{d \text{ (normal curve)}} = \pm\ .67449 \sqrt{\frac{3}{2N}}\,\sigma.$$

Consequently, unless d amounts to three or four times this probable error, the mode cannot be regarded as significantly different from the mean.

In our present example we have

$$d = .3289 \times 11.0668 = 3.640$$

$$\text{P. E.}_{d \text{ (n. c.)}} = \pm\ 0.301.$$

We see that d is more than ten times as large as the probable error. Hence we may conclude that the point of maximum frequency in the variation curve, the mode, is significantly different from the mean. The value of the mode is

$$\text{Mode} = 74.200 - 3.640 = 70.560 \text{ beats per minute.}$$

CONSTANTS MEASURING DISPERSION OR DEGREE OF VARIATION

After having defined and measured the typical condition about which variation is occurring, the next thing wanted is a measure of the degree or extent of the variation itself. In absolute terms the
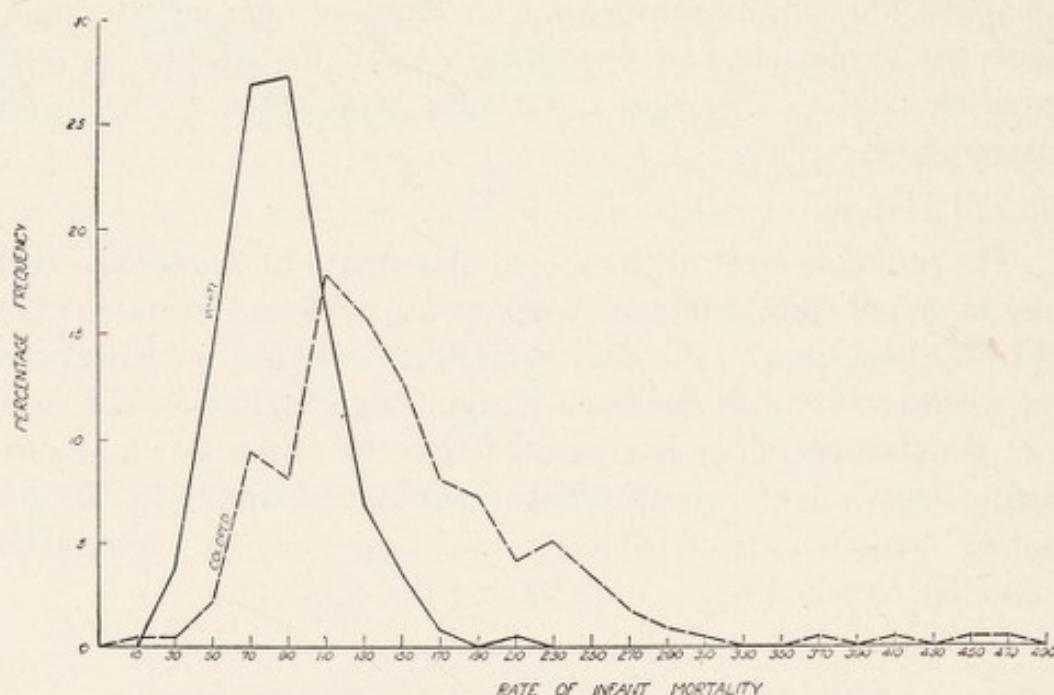
Fig. 79.—Frequency polygons showing variation in infant mortality rate in 1918 of (a) the white population and (b) the colored population of rural counties.

best measure of variation will be one which describes with precision the extent of the "scatter" of the variates about the mean. If values of the varying character widely different from the mean or typical condition are found to occur with considerable frequency, it is common sense to say that the character shows a high degree of variation. In general, the more scattered the variates away from the typical condition, the more variable is the character and *vice versa*.

Thus from Fig. 79 it is apparent that the infant mortality rate in rural areas varies much more in the colored than in the white

population. The broken line polygon is much more "scattered" or spread out than the solid line one.

THE STANDARD DEVIATION

The constant which has been adopted by biometricians to measure in absolute terms the degree of scatter or dispersion of the variates is called the standard deviation. It is the same quantity which in theoretic mechanics is called the radius of gyration. It is a parameter of the variation curve, representing a distance on the x axis such that if the total frequency were concentrated at that point and connected by a rigid bar with the mean, the system would have the same rotational properties about the mean in a frictionless medium as would the whole distribution in its actual form if it were rotated in the same medium about the mean as an axis. Roughly, three times the standard deviation on either side of the mean will include all the variates, as is shown in Fig. 76, Chapter XI. This is the same quantity which in the discussion of the point binomial was called $\sigma = \sqrt{n\,p\,q}$.

The calculation of the standard deviation is done from the following simple relation, σ denoting the standard deviation.

$$\sigma = \sqrt{\mu_2}.$$

The probable error of σ, in distributions of 15 or more individuals, is

$$P.\ E._\sigma = \pm\ \chi_2\sigma,$$

where $\chi_2 = .67449/\sqrt{2N}$, and is tabled in Pearson's "Tables for Statisticians and Biometricians." Where the distribution contains fewer than 15 individuals the same caution should be observed in judging its reliability as has been emphasized for the mean above.

For our pulse-rate example we have

$$\sigma = \sqrt{7.654593} = 2.766694$$

in units of grouping.

The unit of grouping is 4 pulse-beats per class. Whence

$$\text{S. D.} = 4 \times 2.7667 = 11.067 \pm .174$$

pulse-beats per minute.

THE COEFFICIENT OF VARIATION

Since the standard deviation measures degree of variation in concrete units, inches, pounds, beats, degrees, or whatever unit the varying character is measured in, it is evident that its utility for comparative purposes is much restricted. One cannot directly compare inches and degrees of temperature. Obviously, there is needed some comparative or relative measure of variation, which will make it possible to discuss whether, for example, men are more or less variable in respect of the weight of the brain than in respect of pulse-rate. Such a relative measure is furnished by the constant called the coefficient of variation. It expresses the standard deviation as a percentage of the mean. Symbolically we have

$$\text{C. of V.} = \frac{100 \, \sigma}{\text{Mean}}.$$

The probable error of the coefficient of variation is

$$\text{P. E.}_{c.v.} = \pm .67449 \frac{V}{\sqrt{2 N}} \left\{ 1 + 2 \left(\frac{V}{100} \right)^2 \right\}^{\frac{1}{2}} = \chi_2 \times \psi,$$

where both χ_2 and ψ are quantities tabled in Pearson's "Tables for Statisticians and Biometricians." Some caution, which will be, and can only be, acquired by experience, needs to be used in interpreting coefficients of variation. In general, one should always remember that this constant simply measures the degree of scatter of the distribution in relation to the mean value of the thing varying. Usually such a relation has real and significant meaning, but sometimes it does not for reasons inherent in the facts themselves. While space will not permit of going into details here, it may be pointed out that one source of the difficulty referred to arises from the consideration that the mean and the standard deviation are correlated. We have

$$r_{h\mu_2} = \frac{\mu_3}{n \sigma_h \sigma_{\mu_2}},$$

where $r_{h\mu_2}$ denotes the coefficient of correlation between mean and second moment, and σ_h is the standard deviation of the mean, and σ_{μ_2} the standard deviation of the second moment.

In our present example the coefficient of variation is

$$\text{C. V.} = \frac{11.0668 \times 100}{74.200} = 14.915 \pm .239 \text{ per cent.}$$

It is of considerable interest to see how this value measuring the comparative variability of pulse-rate compares with coefficients for variation in other characters of medical interest. To this end Table 57 has been inserted. This gives, in descending order, coefficients of variation for a wide range of physiologic, anatomic, and pathologic characteristics. These records are taken from the general literature of biometry.

TABLE 57

COEFFICIENTS OF VARIATION FOR MAN

	♂	♀
Weight of spleen (General Hospital population)[1]	50.58	
Steadiness of hand (English)[17]	.....	48.54–69.59
Visual acuity (English)[15]	39.12	
Weight of spleen (healthy)[2]	38.21	
Dermal sensitivity[3]	35.70	45.70
Weight of heart (General Hospital population)[1]	32.39	
Interlabral height (American), white[9]	32.15	
Keenness of sight[3]	28.68	32.21
Strength of grip, all ages, left hand[16]	26.85	
Strength of grip, all ages, right hand[16]	25.93	
Thyroid, area (English, age 13, normal thyroids)[11]	.....	25.2
Thyroid, area (English, age 13, definite goiters)[11]	.....	24.9
Weight of kidneys (General Hospital population)[1]	24.63	
Rapidity of hand, females only[17]	.....	23.91–29.49
Interlabral height (American), negro[9]	23.42	
Body weight (Bavarians)[20]	21.32	24.715
Weight of liver (General Hospital population)[1]	21.12	
Swiftness of blow[3]	19.4	17.1
Reaction time to sound (English)	19.149[14]	20.20[13]
Reaction time to sight (English)	19.016[14]	19.02[13]
Nasal depth (American), negro[9]	18.34	
Intelligence quotient, both sexes[18]	18.01	18.01
Vital capacity (English, corrected for age)[15]	17.904	
Respiration rate per minute[19]	17.80	
Weight of heart (healthy)[2]	17.71	
Weight of kidneys (healthy)[2]	16.80	
Breathing capacity[3]	16.6	20.4
Strength of left hand grip (English, age corrected)[16]	16.27	
Auditory acuity[15]	15.84	
Thyroid, breadth (English, age 13, definite goiters)[11]	.....	15.5
Strength of right hand grip (English, age corrected)[16]	15.43	
Strength of pull[3]	15.0	19.3
Pulse-rate per minute[19]	14.89	

COEFFICIENTS OF VARIATION FOR MAN—*Continued*

	♂	♀
Weight of liver (healthy)[2]	14.80	
Nasal depth (American), white[9]	14.66	
Thyroid, breadth (English, age 13, normal thyroids)[11]		14.6
Body weight (American, active tuberculosis)[10]	13.69	
Body weight (American, age 20–49 years, normal health)[10]	13.16	
Pigmentation (American), white[9]		12.94
Body weight (American, arrested tuberculosis)[10]	12.78	
Internipple breadth (American), white[9]	12.01	
Height of mandible (English, both sexes)[4]	11.73	11.73
Blood, relative cell volume (American, active tuberculosis)[10]	11.13	
Lower-nasal breadth (American), white[9]	10.53	
Body weight (English)[3]	10.37	13.37
Skull capacity (Etruscan)[5]	9.58	8.54
Chest circumference (American), negro[9]	9.45	
Skull capacity (Australians)[12]	9.27	6.98
Brain weight (French)[3]	9.16	9.14
Internipple breadth (American), negro[9]	8.93	
Mouth breadth (American), white[9]	8.69	
Pigmentation (American), negro[9]		8.68
Lower-nasal breadth (American), negro[9]	8.67	
Chest circumference (American), white[9]	8.45	
Skull capacity (modern Italian)[5]	8.34	8.99
Skull capacity (English)[6]	8.28	8.68
Skull capacity (Egyptian mummies)[5]	8.13	8.29
Brain weight (Bavarian)[20]	8.118	8.340
Brain weight (Hessian)[20]	8.096	8.125
Chest breadth (American), white[9]	8.06	
Ventral torso length (American), white[9]	8.02	
Skull capacity (Egyptians)[12]	7.89	7.08
Brain weight (Bohemian)[20]	7.809	7.382
Skull capacity (modern German)[5]	7.74	8.19
Skull capacity (Naqada)[5]	7.72	6.92
Brain weight (Swedish)[20]	7.592	8.043
Skull capacity (Parisian, French)[5]	7.36	7.10
Mouth breadth (American), negro[9]	7.17	
Skull capacity (Aino)[5]	7.07	6.90
Chest breadth (American), negro[9]	6.73	
Upper arm length (American), negro[9]	6.42	
Mandible, distance between foramina mentalia (English, both sexes)[4]	6.23	6.23
Head neck length (American), negro[9]	6.20	
Hand length (American), negro[9]	6.15	
Blood, relative cell volume (American, arrested tuberculosis)[10]	5.73	
Arm length without hand (American), negro[9]	5.59	
Blood, relative cell volume (American, age 20–49, normal health)[10]	5.42	
Length of forearm[8]	5.24	5.21
Entire arm length (American), negro[9]	5.16	
Length of femur (French)[3]	5.05	5.04
Length of tibia (French)[3]	4.975	5.365
Hand length (American), white[9]	4.97	
Upper arm length (American), white[9]	4.93	
Length of humerus (French)[3]	4.89	5.61
Head neck length (American), white[9]	4.88	
Length of radius (French)[3]	4.87	5.23
Skull, height to breadth index (English)[6]	4.86	4.16
Skull, breadth to height index (English)[6]	4.83	4.17

	♂	♀
Ventral torso length (American), negro[9]	4.81	
Length of finger (English criminals)[7]	4.74	
Skull, ratio of height to horizontal length (English)[6]	4.61	4.10
Length of foot (English)[7]	4.59	
Skull, cephalic index for horizontal length (English)[6]	4.38	3.99
Length of cubit (English criminals)[7]	4.36	
Skull, least breadth of forehead (English)[6]	4.29	4.55
Skull, height (English)[6]	4.21	3.96
Arm length without hand (American), white[9]	4.20	
Skull, length of base (English)[6]	4.07	4.11
Stature (English)[8]	3.99	3.83
Entire arm length (American), white[9]	3.97	
Skull, cephalic index for greatest length (English)[6]	3.95	4.03
Stature (American, active tuberculosis)[10]	3.86	
Skull, ratio of height to greatest length (English)[6]	3.80	4.21
Stature (American, arrested tuberculosis)[10]	3.77	
Skull, greatest breadth (English)[6]	3.75	3.54
Skull, auricular height (English)[6]	3.73	4.12
Skull, face breadth (English criminals)[7]	3.707	
Skull, cross circumference (English)[6]	3.70	3.97
Skull, sagittal circumference (English)[6]	3.63	3.90
Stature (American, age 20–49 years, normal health)[10]	3.60	
Head, breadth (English criminals)[7]	3.333	
Skull, length (English)[6]	3.31	3.45
Head, length (English criminals)[7]	3.154	
Skull, horizontal circumference (English)[6]	2.87	2.92
Oral temperature[19]	0.49	

[1] Greenwood, M.: Biometrika, 3, 66, 1904.
[2] Ibid., p. 67.
[3] Pearson, Karl: The Chances of Death, vol. 1, 293.
[4] Macdonell, W. R.: Biometrika, 3, 225, 1904.
[5] Ibid., p. 221.
[6] Ibid., p. 222.
[7] Macdonell, W. R.: Biometrika, 1, 202, 1901–02.
[8] Pearson, Karl, and Lee, Alice: Biometrika, 2, 370, 1902–03.
[9] Todd, T. W.: Human Biology, 1, 65, 1929.
[10] Pearl, R., and Miner, J. R.: Bull. Johns Hopkins Hosp., 40, 3–32, 1927.
[11] Stocks, P.: Biometrika, 19, 299, 1927.
[12] Morant, G. M.: Ibid., 19, 430, 1927.
[13] Musselman, J. R.: Ibid., 18, 196, 1926.
[14] Harmon, G. E.: Ibid., 18, 210, 1926.
[15] Holzinger, K. J.: Ibid., 16, 155, 1924.
[16] Arthur, W.: Ibid., 16, 324, 1924.
[17] Tildesley, M. L.: Ibid., 12, 170–177, 1919.
[18] Pearson, Karl: Ibid., 12, 367–372, 1919.
[19] Whiting, M. H.: Ibid., 11, 1–37, 1915.
[20] Pearl, R.: Ibid., 4, 13–104, 1905.

THE GRAPHIC REPRESENTATION OF RELATIVE VARIABILITY

It has been the generally accepted biometric practice to use the coefficient of variation just discussed as the measure of the *relative* variability or scatter of frequency distributions. This constant is, as we have seen,

$$C.\ of\ V. = \frac{100\ (\text{standard deviation})}{\text{Mean}}$$

It gives the standard deviation of the distribution in terms of the mean value of the varying character. By expressing the scatter of the distribution in this way it becomes possible to compare the relative variabilities of characters measured in different absolute units.

But the coefficient of variation has never been an entirely satisfactory constant, to biologists at least. While formally correct enough, within the limits of its definition, it does not readily or instantly call up in the mind an adequate picture of the real degree of scatter of the distribution. This is, in part at least, because two things, the mean and the standard deviation, are involved in it. When one reads the value of the standard deviation of a particular distribution it is recalled that roughly three times this quantity on either side of the mean includes the entire frequency and this gives at once some concept of the biological extent and meaning of the variation, in the particular case.

There would seem to be a place of usefulness for an adequate graphical method of depicting relative variability for comparative purposes, so that one may *see* the difference or likeness in the variation of a man and a mouse, for example, in respect of body-weight. It is the purpose of this section to describe such a graphic method, and to illustrate its applications.

The method may best be approached through a concrete illustrative example. A study[9] was made of the normal variation and correlation of the relative cell volume of human blood, in relation to age, body-weight and stature. The present situation regarding the measurement and graphical depiction of variation in these four characters, in a series of 272 normal males, is fairly exhibited in Table 58 and Figs. 80 to 82.

TABLE 58
VARIATION CONSTANTS

Character.	Mean.	Standard deviation.	Coefficient of variation (per cent.).
(a) Age....................	30.59 ± .21 yrs.	5.22 ± .15 yrs.	17.06 ± .51
(b) Body-weight...........	151.56 ± .82 lbs.	19.95 ± .58 lbs.	13.16 ± .39
(c) Stature................	68.13 ± .10 in.	2.45 ± .07 in.	3.60 ± .10
(d) Relative cell volume.....	45.59 ± .10 %	2.47 ± .07 %	5.42 ± .16
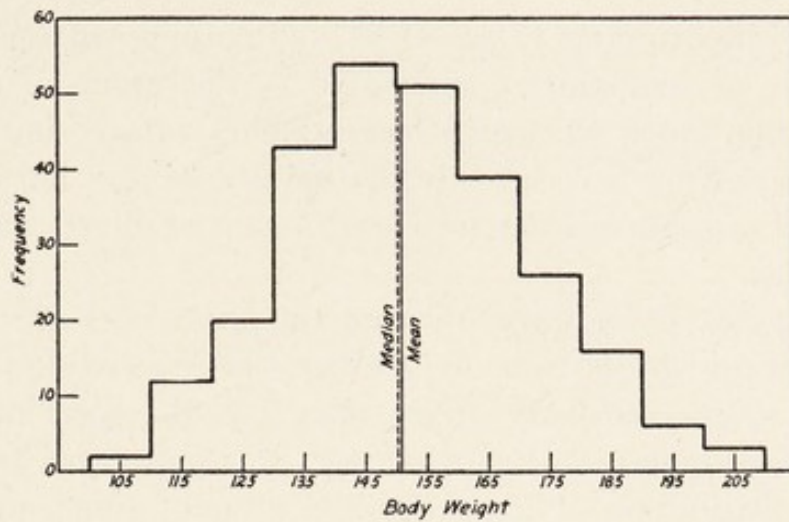
Fig. 80.—Histogram showing variation in body-weight in a group of 272 normal males.
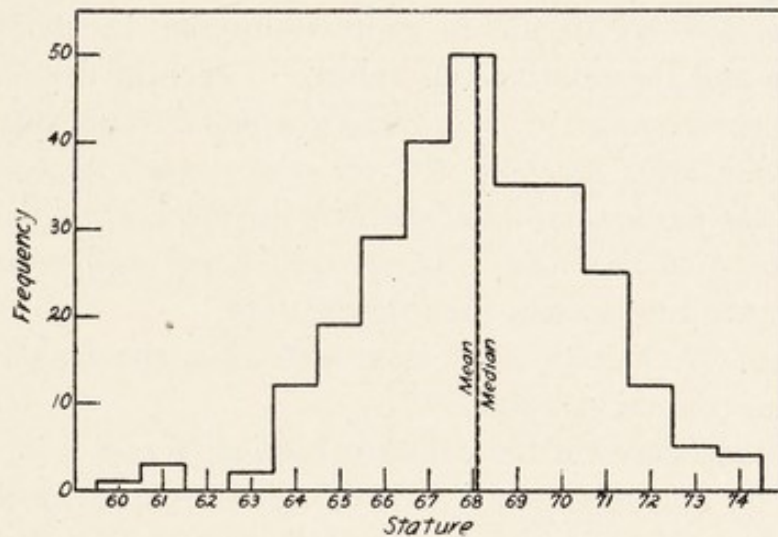
Fig. 81.—Histogram showing variation in stature in a group of 272 normal males.
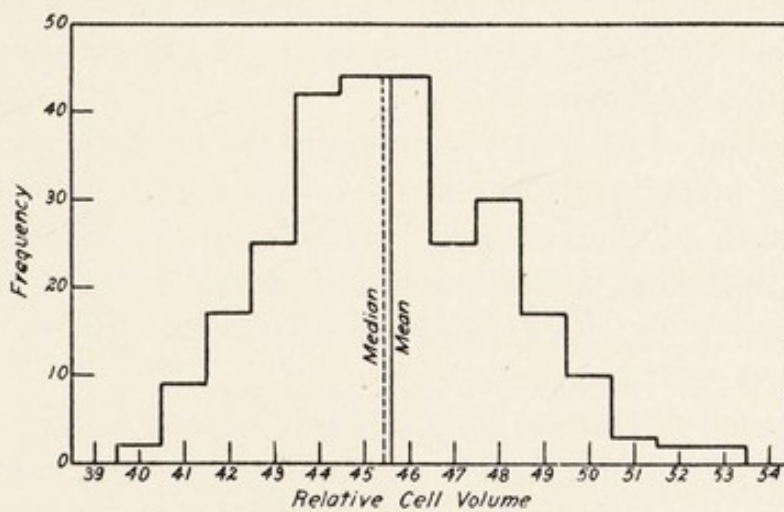
Fig. 82.—Histogram showing variation in relative cell volume of the blood in a group of 272 normal males.

Plainly the diagrams (Figs. 80–82) tell nothing whatever about the relative or comparative variability in this group of males in respect of the three characters, body-weight, stature, and relative cell volume. They are correctly plotted histograms, but the unit of abscissal measure is different in each case and direct comparison is impossible.

From Table 58 we learn, through the coefficients of variation, that the group is from three to five times more variable relatively in respect of age and body-weight than it is in respect of stature or relative cell volume. But what does this mean translated into terms of distribution of frequency? A simple, direct and easily interpreted answer is not forthcoming.

Suppose now we decide to express the age, the body-weight, the stature and the relative cell volume of each of these 272 individuals *as a percentage of their respective mean values, the mean of each character being take nas 100 per cent. And, further, suppose we express the frequencies as respectively so much per 1 per cent. of the mean of each character.* These are simple and entirely permissible transformations of the original data.

The data in their original form and after the transformation described are shown in Table 59.

If now the figures in the columns headed A and B in Table 59 be plotted on arithmetically ruled co-ordinate paper we shall have a true picture of the relative variability of the four characters considered. This is done in Fig. 83. Each of the four frequency polygons has the same area, as a result of the transformations effected in the B columns.

This method of plotting superimposes the different polygons of variation on a common Cartesian co-ordinate grid, with the mean value for each of the compared variables at the same abscissal point. It constitutes no new method of *measuring* biological variation, but merely visualizes effectively what the coefficient of variation measures.

The method of plotting used in Fig. 83 shows at a glance that the 272 men of this group differ among themselves far more widely in respect of age and body-weight than they do in respect of stature or relative cell volume. The variation polygon for stature shows

TABLE 59

ABSOLUTE AND RELATIVE FREQUENCY DISTRIBUTIONS FOR VARIATION IN (a) AGE, (b) BODY-WEIGHT, (c) STATURE, AND (d) RELATIVE CELL VOLUME OF THE BLOOD IN 272 NORMAL MALES

Age.

Class unit in years.	Observed absolute frequency.	A — Per cent, which mid-point of class is of mean.	B — Absolute frequency per 1 per cent, of mean.
20–21.9	9	68.6	1.4
22–23.9	12	75.2	1.8
24–25.9	34	81.7	5.2
26–27.9	41	88.3	6.3
28–29.9	35	94.8	5.4
30–31.9	44	101.3	6.7
32–33.9	31	107.9	4.7
34–35.9	24	114.4	3.7
36–37.9	15	121.0	2.3
38–39.9	12	127.5	1.8
40–41.9	10	134.0	1.5
42–43.9	3	140.6	0.5
44–45.9	1	147.1	0.2
46–47.9	…	153.6	…
48–49.9	1	160.2	0.2
Totals	272	…	…

Body-weight.

Class unit in pounds.	Observed absolute frequency.	A — Per cent, which mid-point of class is of mean.	B — Absolute frequency per 1 per cent, of mean.
99.5–109.4	2	68.9	0.3
109.5–119.4	12	75.5	1.8
119.5–129.4	20	82.1	3.0
129.5–139.4	43	88.7	6.5
139.5–149.4	54	95.3	8.2
149.5–159.4	51	101.9	7.7
159.5–169.4	39	108.5	5.9
169.5–179.4	26	115.1	3.9
179.5–189.4	16	121.7	2.4
189.5–199.4	6	128.3	0.9
199.5–209.4	3	134.9	0.5
Totals	272	…	…

Stature.

Class unit in inches.	Observed absolute frequency.	A — Per cent, which mid-point of class is of mean.	B — Absolute frequency per 1 per cent, of mean.
59.5–60.4	1	88.1	0.7
60.5–61.4	3	89.5	2.0
61.5–62.4	…	91.0	…
62.5–63.4	2	92.5	1.4
63.5–64.4	12	93.9	8.2
64.5–65.4	19	95.4	12.9
65.5–66.4	29	96.9	19.8
66.5–67.4	40	98.3	27.3
67.5–68.4	50	99.8	34.1
68.5–69.4	35	101.3	23.8
69.5–70.4	35	102.7	23.8
70.5–71.4	25	104.2	17.0
71.5–72.4	12	105.7	8.2
72.5–73.4	5	107.1	3.4
73.5–74.4	4	108.6	2.7
Totals	272	…	…

Relative cell volume.

Class unit in per cent, of total volume.	Observed absolute frequency.	A — Per cent, which mid-point of class is of mean.	B — Absolute frequency per 1 per cent, of mean.
39.5–40.4	2	87.7	0.9
40.5–41.4	9	89.9	4.1
41.5–42.4	17	92.1	7.8
42.5–43.4	25	94.3	11.4
43.5–44.4	42	96.5	19.1
44.5–45.4	44	98.7	20.1
45.5–46.4	44	100.9	20.1
46.5–47.4	25	103.1	11.4
47.5–48.4	30	105.3	13.7
48.5–49.4	17	107.5	7.8
49.5–50.4	10	109.7	4.6
50.5–51.4	3	111.9	1.4
51.5–52.4	2	114.1	0.9
52.5–53.4	2	116.3	0.9
Totals	272	…	…

23

the least scatter. That for relative cell volume is somewhat, but not greatly, more spread. Those for age and body-weight are wide, flat distributions, indicating a relatively high variation in the group in respect of these characters.

One more example will be given. What is the comparative individual variability of cows in respect of milk production and of hens in respect of egg production? Table 60 gives the necessary data regarding (a) milk yield in gallons per week in three-year-old
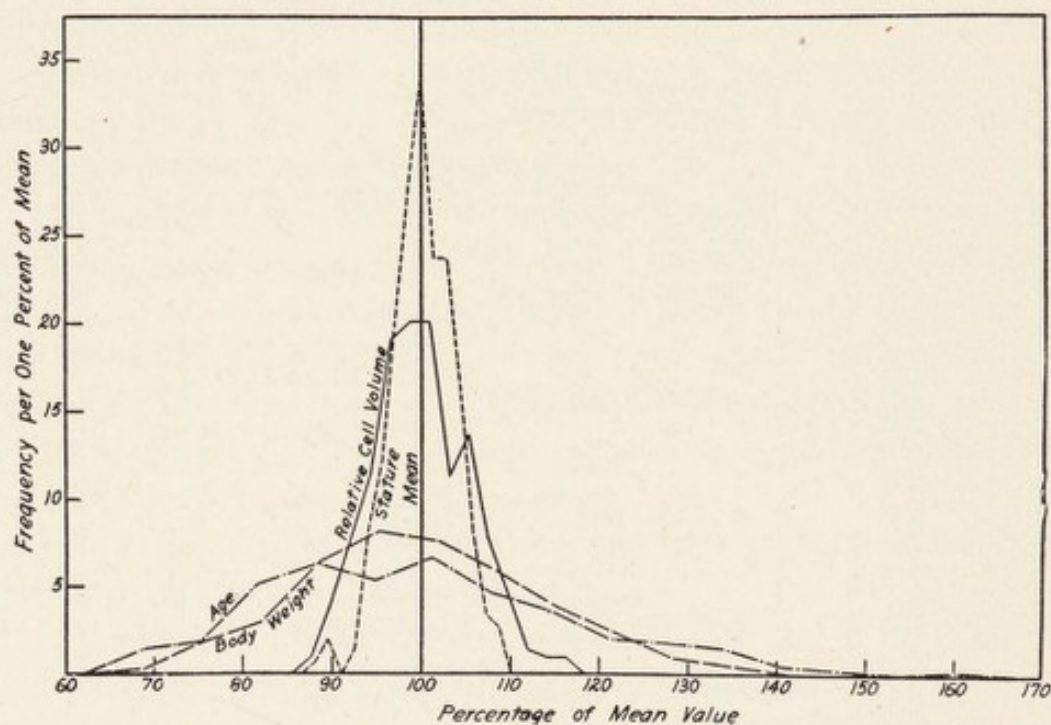
Fig. 83.—Superimposed variation polygons for (1) relative cell volume, (2) stature, (3) body-weight, and (4) age, in 272 normal males. See text for further explanation.

