Introduction to
Medical Biometry and Statistics

By

Raymond Pearl

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

The Johns Hopkins University

Illustrated

Philadelphia and London
W. B. Saunders Company
1923

MADE IN U. S. A.

PRESS OF
W. B. SAUNDERS COMPANY
PHILADELPHIA

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

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 approach 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 (indeed, I hope myself to follow this with a more advanced treatise later) 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 43.

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

CHAPTER XVI

APPENDIX I

APPENDIX II

APPENDIX III

APPENDIX IV

APPENDIX V

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 very 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 differences in things for purposes of statistical reasoning may be a function of discrete separation in either time or space, or both. 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 tool 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.

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

"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[6] 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:

"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."

General experience with other branches of science would make it seem reasonable that the following propositions are true, and should be emphasized in the teaching of medical students, and in the practice and writing of medical men generally:

1. That there is no inherent reason why medicine in every one of its phases should not ultimately become in respect of its methods an *exact* science, in the same sense that physics, chemistry, or astronomy are today exact sciences.

2. That this goal will be reached in exact ratio to the extent to which quantitative methods of thought and action are made an integral part of the training in every sort of medicine.

3. That no number or figure can be said to have any final scientific validity or meaning until we know its probable error, the "probable error" being the measure of the extent to which the number will vary in its value as the result of chance alone.

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, Citta di Castello (Soc. Tipograf ica, "Leonardo da Vinci"), 1922.

4. Czuber, E.: Die statistischen Forschungsmethoden. Wien (L. W. Seidel & Sohn), 1921.

(References 1 to 4 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 most of the books which have been especially prepared for him.)

5. 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.)

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

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

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

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. It has not been altered from its original form.[7]

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 presenting the matter to my classes I have endeavored to justify in more detail than is possible in the table itself the particular items which appear. In any event, it seems clear that any historic review of vital statistics would be

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.
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. lxxvii and lxxviii.
1718	Publication of the "Doctrine of Chances."	A. DeMoivre (1667–1754).	Art. DeMoivre, Encyc. Brit.
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 "Theorie analytique des probabilités."	Pierre Simon Laplace (1749–1827).	Encyc. Brit. Art., Laplace.
1812	Inauguaration 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.*
1825	Publication of "Mémoire sur les lois des naissances et de la mortalité a Bruxelles," Quetelet's first statistical paper.	Lambert Adolph Jacques Quetelet (1796–1874).	Lottin, Quetelet, p. xx.