Ayrshire cows (combined years 1908–09),* and (b) annual egg production of Barred Plymouth Rock hens (1905–06, 150 bird pens).†

The coefficients of variation for the distributions of Table 60 are as follows:

Milk yield: C. of V. = 17.690 ± .229.
Egg production: C. of V. = 31.72 ± 1.00.

* Pearl, R., and Miner, J. R.: Variation of Ayrshire Cows in the Quantity and Fat Content of Their Milk, Jour. Agr. Research, vol. 17, pp. 285–322, 1919.

† Pearl, R., and Surface, F. M.: A Biometrical Study of Egg Production in the Domestic Fowl. I. Variation in Annual Egg Production, U. S. Dept. Agr. Bur. Anim. Ind. Bulletin 110, Part I, pp. 1–80, 1909.

TABLE 60

	Milk yield.				Egg production.				
		A	B	C			A	B	C
Class limits in gallons.	Observed absolute frequency.	Per cent. which mid-point is of mean.	Absolute frequency per 1 per cent. of mean.	Per mille frequency per 1 per cent. of mean.	Class limits (number of eggs).	Observed absolute frequency.	Per cent. which mid-point is of mean.	Absolute frequency per 1 per cent. of mean.	Per mille frequency per 1 per cent. of mean.
6.50– 6.99	2	48.8	0.6	0.4	0– 14	1	6.3	0.08	0.3
7.00– 7.49	6	52.4	1.7	1.2	15– 29	1	18.8	0.08	0.3
7.50– 7.99	7	56.0	1.9	1.3	30– 44	4	31.4	0.32	1.2
8.00– 8.49	7	59.6	1.9	1.3	45– 59	10	44.0	0.80	2.9
8.50– 8.99	5	63.2	1.4	1.0	60– 74	21	56.5	1.67	6.1
9.00– 9.49	24	66.8	6.6	4.6	75– 89	23	69.1	1.83	6.7
9.50– 9.99	28	70.4	7.8	5.4	90–104	35	81.6	2.79	10.1
10.00–10.49	35	74.1	9.7	6.7	105–119	46	94.2	3.66	13.3
10.50–10.99	56	77.7	15.5	10.8	120–134	40	106.8	3.18	11.6
11.00–11.49	68	81.3	18.8	13.0	135–149	35	119.3	2.79	10.1
11.50–11.99	70	84.9	19.4	13.5	150–164	25	131.9	1.99	7.2
12.00–12.49	107	88.5	29.6	20.5	165–179	19	144.4	1.51	5.5
12.50–12.99	118	92.1	32.7	22.7	180–194	8	157.0	0.64	2.3
13.00–13.49	124	95.7	34.3	23.8	195–209	6	169.6	0.48	1.7
13.50–13.99	119	99.3	32.9	22.8	210–224	1	182.1	0.08	0.3
14.00–14.49	133	103.0	36.8	25.5	....	..	.....	....	....
14.50–14.99	87	106.6	24.1	16.7	....	..	.....	....	....
15.00–15.49	102	110.2	28.2	19.6	....	..	.....	....	....
15.50–15.99	78	113.8	21.6	15.0	....	..	.....	....	....
16.00–16.49	76	117.4	21.0	14.6	....	..	.....	....	....
16.50–16.99	43	121.0	11.9	8.3	....	..	.....	....	....
17.00–17.49	43	124.6	11.9	8.3	....	..	.....	....	....
17.50–17.99	28	128.2	7.8	5.4	....	..	.....	....	....
18.00–18.49	20	131.9	5.5	3.8	....	..	.....	....	....
18.50–18.99	22	135.5	6.1	4.2	....	..	.....	....	....
19.00–19.49	14	139.1	3.9	2.7	....	..	.....	....	....
19.50–19.99	5	142.7	1.4	1.0	....	..	.....	....	....
20.00–20.49	6	146.3	1.7	1.2	....	..	.....	....	....
20.50–20.99	3	149.9	0.8	0.6	....	..	.....	....	....
21.00–21.49	2	153.5	0.6	0.4	....	..	.....	....	....
21.50–21.99	2	157.1	0.6	0.4	....	..	.....	....	....
22.00–22.49	..	160.8	....	....	....	..	.....	....	....
22.50–22.99	1	164.4	0.3	0.2	....	..	.....	....	....
Totals....	1441	.....	....	....	....	275	.....	....	....

Using the data as given in columns A and C of Table 60, Fig. 84 has been plotted. The transformation of the absolute frequencies per 1 per cent. of the means given in the B columns to the relative or per mille frequencies of the C columns is necessary in order to bring the two polygons to the same area, since the total observed frequency in one is 1441 and in the other only 275.

The greater relative variability in egg production is apparent.

This method of exhibiting relative variability on an accurately comparative basis may be summarized in the following formulæ:

Let A, B, and C denote the figures in the columns so headed in Tables 59 and 60. Then

$$A = \frac{100\,h}{M}$$

$$B = \frac{Mfd}{100}$$

$$C = \frac{1000\,B}{N}$$

where M is the mean, h is the midpoint of a class interval, f is the absolute frequency of a class, d is the factor by which the class
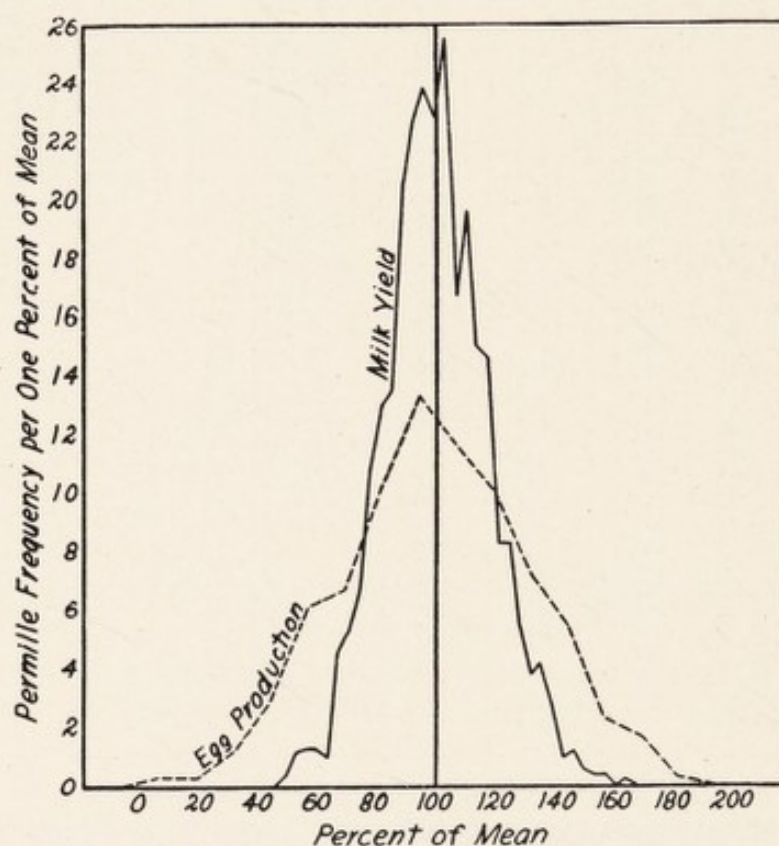
Fig. 84.—Polygons showing the relative variability of cows in milk yield (solid line), and of hens in egg production (dash line). For further explanation see text.

interval must be multiplied to reduce it to unity (*i. e.*, the reciprocal of the class interval), and N is the total absolute frequency.

CONSTANTS MEASURING THE SHAPE OF THE VARIATION CURVE

The Skewness

So far as any *a priori* reason is concerned, it is obvious that variation curves might be symmetric about the mean as a center,

or they might exhibit any degree of asymmetry, or skewness, the variates tailing off farther and more gradually on one side of the curve than on the other. As a matter of fact, a wide range of asymmetry is found in the variation curves of actual natural phenomena. It is important to have an exact measure of the degree or kind of asymmetry exhibited by the curve. Such a constant has been provided by Pearson and called the skewness. Its value, χ denoting skewness is

$$\chi = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}.$$

The larger the value of χ, the greater is the departure of the curve from the symmetric "cocked hat" type. The sign of the expression which indicates the direction of the skewness or asymmetry, whether toward large or toward small values of the variates, is determined generally by giving to $\sqrt{\beta_1}$ the same sign as that of μ_3. There are certain rare types of curve (J-shaped or U-shaped), in which this rule fails. The conventional usage as to the direction of the skewness is as follows: If the curve is skew in the positive direction $(\chi +)$, the median will be smaller than the mean, that is lie to the left of it as ordinarily plotted, and the curve will tail off more on the side of high values. If, on the other hand, the median has larger value than the mean, the curve is negatively skew $(\chi -)$ and tails off more on the side of low values.

In the case of the normal or Gaussian curve $\chi = 0$, the curve being symmetric about the mean. The probable error of χ for the Gaussian curve is

$$\text{P. E. } \chi \text{ (Normal curve)} = \pm .67449 \sqrt{\frac{3}{2N}}.$$

Consequently, unless the skewness χ has a value at least four times as large as this probable error, it cannot safely be asserted that the curve significantly departs from the symmetric Gaussian condition. The probable error of the skewness in the general case may be calculated directly from tables given in Pearson's "Tables for Statisticians and Biometricians."

For the pulse-rate example we have

$$\chi = \frac{.616768 \times 6.469916}{2 (17.349580 - 2.282418 - 9)} = \frac{3.990437}{12.134324} = + .3289.$$

The probable error of the skewness for the normal curve of the same area is

$$\text{P. E. } \chi \text{ (Normal curve)} = \pm .0272.$$

The skewness is, therefore, more than ten times as large as this probable error, and we may safely conclude that this curve of variation in pulse-rate is significantly skew in the positive direction.

Kurtosis

It was shown by Pearson[5] that an important shape characteristic of variation curves is the relative degree of flatness (or peakedness) in the region about the mode, as compared to the condition found in a normal curve. To this attribute of the curve he gave the name *kurtosis*. A curve is said to be *platykurtic* when it is more flat-topped (less peaked) than the Gaussian curve. It is said to be *leptokurtic* when it is less flat topped (more peaked). The Gaussian curve is *mesokurtic*. If η denotes kurtosis, then

$$\eta = \beta_2 - 3.$$

If η is positive (*i. e.*, $\beta_2 > 3$) the curve is leptokurtic. If η is negative ($\beta_2 < 3$) the curve is platykurtic. In the normal or Gaussian curve $\beta_2 = 3$ with a probable error.

$$\text{P. E. } \beta_2 \text{ (normal curve)} = \pm .67449 \sqrt{\frac{24}{N}}.$$

An illustration of a leptokurtic curve is given in Fig. 85 in order that the reader may grasp what is meant by the kurtosis of a curve.

For our pulse-rate example we have:

$$\eta = 3.469916 - 3 = +.4699.$$

The probable error for a normal curve with 924 observations is

$$\text{P. E. } \beta_2 = \pm .1087.$$

The kurtosis is, then, in this case more than four times the probable error, and the curve of pulse-rate variation may be regarded as significantly leptokurtic.

We have now determined the chief physical constants which describe variation. If it is desired to proceed further with the

mathematical analysis what remains to be done is to fit a theoretic curve to the observed distribution, and calculate the ordinates of this curve. The methods for doing this are given in detail in Elder-
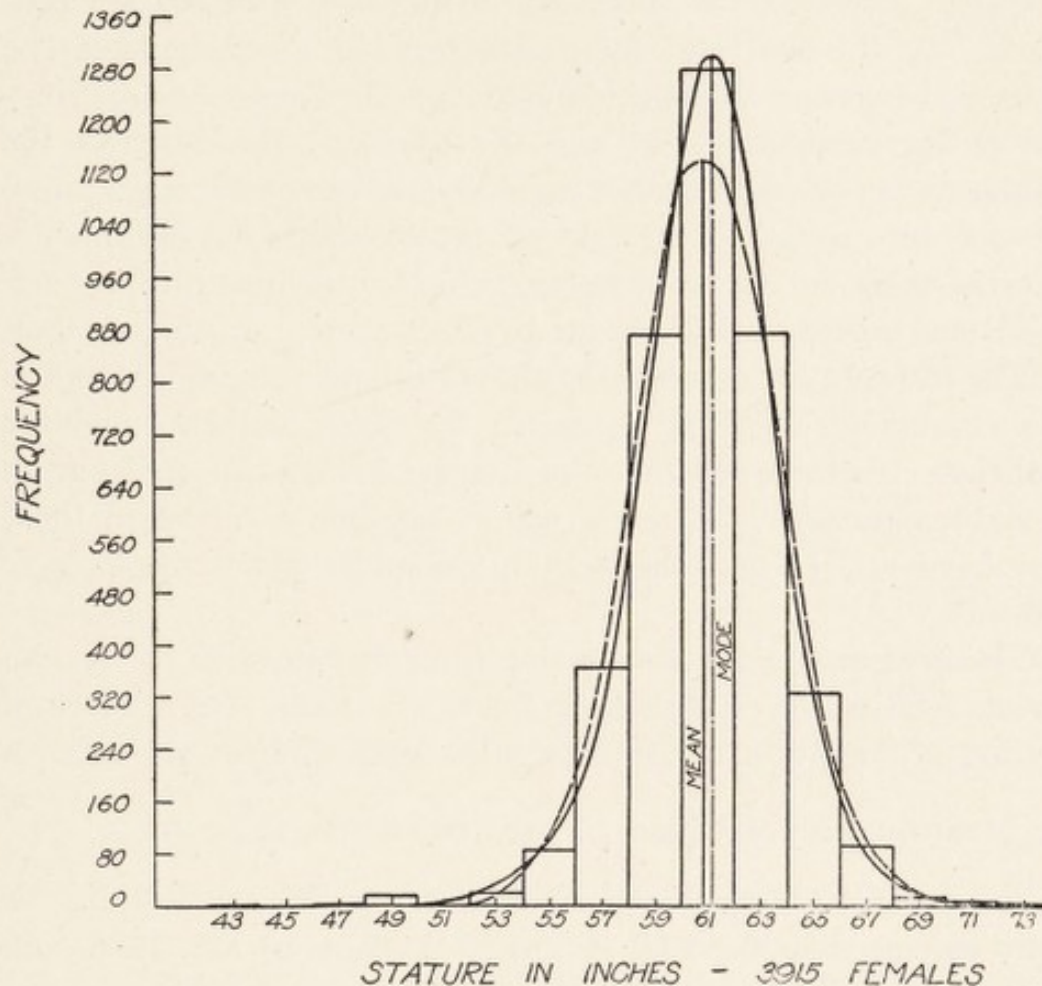
Fig. 85.—Histogram and fitted curves for variation in stature of 3915 Scottish females (insane). The solid curve is the skew curve appropriate to the distribution. The broken curve is the corresponding normal or Gaussian curve. The skew curve is leptokurtic. (Plotted from data of Tocher, Biometrika, 5, pp. 298–350.)

ton's "Frequency Curves and Correlation." Here space is lacking to go further into this phase of the matter.

THE FREQUENCY CONSTANTS OF A VARIABLE $z = f(x_1, x_2)$[*]

It often happens in practical biometric work that one desires to find the frequency constants of a compound character, from a previous knowledge of the constants of the separate components.

[*] Cf. Pearl, R.: Biometrika, vol. 6, pp. 437, 438, 1909; Reed, L. J.: Jour. Washington Acad. Sci., vol. 11, pp. 449–455, 1921.

Thus, for example, one measures the length, the breadth, and the height of each of a series of skulls. He wishes to know at least the mean and the standard deviation of the diametral product ($L \times B \times H$). There are two ways open to find the values of these constants. On the one hand, the length, breadth, and height may be multiplied together for each individual skull, a frequency distribution of the products made, and the constants calculated in the ordinary way; or, on the other hand, by the use of the appropriate formulæ one can deduce straight off the constants for the product knowing those for the components which enter into the product. The latter procedure will obviously effect a great saving of labor.

The formulæ for determining the mean and standard deviation of a character $z = f(x_1, x_2)$ when the same constants and the coefficient of correlation for x_1 and x_2 are known, are well known to mathematicians. They are not so familiar to many of those who have approached the field of biometry along the biologic pathway.

The general method of deducing these formulæ will be clear to anyone who will carefully study Pearson's paper "On a Form of Spurious Correlation which may arise when Indices are used in the Measurement of Organs,"* wherein the formulæ for $z = \dfrac{x_1}{x_2}$ are discussed. The general formulæ for $z = f(x, y)$ will also be found discussed in the Phil. Trans., vol. 187a, p. 278, 1896, and by Reed (*loc. cit.*).

In the formulæ given in Table 61 the various letters have the following meanings:

x_1, x_2, and x_3 the separate characters involved in the compound character z.

m_1, m_2, and m_3 the means of the characters x_1, x_2, and x_3.

σ_1, σ_2, and σ_3 the standard deviations of x_1, x_2, and x_3.

$v_1 = \dfrac{\sigma_1}{m_1}, v_2 = \dfrac{\sigma_2}{m_2}, v_3 = \dfrac{\sigma_3}{m_3}.$ (The v's are the ordinary coefficients of variation *divided by* 100.)

r denotes the coefficient of correlation (see next chapter) between the two characters designated by the subscripts.

* Proc. Roy. Soc., vol. 60, pp. 489–498, 1897.

The table gives the formulæ for the mean and standard deviation of

(a) the sum of two and three variables,
(b) the difference of two variables,
(c) the product of two and of three variables,
(d) the quotient of two variables (index).

In certain of the cases the formulæ are approximations, but very close ones. The nature of the approximations made is indicated in the table.

TABLE 61

Constants of $z = f(x_1, x_2)$.

$z = f(x_1, x_2)$.	Mean of z.	Standard deviation of z.
$z = x_1 + x_2$	$m_1 + m_2$	$\sqrt{(\sigma_1^2 + \sigma_2^2 + 2r_{12}\sigma_1\sigma_2)}$
$z = x_1 + x_2 + x_3$	$m_1 + m_2 + m_3$	$\sqrt{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2 + 2r_{12}\sigma_1\sigma_2 + 2r_{13}\sigma_1\sigma_3 + 2r_{23}\sigma_2\sigma_3)}$
$z = x_1 - x_2$	$m_1 - m_2$	$\sqrt{(\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2)}$
$z = x_1 \cdot x_2$	$m_1 m_2 + r_{12}\sigma_1\sigma_2$	$m_1 m_2[v_1^2 + v_2^2 + 2r_{12}v_1v_2 + v_1^2v_2^2(1 + r_{12})]^{\frac{1}{2}}$,* or approximately $m_1 m_2[v_1^2 + v_2^2 + 2r_{12}v_1v_2]^{\frac{1}{2}}$
$z = x_1 \cdot x_2 \cdot x_3$	$m_1 m_2 m_3[1 + r_{12}v_1v_2 + r_{13}v_1v_3 + r_{23}v_2v_3]$	$m_1 m_2 m_3[v_1^2 + v_2^2 + v_3^2 + 2r_{12}v_1v_2 + 2r_{13}v_1v_3 + 2r_{23}v_2v_3]^{\frac{1}{2}}$ approximately
$z = \dfrac{x_1}{x_2}$	$\dfrac{m_1}{m_2}(1 + v_2^2 - r_{12}v_1v_2)$	$\dfrac{m_1}{m_2}\sqrt{(v_1^2 + v_2^2 - 2r_{12}v_1v_2)}$

* This formula, due to J. F. Tocher, depends on the assumption of normal correlation, see *Biometrika*, vol. iv, p. 320. The approximate value depends on neglecting higher powers of the coefficients of variation. The formula for the mean of the double product (Tocher, *loc. cit.*) is exact. The formula for the mean of the triple product is not exact, any more than the formula for the s. d. of the triple product (see Tocher, *loc. cit.*, p. 321). The formulæ for the mean and s. d. of an index are only true to the lowest powers in v_1 and v_2, and must not be applied if v_1 and v_2 are large. The formulæ for $z = x_1 \pm x_2$, the sums or differences of any number of variables, are exact for both mean and s. d.

CLASS LIMITS

A practical question which frequently arises to vex the beginning statistician in making tables, for the purpose of computing variation constants, is as to how fine the grouping shall be in a table based upon a linear classification. Or, to put it in another way, shall the class limits be narrow or broad? The only general statement which can be made on this point is this: The degree of

fineness of grouping which is permissible depends upon the total magnitude of the experience. It is idle to expand a small observed universe into fine categories, leaving many cells with no frequency or a frequency of only 1. A safe working rule in setting up tables of frequency is: (*a*) to arrange the class limits so as to have from 8 to 15 classes, depending upon the absolute magnitude of the total experience, and (*b*) never to have fewer than 5 classes or more than 20 to 25. As a matter of fact the coarseness or fineness of the elemental class units of grouping makes (within wide limits) extremely little difference in the values of derived biometric constants.

The statement is frequently made, either in comment or criticism upon biometric work, that such work is often caused to take on an unwarranted appearance of precision and exactness by the keeping of a larger number of decimal places in the tabled constants than the character of the original data justifies. The contention is made that under no circumstances whatsoever can any statistical constant be more accurate than the data on which it is based. It is held that if one makes a series of measurements accurate to a tenth of a millimeter, it is a logical absurdity to table the mean and standard deviation deduced from these measurements to hundredths of a millimeter. Not only is this contention made from time to time by biologists, but occasionally even by a mathematician who ought to know better, a fact which, of course, tends strongly to confirm the biologist in his opinion.

The reply which the statistician makes to the criticism that constants cannot be more accurate than the data on which they are based is, in general terms, that the accuracy of a statistical constant depends not alone on the accuracy of the original measurements but also upon the number of such measurements. Further, it is pointed out that, because of this fact, it is possible to deduce from measurements known to be individually inaccurate constants of a high degree of accuracy, provided that the errors in the measurements are unbiased (that is, as often in excess as in defect of the true value) and that there are enough of the data. Finally the statistician contends that the only proper measure of the accuracy of a statistical constant (always assuming that the original data are not collected in a deliberately dishonest or biased manner) is

its "probable error." Unfortunately this statement of the case appears not to carry conviction to the non-statistical worker. It has seemed to the writer that if the assertion made by the statistician regarding the point under discussion is true, it ought to be possible to demonstrate it in such a manner as to carry conviction to anybody.

With this object in view the experiment to be described was tried.[10] Some time ago the writer measured for another purpose the lengths of 450 hens' eggs. The measurements were made with a large steel micrometer caliper manufactured by Browne-Sharpe & Co., reading directly to hundredths of a millimeter. The utmost care was exercised in the making of the measurements; they were all made under the same conditions as to light, temperature, etc.; the caliper was held in a specially constructed stand to get rid of the error arising from expansion and contraction if it is held in the hand; the micrometer screwhead was fitted with a ratchet which mechanically insures that the same pressure shall be exerted on the object in every case; all measurements were made by the same observer who had had considerable experience in close micrometer measuring. The maximum length was the thing measured. There is every reason to believe that these measurements to hundredths of a millimeter are as accurate as it is possible to make them with the instrument used. This being the case all will agree that any statistical constant deduced from them can be held to be accurate to hundredths of a millimeter at least. Now let it be supposed that these eggs had been measured only to the nearest millimeter instead of the nearest hundredth of a millimeter. By how much would the statistical constants deduced from the "millimeter" data differ from those deduced from the "hundredth millimeter data"?

It will be recognized that the problem involved in this question is identical with that of the influence of fineness of grouping in statistical series upon the values of derived constants.

To answer this question it is necessary to calculate some statistical constant for the two sets of data. The mean was chosen as the simplest possible constant. The actual measurements to hundredths of a millimeter were used as one set of data. The "millimeter" data were obtained by discarding the decimals of the

original measurements. In this discarding a record was raised 1 mm. whenever the decimal portion of the original figure was .51 or greater. When the decimal part of the record was .49 or less the integral part stood unchanged. In the 450 measurements there were 6 cases in which the decimal portion of the record was exactly .50. In one-half of these cases the record was raised 1 mm. and in the other half was left unchanged, when the decimals were discarded. This is obviously the only fair way of dealing with such cases since, for example, 51.50 is exactly as near to 51 as to 52.

The original measurements and the "millimeter" data after discarding the decimals were then each added and re-added with a calculating machine. The resulting sums were:

When the measurements were kept to the nearest hundredth of a mm.	When the mensurements were kept to the nearest whole mm.
25,341.95	25,346

Dividing each of these figures by the total number of cases, 450, we get for the means the following:

Mean from "hundredth mm. data"	Mean from "millimeter data"
56.3154	56.3244

The difference between these two figures is .009. That is, there is no difference between the two averages until the third decimal place is reached. To two places of figures both means are 56.32. But this can only mean that the mean or average obtained when the records are made only to the nearest millimeter is more accurate, by two places of decimals, than the data on which it is based.

In interpreting this statement of fact it must not be held to signify that biometric measurements should not be made with the greatest attainable degree of accuracy. Because statistical constants, when the number of cases dealt with is large, are more accurate than the data on which they are based gives no excuse for rough measuring. The reason for this, of course, lies in the principle which actual experience shows to be correct, that the finer and more accurate the measuring, the less chance of the data being unconsciously biased. Statistical constants can only be more accurate than the original data when the data are strictly unbiased.

The "applied psychology" of practical measuring teaches that unconscious bias goes out of the records just in proportion as the measurements are made finer.

SUGGESTED READING

1. Elderton, W. P.: Frequency Curves and Correlation, London (C. and E. Layton), 1906.
2. Sheppard, W. F.: The Calculation of Moments of a Frequency-distribution, Biometrika, vol. 5, pp. 450–459, 1907.
3. Student: The Probable Error of a Mean, Biometrika, vol. 6, pp. 1–25, 1908.
4. Pearl, R.: Biometric Data on Infant Mortality in the United States Birth Registration Area, 1915–18, Amer. Jour. Hygiene, vol. 1, pp. 419–439, 1921. (An illustration of the practical application of the methods of this chapter to a public health problem.)
5. Pearson, K.: Das Fehlergesetz und seine Verallgemeinerungen durch Fechner und Pearson. A Rejoinder, Biometrika, vol. iv, pp. 159–212, 1905. (There is an unfortunate misprint in this paper, p. 174, in which the relations of leptokurtosis and platykurtosis to the value of η are exactly reversed.)
6. Pearson, K.: Tables for Statisticians and Biometricians, Cambridge, 1914. (The introductory text to these tables will be found very useful to the student in connection with the subjects discussed in this chapter.)
7. Venn, J.: On the Nature and Use of Averages, Jour. Roy. Stat. Soc., vol. 54, pp. 429–448, 1891.
8. Yule, G. U.: Introduction to the Theory of Statistics, Chapters VI, VII, and VIII.
9. Pearl, R., and Miner, J. R.: A Biometric Study of the Relative Cell Volume of Human Blood in Normal and Tuberculous Males, Bull. Johns Hopkins Hosp., vol. 40, pp. 3–32, 1927.
10. Pearl, R.: A Note on the Degree of Accuracy of Biometric Constants, Amer. Nat., vol. 43, pp. 238–240, 1909.

CHAPTER XIV

THE MEASUREMENT OF CORRELATION

A PHASE of biometric technic which is of the highest importance and usefulness is that of *correlation* in variation. By the use of this technic complicated problems, which could be efficiently attacked in no other way, may be solved. Pearson defines correlation in the following terms: "Two organs in the same individual, or in a connected pair of individuals, are said to be correlated when, a series of the first organ of a definite size being selected, the mean of the sizes of the corresponding second organs is found to be a function of the size of the selected first organ. If the mean is independent of this size, the organs are said to be non-correlated. Correlation is defined mathematically by any constant, or series of constants, which determine the above function."