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

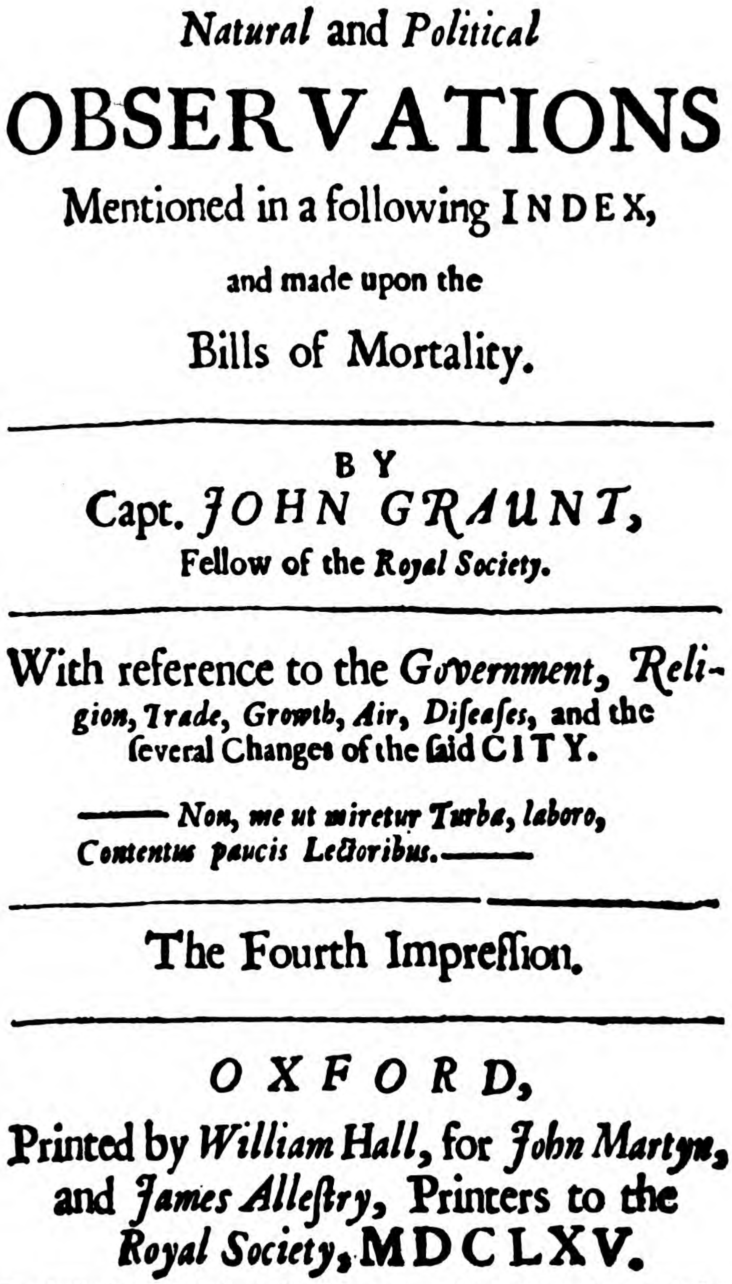
Year.	Event.	Personality concerned.	Authority for record.
1826	Establishment of statistical commission in Belgium.	Ed. Smits.	Julin, A., Hist. Stat., p. 126.
1829	First official Census of Belgium.	Ed. Smits.	Julin, A. Hist. Stat., p. 128.
1832	Publication of "Recherches sur la reproduction et sur la mortalité de l'homme aux different ages et sur la population 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 dévéloppement 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.
✓ 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 mathematique sur les probabilites des erreurs de situation d'un point." Acad. des Sci. Mem. par div. sav. IIe. Ser. t. ix (Correlation).	A. Bravais.	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 Sociéte dé 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."

bound to contain at 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. I am hopeful that by so choosing them I may perchance 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 I refer him 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 "Theorie 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 INDEX,

and made upon the

Bills of Mortality.

BY

Capt. *JOHN GRAUNT*,

Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diſeaſes*, and the
ſeveral Changes of the ſaid CITY.

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

The Fourth Impreſſion.

OXFORD,

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.