This definition will be more intelligible if we look at the matter a little from the standpoint of probability.

THE GENESIS OF CORRELATION

Suppose we carry out some experiments in tossing 12 pennies together, in this manner; make a first toss and record the number of heads, then pick up the pennies and make a second toss. Then enter the results of both tosses in a double entry table. Thus if on the first toss there fell 7 heads and on the second toss 5 heads, these would be entered a frequency of 1 in the cell of Table 62 where the 7 column (first toss) crosses the 5 row (second toss). Continue this process till 500 pairs of throws have been made. The result will be similar to that exhibited in Table 62.*

* This and the following similar tables are taken from Darbishire.[1] His experiments were actually made with dice, but the method of recording was such as to make them precisely equivalent to penny-tossing, and they are capable of more simple statement in the latter form.

Now, plainly, any particular number of heads in the second toss is in this table associated with any given number in the first toss only about as frequently as would be expected from the proportion of that number of heads in the whole experience of first tosses. In other words, the distribution of second toss heads is about random relative to first toss heads. This is what would be expected *a priori* because there is no way in which the result of the first toss

TABLE 62

RELATION BETWEEN THE NUMBER OF HEADS FALLING IN SUCCESSIVE RANDOM TOSSES OF 12 PENNIES TOGETHER

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1						1			1					2
2			1		4				1					6
3			1	4	7	8	5	4	1	1				31
4		4	4	7	9	6	12	5	5					52
5		3	5	13	26	14	14	12	6	1	1			95
6		1	6	15	25	24	28	15	6	2	1			123
7		1	5	7	16	22	15	13	6	1		1		87
8		1	7	15	19	12	6	6						66
9	1	1	2	9	7	6	6		1					33
10			2		1	2								5
11														
12														
Totals	1	9	24	57	112	101	94	62	31	6	2	1		500

(row/column label: Heads in second toss)

r (calc.) $= +.055 \pm .030$ r (theory) $= 0$

can affect the result of the second. The two tosses are *independent* random events. Therefore their results cannot show any sensible quantitative association or correlation with each other.

But now suppose matters to be arranged so that the result of the first toss *can* influence the result of the second. This can easily be done by marking one of the pennies so that it can always be recognized, and then after the first throw *leaving this marked penny*

on the table while the remaining 11 pennies are picked up and tossed at random in order to give, *together with the marked penny left undisturbed,* the second toss. The consequence of this procedure will be that one penny, the marked one, contributes the *same* element (head or tail as the case may be) to *both* tosses. The general result of proceeding in this way is shown in Table 63.

TABLE 63

HEADS IN SUCCESSIVE TOSSES WHERE 11 PENNIES ARE TOSSED IN THE SECOND THROW AND 1 REMAINS AS IT FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

Heads in second toss	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1														
2						2	3	2		1				8
3		1		1	4	4	5	8	4	2				29
4		1		2	3	10	11	6	8	5	2			48
5		1		9	11	13	15	22	11	6				88
6			2	5	8	40	25	32	8	7	1	1		129
7				7	8	13	14	14	14	9	3			82
8			1	2	7	9	12	10	13	2	2	1		59
9					5	10	8	12	7	5				47
10						1	1	3	1	2				8
11				1					1					2
12														
Totals		3	3	26	47	102	94	109	67	39	8	2		500

r (calc.) $= +.073 \pm .030$ r (theory) $= .083$

This Table 63 is, in theory, not quite like Table 62, although to the eye it still is very similar.

If the process be now continued, leaving down successively more and more of the pennies and having them pass over undisturbed from first to second toss, we shall get the results shown in the tables which follow. Table 64 shows the result of marking 2 pennies and leaving them down; Table 65, of marking 3 pennies and leaving them down, and so on up to all 12 pennies.

TABLE 64

HEADS IN SUCCESSIVE TOSSES WHERE 10 PENNIES ARE TOSSED IN THE SECOND THROW AND 2 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1						1								1
2			2			1								3
3			2	2	4	7	7	3						25
4				5	7	7	19	10	4	5				57
5		1	4	3	10	20	26	21	9	3				97
6			1	3	6	30	26	18	11	6	2	3		106
7			1	2	12	13	15	17	30	3	3			96
8		1	1	4	8	7	10	16	10	6		1		64
9			2	6	2	9	8	6	2					35
10						2	1	3	4	2				12
11							1	1		1				3
12						1								1
Totals		2	9	25	53	91	114	97	74	28	5	4		500

(Heads in second toss on vertical axis.)

$$r \text{ (calc.)} = +.194 \pm .029 \qquad r \text{ (theory)} = .167$$

TABLE 65

HEADS IN SUCCESSIVE TOSSES WHERE 9 PENNIES ARE TOSSED IN THE SECOND THROW AND 3 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1				1		1								2
2							6	1						7
3		1	1	5	2	2	4	5						20
4			1	8	6	21	16	6	6					64
5			4	3	12	15	23	22	9	3	1			92
6		1	10	16	17	23	28	22	5	1				123
7			1	4	9	17	18	24	16	5	3			97
8			1	5	6	10	14	8	7	2	1			54
9					4	3	9	6	6	2				30
10						1	1	1	4	3				10
11								1						1
12														
Totals		2	7	31	55	82	111	108	71	25	7	1		500

(Heads in second toss on vertical axis.)

$$r \text{ (calc.)} = +.278 \pm .028 \qquad r \text{ (theory)} = .250$$

24

369

TABLE 66

Heads in Successive Tosses Where 8 Pennies Are Tossed in the Second Throw and 4 Remain as They Fell in the First Throw of 12 Together

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1					1	1								2
2			1	1	1	2	1	1						7
3			2	4	3	4	5	2						20
4			1	2	8	18	8	12	6	1	1			57
5			3	8	16	16	19	21	7	5	2			97
6		1	1	5	19	25	25	20	12	2				110
7			2	1	6	22	17	32	17	12	3			112
8					5	6	16	18	14	7	2			68
9						3	2	6	5	2				18
10							3		2	2				7
11								1	1					2
12														
Totals		1	10	21	59	97	96	113	64	31	8			500

r (calc.) $= +.342 \pm .026$ r (theory) $= .333$

TABLE 67

Heads in Successive Tosses Where 7 Pennies Are Tossed in the Second Throw and 5 Remain as They Fell in the First Throw of 12 Together

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1					1	1								2
2			3	1	5	1	1							11
3			3	3	8	4	4	4						26
4			3	6	9	21	14	10	5	1				69
5				4	11	23	21	15	9					83
6			1	3	9	18	27	29	16	3	2	1		109
7			1	2	5	14	24	28	10	7	4			95
8				1	5	9	10	18	14	4	2			63
9						2	9	13	4	3				31
10					1		2		2	3	1	1		10
11								1						1
12														
Totals			11	20	54	93	112	118	60	21	9	2		500

r (calc.) $= +.432 \pm .025$ r (theory) $= .417$

370

TABLE 68

HEADS IN SUCCESSIVE TOSSES WHERE 6 PENNIES ARE TOSSED IN THE SECOND THROW
AND 6 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1			1	1	1									3
2			1		2	3	2							8
3			2	3	5	6	2	6						24
4			5	9	8	11	16	7	6	1				63
5			2	5	17	24	19	25	11	2				105
6			1	5	14	25	24	24	17	4	3			117
7				2	2	13	16	27	12	4	2			78
8					2	7	13	22	14	5	3			66
9						3	5	6	9	5	2			30
10								2	1	2				5
11									1					1
12														
Totals			12	25	51	92	97	119	71	23	10			500

(Row label: Heads in second toss.)

$$r \text{ (calc.)} = +.449 \pm .024 \qquad r \text{ (theory)} = .500$$

TABLE 69

HEADS IN SUCCESSIVE TOSSES WHERE 5 PENNIES ARE TOSSED IN THE SECOND THROW
AND 7 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1			1											1
2			1	2	4	2								9
3			4	2	6	5	4	3						24
4			1	7	10	19	13	8	1					59
5			1	5	16	14	24	14	2	1				77
6				3	13	17	28	22	9	4	1			97
7				1	3	15	26	40	18	8	2			113
8						8	14	16	16	12	3			69
9						2	3	10	10	9	4			38
10							4	2	3	2		1		12
11							1							1
12														
Totals			7	21	52	82	117	115	59	36	10	1		500

(Row label: Heads in second toss.)

$$r \text{ (calc.)} = +.578 \pm .020 \qquad r \text{ (theory)} = .583$$

TABLE 70

Heads in Successive Tosses Where 4 Pennies Are Tossed in the Second Throw and 8 Remain as They Fell in the First Throw of 12 Together

Heads in first toss.

Heads in second toss.	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1														
2			4	2	3	2								11
3			2	8	9	7	4							30
4			3	1	12	20	9	5	2					52
5				4	21	30	26	13	4					98
6				4	12	30	36	19	10	7	1			119
7				1	1	15	29	27	13	7	1			94
8						1	8	19	13	10	1			52
9						1	1	4	13	5	5	3	1	33
10								1	3	3	3			10
11									1					1
12														
Totals			9	20	58	106	113	88	59	32	11	3	1	500

r (calc.) $= +.676 \pm .016$ r (theory) $= .667$

TABLE 71

Heads in Successive Tosses Where 3 Pennies Are Tossed in the Second Throw and 9 Remain as They Fell in the First Throw of 12 Together

Heads in first toss.

Heads in second toss.	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1			1	1										2
2			2	5	1	1								9
3				5	7	3	1							16
4			1	8	18	19	5	1						52
5				6	17	30	32	13	1					99
6				1	10	18	34	26	10	1				100
7					4	17	26	30	18	7				102
8							7	28	16	11	5			67
9							3	6	15	9	7	1		41
10								1		4	3	2		10
11											1		1	2
12														
Totals			4	26	57	88	108	105	60	32	16	3	1	500

r (calc.) $= +.765 \pm .012$ r (theory) $= .750$

372

TABLE 72

HEADS IN SUCCESSIVE TOSSES WHERE 2 PENNIES ARE TOSSED IN THE SECOND THROW AND 10 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

Heads in second toss	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0	1													1
1		1												1
2			2	5										7
3		1	3	8	9	3								24
4			2	10	18	19	6							55
5				1	24	43	32	10						110
6					4	22	37	24	6					93
7						6	27	39	19	5				96
8							9	17	24	9	1			60
9								10	14	11	7			42
10									1	6	2	1		10
11											1			1
12														
Totals	1	2	7	24	55	93	111	100	64	31	11	1		500

r (calc.) $= +.840 \pm .009$ r (theory) $= .833$

TABLE 73

HEADS IN SUCCESSIVE TOSSES WHERE 1 PENNY IS TOSSED IN THE SECOND THROW AND 11 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

Heads in first toss.

Heads in second toss	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1														
2			5	3										8
3			2	7	10									19
4				7	21	23								51
5					19	44	33							96
6						22	56	30						108
7							31	49	16					96
8								25	38	13				76
9									15	18	5			38
10										1	5	1		7
11											1			1
12														
Totals			7	17	50	89	120	104	69	32	11	1		500

r (calc.) $= +.910 \pm .005$ r (theory) $= .917$

TABLE 74

HEADS IN SUCCESSIVE TOSSES WHERE NO PENNY IS TOSSED IN THE SECOND THROW
AND 12 REMAIN AS THEY FELL IN THE FIRST THROW OF 12 TOGETHER

	Heads in first toss.													
	0	1	2	3	4	5	6	7	8	9	10	11	12	Totals
0														
1		3												3
2			8											8
3				24										24
4					63									63
5						104								104
6							117							117
7								81						81
8									64					64
9										30				30
10											5			5
11												1		1
12														
Totals		3	8	24	63	104	117	81	64	30	5	1		500

(Row label axis: Heads in second toss.)

r (calc.) $= 1$ r (theory) $= 1$

In this series of tables is seen the genesis of correlation. In Table 62 the results of the first toss have no influence on the results of the second. There is no correlation between them. In Table 74 the results of the first toss completely determine, *or cause*, the results of the second. This gives perfect correlation—or, in this particular case, causation—between the two.

In all the tables the diagonal lines cut off the cells in which events cannot possibly happen.

Just below each of these tables there have been placed two values of r (the coefficient of correlation, which is discussed in detail in a later section). The first one of these is the value calculated directly from the table itself. The other is the theoretical value which is a consequence of the number of pennies left down from the first toss, according to the theory of probability.*

* Rietz, H. L.: Urn Schemata as a Basis for the Development of Correlation Theory, Annals of Math., vol. 21, pp. 306–322, 1920.

THE CORRELATION TABLE AND REGRESSION

Suppose one wished an answer to this question: What quantitative relation, if any, exists between brain weight and skull length? One knows from general anatomic relations that there must be some association between these phenomena. A long head and a heavy brain are often observed together in the same individual. But in a statistical sense, how close is this association in general? What is its quantitative degree of intensity?

Quite obviously the way to start getting an answer to this question is to collect information, on as many persons as possible, as to the brain weight and the skull length in the same individual. Having this information, one may set up a table like Table 75. This table is taken from a paper by the present writer,* the original data having been collected by Matiegka.†

TABLE 75

CORRELATION BETWEEN BRAIN-WEIGHT AND SKULL LENGTH. BOHEMIAN MALES, TWENTY TO FIFTY-NINE YEARS OF AGE

Skull length (mm.)	Brain-weight (grams).									Totals.	Midpoints of class ranges of skull length.	Means of brain-weight arrays.
	1000–1099	1100–1199	1200–1299	1300–1399	1400–1499	1500–1599	1600–1699	1700–1799	1800–1899			
155–159....	..	..	1	1	..	..	..	..	..	2	157.5	1300
160–164....	..	..	2	6	4	2	..	..	..	14	162.5	1393
165–169....	1	..	9	10	18	3	1	..	..	42	167.5	1386
170–174....	..	..	5	19	28	11	4	1	..	68	172.5	1440
175–179....	..	..	4	19	29	23	4	..	..	79	177.5	1455
180–184....	..	..	..	10	19	23	8	1	..	61	182.5	1502
185–189....	..	..	..	1	2	12	4	..	..	19	187.5	1550
190–194....	..	..	..	..	1	2	3	4	..	10	192.5	1650
195–199....	..	..	..	..	..	1	1	..	2	4	197.5	1725
Totals. .	1	..	21	66	101	77	25	6	2	299		
Midpoints of class ranges of brain-weight...	1050	1150	1250	1350	1450	1550	1650	1750	1850			
Means of skull length arrays....	167.5	..	169.6	173.8	175.0	179.7	182.1	187.5	197.5			

* Pearl R.: Biometrical Studies on Man. I. Variation and Correlation in Brain-weight, Biometrika, vol. 4, pp. 13–104, 1905.

† Matiegka, H.: Über das Hirngewicht, die Schädelkapacität und die Kopfform. Sitzber. des kön. böhmischen Gesellsch. d. Wiss., Math.-Nat. Cl., Jahrg., 1902, No. xx, pp. 1–75.

A table of this sort is known as a *correlation table*. It is a table of double entry, which enables one to read off, for example, that there were in the total experience 18 persons who had a brain-weight of 1400–1499 grams, and a skull length of 165–169 mm. It is made up of a series of rows and columns, each of which is, of itself, a frequency distribution. Each row and each column is called technically an *array*. Thus there is an array of skull lengths (a column) associated with a midrange brain-weight of 1450, and similarly there is an array of brain-weights (a row) associated with a skull length of 172.5, and so on.

Geometrically the table may be represented best as a surface. Call brain-weight the *x* coördinate, and skull length the *y* coördinate. Then the frequencies in each cell must be represented by the *volumes* (instead of areas as in simple frequency distributions) of rectangular solids with one end of each one covering the cell on which it stands, and their heights reading on the *z* coördinate. Now suppose the tops of these cells to be connected with each other and covered by a smooth surface. The general shape of the resulting surface will usually be quite strikingly similar to that of the "tin hats" worn by the United States soldiers in the late war.

Each array may be treated biometrically as an independent frequency distribution, and the mean, standard deviation, etc., determined. The first step in this direction leads to the array means given on the margins of Table 75. These array means, taken in connection with the midpoints of the class ranges of the other variable set next to them, at once bring out an interesting point. It is that as the midpoints of the brain-weight class range (let us say) increase as we pass from left to right, there is a slightly irregular but still perfectly definite tendency for the means of the corresponding skull length arrays to increase.

This fact can be made more apparent graphically as seen in Fig. 86.

The lines formed by plotting the means of the arrays are called *observed regression lines*, regression being a term introduced into statistical usage by Galton. The manner in which the calculated regression lines are derived will be explained in the next section.

It is apparent from Fig. 86 that the slope of the regression lines gives a means of measuring the degree of correlation or association of variation between the variables. For suppose AB to be rotated about 0 as an axis until it exactly coincided with YY, and CD to be rotated about 0 until it exactly coincided with XX. Then there
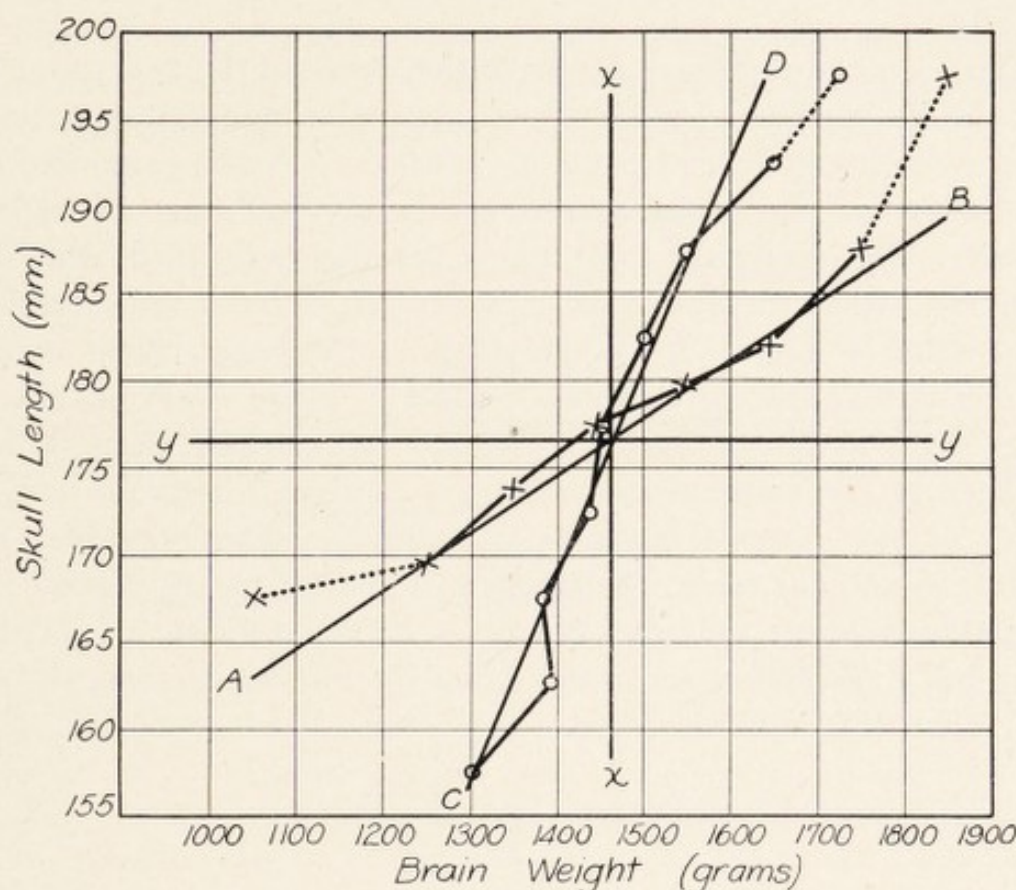
Fig. 86.—Observed and calculated regressions for brain-weight and skull length from Table 75. The crosses are the means of the observed skull length arrays (observed regression of skull length on brain-weight). AB is the calculated regression line of skull length on brain-weight. The circles are the means of the observed brain-weight arrays (observed regression of brain-weight on skull length). CD is the corresponding calculated regression line. XX gives the location on the brain-weight scale of the mean of all 299 brain-weights. YY gives the mean of all skull lengths on the skull length scale.

would be no increase in brain-weight associated with an increase in skull length, or *vice versa*. Actually the method used for measuring correlation, as will be shown in the next section, does make use of just this principle.

THE MEASUREMENT OF SIMPLE CORRELATION WITH LINEAR REGRESSION. THE CORRELATION COEFFICIENT

In the simplest and fundamental case correlation between two variables is measured by a coefficient

$$r_{12} = \frac{S(x_1 x_2)}{N\sigma_1 \sigma_2},$$

where r_{12} is the coefficient of correlation between the two variables X_1 and X_2, of which σ_1 and σ_2 are the respective standard deviations and N is the number of pairs of variates. S denotes summation, and x_1 and x_2 are deviations from the means of X_1 and X_2 respectively. This coefficient may take any value between 0, which is the result when there is no correlation at all between the variables, and either $+1$ or -1. When either of the latter values occurs it means that the correlation is perfect, *i. e.*, for every change in one of the variables there is a definite and constant proportional change in the value of the other. A positive correlation means that as one variable increases in value the other variable also increases and *vice versa*. A negative correlation means that as one variable increases the other decreases. The coefficient of correlation has a probable error, which takes the following value:

When N is say 25 or more

$$P.\ E._r = .67449\frac{1-r^2}{\sqrt{N}}.$$

When it is desired to test whether an observed correlation coefficient is significantly different from zero, r in the above formula should be put $= 0$ in calculating the P. E. For very small numbers ($N < 25$) special caution must be used in estimating the reliability of a correlation coefficient. Here the section in R. A. Fisher's "Statistical Methods for Research Workers" on "The Significance of an Observed Correlation" (pp. 157–161) will be found helpful.

The method of calculating the coefficient of correlation r will now be described. The method here given is a short one worked out as to its details in this laboratory. In principle it is the same as short methods which have been described by other workers, but possesses some advantages in practical computation

over any that have come to the writer's notice. For a detailed account of the arithmetic of the old direct product-moment method of determining a coefficient of correlation, see Yule.[2]

As an example we may take Table 75 giving the correlation between skull length and brain-weight. This table is repeated, with the arithmetic of the first steps in the computations, as Table 76.

First we may consider the notation used, which is identical with that in the preceding chapter on the measurement of variation. The marginal total arrays of the table are designated

Z_x = frequency in the several brain-weight classes.

Z_y = frequency in the several skull length classes.

TABLE 76

SHOWING THE STEPS IN THE CALCULATION OF A CORRELATION COEFFICIENT

		Brain-weight (grams).									Totals, Z_y	y	$Z_y y$	$Z_y y^2$	$z_{xy}x$	$z_{xy}xy$
		1000–1099	1100–1199	1200–1299	1300–1399	1400–1499	1500–1599	1600–1699	1700–1799	1800–1899						
Skull length (mm.)	155–159	...	...	1	1	...	...	...	...	...	2	−3	− 6	18	− 3	+ 9
	160–164	...	...	2	6	4	2	...	...	...	14	−2	− 28	56	− 8	+ 16
	165–169	1	...	9	10	18	3	1	...	...	42	−1	− 42	42	−27	+ 27
	170–174	...	...	5	19	28	11	4	1	...	68	0	0	0	− 7	0
	175–179	...	...	4	19	29	23	4	...	...	79	1	79	79	+ 4	+ 4
	180–184	...	...	...	10	19	23	8	1	...	61	2	122	244	+32	+ 64
	185–189	...	...	...	1	2	12	4	...	...	19	3	57	171	+19	+ 57
	190–194	...	...	...	...	1	2	3	4	...	10	4	40	160	+20	+ 80
	195–199	...	...	...	...	...	1	1	...	2	4	5	20	100	+11	+ 55
Totals Z_x....		1	...	21	66	101	77	25	6	2	299	...	+242	870	+41	+312
x........		−4	−3	−2	−1	0	1	2	3	4						
$Z_x x$.....		−4	...	−42	−66	0	77	50	18	8	+41					
$Z_x x^2$....		16	...	84	66	0	77	100	54	32	429					

x denotes deviations, in class units of 100 grams each, of each brain-weight class from the arbitrary origin ($x = 0$) at the mid-point (1450) of the brain-weight class 1400–1499.

y denotes deviations in class units of 5 mm. each, of each skull length class from its arbitrary origin at 172.5 mm.

We need as the first step to get the means and standard devia-

tions for the two variables. Proceeding just as in Chapter XIII, we have:

$$v_{1x} = \frac{S\,(Z_x x)}{S\,(Z_x)} = \frac{41}{299} = .137124$$

$$v_{2x} = \frac{S\,(Z_x x^2)}{S\,(Z_x)} = \frac{429}{299} = 1.434783$$

Omitting Sheppard's corrections for the sake of simplicity, we then have

$$\pi_{2x} = 1.434783 - (.137124)^2 = 1.415980,$$

whence

$$\sigma_x = \sqrt{\overline{\pi_{2x}}} = 1.189950 \text{ in class units.}$$

We then have

Mean brain-weight $= 1450 + (100 \times .1371) = 1463.71 \pm 4.64$ grams.

Standard deviation (in brain-weight) $= 100 \times 1.18995 = 119.00 \pm 3.28$ grams.

Similarly for skull length we have:

$$v_{1y} = \frac{S\,(Z_y y)}{S\,(Z_y)} = \frac{242}{299} = .809365$$

$$v_{2y} = \frac{S\,(Z_y y^2)}{S\,(Z_y)} = \frac{870}{299} = 2.909699$$

$$\pi_{2y} = 2.909699 - (.809365)^2 = 2.254627$$

$$\sigma_y = \sqrt{2.254627} = 1.501542 \text{ in class units.}$$

Mean skull length $= 172.5 + (.809365 \times 5) = 176.55 \pm .29$ mm.

Standard deviation (in skull length) $= 5 \times 1.501542 = 7.51 \pm .21$ mm.

Now in proceeding to get the coefficient of correlation we may first break it up into this form,

$$r_{12} = \frac{S\,(xy)}{N \sigma_1\,\sigma_2} = \frac{S\,(xy)}{N} \times \frac{1}{\sigma_1\,\sigma_2},$$

and determine first $S\,(xy)/N$. Call $\frac{S\,(xy)}{N} = A$, and $\sigma_1\,\sigma_2 = B$.

Suppose the row designated x at the bottom of the table and surrounded by a heavy line frame to be movable. Then suppose it to be moved up on the table till it rests just under the first brain-weight array (the first frequency row in the table, corresponding to a skull length of 157.5). Then multiply each cell frequency (z_{xy}) in that array by the number in the x row which

falls directly under that cell, *having regard to the sign of the x always*. We shall have

$$1 \times (-2) = -2$$
$$1 \times (-1) = -1$$
$$\text{Sum} = -3$$

This -3 is the first entry in the marginal column to the right of the table headed $z_{xy}x$.

Now slide the movable x row down one array till it is just below the brain-weight array corresponding to skull-length 162.5, and repeat the same process as before. We have:

$$2 \times (-2) = -4$$
$$6 \times (-1) = -6$$
$$4 \times 0 \quad = \quad 0$$
$$2 \times (+1) = \quad 2$$
$$\text{Sum} = -8$$

This -8 is the second entry in the $z_{xy}x$ column.

Let this process be repeated for each of the brain-weight arrays. The results will be those seen in the $z_{xy}x$ column. When completed the algebraic sum of this is found to be $+41$. This will be seen to agree with the sum of the *row* at the bottom of the table headed $Z_x x$. This agreement between these two sums must always be exact, and furnishes an important check on the correctness of the work. If they do not agree a mistake has been made and one should proceed no farther till it has been found and corrected.

Now what we have so far is the product of each elemental cell frequency (z_{xy}) by the deviation of its position from the arbitrary origin of the x variable. The next step is to multiply in the deviation of the cell from the arbitrary origin of the y variable. This is done in the last column to the right, headed $z_{xy}xy$.

Thus we have

$$(-3) \times (-3) = +9$$
$$(-2) \times (-8) = +16$$
$$(-1) \times (-27) = +27$$
$$0 \quad \times (-7) = \quad 0$$
$$(+1) \times (+4) = +4$$
$$\text{and so on.}$$

The sum of this column ($S(z_{xy}xy)$) is the product moment of the table, referred to the arbitrarily chosen axes of origin. We

need, just as with simple frequency distributions, to transfer this to the mean as origin, and the method of doing so is in principle just the same, namely, by shifting its value by an amount equal to the product of the two first moments (ν_{1x} and ν_{1y}) about the arbitrary origin. Remembering that, in the notation used above,

$$r_{12} = \frac{S(xy)}{N\sigma_1\sigma_2} = \frac{A}{B},$$

we have the rule for transferring to the mean that

$$A = \frac{S(z_{xy}xy)}{N} - \nu_{1x}\,\nu_{1y}.$$

In the present example

$$A = \frac{S(z_{xy}xy)}{N} - \nu_{1x}\nu_{1y} = \frac{+312}{299} - (.137124 \times .809365)$$
$$= 1.043478 - .110983$$
$$= +.932495$$

Remembering always that we are computing in terms of class units of grouping

$$B = \sigma_1\sigma_2 = 1.189950 \times 1.501542 = 1.786760$$

Whence finally

$$r_{12} = \frac{+.932495}{1.786760} = +.522 \pm .028.$$

The probable error of $\pm .028$ is the one to be used in comparing this coefficient $+.522$ with other observed correlation coefficients. If one wished to test whether $+.522$ is itself significantly different from zero the proper probable error for the purpose is $\pm .039$.

While it has taken a good deal of space to describe this process, it is, in fact, a very simple matter to calculate a correlation coefficient, and by the method here described takes but a short time.

Let us consider now the *regression coefficients*. These are two quantities defined as follows:

$$b_1 = r_{12}\frac{\sigma_1}{\sigma_2}$$

$$b_2 = r_{12}\frac{\sigma_2}{\sigma_1}$$

These quantities measure the slopes of the regression lines (cf. Fig. 86 *supra*).   That is

$$\bar{x} = b_1 \, y$$
$$\bar{y} = b_2 \, x$$

Let subscript 1 denote the brain-weight or x variable, and subscript 2 denote the skull length or y variable, and $\bar{x}$ denote the deviation of the mean of a brain-weight array from the mean brain-weight of the whole sample, and $\bar{y}$ the deviation of the mean of a skull length array from the mean skull length of the whole sample.

Then in our example

$$b_1 = r_{12} \frac{\sigma_1}{\sigma_2} = .521892 \frac{118.995}{7.508} = 8.272$$

Whence

$$\bar{x} = 8.272 \, y.$$

But $\bar{x}$ and y are deviations from the means of brain-weight and skull length respectively.   We shall do better to work with absolute values rather than deviations.   Doing so, we have,

$$\bar{x} = (\overline{X} - 1463.7)$$
$$y = (Y - 176.5)$$

So then,

$$\overline{X} - 1463.7 = 8.272 \, (Y - 176.5).$$

Simplifying, we get

Mean brain-weight (in grams) $= 3.7 + 8.272$ skull length (in mm.).

This is the equation of the regression line *CD* of Fig. 86.   It expresses the regression of brain-weight on skull length.

Proceeding in the same way for the regression of skull length on brain-weight we have

$$b_2 = r_{12} \frac{\sigma_2}{\sigma_1} = .521892 \frac{7.508}{118.995} = .033.$$

$$\bar{y} = .033 \, x$$
$$\overline{Y} - 176.5 = .033 \, (X - 1463.7)$$

Mean skull length (in mm.) $= 128.2 + .033$ brain-weight (in grams).

This is the equation of the line AB in Fig. 86.

This completes the essential mathematical treatment of simple two-variable correlation with linear regression.

ILLUSTRATION OF CORRELATION IN HUMAN MATERIAL

In order to give some idea of the extent to which various human characteristics are correlated Table 77 is presented. It gives the values of the coefficient of correlation for a number of representative characters. It represents only a small fraction of the large number of correlations for human characters which are now known. In considering the values in this table it must be remembered, from principles already stated, that if a correlation coefficient is not 4 or more times its probable error it cannot be asserted to be *certainly* different from zero, though if it is 3 times the probable error it is *probably* so.

TABLE 77

CORRELATION IN MAN

Correlated Characters.	Coefficient of correlation.
Age (adults) and temperature (oral) [1]	$-.150 \pm .022$
Age (adults) and pulse rate [1]	$+.121 \pm .022$
Age (adults) and respiration rate [1]	$+.077 \pm .022$
Age (adults) and body weight [1]	$+.136 \pm .030$
Temperature (oral) and pulse rate [1]	$+.288 \pm .020$
Temperature (oral) and respiration rate [1]	$+.142 \pm .022$
Temperature (oral) and height [1]	$+.003 \pm .022$
Temperature (oral) and body weight [1]	$+.043 \pm .022$
Pulse rate and respiration rate [1]	$+.060 \pm .022$
Pulse rate and height [1]	$-.078 \pm .022$
Pulse rate and body weight [1]	$+.114 \pm .022$
Respiration rate and height [1]	$-.144 \pm .022$
Respiration rate and body weight [1]	$-.089 \pm .022$
Corrected death rates from (a) cancer of the liver, and (b) cancer of the stomach (Switzerland) [2]	$+.161 \pm .140$
Corrected death rates from (a) cancer of the stomach, and (b) cancer of the rectum and intestines (Switzerland) [2]	$+.263 \pm .134$
Occupation and cancer mortality (occupied and retired males, 1900–2, weighted) [3]	$+.40 \pm .06$
Weight and length of infants at birth (males) [4]	$+.644 \pm .012$
Body weight and height (adult males) [4]	$+.486 \pm .016$
Strength of pull and height (adult males) [4]	$+.303 \pm .019$
Strength of pull and body weight (adult males) [4]	$+.545 \pm .015$
Length of first joint of forefinger in (a) right hand, and (b) left hand [5]	$+.925 \pm .004$
Stature in (a) brother and (b) sister [6]	$+.375 \pm .017$
Cephalic index in (a) brother, and (b) sister [6]	$+.340 \pm .050$
Birth rate and infant death rate (London, 1901) [7]	$+.51 \pm .10$
Birth rate and poverty rate [8]	$+.420 \pm .047$
Infant mortality and artificial feeding rate [8]	$+.760 \pm .029$
Heart weight and body weight [9]	$+.65 \pm .04$
Heart weight and kidney weight [9]	$+.56 \pm .05$
Heart weight and liver weight [9]	$+.52 \pm .06$
Heart weight and brain weight [9]	$+.08 \pm .08$
Obstetric conjugate and inter-crests diameters of pelvis [10]	$+.17 \pm .04$
Obstetric conjugate and inter-spines diameters of pelvis [10]	$+.13 \pm .05$
Obstetric conjugate and transverse diameters of pelvis [10]	$+.07 \pm .05$
Obstetric conjugate and diagonal conjugate diameters of pelvis [10]	$+.91 \pm .01$
Obstetric conjugate and antero-posterior diameters of pelvis [10]	$+.30 \pm .04$
Duration of life of (a) father, and (b) adult son [11]	$+.135 \pm .021$
Duration of life of (a) father and (b) minor son [11]	$+.087 \pm .022$
Duration of life of (a) father, and (b) adult daughter [11]	$+.130 \pm .020$
Duration of life of (a) mother, and (b) adult son [11]	$+.131 \pm .019$
Duration of life of (a) mother, and (b) adult daughter [11]	$+.149 \pm .020$
Duration of life of (a) adult brother and (b) adult brother [11]	$+.285 \pm .020$
Duration of life of (a) adult sister and (b) adult sister [11]	$+.332 \pm .019$
Vaccination and recovery from smallpox [12]	$+.656 \pm .009$
Lung capacity and body weight (age 19, males) [13]	$+.62 \pm .02$
Number of decayed teeth and use of tooth-brush (boys) [14]	$+.074 \pm .030$
Mean age at death of (a) husband, and (b) wife [15]	$+.224 \pm .022$

[1] Whiting, M. H.: Biometrika 11:11, 1915–17.

[2] Brown, J. W., and Lal, Mohan: J. Hyg. 14:192, 1914.

[3] Greenwood, M., and Wood, Frances: Proc. Roy. Soc. Med. 8 (Sect. Epidemiology):119, 1914.

[4] Pearson, Karl: Proc. Roy. Soc. Lond. 66:25, 1899–90.

[5] Whiteley, M. A., and Pearson, Karl: Proc. Roy. Soc. Lond. 65:130, 1899.

[6] Fawcett, Cicely D., and Pearson, Karl: Proc. Roy. Soc. Lond. 62:415, 1898.

[7] Heron, David: On the Relation of Fertility in Man to Social Status, London, Dulau & Co., 1906.

[8] Greenwood, M.: Eugenics Rev. 4:248, 1912–1913.

[9] Greenwood, M., and Brown, J.: Biometrika 9:478, 1913.

[10] De Souza, D. H.: Biometrika 9:490, 1913.

[11] Beeton, Mary, and Pearson, Karl: Biometrika 1:60, 1901–02.

[12] Macdonell, W. R.: Biometrika 1:376, 1901–1902.

[13] Schuster, E.: Biometrika 8:51, 1911–12.

[14] Rock, Frank: Biometrika 8:238, 1911–12.

[15] Assortive Mating in Man, Biometrika 2:485, 1902–03.

SKEW CORRELATION AND NON-LINEAR REGRESSION. THE CORRE-LATION RATIO

So far we have dealt only with two-variable correlation where the means of the arrays fall upon a straight line, within the errors of sampling. It will be at once obvious to any biologist that there are many cases in nature in which this condition is not at all closely approached. An example is the correlation between a bodily characteristic and age during the growing period of the organism; the data, in short, which lead to a growth curve.

Pearson[3] has dealt with non-linear regression under the designation of *skew correlation*, and devised a satisfactory method of measuring the correlation or association in such cases. In the first place it is apparent that such a constant as

$$r_{12} = \sqrt{b_1 \cdot b_2}$$

fails wholly in such a case as that of a growth curve, because b_1 and b_2 no longer have the simple meaning they did in linear regression.

Pearson, therefore, proposed a new constant, the *correlation ratio*, conventionally denoted by the Greek letter eta (η). Let us now try to explain, with a minimum of mathematical notation, just what this constant means.

Going back to Table 75 it must be apparent to anyone that each array of such a table may be treated biometrically as a separate frequency distribution. Thus the array of brain-weights associated with skull lengths 170–174 mm. is as follows:

Brain-weight.	Frequency.
1200–1299	5
1300–1399	19
1400–1499	28
1500–1599	11
1600–1699	4
1700–1799	1
Total	68

For this, or any other similar array distribution, we can, if it is desired, compute in the regular way the mean and the standard deviation. The former will measure the type *of the array*, and the latter the variability *of the array*. Suppose we calculate in this

way the standard deviation, measuring the variability, of each brain-weight array in the table. We shall then have a series of 9 standard deviations. If we add these together and divide by 9 we shall have as the result the *unweighted* mean variability of brain-weight arrays associated with particular skull lengths. If we multiply each standard deviation of an array by the total frequency in that array, add up the results and divide by 299, the sum of all the frequencies in all arrays, the result will be the *weighted* mean variability of arrays of brain-weight associated with particular skull lengths.

Plainly, from mere inspection of the table, this weighted mean variability of brain-weight *arrays* will be *smaller* than the variability of brain-weight in general over the whole table, provided there is any correlation or association between brain-weight and skull length. One can see at once that no single *row* (*i. e.*, brain-weight array) of Table 75 shows as great a scatter or variability, as does the total row for all brain-weights at the bottom of the table. It follows that if no single row is as variable as the total, the average variability of all single rows must be less than the variability of the total.

Suppose now we define a quantity η as follows:

$$\sigma_{ax}^2 = (1 - \eta^2)\, \sigma_x^2, \ldots\ldots\ldots \text{(i)}$$

where σ_{ax} is the weighted mean variability of the single arrays, of which we have just been speaking, and σ_x is the total variability of the same variable.

Thus η^2 plainly is the ratio of reduction of average variability of an array below the variability of the sample as a whole when these variabilities are expressed as squared standard deviations. Now one can see by studying again Table 62 to 74 *supra* that when the correlation or association between the two variables is *high* σ_{ax} is bound to be small as compared with σ_x, and consequently η will be large. When, on the other hand, the correlation is *low*, σ_{ax} will be of the same order of magnitude as σ_x, and η will necessarily be small. Therefore it follows that η may be used as a measure of the degree of correlation existing in a particular case, quite

regardless of whether the regression is linear or not. When the regression is strictly linear η will be equal to r.

The value of the correlation ratio may be computed in either of two ways. One may proceed in just the manner outlined above, getting the standard deviation, or rather the second moment about the mean of each array, determining their weighted average, and then applying in equation (i) to determine η.

A shorter method is, however, more commonly used. From equation (i)

$$\eta^2 = \frac{\sigma_x^2 - \sigma_{ax}^2}{\sigma_x^2}$$

Take a new quantity

$$\sigma_{mx}^2 = \sigma_x^2 - \sigma_{ax}^2$$

It can be shown that this quantity σ_{mx} is the *standard deviation of the means of arrays*, and therefore easily determined because we already have the means of the arrays for the purpose of plotting regression lines. So then we have

$$\eta = \frac{\sigma_{mx}}{\sigma_x}$$

Let us take as a first numerical example of the computation of the correlation ratio the brain-weight skull length case of Table 75. The work is shown in Table 78.

TABLE 78

CALCULATION OF CORRELATION RATIO FROM DATA OF TABLE 75

Skull length classes.	Means of the x arrays (brain-weight).	x	x^2	Z_y	$Z_y x^2$
155–159.	1300	-164	26,896	2	53,792
160–164.	1393	-71	5,041	14	70,574
165–169.	1386	-78	6,084	42	255,528
170–174.	1440	-24	576	68	39,168
175–179.	1455	-9	81	79	6,399
180–184.	1502	$+38$	1,444	61	88,084
185–189.	1550	$+86$	7,396	19	140,524
190–194.	1650	$+186$	34,596	10	345,960
195–199.	1725	$+261$	68,121	4	272,484
Totals.	. . . .	. . . . .	. . . . .	299	1,272,513

$$\sigma_{mx} = \sqrt{\frac{1272513}{299}} = \sqrt{4255.896} = 65.237$$

$$\sigma_x = 118.995 \text{ (from p. 380 } supra)$$

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x} = \frac{65.237}{118.995} = .548$$

It is evident that the whole process of getting η might equally well have been carried out on the skull-length variabilities. Would the result have been the same? There is no way to find out equal to trying, which is done in Table 79.

TABLE 79

ALTERNATIVE CALCULATION OF CORRELATION RATIO FROM DATA OF TABLE 75

Brain-weight classes.	Means of the y arrays.	y	y^2	Z_x	$Z_x y^2$
1000–1099	167.5	9.0	81.00	1	81.00
1100–1199					
1200–1299	169.6	6.9	47.61	21	999.81
1300–1399	173.8	2.7	7.29	66	481.14
1400–1499	175.0	1.5	2.25	101	227.25
1500–1599	179.7	3.2	10.24	77	788.48
1600–1699	182.1	5.6	31.36	25	784.00
1700–1799	187.5	11.0	121.00	6	726.00
1800–1899	197.5	21.0	441.00	2	882.00
Totals	......	...	....	299	4969.68

$$\sigma_{my} = \sqrt{\frac{4969.68}{299}} = \sqrt{16.621} = 4.077$$

$$\sigma_y = 7.508$$

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y} = \frac{4.077}{7.508} = .543$$

It is seen that η_{yx} is substantially the same as η_{xy} and that both are practically the same as r_{xy} from the same data, its value being $.522 \pm .028$. Thus it appears from analytic as well as visual evidence that the regressions of Table 75 are linear.

Let us take another example where the regression is more evidently non-linear. Such a case is furnished in Table 80, the data of which are taken from Streeter,* using only embryos below 400 grams in weight.

* Streeter, G. L.: Weight, Sitting Height, Head Size, Foot Length, and Menstrual Age of the Human Embryo, Carnegie Institution of Washington Publication No. 274, pp. 143–170.

TABLE 80

CORRELATION BETWEEN WEIGHT AND SITTING HEIGHT OF EMBRYOS BELOW 400 GRAMS IN WEIGHT

Sitting height in mm.	\ Weight in grams → 0–19	20–39	40–59	60–79	80–99	100–119	120–139	140–159	160–179	180–199	200–219	220–239	240–259	260–279	280–299	300–319	320–339	340–359	360–379	380–399	Total's
30–44	5	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	5
45–59	38	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	38
60–74	40	22	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	62
75–89	·	32	17	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	49
90–104	·	·	23	27	8	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	98
105–119	·	·	·	3	19	19	8	·	·	·	·	·	·	·	·	·	·	·	·	·	49
120–134	·	·	·	·	·	4	13	13	11	2	1	·	·	·	·	·	·	·	·	·	44
135–149	·	·	·	·	·	·	1	2	11	8	7	9	5	3	2	1	·	·	·	·	49
150–164	·	·	·	·	·	·	·	·	·	1	2	6	11	12	9	10	4	1	2	2	60
165–179	·	·	·	·	·	·	·	·	·	·	·	1	·	1	3	5	1	9	8	9	37
180–194	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	3	3
Totals	85	54	40	30	27	23	22	15	22	11	10	16	16	16	14	16	5	10	10	14	454

From this table it is at once evident that sitting height does not increase in a linear manner as weight increases.

Calculated in the manner described earlier in this chapter, the correlation coefficient is

$$r = .9440 \pm .0034.$$

The computation of the correlation ratio η from the same data is given in Table 81.

TABLE 81

CORRELATION RATIO: WEIGHT AND SITTING HEIGHT OF EMBRYOS

Type of array (weight).	Mean of array m_x (sitting height).	$m_x - M_x$	$(m_x - M_x)^2$	Z_v	$Z_y \times (m_x - M_x)^2$
10	1.4217	−3.5034	12.2738	83	1018.73
30	2.5926	−2.3325	5.4406	54	293.79
50	3.5750	−1.3501	1.8228	40	72.91
70	4.1000	−0.8251	.6808	30	20.42
90	4.7037	−0.2214	.0490	27	1.32
110	5.1739	+0.2488	.0619	23	1.42
130	5.6818	+0.7567	.5726	22	12.60
150	6.1333	+1.2082	1.4597	15	21.90
170	6.5000	+1.5749	2.4803	22	54.57
190	6.9091	+1.9840	3.9363	11	43.30
210	7.1000	+2.1749	4.7302	10	47.30
230	7.5000	+2.5749	6.6301	16	106.08
250	7.6875	+2.7624	7.6309	16	122.09
270	7.8750	+2.9499	8.7019	16	139.23
290	8.0714	+3.1463	9.8992	14	138.59
310	8.2500	+3.3249	11.0550	16	176.88
330	8.2000	+3.2749	10.7250	5	53.63
350	8.9000	+3.9749	15.7998	10	158.00
370	8.8000	+3.8749	15.0149	10	150.15
390	9.0714	+4.1463	17.1918	14	240.69
Totals	Mean = 4.9251 = M_x			454	2873.60

$$\sigma_{mx} = \sqrt{\frac{2873.60}{454}} = \sqrt{6.3295} = 2.5158$$

$$\sigma_x = 2.5661$$

$$\eta_{xy} = \frac{2.5158}{2.5661} = .9804$$

The question will arise in the reader's mind: Is η significantly different from r? To the eye the regression is plainly non-linear, but we have

$$\eta = .9804$$
$$r = .9440$$
$$\text{Difference} = .0364$$

This is absolutely a small difference. Is it significant in comparison with its probable error? To answer this question resort is necessary to the methods developed by Blakeman[4] for testing the

significance of the difference between η and r. Of the several tests proposed by Blakeman we may take as the most useful, considering ease of computation,

$$\text{P. E.}_\zeta = 2\,\chi_1\,\sqrt{\zeta}\,\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1},$$

where

$$\zeta = \eta^2 - r^2$$

$$\chi_1 = .67449/\sqrt{N}, \text{ and is given in Pearson's Tables.}$$

In the present example we have:

$$\text{P. E.}_\zeta = 2 \times .03166 \times .2648\,\sqrt{.0388^2 - .1089^2 + 1}$$

$$= .01677\,\sqrt{.9896} = .01677 \times .9948 = .017$$

$$\zeta = \eta^2 - r^2 = .961 - .891 = .070 \pm .017$$

ζ is 4.1 times its probable error, and therefore to be regarded as significant. We may then conclude that the regression of sitting height on weight is non-linear.

CORRECTION FOR CORRELATION RATIO

It is important to remember when using the correlation ratio η that, as shown by Pearson,[5] in samples from material in which η is actually zero, the mean value of η from samples will be $\sqrt{\dfrac{\kappa-1}{N}}$, where κ is the number of arrays involved in calculating η and N is the size of the sample. It is evident, therefore, that in any value of η actually obtained from a sample, there needs to be some correction to allow for the influence of number of arrays. Pearson[6] has suggested that

$$\frac{\text{Observed }\eta^2 - (\kappa - 3)/N}{1 - (\kappa - 3)/N}$$

is a reasonable value for the η^2 of the sampled population, provided N is fairly large.

"Of course the first consideration in any investigation of η^2 is to determine whether it is comparable with $(\kappa - 1)/N$. If it be less than this value we cannot assert significant association. If it be greater than this value we have to consider whether η as observed differs considerably from

$$\sqrt{\frac{\kappa-1}{N}} + .67449\,\frac{1}{\sqrt{N}},$$

and for general purposes we must settle whether η differs from $\sqrt{(\kappa-1)/N}$ by, say, $1.7/\sqrt{N}$."

SUGGESTED READING

1. Darbishire, A. D.: Some Tables for Illustrating Statistical Correlation, Mem. and Proc. Manchester Lit. and Phil. Soc., vol. 51, pp. (of reprint) 1–20, 1907.
2. Yule, G. U.: Introduction to the Theory of Statistics, Chapters IX, X, and XI.
3. Pearson, K.: Mathematical Contributions to the Theory of Evolution. XIV. On the General Theory of Skew Correlation and Non-linear Regression, Draper's Company Research Mem. Biometric Series II, Cambridge (University Press), 1905.
4. Blakeman, J.: On Tests for Linearity of Regression in Frequency Distribution, Biometrika, vol. 4, pp. 332–350, 1905.
5. Pearson, K.: On a Correction to Be Made to the Correlation Ratio, Biometrika, vol. 8, pp. 254–256, 1911.
6. Pearson, K.: On the Correction Necessary for the Correlation Ratio, Ibid., vol. 14, pp. 412–417, 1923.
7. Pearson, K.: Notes on the History of Correlation, Ibid., vol. 13, pp. 25–45, 1920 (An excellent account of the early history of the subject of correlation.)

CHAPTER XV

PARTIAL CORRELATION

By a simple extension of the principle of two-variable correlation, described in the last chapter, multiple and net or partial correlations may be determined. Multiple correlation is the correlation between one variable and a series of other variables taken together. A net or partial correlation is the correlation between two variables when a whole series of other variables are held constant. The epistemologic value of the method of partial correlation is great. This is evident from the following considerations.

The most useful general method of acquiring knowledge of dynamic phenomena is unquestionably the experimental method. When we deal with phenomena of human biology, there is a wide range of matters in which the laboratory experimental method is, in the nature of the case, ruled out. Unfortunately, one cannot breed homozygous strains of men at will for experimental purposes, nor subject them methodically to desired environmental conditions. In studying most problems of human biology, resort must be had to some form of the statistical method. This is fundamentally a descriptive method, and hence, in many of its phases, ill-adapted to the analysis of dynamically active events.

The essence of the experimental method, as practised in the laboratory, and in theory, is that, of the multitude of variables conditioning a phenomenon, as many as possible are, by appropriate methods, held constant while one or at most a very few selected variables are allowed to vary and the results noted. One may then deduce the relative significance of the selected variable in determining the phenomenon under observation. Now we frequently hear in scientific discussions about the experiments that nature makes. Actually the true conditions of an experiment are rarely if ever realized in the course of natural events. It is just because nature permits manifold and haphazard changes in

all variables at the same time that recourse must be had to the method of experimental control in the laboratory. What is needed in order to interpret the results, in the experimental sense, and determine the meaning of the manifold and ceaseless changes and variations in the flow of naturally determined events, is some method of picking out of the manifold some selected *constant* conditions of a series of variables, and then measuring the extent and character of the variations in a *single* selected variable, whose true relative influence upon the phenomenon it is desired to know, while all these other variables are held constant. If this can be done we shall have realized some of the epistemologic advantages of the experimental method as practised in the laboratory, and have freed ourselves at the same time of the limitations which in so many cases inhere in the material itself, and make the laboratory type of experimental inquiry impossible. In other words, we shall have let nature perform the experiment, in the sense of determining the phenomena, in her own way, while we evaluate the results in critically analytic terms of similar sort and meaning to those in which we evaluate the results of a laboratory experiment.

Now exactly this epistemologic boon is actually afforded in the method of partial or net correlation, if properly handled. This calculus enables one, out of a manifold complex of variables operating in an entirely uncontrolled and natural manner, to determine the variation of any selected single variable, or the correlation of any selected pair of variables for constant conditions or values of the other variables in the complex.

The fundamental theorems in partial correlation were developed in Pearson's biometric laboratory (cf. Pearson[1]). The notation now almost universally used in this field is due to Yule,[2] whose paper should be carefully studied for the full mathematical development of the subject, which cannot be gone into here. It is as follows (Yule, *loc. cit.*, p. 182):

"Let x_1 x_2 ... x_n denote deviations in the values of the n variables from their respective arithmetic means. Then the regression equation may be written:

$$x_1 = b_{12\cdot34\ldots n}x_2 + b_{13\cdot24\ldots n}x_3 + \cdots + b_{1n\cdot23\ldots n-1}x_n \qquad (1)$$

In this notation the suffix of each regression coefficient completely

defines it.   The first subscript gives the dependent variable, the second the variable of which the given regression is the coefficient, and the subscripts after the period show the remaining independent variables which enter into the equation.   It is convenient to distinguish the subscripts before and after the period as 'primary' and 'secondary' subscripts respectively.   The order in which the secondary subscripts are arranged is indifferent, but the order of the two primary subscripts is material; e. g., $b_{12.3...n}$ and $b_{21.3...n}$ denote two quite distinct coefficients.   A coefficient with p secondary subscripts may be termed a regression of the pth order, the total regression b_{12}, b_{13}, b_{23}, etc. being thus regarded as of order zero.

"The correlation coefficients may be distinguished by subscripts in precisely the same manner.   Thus the correlation $r_{12.34...n}$ is defined by the relation

$$r_{12.34...n} = (b_{12.34...n} \cdot b_{21.34...n})^{\frac{1}{2}}. \tag{2}$$

In the case of the correlations, the order of both primary and secondary subscripts is indifferent.   A correlation with p secondary subscripts may be termed a correlation of order p, the total correlations r_{12}, r_{13}, r_{23}, etc., being regarded as of order zero."

Now the essence of the partial correlation calculus is that in the expression

$$r_{12.34...n}$$

the variables represented by the secondary subscripts $34....n$ *are held constant* (and therefore their effect upon the total variation or correlation in the original, unrestrained conditions is corrected or allowed for), while those represented by the primary subscripts 1 and 2 are allowed to vary as much as they will under the restriction that all the others are constant, and the correlation between variables 1 and 2 under those circumstances is measured. What this means in terms of biologic realities is this: In the last chapter it was seen that there was less variation in brain-weight among the persons composing a single array than among all the persons in the sample taken together.   But this is precisely what would be expected biologically.   For what is a brain-

weight array? It is in this case simply a group of persons so picked out as to be all alike (within certain narrow limits) in respect of skull length. Naturally, if they are all alike in skull length they cannot differ (or vary) very much among themselves in respect of brain-weight, because of the biologic correlation which exists between skull-size and brain-weight. Now consider an extension of the same process. Suppose a group of persons to be selected all of the same stature, and let measurements be made of the skull length and brain-weight of each. Plainly, a correlation table can be set up between skull length and brain-weight *in this group*. The resulting coefficient of correlation will be of the sort $r_{12.3}$, where 1 denotes skull length, 2 denotes brain-weight, and 3 • stature. The coefficient will measure the correlation between skull length and brain-weight for the one particular *constant stature*, to which the persons were selected. So, similarly, there might be picked a group of persons in which all were alike in respect of both stature and body-weight, let us say, and the correlation between skull length and brain-weight determined for this group. This would lead to a correlation of the sort $r_{12.34}$. And so, theoretically, the process might be continued on to any number of characters in respect of all of which the persons in the group were so selected as to be all just alike.

For the arithmetic work of the following numerical example on this point I am indebted to my colleague, Doctor L. J. Reed. Some years ago Pearl and Surface* published detailed measurements of length, breadth, and weight of 453 hens' eggs. Now from all these eggs

$$r_{12.3} = - .8955.$$

This coefficient measures for the whole material the net correlation between length and breadth when weight is held constant by the application of equation (3) *infra*.