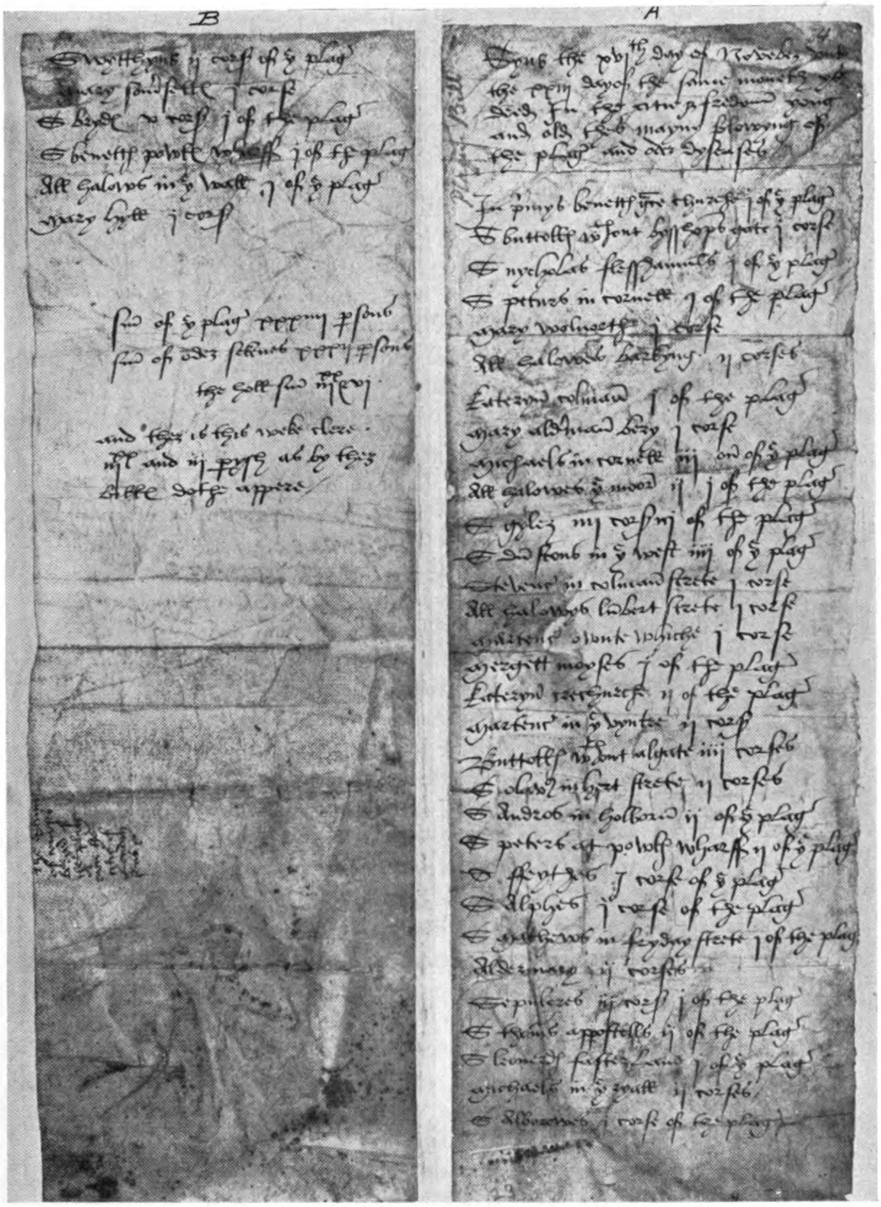
the sexes in the population. Third, he demonstrated the relatively *high rate of mortality in the earliest years of life*, and finally he discovered that the *urban is higher than the rural death-rate* normally.

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 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.

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.

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. 3.



1532

Sens the xvj th day of Novembr unto the xxiij day of the same moneth ys dede in the citie & fredom yong and old thes many folowing of the plage and odr dyseasis

In primis benett [illegible] j of ye plag
S buttolff wt out bysshopsgate j corse
S nycholas flesshamuls j of ye plag
S petirs in cornell j of the plag
mary wolnethe j corse
All halows barkyng ij corses
Kateryn colman j of the plag
mary aldrmanbery j corse
myghaels in cornell iij on of ye plag
All halowes ye more ij j of the plag
S [illegible] iiij corsis iij of the plag
S dunstons in ye west iiij of ye plag
S lawrens in colmand strete j corse
All halows lubert strete ij corse
martyns owtwhiche j corse
mergarett moysis j of the plag
Kateryn crechurch ij of the plag
martens in ye vyntre ij corse
Buttolff wt out alsgate iiij corsis
S olyff in hert strete ij corsis
S andros in holborn ij of ye plag
S peters at powls wharff ij of ye plag
S ffaytis j corse of ye plag
S alphes j corse of the plag
S gabriells in fendiu strete j of the plag
Aldermary ij corsis
S pulcres iij corse j of the plag
S thomas appostells ij of the plag
S leonerds fasterlane j of ye plag
myghaels in ye ryall ij corsis
S albones j corse of the plag

S [illegible] ij corse of ye plag
mary [illegible] j corse
S bryde x corse j of the plag
S benettf powls wharff j of the plag
All halows in ye wall j of ye plag
mary hyll j corse

Sm of ye plag xxxiiij persons
Sm of odr sekenes xxxj persons
the holl sm lxv

and ther is this weke clere [illegible] and iij parych as by ther bille doth appere

Fig. 3.—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.)