But now suppose from the table of individual measurements given as an appendix to the paper cited there are picked out all those eggs that weighed 53 to 53.9 grams, and a correlation

* Pearl, R., and Surface, F. M.: A Biometrical Study of Egg Production in the Domestic Fowl. III. Variation and Correlation in the Physical Characters of the Egg, U. S. Dept. Agr. Bur. Anim. Ind., Bulletin 110, pp. 171–241, 1914.

table then constructed, *for these selected eggs*, between length and
breadth. There were 42 such eggs and the table is shown as
Table 82.

TABLE 82

CORRELATION BETWEEN EGG LENGTH AND BREADTH, FOR EGGS WEIGHING 53 TO
53.9 GRAMS

Egg breadth (mm.)

Egg length (mm.)	40.0	40.5	41.0	41.5	42.0	42.5	43.0	Totals
51	-	-	-	-	-	1	1	2
52	-	-	-	-	1	1	1	3
53	-	-	-	-	1	1	-	2
54	-	-	-	6	3	-	-	9
55	-	-	2	3	-	-	-	5
56	1	1	6	-	-	-	-	8
57	2	3	2	1	-	-	-	8
58	-	1	1	-	-	-	-	2
59	2	1	-	-	-	-	-	3
Totals	5	6	11	10	5	3	2	42

From this table the coefficient of correlation calculated in the
usual manner described in the preceding chapter is

$$r_{12} = - \ .9117.$$

It will be noted that this is very close indeed to the value of
$r_{12.3}$ given above. But let us take another array and see what the
result is. Table 83 gives the correlation between length and
breadth of 46 eggs picked out of the whole lot, each having a
weight between 56 and 56.9 grams.

Here the coefficient worked out in the usual way is

$$r_{12} = - \ .8911,$$

a result still closer to the $r_{12.3}$ value given above.

Let us take one more example, choosing this time eggs which
are near the extreme of weight, instead of arrays near the middle

TABLE 83

CORRELATION BETWEEN EGG LENGTH AND BREADTH FOR EGGS WEIGHING 56 TO 56.9 GRAMS

Egg breadth (mm.)

Egg length (mm)	40.0	40.5	41.0	41.5	42.0	42.5	43.0	43.5	Totals
52	-	-	-	-	-	-	-	1	1
53	-	-	-	-	-	-	-	1	1
54	-	-	-	-	-	2	4	1	7
55	-	-	-	-	3	2	2	-	7
56	-	-	-	2	3	4	-	-	9
57	-	-	-	6	8	-	-	-	14
58	-	-	-	1	-	-	-	-	1
59	-	-	2	-	1	-	-	-	3
60	1	-	-	-	-	-	-	-	1
61	-	2	-	-	-	-	-	-	2
Totals	1	2	2	9	15	8	6	3	46

TABLE 84

CORRELATION BETWEEN EGG LENGTH AND BREADTH FOR EGGS WEIGHING 62 TO 62.9 GRAMS

Egg breadth (mm.)

Egg length (mm.)	42.5	43.0	43.5	44.0	44.5	45.0	Totals
55	-	-	-	1	-	2	3
56	-	-	1	-	1	-	2
57	-	-	1	2	-	-	3
58	-	1	1	-	-	-	2
59	1	2	-	-	-	-	3
Totals	1	3	3	3	1	2	13

value. Table 84 gives the length-breadth correlation for 13 eggs each having a weight between 62 and 62.9 grams, that is, heavy eggs.

Here, with such a small array, the length-breadth correlation is

$$r_{12} = -\ .8739.$$

Let us now take a weighted mean of these three length-breadth correlations (r_{12}). We have:

$$-.9117 \times 42 = -38.2914$$
$$-.8911 \times 46 = -40.9906$$
$$-.8739 \times 13 = -11.3607$$

Totals $\overline{101}$ -90.6427

Whence

Mean r_{12} $= -\ .8975$
(By partial correlation) $r_{12 \cdot 3}$ $= -\ .8955$
Difference $=$ $.0020$

Thus it is seen, by this process of actual trial, that if we physically select individuals so that they are all alike relative to one variable (3) and then directly measure their correlation in respect of two other variables (1 and 2), the average correlation (r_{12}) so obtained is substantially identical with the result which we get mathematically when we calculate the partial correlation $r_{12 \cdot 3}$.

The only difference between the perfectly simple biologic procedure, which anyone can understand, of selecting individuals alike in respect of n variables and then measuring the correlation between two other variables, and the processes implicit in the arithmetic working out of the equation for a partial correlation coefficient,

$$r_{12 \cdot 34 \ldots\ldots n} = \frac{r_{12 \cdot 34 \ldots\ldots (n-1)} - r_{1n \cdot 34 \ldots\ldots (n-1)} \cdot r_{2n \cdot 34 \ldots\ldots (n-1)}}{(1 - r^2_{1n \cdot 34 \ldots\ldots (n-1)})^{\frac{1}{2}} (1 - r^2_{2n \cdot 34 \ldots\ldots (n-1)})^{\frac{1}{2}}}, \quad (3)$$

is simply that the mathematical procedure operates upon the basis of the weighted *average* variability *of all arrays* in the manifold space involved by the variables held constant. In the process of concrete physical selection of individuals described above one set of arrays only can be dealt with at one time.

Not only can the correlation between two variables be determined from equation (3) when a whole series of other characters are constant, but also the reduction in the variability of any

character as 1, 2, 3...n other variables are held constant can be measured. The expression for this is

$$\sigma^2_{1 \cdot 23 \ldots n} = \sigma^2_1 (1 - r^2_{12}) (1 - r^2_{13 \cdot 2}) (1 - r^2_{14 \cdot 23}) \ldots (1 - r^2_{1n \cdot 23 \ldots n-1}) \quad (4)$$

The arithmetic of the whole process is extremely simple. For 3 variables equation (3) is, obviously,

$$r_{12 \cdot 3} = \frac{r_{12} - r_{13} \cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}} (1 - r^2_{23})^{\frac{1}{2}}} \quad (5)$$

The zero order correlations r_{12}, r_{13}, and r_{23} will be calculated from the observed correlation tables like Table 75 in the preceding chapter. If we have in the whole system under consideration say 5 variables there will obviously be 29 other possible first order coefficients as follows: $r_{12 \cdot 4}, r_{12 \cdot 5}, r_{13 \cdot 2}, r_{13 \cdot 4}, r_{13 \cdot 5}, r_{14 \cdot 2}, r_{14 \cdot 3}, r_{14 \cdot 5},$ $r_{15 \cdot 2}, r_{15 \cdot 3}, r_{15 \cdot 4}, r_{23 \cdot 1}, r_{23 \cdot 4}, r_{23 \cdot 5}, r_{24 \cdot 1}, r_{24 \cdot 3}, r_{24 \cdot 5}, r_{25 \cdot 1}, r_{25 \cdot 3}, r_{25 \cdot 4}, r_{34 \cdot 1},$ $r_{34 \cdot 2}, r_{34 \cdot 5}, r_{35 \cdot 1}, r_{35 \cdot 2}, r_{35 \cdot 4}, r_{45 \cdot 1}, r_{45 \cdot 2}, r_{45 \cdot 3}.$ Each one of these can be determined from the zero order coefficient just as $r_{12 \cdot 3}$ was in (5) above.

For the second order coefficients (3) becomes, for example,

$$r_{12 \cdot 34} = \frac{r_{12 \cdot 3} - r_{14 \cdot 3} \cdot r_{24 \cdot 3}}{(1 - r^2_{14 \cdot 3})^{\frac{1}{2}} (1 - r^2_{24 \cdot 3})^{\frac{1}{2}}}$$

But we may equally well write

$$r_{12 \cdot 34} = \frac{r_{12 \cdot 4} - r_{13 \cdot 4} \cdot r_{23 \cdot 4}}{(1 - r^2_{13 \cdot 4})^{\frac{1}{2}} (1 - r^2_{23 \cdot 4})^{\frac{1}{2}}}$$

These two methods of calculation should give the same result, and, in fact, do, thus furnishing in actual practice a most useful check on the arithmetical work.

For the third order coefficients (3) takes such forms as

$$r_{12 \cdot 345} = \frac{r_{12 \cdot 34} - r_{15 \cdot 34} \cdot r_{25 \cdot 34}}{(1 - r^2_{15 \cdot 34})^{\frac{1}{2}} (1 - r^2_{25 \cdot 34})^{\frac{1}{2}}}$$

And so on, indefinitely, except for the two following limitations:

(a) All the zero order correlations must have linear regressions, or the method is not valid. Therefore before embarking on an

26

extensive partial correlation project we should always test the zero order correlations for linearity in the manner described in the preceding chapter.

(b) The number of observations in each of the zero order tables must be fairly large, as compared with the number of variables dealt with, if the partial correlation results are to be in any degree conclusive.

It will be noted from the form of equation (3) that if one had available tables of $\sqrt{1 - r^2}$, sufficiently detailed so that interpolation would be unnecessary, the computation of partial correlation coefficients would become a very simple matter indeed. Such tables have, in fact, been provided by my colleague, Dr. John Rice Miner,[3] and can be obtained from the Johns Hopkins Press at a nominal price.

ILLUSTRATION OF PARTIAL CORRELATION

In order that the reader may become thoroughly familiar with the operation of the useful partial correlation technic, a numerical example will now be presented in detail. The example is drawn from the writer's (Pearl[4]) studies on the epidemiology of influenza.

The problem set is this: What is the net correlation between the destructiveness of the 1918–1919 influenza epidemic in large American cities and the normal death-rate in the same cities from organic diseases of the heart, when all the cities are made constant in respect of (a) the age constitution of the population, (b) the sex ratio of the population, and (c) the density of population?

The data are taken from Pearl.[4] The subscripts have the following significance:

Subscript 2 denotes the destructiveness of the epidemic, measured by the twenty-five-week excess mortality rates calculated and published by the Bureau of the Census. These twenty-five-week excess rates indicate the number of people dying from all causes, during the twenty-five weeks following the initial outbreak of the epidemic in this country in the autumn of 1918, in excess of the number who probably would have died in the same period had no epidemic occurred. The rates for the 34 cities are

given in Table 1 (p. 12) of Influenza Studies I, and hence need not be reprinted here.

Subscript 3 denotes the normal death-rate in each city from organic diseases of the heart, averaged for the three years 1915, 1916, and 1917.

Subscript 4 denotes the age distribution of the population, as measured by an age-constitution index having the form

$$\phi = S\left\{\frac{\Delta^2}{P}\right\}(M - M_p)$$

where Δ is the deviation for each of six age groups (viz., 0–4, 5–14, 15–24, 25–44, 45–64, 65 and over) of the percentage of the actual population of each city in 1910 in each age group, from the percentage in the same group in the standard population of Glover's life table, denoted in the formula by P; S denotes summation of all six values; M = mean age of living population in any community; M_p = mean age of persons in a stationary population unaffected by migration and which, assuming the mortality rates of Glover's life table, would result if 100,000 persons were born alive uniformly throughout each year (M_p calculated from L_x line of Glover's table (p. 16) = 33.796 years).

Subscript 5 denotes the ratio of males to 100 females in each of the cities in 1910.

Subscript 6 denotes density of population calculated from data furnished in the "Financial Statistics of Cities," issued annually by the Bureau of the Census, and was expressed as the number of persons per acre of land area within the legally defined limits of the city.

The values of the zero-order correlations and the first order coefficients derived from them are given in Table 85, which includes all the figures set down in making the calculations, the multiplications and divisions having been made on a calculating machine.

The computations go in this way, taking the upper block of Table 85. To get the product term of the numerator of equation (3) r_{24} = .0238 is multiplied by r_{34} = .6093, giving the result .0145, set down in the column headed "Product term of numerator."

TABLE 85

PARTIAL CORRELATIONS. INFLUENZA. ZERO AND FIRST ORDER COEFFICIENTS

r 0 Order Subscript.	Coefficient.	$(1 - r^2)^{\frac{1}{2}}$.	Product term of numerator.	Whole numerator.	Denominator.	r First order. Subscript.	Coefficient.
23.........	+.4874	.....	+.0145	+.4729	.7928	23.4	+.5965
24.........	+.0238	.9997					
34.........	+.6093	.7930					
23.........	+.4874	.....	+.0050	+.4824	.9853	23.5	+.4896
25.........	−.0295	.9996					
35.........	−.1682	.9857					
23.........	+.4874	.....	−.0177	+.5051	.9811	23.6	+.5148
26.........	+.1108	.9938					
36.........	−.1595	.9872					
24.........	+.0238	.9997	+.0035	+.0203	.9926	24.5	+.0205
25.........	−.0295	.9996	−.0028	−.0267	.9927	25.4	−.0269
45.........	−.1184	.9930					
24.........	+.0238	.9997	−.0259	+.0497	.9663	24.6	+.0514
26.........	+.1108	.9938	−.0056	+.1164	.9720	26.4	+.1198
46.........	−.2338	.9723					
25.........	−.0295	.9996	+.0017	−.0312	.9937	25.6	−.0314
26.........	+.1108	.9938	−.0005	+.1113	.9995	26.5	+.1114
56.........	+.0155	.9999					
34.........	+.6093	.7930	+.0199	+.5894	.9788	34.5	+.6022
35.........	−.1682	.9857	−.0721	−.0961	.7874	35.4	−.1220
45.........	−.1184	.9930					
34.........	+.6093	.7930	+.0373	+.5720	.9598	34.6	+.5960
36.........	−.1595	.9872	−.1425	−.0170	.7710	36.4	−.0220
46.........	−.2338	.9723					
35.........	−.1682	.9857	−.0025	−.1657	.9871	35.6	−.1679
36.........	−.1595	.9872	−.0026	−.1569	.9856	36.5	−.1592
56.........	+.0155	.9999					
45.........	−.1184	.9930	−.0036	−.1148	.9722	45.6	−.1181
46.........	−.2338	.9723	−.0018	−.2320	.9929	46.5	−.2337
56.........	+.0155	.9999	+.0277	−.0122	.9655	56.4	−.0126

The two elements in the denominator $\sqrt{(1 - .0238^2)}$, and $\sqrt{(1 - .6093^2)}$, are read off from Miner's Tables, as .9997 and .7930 respectively. The whole numerator is .4874 − .0145 = .4729, while the denominator is .9997 × .7930 = .7928. Finally $r_{23.4} = \frac{.4729}{.7928} = .5965$. And so on for the other cases.

The calculation of the second order coefficients is given in Table 86, which is of exactly the same form as Table 85, except that each second order coefficient is calculated in two different ways (i. e., with two different sets of first-order coefficients) as a check on the arithmetic.

Finally, Table 87 gives the third order coefficient in which we are interested, again calculated in two ways as a check.

TABLE 86

PARTIAL CORRELATIONS. INFLUENZA. FIRST AND SECOND ORDER COEFFICIENTS

r First order.		$(1-r^2)^{\frac{1}{2}}$	Product term of numerator.	Whole numerator.	Denominator.	r Second order.	
Subscript.	Coefficient.					Subscript.	Coefficient.
23.4........	+.5965	.....	+.0033	+.5932	.9921	23.45	+.5979
25.4........	−.0269	.9996					
35.4........	−.1220	.9925					
23.5........	+.4896	.....	+.0123	+.4773	.7981	23.45	+.5980
24.5........	+.0205	.9998					
34.5........	+.6022	.7983					
23.4........	+.5965	.....	−.0026	+.5991	.9926	23.46	+.6036
26.4........	+.1198	.9928					
36.4........	−.0220	.9998					
23.6........	+.5148	.....	+.0306	+.4842	.8019	23.46	.+6038
24.6........	+.0514	.9986					
34.6........	+.5960	.8030					
23.5........	+.4896	.....	−.0177	+.5073	.9811	23.56	+.5171
26.5........	+.1114	.9938					
36.5........	−.1592	.9872					
23.6........	+.5148	.....	+.0053	+.5095	.9853	23.56	+.5171
25.6........	−.0314	.9995					
35.6........	−.1679	.9858					
25.4........	−.0269	.9996	−.0015	−.0254	.9927	25.46	−.0256
26.4........	+.1198	.9928	+.0003	+.1195	.9995	26.45	+.1196
56.4........	−.0126	.9999					
24.5........	+.0205	.9998					
26.5........	+.1114	.....	−.0048	+.1162	.9721	26.45	+.1195
46.5........	−.2337	.9723					
24.6........	−.0514	.9986					
25.6........	−.0314	.....	−.0061	−.0253	.9916	.25.46	−.0255
45.6........	−.1181	.9930					
35.4........	−.1220	.9925	+.0003	−.1223	.9997	35.46	−.1223
36.4........	−.0220	.9998	+.0015	−.0235	.9924	36.45	−.0237
56.4........	−.0126	.9999					
34.5........	+.6022	.7983					
36.5........	−.1592	.....	−.1407	−.0185	.7762	36.45	−.0238
46.5........	−.2337	.9723					
34.6........	+.5960	.8030					
35.6........	−.1679	.....	−.0704	−.0975	.7974	35.46	−.1223
45.6........	−.1181	.9930					

TABLE 87

PARTIAL CORRELATIONS. INFLUENZA. SECOND AND THIRD ORDER COEFFICIENTS

r Second order.		$(1-r^2)^{\frac{1}{2}}$.	Product term of numerator.	Whole numerator.	Denominator.	r Third order.	
Subscript.	Coefficient.					Subscript.	Coefficient.
23.45......	+.5979	.....	−.0028	+.6007	.9925	23.456	+.6052
26.45......	+.1195	.9928					
36.45......	−.0237	.9997					
23.46......	+.6037	.....	+.0031	+.6006	.9922	23.456	+.6053
25.46......	−.0255	.9997					
35.46......	−.1223	.9925					

From this we see that there was a relatively high net or partial correlation between destructiveness of the epidemic outbreak and normal cardiac death-rate, the coefficient being

$$r_{23 \cdot 456} = +.605 \pm .073,$$

when the demographic variables of age, sex, and density are held constant.

It should be noted that the probable error of a partial correlation of higher order is of the same form as that of a zero order coefficient (see Chapter XIV).

The student should read some of the extended investigations which have been made by the partial correlation method, particularly that of Miner.[5]

SUGGESTED READING

1. Pearson, K.: Mathematical Contributions to the Theory of Evolution. XI. On the Influence of Natural Selection on the Variability and Correlation of Organs, Phil. Trans. A., vol. 200, pp. 1–66, 1902.
2. Yule, G. U.: On the Theory of Correlation for Any Number of Variables, Treated by a New System of Notation, Proc. Roy. Soc. A., vol. 79, pp. 182–193, 1907.
3. Miner, J. R.: Tables of $\sqrt{1-r^2}$ and $1-r^2$ for Use in Partial Correlation and Trigonometry, Baltimore (The Johns Hopkins Press), 1922, pp. 49.
4. Pearl, R.: Influenza Studies. I. On Certain General Statistical Aspects of the 1918 Epidemic in American Cities, Public Health Reports, vol. 34, pp. 1743–1783, 1919. II. Further Data on the Correlation of Explosiveness of Outbreak of the 1918 Epidemic, Ibid., vol. 36, pp. 273–289, 1921. III. On the Correlation of Destructiveness of the 1918 Epidemic, Ibid., vol. 36, pp. 289–294, 1921. IV. On the Correlation Between Explosiveness and Total Destructiveness of the Epidemic Mortality, Ibid., vol. 36, pp. 294–298, 1921.
5. Miner, J. R.: Suicide and Its Relation to Climatic and Other Factors, Amer. Jour. Hyg., Monograph No. 2, pp. 1–146, 1922.

CHAPTER XVI

SIMPLE CURVE FITTING

The worker in practically any branch of science is more or less frequently confronted with this sort of problem: he has a series of observations in which there is clear evidence of a certain orderliness, on the one hand, and evident fluctuations from this order, on the other hand. What he obviously wishes to do, on the basis of a quite sound instinct, is to emphasize the orderliness and minimize the fluctuations about it. His reasoning, deeply rooted in racial experience of more or less scientific matters, is that the orderliness of which he sees traces, if really there, depends upon a true lawful relation between the variables he is studying, and that the fluctuations are in general merely accidents of random sampling. He would like an expression, exact if possible, or, failing that, approximate, of the law if there be one. This means a mathematical expression of the functional relation between the variables.

It seems desirable to give the medical man some little introduction to the methods which the followers of the sciences at the moment more exact than medicine, use in fitting together mathematical expressions and observational data. It should be made clear at the start that there is, unfortunately, no method known to mathematics which will tell anyone in advance of the trial what is either the correct or even the best mathematical function with which to graduate a particular set of data. The choice of the proper mathematical function is essentially, at its very best, only a combination of good judgment and good luck. In this realm, as in every other, good judgment depends in the main only upon extensive experience. What we call good luck in this sort of connection has also about the same basis. The experienced person in this branch of applied mathematics knows at a glance what general class of mathematical expression will take a course, when plotted, on the whole like that followed by

the observations. He furthermore knows that by putting as many constants into his equation as there are observations in the data he can make his curve hit all the observed points exactly, but in so doing will have defeated the very purpose with which he started, which was to emphasize the law (if any) and minimize the fluctuations, because actually if he does what has been described he emphasizes the fluctuations and probably loses completely any chance of discovering a law.

Of mathematical functions involving a small number of constants there are but relatively few. If one takes account of that group of curves which in his youth he studied under the name of "conic sections," adds to it the curves which derive from the trigonometrical functions, and fills out the equipment with the logarithmic-exponential family, he will not have exhausted the possibilities of curves with few constants, but he will have included the great bulk of the mathematical functions which have so far been found to be of wide utility in expressing the laws of nature. In short, we live in a world which appears to be organized in accordance with relatively few and relatively simple mathematical functions. Which of these one will choose in starting off to fit empirically a group of observations depends fundamentally, as has been said, only on good judgment and experience. There is no higher guide.

Of the observational data which the medical man has occasion or desire to graduate (which means fit a curve to) perhaps the most frequent will be those in which there is a definite trend up or down, or first in one direction and then in the other. It is proposed now to show briefly how to fit three simple functions, namely, a straight line, a second-order parabola, and a logarithmic curve, to such data. The method which will be used is that known in mathematics as the "method of least squares," but the reader should not let this discourage him. It is really very simple. If he wants to know about its foundation perhaps the best thing to read is a short paper by Ellis.[1] If he prefers a more detailed mathematical approach than the present one, both specifically and in general, to curve fitting problems, Running's[2] book, or the excellent text on least squares of Brunt[3] can be recommended, or,

perhaps most useful of all, Whittaker and Robinson's[4] comprehensive treatise.

After one has, on the basis of his general judgment of the whole situation, *chosen* a particular function with which to graduate a set of data, the theory of least squares says that "the best fitting" curve is that particular one, out of the whole range given by the chosen function, which makes the sum of the squares of the differences between the observed points and the corresponding points on the fitted curve *a minimum*. This, it should clearly be understood, is simply a convention. Other conventions quite as sound and well justified could be, and have been, used. For example, it may be said that, under the same initial premise as before, the "best fitting curve" shall be that one having its area and moments equal to the area and moments of the observations. If one follows this definition he fits by the method of moments; if he follows the first definition he fits by the method of least squares. We have chosen for discussion here the least square definition.

Take as the equation to a straight line

$$y = a + bx.$$

Now, plainly, the difference between any observation and this curve (for a straight line is a curve of zero curvature) will be

$$(y - a - bx).$$

There will be as many of such differences as there are observations. The theory of least squares insists that values for the constants a and b be so chosen that

$$S (y - a - bx)^2,$$

where S denotes summation, shall be a minimum. How shall we determine from the observations the values of a and b which will fulfil this requirement?

This is done by solving two equations (since there are two constants to be determined) which are known technically as the *normal* equations. How it is known that they are the right equations, in respect of their form, comes about from an application

of certain principles of the differential calculus, which need not be gone into here. The normal equations for fitting a straight line are

$$S(y) - na - bS(x) = 0$$
$$S(xy) - aS(x) - bS(x^2) = 0$$

Transposing terms in form for computation these become

$$na + bS(x) = S(y)$$
$$aS(x) + bS(x^2) = S(xy),$$

where n is the number of observed points.

The location of the points on the abscissal scale can, of course, take origin from any place one pleases. It is convenient, since usually the observations are equally spaced on the x axis, to take origin of x at one abscissal unit below the first observation. Then the x of the first observation is 1, that of the second 2, and so on; and the sum of the x's ($S(x)$) and $S(x^2)$ can be read directly from tables of the sums of the powers of the natural numbers (as in Pearson's Tables). All of this is merely another way of saying that in curve fitting just as in the calculation of frequency constants (cf. earlier chapters) it is convenient to work in abscissal

TABLE 88
MEAN SITTING HEIGHTS OF EMBRYO. CURVE FITTING

Weight of embryo in grams.	Mean sitting height in mm. y	x	xy	x^2y	$y \log x.$	Calculated y from parabola.	Calculated y from log curve.
0– 19	58.8	1	58.8	58.8	0	66.9	55.9
20– 39	76.4	2	152.8	305.6	22.9987	77.3	78.1
40– 59	91.1	3	273.3	819.9	43.4658	87.1	91.7
60– 79	99.0	4	396.0	1,584.0	59.6039	96.3	101.8
80– 99	108.1	5	540.5	2,702.5	75.5587	105.0	110.0
100–119	115.1	6	690.6	4,143.6	89.5652	113.2	117.0
120–139	122.7	7	858.9	6,012.3	103.6935	120.7	123.2
140–159	129.5	8	1,036.0	8,288.0	116.9502	127.8	128.7
160–179	135.0	9	1,215.0	10,935.0	128.8227	134.3	133.7
180–199	141.1	10	1,411.0	14,110.0	141.1000	140.2	138.4
200–219	144.0	11	1,584.0	17,424.0	149.9605	145.5	142.8
220–239	150.0	12	1,800.0	21,600.0	161.8772	150.3	147.0
240–259	152.8	13	1,986.4	25,823.2	170.2106	154.6	150.9
260–279	155.6	14	2,178.4	30,497.6	178.3375	158.3	154.7
280–299	158.6	15	2,379.0	35,685.0	186.5281	161.4	158.3
300–319	161.3	16	2,580.8	41,292.8	194.2246	164.0	161.8
320–339	160.5	17	2,728.5	46,384.5	197.4870	166.0	165.1
340–359	171.0	18	3,078.0	55,404.0	214.6516	167.5	168.4
360–379	169.5	19	3,220.5	61,189.5	216.7487	168.4	171.5
380–399	173.6	20	3,472.0	69,440.0	225.8588	168.8	174.6
Totals....2673.7		....	31,640.5	453,700.3	2677.6433		

units of grouping rather than in concrete units such as pounds, feet, etc. $S(y)$ will be readily got simply by summing the observed points (the numerical values of the ordinates). $S(xy)$ involves multiplying each x by its y and summing.

The best way to show how simple this all is will be to work out an example. This is done in Table 88. The data are drawn from Table 80 in Chapter XIV, and consist of the mean sitting heights of human embryos. The figures constitute the observed regression line of sitting height on weight.

From Table 88, and a table of the sums of the powers of the natural numbers, we have,

$$n = 20$$
$$S(x) = 210$$
$$S(x^2) = 2870$$
$$S(y) = 2673.7$$
$$S(xy) = 31640.5$$

Whence the equations are

$$20\,a + 210\,b = 2673.7$$
$$210\,a + 2870\,b = 31640.5$$

Solving, we get

$$a = 77.37$$
$$b = 5.36$$
$$y = 77.37 + 5.36\,x.$$

We next proceed to calculate the value of y (sitting height) for two values of x as follows:

When

$$x = 1, y = 82.73$$
$$x = 20, y = 184.64$$

The line can then be drawn. The result is shown graphically in Fig. 87.

It is apparent that a straight line is not the mathematical function best adapted to fit these observations. This was already known from the value of $\eta^2 - r^2$ in this case, which proved that this was non-linear regression (cf. p. 392).

A parabola may be fitted next to the data. Its equation is

$$y = a + b\,x + c\,x^2$$

The normal equations now are three in number, since this is a three constant equation, as follows:

$$n\,a + b\,S\,(x) + c\,S\,(x^2) = S\,(y)$$
$$a\,S\,(x) + b\,S\,(x^2) + c\,S\,(x^3) = S\,(xy)$$
$$a\,S\,(x^2) + b\,S\,(x^3) + c\,S\,(x^4) = S\,(x^2y)$$

Filling in the values from Table 88 these become

$$20\,a + 210\,b + 2870\,c = 2673.7$$
$$210\,a + 2870\,b + 44{,}100\,c = 31{,}640.5$$
$$2870\,a + 44{,}100\,b + 722{,}666\,c = 453{,}700.3$$
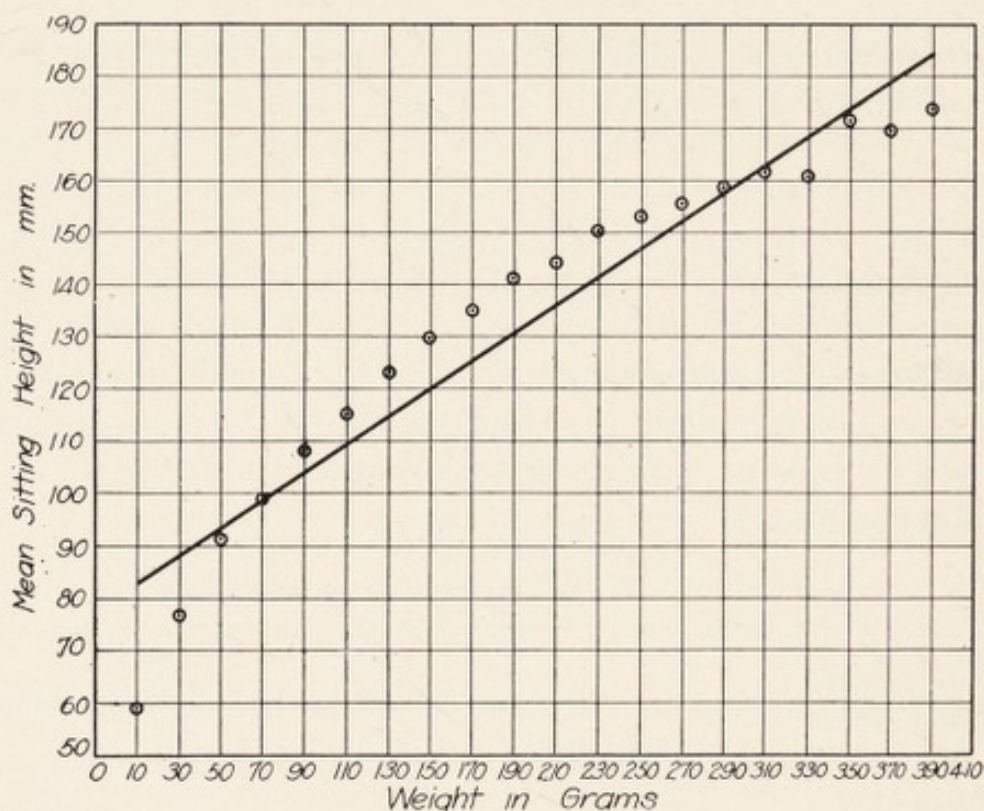
Fig. 87.—Observed mean sitting heights of embryos (circles) and straight line fitted by least squares.

Solving,

$$y = 55.986 + 11.195\,x - .278\,x^2$$

Substituting successive values of x and solving for y gives the values of the ordinates of the curve exhibited in the last column but one of Table 88. It is at once apparent that the parabola comes closer to the observation than the straight line, but it still is a poor fit.

The result is shown graphically in Fig. 88.

Turning to the logarithmic curve the equation we shall use is

$$y = a + b x + c \log x$$

It may be well at this point to say a word as to the reasoning which leads to the choice of this particular form of a logarithmic curve. If one had had no pedagogic purpose in mind, this is the one of the three curves which would have been chosen in the first instance, and no straight line or parabola would have been fitted. It is apparent to anyone of experience in such matters that the
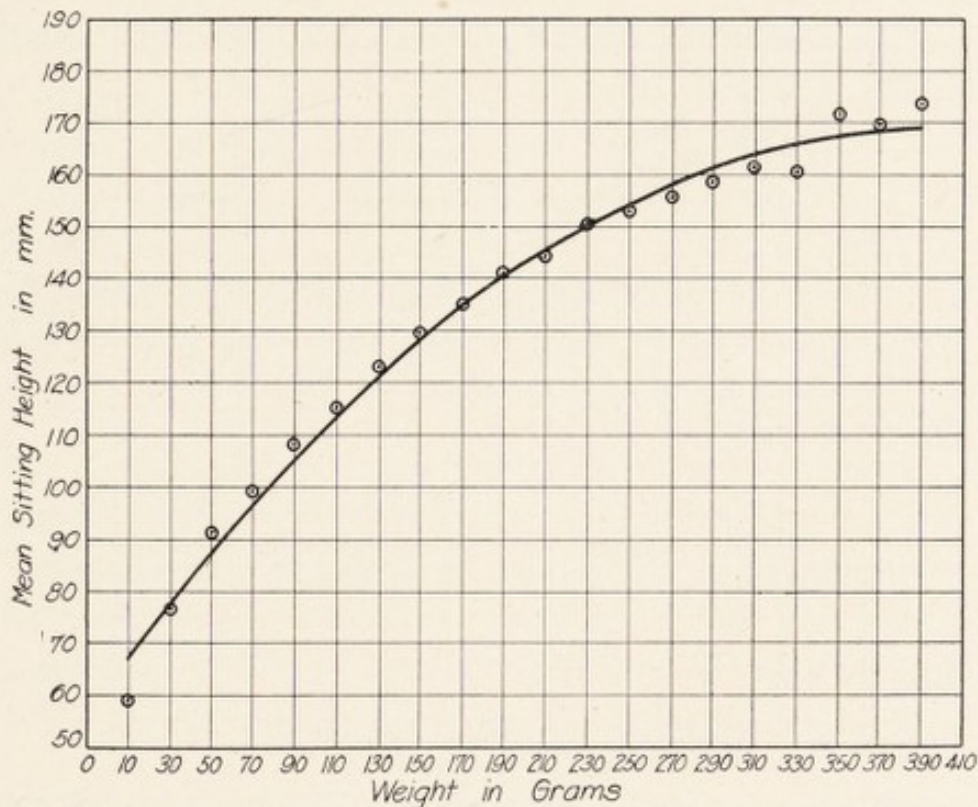
Fig. 88.—Observed mean sitting height of embryo (circles) and parabola of the second order fitted by least squares.

first 6 or 8 observations are curving too rapidly to be capable of representation by a second order parabola, if the same parabola is to come anywhere near the remaining observations. At the low values of x a logarithmic curve is curving relatively rapidly as compared with what it does at higher values of x. But this is precisely what the observations in this case actually do. Hence one perceives that there is needed in the equation a term in $\log x$. But it is further seen that the observations are more spread out

horizontally, that is, the whole series is flatter, than could be represented by

$$y = c \log x$$

whatever value might be given to c. So there is put in a line term, $b\,x$, which has the effect of stretching the curve horizontally. Finally, since all the observations have fairly considerable values (starting at 58.8) it will be desirable to put in a constant term a to raise the general level, from which the terms in x operate, up to a reasonable point.

For the form of logarithmic curve chosen the normal equations are:

$$n\,a + b\,S\,(x) \qquad\qquad + c\,S\,(\log x) \quad = S\,(y)$$
$$a\,S\,(x) + b\,S\,(x^2) \qquad\quad + c\,S\,(x \log x) = S\,(xy)$$
$$a\,S\,(\log x) + b\,S\,(x \log x) + c\,S\,(\log x)^2 \quad = S\,(y \log x)$$

The numerical values here are again drawn from Table 88 and for $S\,(\log x)$, $S\,(x \log x)$ and $S\,(\log x)^2$ from table of sums of logarithmic functions given as Appendix V of this book.

The final equations are

$$20\,a + 210\,b \qquad\qquad\quad + \quad 18.3861246\,c = 2673.7$$
$$210\,a + 2870\,b \qquad\qquad + \quad 230.0033043\,c = 31640.5$$
$$18.3861246\,a + 230.0033043\,b + \quad 19.2694686\,c = 2677.6433$$

Solving, we have

$$y = 54.347 + 1.555\,x + 68.549 \log x$$

Substituting successive values of x as before and solving for y gives the values in the last column of Table 88, which are shown graphically in comparison with the observations in Fig. 89.

It is at once apparent that we now have a much more satisfactory graduation than any attained in the other trials. We could do still better by introducing another term in the equation, but, on the whole, the present result may be taken as reasonably satisfactory.

A final word may be said as to the writing of normal equations in fitting by least squares. In the first place it must always be remembered that the method cannot be applied directly in any

case where any one of the functions of the independent variable involves an arbitrary constant. If, for example, in fitting a log curve we wish to use a term in the equation of the form log $(a + x)$, which it is often convenient to do because it changes the origin of the log term without correspondingly changing the origin of the terms in simple powers of x, it is necessary to go through a round-about process of trial and error to get a proper value of a. It cannot be determined directly by the least square method.
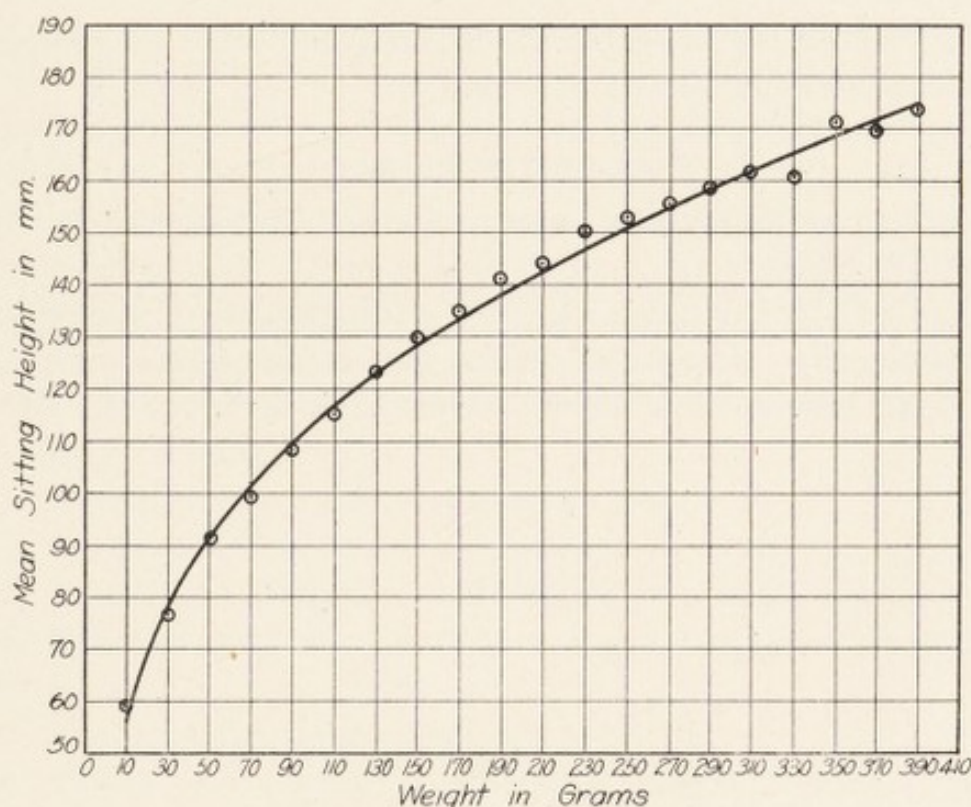
Fig. 89.—Observed mean sitting heights of embryo (circles), and a logarithmic curve fitted by least squares.

But with this caution in mind we can lay down a series of rules as follows:

1. Write the equation of the curve it is proposed to fit with the summation sign S before the variable, in each term which contains a variable (*i. e.*, x or y) and write n before any term which does not contain a variable. Call the equation (i).

2. Multiply each term in (i) by the function of x (x itself, x^2, x^3, log x, etc.) that has for its coefficient the first constant in

(i), writing S before the variable in each case, and dropping the n which appears in (i).

3. Multiply each term in (i) by the function of x, that has for its coefficient the second constant in (i), writing S before the variable in each case as before.

4. Continue this process till (i) has been successively multiplied in this way by each function of x which appears in it. This will make as many equations (including (i)) as there are constants to determine.

5. Perform the indicated summations and solve the system of simultaneous equations for the unknowns.

SUGGESTED READING

1. Ellis, R. L.: On the Method of Least Squares, Trans. Cambridge Phil. Soc., vol. 8, pp. 204–219, 1849.
2. Running, T. R.: Empirical Formulas, New York (John Wiley & Sons), 1917.
3. Brunt, D.: The Combination of Observations, Cambridge (University Press), 1917.
4. Whittaker, E. T., and G. Robinson: The Calculus of Observations, a Treatise on Numerical Mathematics, London and Glasgow (Blackie & Son, Ltd.), 1929.

CHAPTER XVII

THE LOGISTIC CURVE

In 1838 a Belgian mathematician, P. F. Verhulst[1] published a note, later to be followed by two longer memoirs, suggesting the use of a curve which he called the "logistic" to describe the growth of human populations. His work was for many years forgotten. In 1918 Du Pasquier[2] called attention to Verhulst's work. In 1920 Pearl and Reed,[3] without knowing of Verhulst's contribution, independently derived the logistic curve, as an empirical curve to meet certain postulates for a curve to describe the growth of a population. It was held that a curve to describe adequately the growth of population in an area of fixed limits should fulfill the following conditions:

1. Asymptotic to a line $y = k$, when $x = +\infty$.
2. Asymptotic to a line $y = 0$, when $x = -\infty$.
3. A point of inflection at some time $x = a$ and $y = \beta$.
4. Concave upward to left of $x = a$ and concave downward to right of $x = a$.
5. No horizontal slope except at $x = \pm\infty$.
6. Values of y varying continuously from 0 to k as x varies from $-\infty$ to $+\infty$.

These postulates led to the simple, symmetrical logistic curve of Verhulst

$$y = \frac{K}{1 + Ce^{rt}} \qquad\qquad (i)$$

where y denotes population, t denotes time, and K, C, and r are constants.

Shortly after the first paper by Pearl and Reed a rational, as distinguished from an empirical, derivation of the curve was given by Lotka[5, 1b] in his important book "Elements of Physical Biology." Prior to this work on population growth a number of persons, not-

ably the late Dr. T. Brailsford Robertson, on the basis of an assumed analogy between organic growth and chemical autocatalysis, had used the same curve to describe the growth in size of an individual organism.

In practical work with the logistic curve it soon became apparent that Pearl and Reed's postulate 2 was too rigid; that in fact the lower asymptote of the curve often was not zero but was distant from zero by some amount which may be called d. The curve becomes

$$y - d = \frac{K}{1 + Ce^{rt}} \qquad \text{(ii)}$$

Also it became apparent that any complete theory of population growth demanded recognition of its occasionally cyclic character, together with the possibility of skew as well as symmetrical growth. This led to the generalized logistic[4]

$$y = \frac{K}{1 + Ce^{a_1 t + a_2 t^2 + a_3 t^3 + \ldots a_n t^n}} \qquad \text{(iii)}$$

The logistic curve has in recent years been extensively discussed, and its usefulness demonstrated for the description of many sorts of phenomena.[5] This fact appears to justify the inclusion in this text-book of a brief description, with a numerical example, of the method of fitting this curve, a varied and extensive experience having shown the method here described to be simple and accurate.

The equation to the logistic may be written in the form

$$y = \frac{K}{1 + e^{a + rt}} \qquad \text{(iv)}$$

The rate of change of y with respect to t, that is, the increase in mass per unit of time, is given by the equation

$$\frac{dy}{dx} = -\frac{Kre^{a + rt}}{(1 + e^{a + rt})^2} \qquad \text{(v)}$$

By substitution from equation (iv) this may be put in the form

$$y' = \frac{dy}{dx} = -\frac{ry(K - y)}{K} \qquad \text{(vi)}$$

In Fig. 90 are shown the graphs of equations (iv) and (v). Study of these graphs, and a little elementary analysis of equations
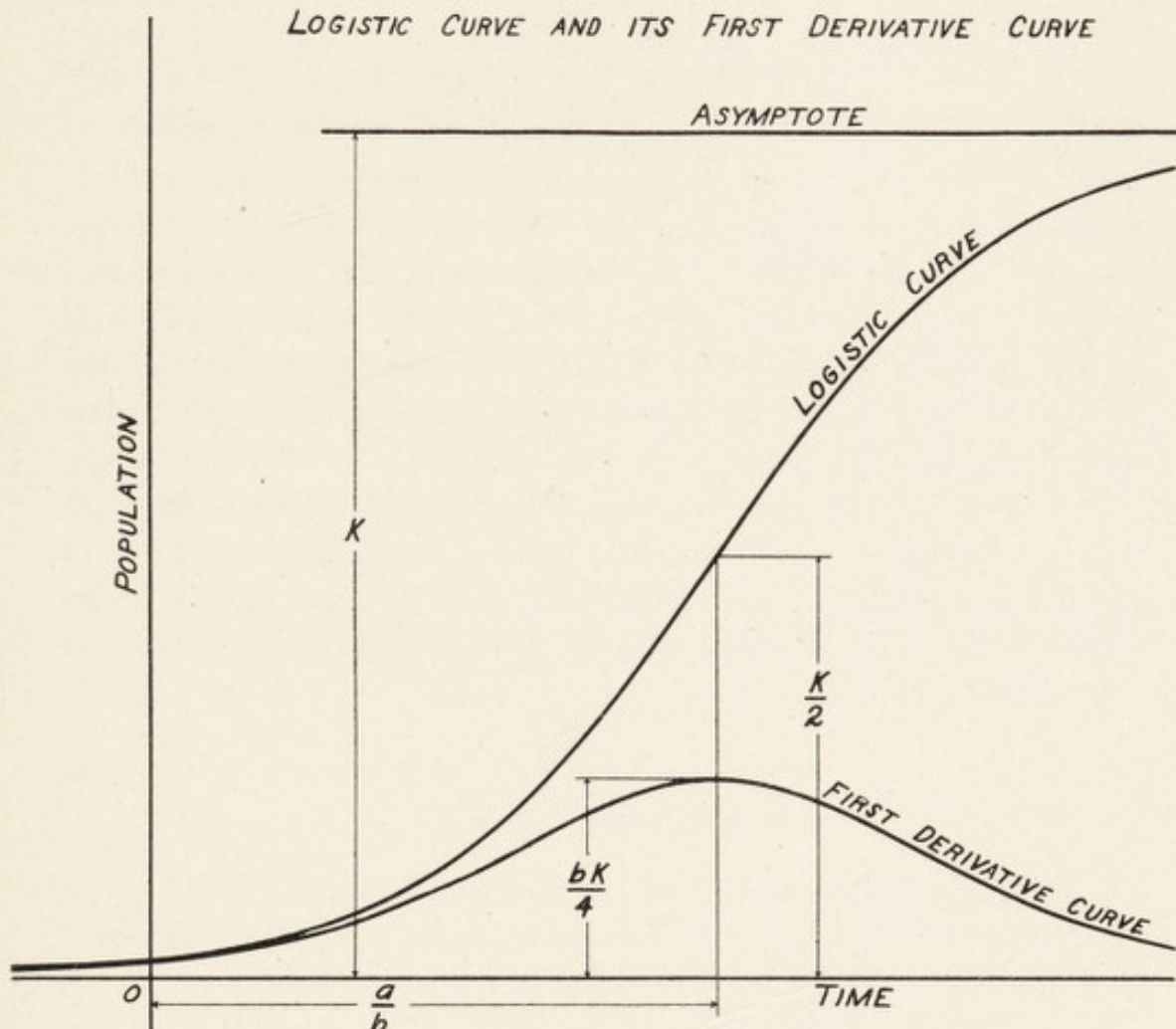
Fig. 90.—Diagram of simple logistic curve.

(iv) and (vi), lead to the following statements which are important in the understanding of the logistic:

(*a*) The logistic is asymptotic to a line K units above the t axis and parallel to it.

(*b*) The curve has a point of inflection at the co-ordinates

$$t = -\frac{a}{r}, \text{ and } y = \frac{K}{2}.$$

(c) The time rate of change of the mass y is greatest at the point of inflection, and the rate at this point is given by

$$\frac{dy}{dx} = -\frac{rK}{4}.$$

(d) When K approaches infinity in equation (vi) we have the limiting equation $\frac{dy}{dx} = -ry$. This is the differential equation for geometric increase, and in this equation r indicates the rate of compounding. Thus we may say that in the logistic curve the constant r is the inherent rate of growth of the population, and that this rate diminishes with time. That the rate does not hold to the inherent value r is a result of the damping effect of the factor $(K - y)$, which measures the aggregate of forces that slow down and finally stop the growth.

(e) The constant a (or C in equation (1)) is obviously the constant of integration and therefore defines the relative positions of the origin and the curve. If the origin on the time axis is transferred to the position of the point of inflection, a becomes zero and the curve takes the form

$$y = \frac{K}{1 + e^{rt}} \tag{vii}$$

FITTING THE LOGISTIC

The simplest method of fitting the symmetrical logistic curve depends on the fact that equation (ii) may be changed into the form

$$\log_e \frac{K - (y - d)}{y - d} = \log_e C + rt \tag{viii}$$

In other words $\log_e [K - (y - d)]/(y - d)$ is a straight line function of time. In fitting a set of observations, therefore, we begin by making as good a guess as we can as to the values of the upper and lower asymptotes. The lower asymptote gives the value of d, while the value of K is obtained by subtracting d from the upper asymptote. We then calculate $Z \equiv [K - (y - d)]/(y - d)$ for each observation and plot each value as an ordinate on arithlog paper against the corresponding time as an abscissa. If we have made good guesses as to the values of the upper and lower asymptotes and if the observations are approximately symmetrical, the

plotted points should fall nearly on a straight line. If they do not, we try new values of d or K or both until the resulting values of Z plotted on arithlog paper are approximately fitted by a straight line, which we determine by eye.* From any two convenient points on this line we determine the slope m and the value of Z when $t = 0$. This value, Z_0, $= C$ while $r = 2.30259m$.

As an example we shall fit the growth of the population of Sweden from 1750 to 1920, as shown in Table 89. As a first assumption we take the upper asymptote equal to 7.66 million population and the lower asymptote equal to 1.56 million. Therefore $d = 1.56$ and $K = 7.66 - 1.56 = 6.10$. The calculation of Z is shown in Table 89. Plotting the values of Z on arithlog paper against the corresponding values of t, as shown in Fig. 91, we see that they are satisfactorily fitted by a straight line. The unit in which t is measured is one year, and the origin the year 1800. To find the values of r

TABLE 89

FITTING OF LOGISTIC CURVE TO POPULATION OF SWEDEN: FIRST APPROXIMATION BY GRAPHIC METHOD

$d = 1.56$; $K = 6.10$

Year.	Population in millions (observed). y	$y - d = y'$	$\dfrac{K - y}{y'} = Z$	$t.$	$e^{rt}.$	$1 + Ce^{rt}.$	y calculated.	Difference calculated −observed.
1750	1.763	.203	29.0	−50	3.245749	23.999994	1.8142	.0512
60	1.893	.333	17.3	−40	2.564791	19.174596	1.8781	−.0149
70	2.030	.470	12.0	−30	2.026698	15.361567	1.9571	−.0729
80	2.118	.558	9.93	−20	1.601497	12.348512	2.0540	−.0640
90	2.158	.598	9.20	−10	1.265503	9.967595	2.1720	.0140
1800	2.347	.787	6.75	0	1.000000	8.086190	2.3144	−.0326
10	2.378	.818	6.46	10	.7902000	6.5995073	2.4843	.1063
20	2.585	1.025	4.95	20	.6244160	5.4247304	2.6845	.0995
30	2.888	1.328	3.59	30	.4934136	4.4964225	2.9166	.0286
40	3.139	1.579	2.86	40	.3898954	3.7628729	3.1811	.0421
50	3.483	1.923	2.17	50	.3080952	3.1832211	3.4763	−.0067
60	3.860	2.300	1.65	60	.2434568	2.7251811	3.7984	−.0616
70	4.168	2.608	1.34	70	.1923796	2.3632384	4.1412	−.0268
80	4.566	3.006	1.03	80	.1520184	2.0772313	4.4966	−.0694
90	4.785	3.225	.891	90	.1201249	1.8512279	4.8551	.0701
1900	5.136	3.576	.706	100	.09492267	1.6726401	5.2069	.0709
10	5.522	3.962	.540	110	.07500790	1.5315202	5.5430	.0210
20	5.904	4.344	.404	120	.05927123	1.4200072	5.8558	−.0482

* Of course a straight line can be fitted exactly by least squares according to the method given in Chapter XVI if this refinement is thought desirable. It should be noted, however, that in fitting this line by least squares, points near either asymptote will have undue influence in determining the slope.

and C, we take the value of Z at $t = -50$, $Z_{-50} = 23$, and the value of Z at $t = +120$, $Z_{120} = 0.42$; log $Z_{-50} = 1.3617278$, log $Z_{120} = 9.6232493 - 10$.
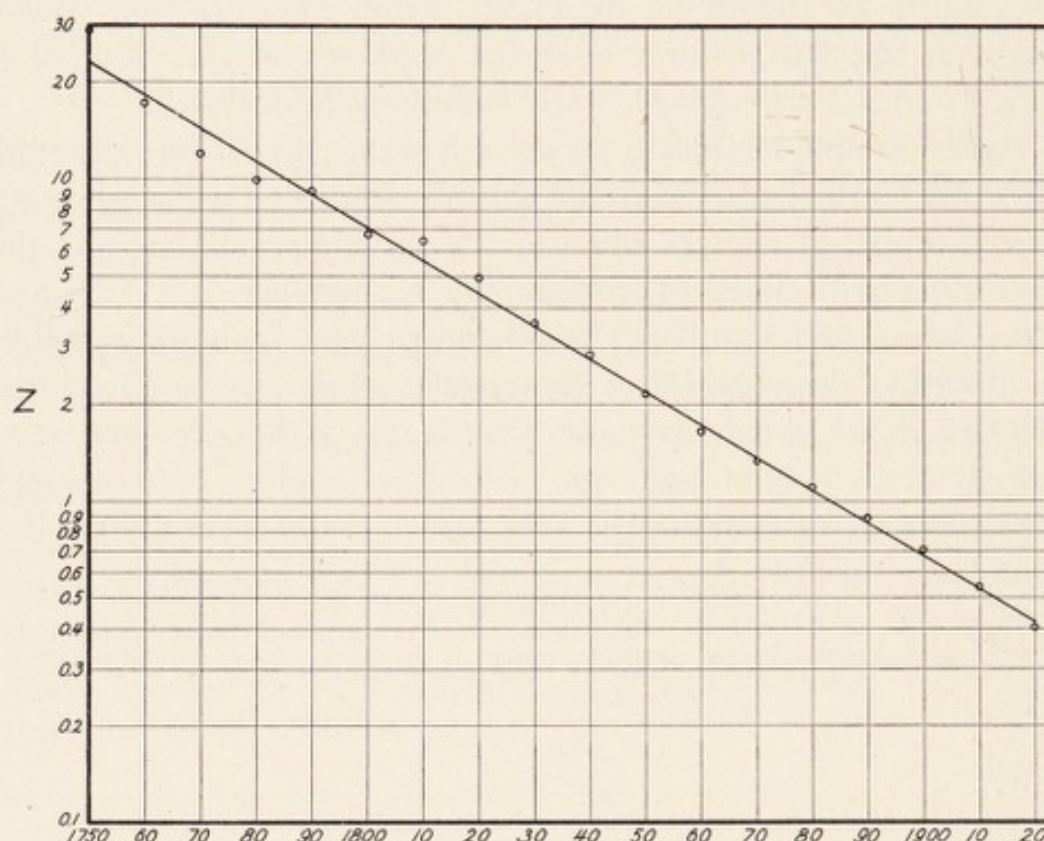
Fig. 91.—Plot of $Z = \dfrac{K - {}'(y - {}''d)}{(y - d)}$ on arithlog paper in fitting logistic curve to the population growth of Sweden.

$$m = \frac{\log Z_{120} - \log Z_{-50}}{120 - (-50)} = \frac{-1.738\ 4785}{170} = -0.010\ 2263$$

$$r = 2.302\ 59m = -0.023\ 5470$$

$$\log Z_0 = \log Z_{-50} + 50m = 0.850\ 4128$$

$$C = Z_0 = 7.086\ 190$$

In other words, the symmetrical logistic curve which we have fitted to the population growth of Sweden is

$$y - 1.56 = \frac{6.10}{1 + 7.086e^{-0.0235t}} \tag{ix}$$

where y is population in millions and t is time in years from the year 1800.

Having got the equation of our curve, we next wish to calculate from it the populations at the different census years to compare with the observed populations. These calculations are shown in Table 89. Since $e^{rt} = 10^{mt}$, we multiply the successive values of t by m, and look up the antilogarithms in a logarithm table. These antilogarithms are the successive values of e^{rt}. K is divided by the successive values of $1 + Ce^{rt}$; the quotient, plus d, is the population value from the curve for a given census year. The last column of Table 89 shows the differences between the calculated and observed values. The largest of these is 0.106 million or about 4 per cent. of the population of that year. We square each difference, add, divide by the number of differences, and get the square root of the quotient. The result, which is called the root-mean-square deviation (and which, be it noted, is closely related to the standard deviation) turns out to be 0.0574 millions.

This is a good fit of the curve to the observations. It is desirable, however, to get the best fitting curve. Now, as has already been seen, the method of least squares cannot be applied to the fitting of a curve in which, as is the case in the logistic curve, the constants to be determined enter the expression in other than a linear manner. However, when we already have a good fit of the curve, this can be treated as a first approximation to the best fitting curve sought and expanded by Taylor's theorem. Neglecting terms of higher order than the first we have an expression linear in the correction terms which are to be determined. Normal equations can then be formed in the manner already explained in Chapter XVI and solved for the unknown correction terms. If the second approximation to the best fitting curve thus obtained is not close enough to suit, we may repeat the process to obtain a third approximation and so on. The details of the work are as follows:

Let ρ be an approximate value of r, and h be the correction to ρ ($r = \rho + h$)
C' be an approximate value of C, and i be the correction to C' ($C = C' + i$)
σ be an approximate value of d, and j be the correction to σ ($d = \sigma + j$)
K' be an approximate value of K, and k be the correction to K' ($K = K' + k$)

Then

$$y = d + \frac{K}{1 + Ce^{rt}} = \sigma + j + \frac{K' + k}{1 + (C' + i)e^{(\rho + h)t}}$$

or, approximately, expanding by Taylor's theorem and neglecting terms of higher order than the first

$$y = \left(\sigma + \frac{K'}{1 + C'e\rho t}\right) + h\left(\frac{-K'e\rho tC't}{(1+C'e\rho t)^2}\right) + i\left(\frac{-K'e\rho t}{(1+C'e\rho t)^2}\right) + j + k\left(\frac{1}{1+C'e\rho}\right) \text{ (xi)}$$

Let

$$\frac{-K'e\rho tC't}{(1+C'e\rho t)^2} = a; \quad \frac{-K'e\rho t}{(1+C'e\rho t)^2} = b; \quad \frac{1}{1+C'e\rho t} = c; \quad \sigma + \frac{K'}{1+C'e\rho t} - y = l \text{ (xii)}$$

(In calculating these quantities we may note that $a = C'tb$; $b = -K'e\rho tc^2$; $l = \sigma + K'c - y$.)

Then

$$ah + bi + j + ck + l = 0$$

From this are derived the following normal equations:

$$\left.\begin{array}{l} h\Sigma(a^2) + i\Sigma(ab) + j\Sigma(a) + k\Sigma(ac) + \Sigma(al) = 0 \\ h\Sigma(ab) + i\Sigma(b^2) + j\Sigma(b) + k\Sigma(bc) + \Sigma(bl) = 0 \\ h\Sigma(a) + i\Sigma(b) + jN + k\Sigma(c) + \Sigma(l) = 0 \\ h\Sigma(ac) + i\Sigma(bc) + j\Sigma(c) + k\Sigma(c^2) + \Sigma(cl) = 0 \end{array}\right\} \text{ (xiii)}$$

TABLE 90

FITTING OF LOGISTIC CURVE TO POPULATION OF SWEDEN: SECOND APPROXIMATION BY LEAST SQUARES—CALCULATION OF PRODUCT-SUMS

$\rho = -0.0235470$; $C' = 7.086190$; $\sigma = 1.56$; $K' = 6.1$

Year.	y.	t.	a.	b.	c.	l.	s.
1750	1.763	−50	12.178822	−.03437340	.04166668	.05116675	12.237283
60	1.893	−40	12.061526	−.04255293	.05215234	−.01487073	12.056254
70	2.030	−30	11.137352	−.052 8994	.06509753	−.07290507	11.077155
80	2.118	−20	9.079659	−.06406587	.08098142	−.06401334	9.032561
90	2.158	−10	5.505860	−.07769845	.10032510	.01398311	5.542470
1800	2.347	0	0	−.09329148	.12366764	−.03262740	−.002250
10	2.378	10	−7.842537	−.11067354	.15152646	.10631141	−7.695374
20	2.585	20	−18.343844	−.12943376	.18434096	.09947986	−18.189457
30	2.888	30	−31.647599	−.14886984	.22239903	.02863408	−31.545436
40	3.139	40	−47.611473	−.16797275	.26575439	.04210178	−47.471590
50	3.483	50	−65.71492	−.1854732	.31414720	−.00670208	−65.59294
60	3.860	60	−85.02076	−.1999682	.36694809	−.06161665	−84.91540
70	4.168	70	−104.22805	−.2101231	.42314817	−.02679616	−104.04182
80	4.566	80	−121.83133	−.2149098	.48141004	−.06939876	−121.63423
90	4.785	90	−136.36356	−.2138174	.54018201	.07011026	−135.96709
1900	5.136	100	−146.65862	−.2069640	.59785724	.07092916	−146.19679
10	5.522	110	−152.05365	−.1950704	.65294601	.02097066	−151.57480
20	5.904	120	−152.47080	−.1793051	.70422178	−.04824714	−151.99414
			−1019.82392	−2.5269531	5.36877209	.10650974	−1016.87561

$\Sigma(a^2)$	127,921.79	$\Sigma(ab)$	207.20940	$\Sigma(ac)$	−543.70734
$\Sigma(ab)$	207.20940	$\Sigma(b^2)$	.42744085	$\Sigma(bc)$	−.96685
$\Sigma(ac)$	−543.70734	$\Sigma(bc)$	−.96685321	$\Sigma(c^2)$	2.4332851
$\Sigma(al)$	−5.310004	$\Sigma(bl)$	−.018225622	$\Sigma(cl)$	.031328884
	127,579.98		206.65176		−542.20957
$\Sigma(as)$	127,579.99	$\Sigma(bs)$	206.65174	$\Sigma(cs)$	−542.20955

The calculation of the various sums is shown in Table 90. Each item in column s is the sum of the corresponding items in columns a, b, c, and l. In this way we have a check on the calculation of the various product-sums entering into the normal equations.

The values of the correction terms may be found from the normal equations by any of the algebraic methods for solving simultaneous equations. However, it is usually most advantageous to solve by a method developed by the great astronomer and mathe-

TABLE 91

FITTING OF LOGISTIC CURVE TO POPULATION OF SWEDEN: SECOND APPROXIMATION BY LEAST SQUARES—SOLUTION OF NORMAL EQUATIONS BY DOOLITTLE METHOD

	h.	i.	j.	k.
	127,921.79	207.20940	−1019.8239	−543.70734	−5.310004
−.0000078172765	$h =$	−.0016198132	.0079722454	.0042503106	.00004150977
		.09180033	−.8750289	−.08614890	−.009624407
−10.893207		$i =$	9.531871	.9384378	.10484066
			1.529051	.2130436	−.02756153
−.6540004			$j=$	−.1393306	.01802525
				.0118311	.00356796
		.42744085	−2.5269531	−.96685321	−.018225622
		−.33564052	1.6519242	.88070431	.008601215
			18.	5.3687721	.10650974
			−8.130286	−4.3345683	−.04233266
			−8.340663	−.8211602	−.09173861
$l. = -.00140006;\ r\ =\rho\ +h = -.02494706$				2.4332851	.03132888
$i = +.394162;\ \ C\ = C'\ +i\ =\ 7.480352$				−2.3109251	−.02256917
$\ = +.0600439;\ d\ =\sigma\ +j\ =\ 1.6200439$				−.0808454	−.00903191
$l = -.301575;\ \ K\ = K'\ +k\ =\ 5.798425$				−.0296835	.00384016

Upper asymptote = 7.418

Check
−543.707 h = .761222
−.966853 i = −.381097
5.36877 j = .322362
2.43329 k = −.733819
.031329
—————
−.000003

matician, Gauss, or by a modification of this method devised by M. H. Doolittle, of the United States Coast and Geodetic Survey. Both of these methods are described by Brunt (see reference 3, Chapter XVI). The solution of the normal equations by the Doolittle method is shown in Table 91, and the calculation of ordinates from the new curve in Table 92. This does not differ in principle from the calculation of ordinates of the first approximation curve already shown in Table 89. For the new curve the root-mean-

TABLE 92

FITTING OF LOGISTIC CURVE TO POPULATION OF SWEDEN: SECOND APPROXIMATION
BY LEAST SQUARES—CALCULATION OF ORDINATES

Year.	t.	e^{rt}.	$1 + Ce^{rt}$.	y calculated.	y observed.	Difference.
1750	−50	3.48112	27.0400	1.8344	1.763	+.0714
60	−40	2.71253	21.2907	1.8923	1.893	−.0007
70	−30	2.11364	16.8108	1.9649	2.030	−.0651
80	−20	1.64698	13.3200	2.0553	2.118	−.0627
90	−10	1.28335	10.5999	2.1670	2.158	+.0090
1800	0	1.00000	8.48035	2.3037	2.347	−.0433
10	10	.779213	6.82879	2.4691	2.378	+.0911
20	20	.607174	5.54188	2.6663	2.585	+.0813
30	30	.473117	4.53908	2.8974	2.888	+.0094
40	40	.368659	3.75770	3.1631	3.139	+.0241
50	50	.287264	3.14884	3.4614	3.483	−.0216
60	60	.223840	2.67440	3.7881	3.860	−.0719
70	70	.174419	2.30472	4.1359	4.168	−.0321
80	80	.135910	2.01665	4.4953	4.566	−.0707
90	90	.105903	1.792192	4.8554	4.785	+.0704
1900	100	.0825207	1.617284	5.2053	5.136	+.0693
10	110	.0643012	1.480996	5.5352	5.522	+.0132
20	120	.0501044	1.374799	5.8377	5.904	−.0663

Root mean square deviation   = .0562
Root mean square deviation of first approximation = .0574
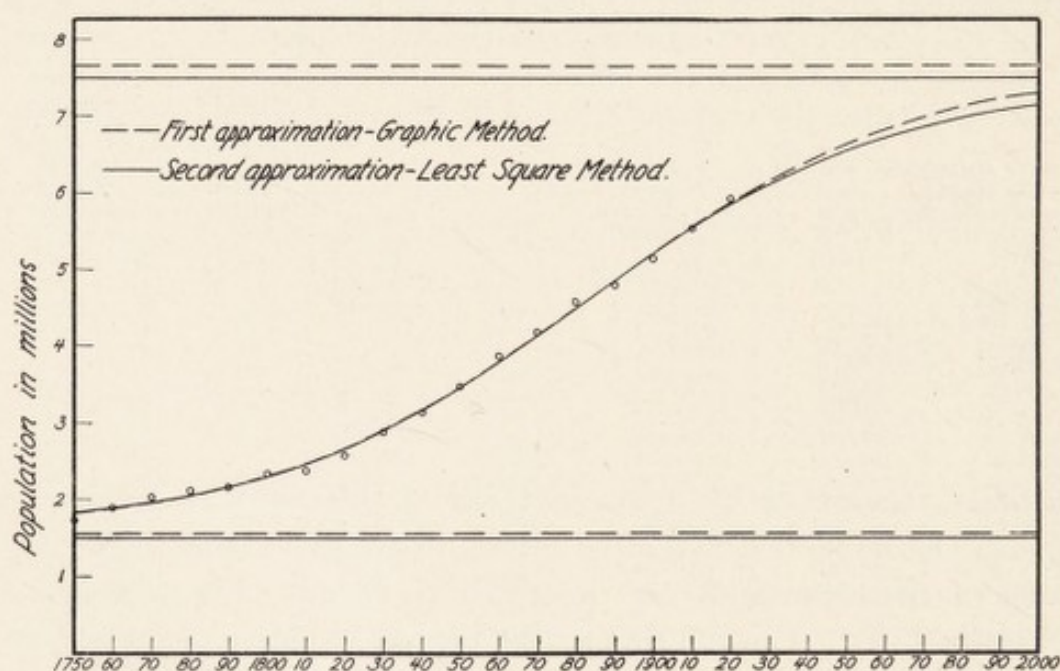
Fig. 92.—The population growth of Sweden fitted with two symmetrical logistic curves.

square deviation is .0562, as compared with the value .0574 already
found for the old curve. In other words, the new curve is a slightly

better fit to the observations than the old curve, but only slightly. This is apparent graphically in Fig. 92.

In some cases a growth curve is skew, *i. e.*, the curvature in the early part of growth is at a different rate from the curvature in the latter part of growth. In these cases Z plotted on arithlog paper cannot be fitted with a straight line, whatever values of d and K are chosen, but must be fitted with a parabola of odd degree; usually a cubic parabola is sufficient. If it is not, a fifth order parabola may be used. The remainder of the work of fitting a skew logistic follows the same principles as have been set forth above in detail for the symmetrical logistic.

SUGGESTED READING

1. Verhulst, P. F.: (*a*) Notice sur la loi que la population suit dans son accroissement. Corr. math. et phys. publ. par A. Quetelet. T. X (also numbered T. II of the third series), pp. 113–121, 1838. (*b*) Recherches mathématiques sur la loi d'accroissement de la population. Nouv. mém. de l'Acad. Roy. des Sci. et Belles-lett. de Bruxelles. T. 18, pp. 1–38, 1845. (Read November 30, 1844.) (*c*) Deuxième mémoire sur la loi d'accroissement de la population. Ibid., T. 20, pp. 1–32, 1847. (Read May 15, 1846.)

2. Du Pasquier, L. G.: Esquisse d'une nouvelle théorie de la population, Viertel-jahrschr. des Naturforsch. Ges. Zürich, Jahrg. 63, pp. 236–249, 1918.

3. Pearl, R., and Reed, L. J.: On the Rate of Growth of the Population of the United States since 1790 and Its Mathematical Representation, Proc. Nat. Acad. Sci., vol. 6, pp. 275–288, 1920.

4. Pearl, R., and Reed, L. J.: (*a*) A Further Note on the Mathematical Theory of Population Growth, Proc. Nat. Acad. Sci., vol. 8, pp. 365–368, 1922; (*b*) On the Mathematical Theory of Population Growth, Metron, vol. 3, pp. 6–19, 1923; (*c*) The Probable Error of Certain Constants of the Population Growth Curve, Amer. Jour. Hygiene, vol. 4, pp. 239–240, 1924; (*d*) Skew Growth Curves, Proc. Nat. Acad. Sci., vol. 11, pp. 16–22, 1925.

5. Some of the more important references to the recent literature on the logistic curve, in addition to those cited above, are:

  I. *Mathematical Aspects:*
  (*a*) Pearl, R.: Studies in Human Biology, Baltimore (Williams and Wilkins), 1924, Chap. XXIV.
  (*b*) Lotka, A. J.: Elements of Physical Biology, Baltimore (Williams and Wilkins), 1925, Chap. VII.
  (*c*) Yule, G. U.: The Growth of Population and the Factors which Control It, Jour. Roy. Stat. Soc., vol. 88, pp. 1–59, 1925.
  (*d*) Reed, L. J., and Pearl, R.: On the Summation of Logistic Curves, Jour. Roy. Stat. Soc., vol. 90, pp. 730–746, 1927.
  (*e*) Reed, L. J., and Berkson, J.: The Application of the Logistic Function to Experimental Data, Jour. Phys. Chem., vol. 33, pp. 760–779, 1929.

(f) Lotka, A. J.: Biometric Functions in a Population Growing in Accordance with a Prescribed Law, Proc. Nat. Acad. Sci., vol. 15, pp. 793–798, 1929.

(g) Schultz, H.: The Standard Error of a Forecast from a Curve, Jour. Amer. Stat. Assoc., vol. 25, pp. 139–185, 1930.

II. *Applications:*

(a) Pearl, R.: Studies in Human Biology, Chap. XXV.

(b) Pearl, R.: The Biology of Population Growth, New York (Alfred Knopf, Inc.), 1925.

(c) Pearl, R., and Reed, L. J.: Predicted Growth of Population of New York and Its Environs; New York (Plan of New York and Its Environs), 1923, pp. 42.

(d) Pearl, R.: The Growth of Populations, Quarterly Rev. Biol., vol. 2, pp. 532–548, 1927.

(e) Gover, M.: Increase of the Negro Population in the United States, Human Biology, vol. 1, pp. 263–273, 1929.

(f) Anderson, D. D.: The Point of Population Saturation, Its Transgression in Mauritius, Human Biology, vol. 1, pp. 528–543, 1929.

(g) Belz, M. H.: Theories of Population and Their Application to Australia, The Economic Record, November, 1929, pp. 253–262.

(h) Monk, A. T., and Jeter, H. R.: The Logistic Curve and the Prediction of the Population of the Chicago Region, Jour. Amer. Stat. Assoc., vol. 23, pp. 361–385, 1928. This paper is discussed by R. Pearl and L. J. Reed: The Population of an Area Around Chicago and the Logistic Curve, Ibid., vol. 24, pp. 66, 67, 1929.

APPENDIX I

AIDS TO BIOMETRIC WORKERS

The following tables are indispensable to the biometric worker.

1. Pearson, K. (Editor): Tables for Statisticians and Biometricians, Cambridge University Press, 1914.
2. Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, Reciprocals, London (E. & F. N. Spon, Ltd.), 1919.
3. Bruhns, C.: Neues logarithmisch-trigonometrisches Handbuch auf sieben Decimalen, Leipzig (Tauchnitz), 1919.   (Any other 7-place table will do, but Bruhns is surpassed by none.)
4. Miner, J. R.: Tables of $\sqrt{1 - r^2}$ and $1 - r^2$ for Use in Partial Correlation and in Trigonometry, Baltimore (The Johns Hopkins Press), 1922.

In addition to the above, the following will be found useful:

Glover, J. W.: Tables of Applied Mathematics in Finance, Insurance, Statistics, Ann Arbor, Mich. (George Wahr), 1923.   (This contains what appears to be a photographic reprint of Bruhns' 7-place logarithms of numbers.)

Carr, G. S.: A Synopsis of Elementary Results in Pure Mathematics: Containing Propositions, Formulæ, and Methods of Analysis, with Abridged Demonstrations, London (Francis Hodgson), 1886.   (This book is out of print and, therefore, difficult to acquire, but to him who has it it is an invaluable desk companion.)

APPENDIX II

MATHEMATICAL FORMULÆ AND CONSTANTS

MULTIPLICATION

(1) $1a = a;\; 3a = a + a + a$

(2) $(a + b)\, c = ac + bc$

(3) $(a - b)\, c = ac - bc$

(4) $(a + b) \cdot (c + d) = (a + b)\, c + (a + b)\, d$
$$= ac + ad + bc + bd$$

(5) $(a - b) \cdot (c + d) = (a - b)\, c + (a - b)\, d$
$$= ac + ad - bc - bd$$

(6) $(a + b)\, (c - d) = (a + b)\, c - (a + b)\, d$
$$= ac - ad + bc - bd$$

(7) $(a - b)(c - d) = (a - b)c - (a - b)d$
$$= ac - ad - bc + bd$$

(8) $(a + 1)b = ab + b$

(9) $(a - 1)b = ab - b$

(10) $(a + b)(c + 1) = ac + bc + a + b$

(11) $(a + b)(c - 1) = ac + bc - a - b$

(12) $(a - b)(c + 1) = ac - bc + a - b$

(13) $(a - b)(c - 1) = ac - bc - a + b$

(14) $ab = ba$

(15) $a \cdot 0 = 0$

(16) $(+ a) \cdot (+ b)$ or $(- a) \cdot (- b) = + ab$

(17) $(+ a) \cdot (- b)$ or $(- a) \cdot (+ b) = - ab$

DIVISION

(1) $\dfrac{a}{b} \cdot b$ or $\dfrac{ab}{b} = a$

(2) $\dfrac{ab}{c} = \dfrac{a}{c} \cdot b = \dfrac{b}{c} \cdot a$

(3) $\dfrac{a}{b} : c = \dfrac{a}{b} \cdot \dfrac{1}{c} = \dfrac{a}{bc}$

(4) $\dfrac{a}{b} = \dfrac{a \cdot c}{b \cdot c} = \dfrac{a : c}{b : c}$

(5) $\dfrac{a}{b} \cdot \dfrac{c}{d} = \dfrac{ac}{bd}$

(6) $\dfrac{a}{b} : \dfrac{c}{d} = \dfrac{ad}{bc}$

(7) $\dfrac{a}{c} + \dfrac{b}{c} = \dfrac{a + b}{c}$

(8) $\dfrac{a}{c} - \dfrac{b}{c} = \dfrac{a - b}{c}$

(9) $a + \dfrac{b}{c} = \dfrac{ac + b}{c}$

(10) $a - \dfrac{b}{c} = \dfrac{ac - b}{c}$

(11) $\dfrac{a}{b} + 1 = \dfrac{a+b}{b}$

(12) $\dfrac{a}{b} - 1 = \dfrac{a-b}{b}$

(13) $\dfrac{1}{a} + \dfrac{1}{b} = \dfrac{a+b}{ab}$

(14) $\dfrac{1}{a} - \dfrac{1}{b} = \dfrac{b-a}{ab} = -\dfrac{a-b}{ab}$

(15) $\dfrac{a}{b} + \dfrac{c}{d} = \dfrac{ad+bc}{bd}$

(16) $\dfrac{a}{b} - \dfrac{c}{d} = \dfrac{ad-bc}{bd}$

(17) $\dfrac{a}{a+b} + 1 = \dfrac{2a+b}{a+b}$

(18) $\dfrac{a+b}{2} + \dfrac{a-b}{2} = a$

(19) $\dfrac{a+b}{2} - \dfrac{a-b}{2} = b$

(20) $\dfrac{\dfrac{1}{a} + \dfrac{1}{b}}{\dfrac{1}{a} - \dfrac{1}{b}} = \dfrac{b+a}{b-a}$

(21) $\dfrac{o}{a} = o$

(22) $\dfrac{a}{o} = \infty$

(23) $\dfrac{+a}{+b}$ or $\dfrac{-a}{-b} = +\dfrac{a}{b}$

(24) $\dfrac{+a}{-b}$ or $\dfrac{-a}{+b} = -\dfrac{a}{b}$

Powers

$$a^4 = aaaa; \quad a^1 = a; \quad 1^a = 1$$

(1) $(+a)^n = +a^n$

(2) $(-a)^n = +a^n$, if n is an even number

(3) $(-a)^n = -a^n$, if n is an odd number

(4) $(ab)^m = a^m b^m$

(5) $(a:b)^m = \left(\dfrac{a}{b}\right)^m = \dfrac{a^m}{b^m}$

(6) $a^m \cdot a^n = a^{m+n}$

(7) $a^m : a^n = \dfrac{a^m}{a^n} = a^{m-n}$

(8) $a^{n+1} : a^n = a$

(9) $a^n : a^{n-1} = a$

(10) $\dfrac{1}{a^n} = \left(\dfrac{1}{a}\right)^n$

(11) $(a^m)^n = a^{mn}$

(12) $3a^0 = 3; (3a)^0 = 1$

(13) $a^{-1} = \dfrac{1}{a}$

(14) $\left(a^{\frac{m}{n}}\right)^{\frac{x}{y}} = a^{\frac{mx}{ny}}$

(15) $(a^{n+1})^2 = a^{2n} \cdot a^2$

(16) $a^x \cdot a^{-y} = a^{x-y}$

(17) $a^{-x} a^{-y} = a^{-(x+y)}$

(18) $\dfrac{a^x}{a^{-y}} = a^{x+y}$

(19) $\dfrac{a^{-x}}{a^y} = \dfrac{1}{a^{x+y}}$

(20) $(a^{-x})^y = a^{-xy}$

(21) $(a^x)^{-y} = a^{-xy}$

(22) $(a^{-x})^{-y} = a^{xy}$

(23) $(a+b)^2 = a^2 + 2ab + b^2$

(24) $(a-b)^2 = a^2 - 2ab + b^2$

(25) $a^2 - b^2 = (a+b)(a-b)$

(26) $(a+b+c)^2 = a^2 + 2ab + b^2 + 2ac + 2bc + c^2$

(27) $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$

(28) $(a - b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$

(29) $a^3 + b^3 = (a + b)(a^2 - ab + b^2)$

(30) $a^3 - b^3 = (a - b)(a^2 + ab + b^2)$

(31) $(a + b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$

(32) $(a + b)^5 = a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$

Roots

(1) $\sqrt[n]{a} = b; \quad \sqrt[n]{a^n} = a; \quad \left(\sqrt[n]{a}\right)^n = a$

(2) $\sqrt[n]{a^{mn}} = a^m$

(3) $\sqrt[n]{ab} = \sqrt[n]{a} \cdot \sqrt[n]{b}$

(4) $\sqrt[n]{\dfrac{a}{b}} = \dfrac{\sqrt[n]{a}}{\sqrt[n]{b}}$

(5) $\sqrt[n]{a^m} = \left(\sqrt[n]{a}\right)^m$

(6) $\sqrt[n]{a^m} = \sqrt[np]{a^{mp}}$

(7) $\sqrt[m]{\sqrt[n]{a}} = \sqrt[n]{\sqrt[m]{a}} = \sqrt[mn]{a}$

(8) $\sqrt{a^2} = \pm a; \quad \sqrt{(a + b)^2} = \pm (a + b)$

Incorrect: $\sqrt{a^2 + b^2} = a + b$; and

$$\sqrt[3]{a^3 + b^3} = a + b$$

(9) $(\sqrt{a} + \sqrt{b})(\sqrt{a} - \sqrt{b}) = a - b$

(10) $(a + \sqrt{b})(a - \sqrt{b}) = a^2 - b$

(11) $(\sqrt{a} + b)(\sqrt{a} - b) = a - b^2$

(12) $\sqrt{1 + x} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + \frac{1}{16}x^3 - \frac{5}{128}x^4 + \frac{7}{256}x^5 - \ldots$

(13) $\sqrt{1 - x} = 1 - \frac{1}{2}x - \frac{1}{8}x^2 - \frac{1}{16}x^3 - \frac{5}{128}x^4 - \frac{7}{256}x^5 - \ldots$

(14) $\sqrt[3]{1 + x} = 1 + \frac{1}{3}x - \frac{1}{9}x^2 + \frac{5}{81}x^3 - \frac{10}{243}x^4 + \ldots$

(15) $\sqrt[3]{1 - x} = 1 - \frac{1}{3}x - \frac{1}{9}x^2 - \frac{5}{81}x^3 - \frac{10}{243}x^4 - \frac{22}{729}x^5 - \ldots$

28

Fractional Powers

(1) $a^{\frac{m}{x}} = \sqrt[x]{a^m}$

(2) $a^{\frac{m}{n}} \cdot a^{\frac{p}{q}} = a^{\frac{m}{n} + \frac{p}{q}}$

(3) $a^{\frac{m}{n}} : a^{\frac{p}{q}} = a^{\frac{m}{n} - \frac{p}{q}}$

(4) $(a\,b)^{\frac{m}{n}} = a^{\frac{m}{n}} \cdot b^{\frac{m}{n}}$

(5) $\left(\dfrac{a}{b}\right)^{\frac{m}{n}} = a^{\frac{m}{n}} : b^{\frac{m}{n}}$

(6) $\left(a^{\frac{m}{n}}\right)^{\frac{p}{q}} = a^{\frac{mp}{nq}}$

Logarithms

(1) $\log_a a = 1; \log 1 = 0.$

(2) $\log MN = \log M + \log N$

(3) $\log \dfrac{M}{N} = \log M - \log N.$

(4) $\log (M)^n = n \log M.$

(5) $\log \sqrt[n]{M} = \dfrac{1}{n} \log M.$

Proportion

From $a : b = c : d$ it follows:

(1) $a : c = b : d$
$b : a = d : c$
$b : d = a : c$
$c : a = d : b$
$c : d = a : b$
$d : b = c : a$
$d : c = b : a$

(2) $ad = bc$

(3) $a = \dfrac{bc}{d}; \quad d = \dfrac{bc}{a}; \quad b = \dfrac{ad}{c}; \quad c = \dfrac{ad}{b}$

(4) $ma : mb = c : d$

$ma : b\ \ = mc : d$, etc.

(5) $\dfrac{a}{n} : \dfrac{b}{n} = c : d$

$\dfrac{a}{n} : b = \dfrac{c}{n} : d$, etc.

(6) $a^n : b^n = c^n : d^n$

(7) $\sqrt[n]{a} : \sqrt[n]{b} = \sqrt[n]{c} : \sqrt[n]{d}$

Differential Coefficients of Simple Functions

(1) $y = x^n,\ \dfrac{dy}{dx} = nx^{n-1}$

$y = ax^n,\ \dfrac{dy}{dx} = nax^{n-1}$

(2) $y = \dfrac{a}{x^n},\ \dfrac{dy}{dx} = -\dfrac{na}{x^{n+1}}$

$= ax^{-n},\ \dfrac{dy}{dx} = -nax^{-n-1}$

(3) $y = a\sqrt[n]{x},\ \dfrac{dy}{dx} = \dfrac{a}{n}\sqrt[n]{x^{1-n}}$

$= ax^{\frac{1}{n}},\ \dfrac{dy}{dx} = \dfrac{1}{n}ax^{\frac{1}{n}-1}$

(4) $y = a\sqrt[n]{x^m},\ \dfrac{dy}{dx} = \dfrac{m}{n}a\sqrt[n]{x^{m-n}}$

$= ax^{\frac{m}{n}},\ \dfrac{dy}{dx} = \dfrac{m}{n}ax^{\frac{m}{n}-1}$

(5) $y = \sqrt{x},\ \dfrac{dy}{dx} = \dfrac{1}{2\sqrt{x}}$

$= x^{\frac{1}{2}},\ \dfrac{dy}{dx} = \dfrac{1}{2}x^{-\frac{1}{2}}$

(6) $y = e^x,\ \dfrac{dy}{dx} = e^x$

(7) $y = a^x,\ \dfrac{dy}{dx} = a^x \log_e a$

$$(8) \quad y = \log_e x, \frac{dy}{dx} = \frac{1}{x}$$

$$y = \log x, \frac{dy}{dx} = M \cdot \frac{1}{x}, \text{ where } M = 0.43429$$

$$(9) \quad y = \sin x, \frac{dy}{dx} = \cos x$$

$$(10) \quad y = \cos x, \frac{dy}{dx} = -\sin x$$

$$(11) \quad y = \text{tg } x, \frac{dy}{dx} = \frac{1}{\cos^2 x}$$

$$(12) \quad y = \text{ctg } x, \frac{dy}{dx} = -\frac{1}{\sin^2 x}$$

$$(13) \quad y = \text{arc sin } x, \frac{dy}{dx} = \frac{1}{\sqrt{1 - x^2}}$$

$$(14) \quad y = \text{arc cos } x, \frac{dy}{dx} = -\frac{1}{\sqrt{1 - x^2}}$$

$$(15) \quad y = \text{arc tg } x, \frac{dy}{dx} = \frac{1}{1 + x^2}$$

$$(16) \quad y = \text{arc ctg } x, \frac{dy}{dx} = -\frac{1}{1 + x^2}$$

SIMPLE INTEGRALS

$$(1) \quad \int a \, dx = ax + C$$

$$(2) \quad \int ax^n \, dx = \frac{ax^{n+1}}{n + 1} + C$$

$$(3) \quad \int e^x \, dx = e^x + C$$

$$(4) \quad \int \frac{1}{x} \, dx = \log_e x + C$$

$$(5) \quad \int a^x \, dx = \frac{a^x}{\log_e a} + C$$

$$(6) \quad \int \sin x \, dx = -\cos x + C$$

$$(7) \quad \int \cos x \, dx = \sin x + C$$

(8) $\int a \cos x \, dx = a \cdot \sin x + C$

(9) $\int \dfrac{1}{\cos^2 x} \, dx = \operatorname{tg} x + C$

(10) $\int \dfrac{1}{\sin^2 x} \, dx = - \operatorname{ctg} x + C$

(11) $\int \dfrac{1}{\sqrt{1 - x^2}} \, dx = \arcsin x + C$

$$= - \arccos x + C'$$

(12) $\int \dfrac{1}{1 + x^2} \, dx = \operatorname{arc\,tg} x + C$

$$= - \operatorname{arc\,ctg} x + C'$$

Constants

		log.
Base of Napierian logarithms.................	$e =$ 2.7182818	0.4342945
Log. $e =$ Modulus of common logarithms......	$M =$ 0.4342945	9.6377843 − 10
Radius reduced to seconds....................	206264.8	5.3144251
Radius reduced to minutes....................	3437.7468	3.5362739
Radius reduced to degrees....................	57.29578	1.7581226
360 degrees expressed in seconds.............	1296000	6.1126050
360 degrees expressed in minutes.............	21600	4.3344538
360 degrees expressed in degrees.............	360	2.5563025
Diameter 1, circumference...................	$\pi =$ 3.14159265	0.4971499
	$\dfrac{1}{\pi} =$ 0.3183099	9.5028501 − 10
	$\pi^2 =$ 9.8696044	0.9942997
	$\sqrt{\pi} =$ 1.7724539	0.2485749
	$\sqrt[3]{\dfrac{\pi}{6}}$	9.9063329 − 10

APPENDIX III

TABLES FOR ESTIMATING THE SIGNIFICANCE OF DEVIATIONS

TABLE A

Showing the Probability of Occurrence of Statistical Deviations of Different Magnitudes Relative to the Probable Error

Deviation / P.E.	Probable occurrence of a deviation as great as or greater than designated one in 100 trials.	Odds against the occurrence of a deviation as great as or greater than the designated one.	Deviation / P.E.	Probable occurrence of a deviation as great as or greater than designated one in 100 trials.	Odds against the occurrence of a deviation as great as or greater than the designated one.
1.0	50.00	1.00 to 1	3.3 ...	2.60	37.42 to 1
1.1	45.81	1.18 to 1	3.4 ...	2.18	44.80 to 1
1.2	41.83	1.39 to 1			
1.3	38.06	1.63 to 1	3.5 ...	1.82	53.82 to 1
1.4	34.50	1.90 to 1	3.6 ...	1.52	64.89 to 1
			3.7 ...	1.26	78.53 to 1
1.5	31.17	2.21 to 1	3.8 ...	1.04	95.38 to 1
1.6	28.05	2.57 to 1	3.9 ...	.853	116.3 to 1
1.7	25.15	2.98 to 1			
1.8	22.47	3.45 to 1	4.0 ...	.698	142.3 to 1
1.9	20.00	4.00 to 1	4.1 ...	.569	174.9 to 1
			4.2 ...	.461	215.8 to 1
			4.3 ...	.373	267.2 to 1
2.0	17.73	4.64 to 1	4.4 ...	.300	332.4 to 1
2.1	15.67	5.38 to 1			
2.2	13.78	6.25 to 1	4.5 ...	.240	415.0 to 1
2.3	12.08	7.28 to 1	4.6 ...	.192	520.4 to 1
2.4	10.55	8.48 to 1	4.7 ...	.152	655.3 to 1
			4.8 ...	.121	828.3 to 1
2.5	9.18	9.90 to 1	4.9 ...	.0950	1,052. to 1
2.6	7.95	11.58 to 1			
2.7	6.86	13.58 to 1	5.0 ...	.0745	1,341. to 1
2.8	5.89	15.96 to 1	6.0 ...	.0052	19,300. to 1
2.9	5.05	18.82 to 1	7.0 ...	.00023	427,000. to 1
			8.0 ...	.0000068	14,700,000. to 1
3.0	4.30	22.24 to 1	9.0 ...	.00000013	730,000,000. to 1
3.1	3.65	26.37 to 1			
3.2	3.09	31.36 to 1	10.0 ...	.0000000015	65,000,000,000. to 1

TABLE B

SHOWING THE PROBABILITY OF OCCURRENCE OF STATISTICAL DEVIATIONS OF DIF-
FERENT MAGNITUDES RELATIVE TO THE STANDARD DEVIATION

Deviation σ	Probable oc-currence of a deviation as great as or greater than desig-nated one in 100 trials.	Odds against the occur-rence of a deviation as great as or greater than the desig-nated one.	Deviation σ	Probable occurrence of a deviation as great as or greater than designated one in 100 trials.	Odds against the occurrence of a deviation as great as or greater than the desig-nated one.
0.67449	50.00	1.00 to 1	2.7	.693	143.2 to 1
0.7	48.39	1.07 to 1	2.8	.511	194.7 to 1
0.8	42.37	1.36 to 1	2.9	.373	267.0 to 1
0.9	36.81	1.72 to 1	3.0	.270	369.4 to 1
1.0	31.73	2.15 to 1	3.1	.194	515.7 to 1
1.1	27.13	2.69 to 1	3.2	.137	726.7 to 1
1.2	23.01	3.35 to 1	3.3	.0967	1,033 to 1
1.3	19.36	4.17 to 1	3.4	.0674	1,483 to 1
1.4	16.15	5.19 to 1	3.5	.0465	2,149 to 1
1.5	13.36	6.48 to 1	3.6	.0318	3,142 to 1
1.6	10.96	8.12 to 1	3.7	.0216	4,637 to 1
1.7	8.91	10.22 to 1	3.8	.0145	6,915 to 1
1.8	7.19	12.92 to 1	3.9	.00962	10,390 to 1
1.9	5.74	16.41 to 1	4.0	.00634	15,770 to 1
2.0	4.55	20.98 to 1	5.0	.0000573	1,744,000 to 1
2.1	3.57	26.99 to 1	6.0	.00000020	500,000,000 to 1
2.2	2.78	34.96 to 1	7.0	.00000000026	400,000,000,000 to 1
2.3	2.14	45.62 to 1			
2.4	1.64	60.00 to 1			
2.5	1.24	79.52 to 1			
2.6	.932	106.3 to 1			

TABLE OF AREAS AND ORDINATES OF THE NORMAL CURVE

x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.
.00	.0000	.3989	.35	.1368	.3752
.01	.0040	.3989	.36	.1406	.3739
.02	.0080	.3989	.37	.1443	.3725
.03	.0120	.3988	.38	.1480	.3712
.04	.0160	.3986	.39	.1517	.3697
.05	.0199	.3984	.40	.1554	.3683
.06	.0239	.3982	.41	.1591	.3668
.07	.0279	.3980	.42	.1628	.3653
.08	.0319	.3977	.43	.1664	.3637
.09	.0359	.3973	.44	.1700	.3621
.10	.0398	.3970	.45	.1736	.3605
.11	.0438	.3965	.46	.1772	.3589
.12	.0478	.3961	.47	.1808	.3572
.13	.0517	.3956	.48	.1844	.3555
.14	.0557	.3951	.49	.1879	.3538
.15	.0596	.3945	.50	.1915	.3521
.16	.0636	.3939	.51	.1950	.3503
.17	.0675	.3932	.52	.1985	.3485
.18	.0714	.3925	.53	.2019	.3467
.19	.0753	.3918	.54	.2054	.3448
.20	.0793	.3910	.55	.2088	.3429
.21	.0832	.3902	.56	.2123	.3410
.22	.0871	.3894	.57	.2157	.3391
.23	.0910	.3885	.58	.2190	.3372
.24	.0948	.3876	.59	.2224	.3352
.25	.0987	.3867	.60	.2257	.3332
.26	.1026	.3857	.61	.2291	.3312
.27	.1064	.3847	.62	.2324	.3292
.28	.1103	.3836	.63	.2357	.3271
.29	.1141	.3825	.64	.2389	.3251
.30	.1179	.3814	.65	.2422	.3230
.31	.1217	.3802	.66	.2454	.3209
.32	.1255	.3790	.67	.2486	.3187
.33	.1293	.3778	.68	.2517	.3166
.34	.1331	.3765	.69	.2549	.3144

AREAS AND ORDINATES OF THE NORMAL CURVE (*Continued*)

x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.
.70	.2580	.3123	1.10	.3643	.2179
.71	.2611	.3101	1.11	.3665	.2155
.72	.2642	.3079	1.12	.3686	.2131
.73	.2673	.3056	1.13	.3708	.2107
.74	.2703	.3034	1.14	.3729	.2083
.75	.2734	.3011	1.15	.3749	.2059
.76	.2764	.2989	1.16	.3770	.2036
.77	.2794	.2966	1.17	.3790	.2012
.78	.2823	.2943	1.18	.3810	.1989
.79	.2852	.2920	1.19	.3830	.1965
.80	.2881	.2897	1.20	.3849	.1942
.81	.2910	.2874	1.21	.3869	.1919
.82	.2939	.2850	1.22	.3888	.1895
.83	.2967	.2827	1.23	.3907	.1872
.84	.2995	.2803	1.24	.3925	.1849
.85	.3023	.2780	1.25	.3944	.1826
.86	.3051	.2756	1.26	.3962	.1804
.87	.3078	.2732	1.27	.3980	.1781
.88	.3106	.2709	1.28	.3997	.1758
.89	.3133	.2685	1.29	.4015	.1736
.90	.3159	.2661	1.30	.4032	.1714
.91	.3186	.2637	1.31	.4049	.1691
.92	.3212	.2613	1.32	.4066	.1669
.93	.3238	.2589	1.33	.4082	.1647
.94	.3264	.2565	1.34	.4099	.1626
.95	.3289	.2541	1.35	.4115	.1604
.96	.3315	.2516	1.36	.4131	.1582
.97	.3340	.2492	1.37	.4147	.1561
.98	.3365	.2468	1.38	.4162	.1539
.99	.3389	.2444	1.39	.4177	.1518
1.00	.3413	.2420	1.40	.4192	.1497
1.01	.3438	.2396	1.41	.4207	.1476
1.02	.3461	.2371	1.42	.4222	.1456
1.03	.3485	.2347	1.43	.4236	.1435
1.04	.3508	.2323	1.44	.4251	.1415
1.05	.3531	.2299	1.45	.4265	.1394
1.06	.3554	.2275	1.46	.4279	.1374
1.07	.3577	.2251	1.47	.4292	.1354
1.08	.3599	.2227	1.48	.4306	.1334
1.09	.3621	.2203	1.49	.4319	.1315

AREAS AND ORDINATES OF THE NORMAL CURVE (*Continued*)

x/σ.	Area from middle of curve $(x/\sigma = 0)$ to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve $(x/\sigma = 0)$ to indicated x/σ.	Ordinate at x/σ.
1.50	.4332	.1295	1.90	.4713	.0656
1.51	.4345	.1276	1.91	.4719	.0644
1.52	.4357	.1257	1.92	.4726	.0632
1.53	.4370	.1238	1.93	.4732	.0620
1.54	.4382	.1219	1.94	.4738	.0608
1.55	.4394	.1200	1.95	.4744	.0596
1.56	.4406	.1182	1.96	.4750	.0584
1.57	.4418	.1163	1.97	.4756	.0573
1.58	.4429	.1145	1.98	.4761	.0562
1.59	.4441	.1127	1.99	.4767	.0551
1.60	.4452	.1109	2.00	.4772	.0540
1.61	.4463	.1092	2.01	.4778	.0529
1.62	.4474	.1074	2.02	.4783	.0519
1.63	.4484	.1057	2.03	.4788	.0508
1.64	.4495	.1040	2.04	.4793	.0498
1.65	.4505	.1023	2.05	.4798	.0488
1.66	.4515	.1006	2.06	.4803	.0478
1.67	.4525	.0989	2.07	.4808	.0468
1.68	.4535	.0973	2.08	.4812	.0459
1.69	.4545	.0957	2.09	.4817	.0449
1.70	.4554	.0940	2.10	.4821	.0440
1.71	.4564	.0925	2.11	.4826	.0431
1.72	.4573	.0909	2.12	.4830	.0422
1.73	.4582	.0893	2.13	.4834	.0413
1.74	.4591	.0878	2.14	.4838	.0404
1.75	.4599	.0863	2.15	.4842	.0395
1.76	.4608	.0848	2.16	.4846	.0387
1.77	.4616	.0833	2.17	.4850	.0379
1.78	.4625	.0818	2.18	.4854	.0371
1.79	.4633	.0804	2.19	.4857	.0363
1.80	.4641	.0790	2.20	.4861	.0355
1.81	.4649	.0775	2.21	.4864	.0347
1.82	.4656	.0761	2.22	.4868	.0339
1.83	.4664	.0748	2.23	.4871	.0332
1.84	.4671	.0734	2.24	.4875	.0325
1.85	.4678	.0721	2.25	.4878	.0317
1.86	.4686	.0707	2.26	.4881	.0310
1.87	.4693	.0694	2.27	.4884	.0303
1.88	.4699	.0681	2.28	.4887	.0297
1.89	.4706	.0669	2.29	.4890	.0290

AREAS AND ORDINATES OF THE NORMAL CURVE (*Continued*)

x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.
2.30	.4893	.0283	2.70	.4965	.0104
2.31	.4896	.0277	2.71	.4966	.0101
2.32	.4898	.0270	2.72	.4967	.0099
2.33	.4901	.0264	2.73	.4968	.0096
2.34	.4904	.0258	2.74	.4969	.0093
2.35	.4906	.0252	2.75	.4970	.0091
2.36	.4909	.0246	2.76	.4971	.0088
2.37	.4911	.0241	2.77	.4972	.0086
2.38	.4913	.0235	2.78	.4973	.0084
2.39	.4916	.0229	2.79	.4974	.0081
2.40	.4918	.0224	2.80	.4974	.0079
2.41	.4920	.0219	2.81	.4975	.0077
2.42	.4922	.0213	2.82	.4976	.0075
2.43	.4925	.0208	2.83	.4977	.0073
2.44	.4927	.0203	2.84	.4977	.0071
2.45	.4929	.0198	2.85	.4978	.0069
2.46	.4931	.0194	2.86	.4979	.0067
2.47	.4932	.0189	2.87	.4979	.0065
2.48	.4934	.0184	2.88	.4980	.0063
2.49	.4936	.0180	2.89	.4981	.0061
2.50	.4938	.0175	2.90	.4981	.0060
2.51	.4940	.0171	2.91	.4982	.0058
2.52	.4941	.0167	2.92	.4982	.0056
2.53	.4943	.0163	2.93	.4983	.0055
2.54	.4945	.0158	2.94	.4984	.0053
2.55	.4946	.0154	2.95	.4984	.0051
2.56	.4948	.0151	2.96	.4985	.0050
2.57	.4949	.0147	2.97	.4985	.0048
2.58	.4951	.0143	2.98	.4986	.0047
2.59	.4952	.0139	2.99	.4986	.0046
2.60	.4953	.0136	3.00	.4987	.0044
2.61	.4955	.0132	3.01	.4987	.0043
2.62	.4956	.0129	3.02	.4987	.0042
2.63	.4957	.0126	3.03	.4988	.0040
2.64	.4959	.0122	3.04	.4988	.0039
2.65	.4960	.0119	3.05	.4989	.0038
2.66	.4961	.0116	3.06	.4989	.0037
2.67	.4962	.0113	3.07	.4989	.0036
2.68	.4963	.0110	3.08	.4990	.0035
2.69	.4964	.0107	3.09	.4990	.0034

AREAS AND ORDINATES OF THE NORMAL CURVE (*Continued*)

x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.
3.10	.4990	.0033	3.50	.4998	.0009
3.11	.4991	.0032	3.51	.4998	.0008
3.12	.4991	.0031	3.52	.4998	.0008
3.13	.4991	.0030	3.53	.4998	.0008
3.14	.4992	.0029	3.54	.4998	.0008
3.15	.4992	.0028	3.55	.4998	.0007
3.16	.4992	.0027	3.56	.4998	.0007
3.17	.4992	.0026	3.57	.4998	.0007
3.18	.4993	.0025	3.58	.4998	.0007
3.19	.4993	.0025	3.59	.4998	.0006
3.20	.4993	.0024	3.60	.4998	.0006
3.21	.4993	.0023	3.61	.4998	.0006
3.22	.4994	.0022	3.62	.4999	.0006
3.23	.4994	.0022	3.63	.4999	.0005
3.24	.4994	.0021	3.64	.4999	.0005
3.25	.4994	.0020	3.65	.4999	.0005
3.26	.4994	.0020	3.66	.4999	.0005
3.27	.4995	.0019	3.67	.4999	.0005
3.28	.4995	.0018	3.68	.4999	.0005
3.29	.4995	.0018	3.69	.4999	.0004
3.30	.4995	.0017	3.70	.4999	.0004
3.31	.4995	.0017	3.71	.4999	.0004
3.32	.4995	.0016	3.72	.4999	.0004
3.33	.4996	.0016	3.73	.4999	.0004
3.34	.4996	.0015	3.74	.4999	.0004
3.35	.4996	.0015	3.75	.4999	.0004
3.36	.4996	.0014	3.76	.4999	.0003
3.37	.4996	.0014	3.77	.4999	.0003
3.38	.4996	.0013	3.78	.4999	.0003
3.39	.4997	.0013	3.79	.4999	.0003
3.40	.4997	.0012	3.80	.4999	.0003
3.41	.4997	.0012	3.81	.4999	.0003
3.42	.4997	.0012	3.82	.4999	.0003
3.43	.4997	.0011	3.83	.4999	.0003
3.44	.4997	.0011	3.84	.4999	.0003
3.45	.4997	.0010	3.85	.4999	.0002
3.46	.4997	.0010	3.86	.4999	.0002
3.47	.4997	.0010	3.87	.4999	.0002
3.48	.4997	.0009	3.88	.4999	.0002
3.49	.4998	.0009	3.89	.4999	.0002

AREAS AND ORDINATES OF THE NORMAL CURVE (*Concluded*)

x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.	x/σ.	Area from middle of curve ($x/\sigma = 0$) to indicated x/σ.	Ordinate at x/σ.
3.90.........	.5000	.0002	4.10.........	.5000	.0001
3.91.........	.5000	.0002	4.11.........	.5000	.0001
3.92.........	.5000	.0002	4.12.........	.5000	.0001
3.93.........	.5000	.0002	4.13.........	.5000	.0001
3.94.........	.5000	.0002	4.14.........	.5000	.0001
3.95.........	.5000	.0002	4.15.........	.5000	.0001
3.96.........	.5000	.0002	4.16.........	.5000	.0001
3.97.........	.5000	.0002	4.17.........	.5000	.0001
3.98.........	.5000	.0001	4.18.........	.5000	.0001
3.99.........	.5000	.0001	4.19.........	.5000	.0001
4.00.........	.5000	.0001	4.20.........	.5000	.0001
4.01.........	.5000	.0001	4.21.........	.5000	.0001
4.02.........	.5000	.0001	4.22.........	.5000	.0001
4.03.........	.5000	.0001	4.23.........	.5000	.0001
4.04.........	.5000	.0001	4.24.........	.5000	.0000
4.05.........	.5000	.0001			
4.06.........	.5000	.0001			
4.07.........	.5000	.0001			
4.08.........	.5000	.0001			
4.09.........	.5000	.0001			

APPENDIX V

SUMS OF LOGARITHMS

Table of the Sums of the Logarithms of the Natural Numbers from 1 to 100

x.	$S (\log x)$.	$S (x \log x)$.	$S (\log x)^2$.
1.	0.0000000	0.0000000	0.0000000
2.	0.3010300	0.6020600	0.0906191
3.	0.7781513	2.0334238	0.3182638
4.	1.3802112	4.4416637	0.6807400
5.	2.0791812	7.9365137	1.1692991
6.	2.8573325	12.6054212	1.7748184
7.	3.7024305	18.5211075	2.4890091
8.	4.6055205	25.7458274	3.3045806
9.	5.5597630	34.3340100	4.2151594
10.	6.5597630	44.3340100	5.2151594
11.	7.6011557	55.7893295	6.2996581
12.	8.6803370	68.7395045	7.4642903
13.	9.7942803	83.2207681	8.7051601
14.	10.9404084	99.2665606	10.0187696
15.	12.1164996	116.9079295	11.4019602
16.	13.3206196	136.1738492	12.8518651
17.	14.5510685	157.0914808	14.3658697
18.	15.8063410	179.6863859	15.9415788
19.	17.0850946	203.9827044	17.5767895
20.	18.3861246	230.0033043	19.2694686
21.	19.7083439	257.7699095	21.0177324
22.	21.0507666	287.3032084	22.8198311
23.	22.4124944	318.6229487	24.6741338
24.	23.7927057	351.7480185	26.5791169
25.	25.1906457	386.6965187	28.5333531
26.	26.6056190	423.4858257	30.5355027
27.	28.0369828	462.1326474	32.5843049
28.	29.4841408	502.6530722	34.6785713
29.	30.9465388	545.0626142	36.8171792
30.	32.4236601	589.3762518	38.9990664
31.	33.9150218	635.6084643	41.2232261
32.	35.4201717	683.7732636	43.4887026
33.	36.9386857	733.8842237	45.7945871
34.	38.4701646	785.9545068	48.1400148
35.	40.0142326	839.9968884	50.5241609

SUMS OF LOGARITHMS (*Continued*)

x.	S (log x).	S (x log x).	S (log x)².
36........................	41.5705351	896.0237784	52.9462384
37........................	43.1387369	954.0472422	55.4054951
38........................	44.7185205	1,014.0790189	57.9012113
39........................	46.3095851	1,076.1305385	60.4326979
40........................	47.9116451	1,140.2129382	62.9992941
41........................	49.5244289	1,206.3370763	65.6003659
42........................	51.1476782	1,274.5135465	68.2353041
43........................	52.7811467	1,344.7526901	70.9035233
44........................	54.4245993	1,417.0646079	73.6044600
45........................	56.0778119	1,491.4591710	76.3375716
46........................	57.7405697	1,567.9460313	79.1023352
47........................	59.4126676	1,646.5346306	81.8982465
48........................	61.0939088	1,727.2342100	84.7248186
49........................	62.7841049	1,810.0538179	87.5815814
50........................	64.4830749	1,895.0023181	90.4680804
51........................	66.1906450	1,982.0883971	93.3838763
52........................	67.9066484	2,071.3205710	96.3285438
53........................	69.6309243	2,162.7071920	99.3016711
54........................	71.3633180	2,256.2564551	102.3028592
55........................	73.1036807	2,351.9764030	105.3317215
56........................	74.8518687	2,449.8749325	108.3878829
57........................	76.6077436	2,549.9597993	111.4709794
58........................	78.3711716	2,652.2386229	114.5806577
59........................	80.1420236	2,756.7188916	117.7165745
60........................	81.9201748	2,863.4079666	120.8783964
61........................	83.7055047	2,972.3130866	124.0657990
62........................	85.4978964	3,083.4413713	127.2784670
63........................	87.2972369	3,196.7998259	130.5160934
64........................	89.1034169	3,312.3953443	133.7783793
65........................	90.9163303	3,430.2347124	137.0650341
66........................	92.7358742	3,550.3246122	140.3757742
67........................	94.5619490	3,672.6716240	143.7103234
68........................	96.3944579	3,797.2822300	147.0684123
69........................	98.2333070	3,924.1628173	150.4497786
70........................	100.0784050	4,053.3196801	153.8541654
71........................	101.9296634	4,184.7590228	157.2813229
72........................	103.7869959	4,318.4869626	160.7310069
73........................	105.6503187	4,454.5095314	164.2029790
74........................	107.5195505	4,592.8326786	167.6970062
75........................	109.3946117	4,733.4622734	171.2128609
76........................	111.2754253	4,876.4041064	174.7503207
77........................	113.1619160	5,021.6638922	178.3091670
78........................	115.0540106	5,169.2472713	181.8891890
79........................	116.9516377	5,319.1598115	185.4901776
80........................	118.8547277	5,471.4070104	189.1119291

SUMS OF LOGARITHMS (*Concluded*)

x.	S (log x).	S (x log x).	S (log x)².
81	120.7632127	5,625.9942970	192.7542442
82	122.6770266	5,782.9270329	196.4169276
83	124.5961047	5,942.2105145	200.0997884
84	126.5203840	6,103.8499746	203.8026391
85	128.4498029	6,267.8505832	207.5252965
86	130.3843013	6,434.2174510	211.2675808
87	132.3238206	6,602.9556260	215.0293157
88	134.2683033	6,774.0701012	218.8103286
89	136.2176933	6,947.5658118	222.6104500
90	138.1719358	7,123.4476376	226.4295137
91	140.1309772	7,301.7204043	230.2673568
92	142.0947650	7,482.3888844	234.1238194
93	144.0632480	7,665.4577986	237.9987445
94	146.0363758	7,850.9318169	241.8919781
95	148.0140994	8,038.8155594	245.8033687
96	149.9963707	8,229.1135977	249.7327590
97	151.9831424	8,421.8304560	253.6800209
98	153.9743685	8,616.9706114	257.6450022
99	155.9700037	8,814.5384957	261.6275620
100	157.9700037	9,014.5384957	265.6275620

INDEX