Creighton[8] 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). This earliest of official reports of vital statistics to be preserved is transcribed by Creighton (retaining the original spelling) as follows:

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 i corse
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 $\overset{\text{xx}}{\text{iii}}$ & vi.
And there is this weke clere $\overset{\text{xx}}{\text{iii}}$ and iii paryshes as by this bille doth appere.

The exec^{u}
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 forerunner of the great work of the famous Belgian vital statistician, Lambert Adolph Jacques Quetelet (1796–1874), entitled *Sur l'homme et le dêveloppment 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 mathematician, he brought to bear upon the data of human vital phenomena 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[11] says: "But if ultimately Graunt has 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

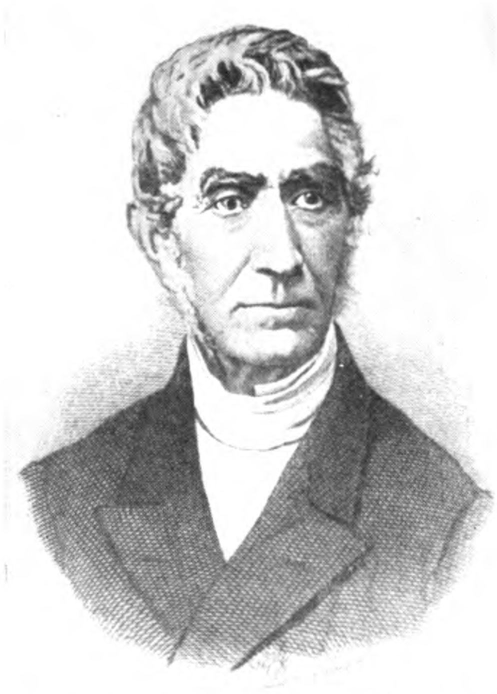


Fig. 4.—Portrait of Lambert Adolph Jacques Quetelet (1796–1874).

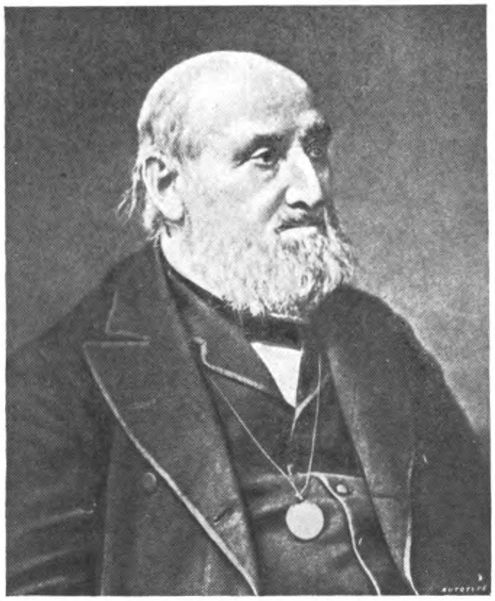


Fig. 5.—Portrait of Dr. William Farr (1807–1883).

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 Brownlee. But it is an idle dream; governments in England, so the newspapers tell us, often spend money in odd ways, but at least 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 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

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

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."

* Simon: English Sanitary Institution, p. 212.

THE HISTORY OF BIOMETRY

In discussing the development of biometry the writer will follow closely an account which he gave of the same matter some time ago.[9] 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*[10] 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 inheritance. 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 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 temperamental 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 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 superstructure 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

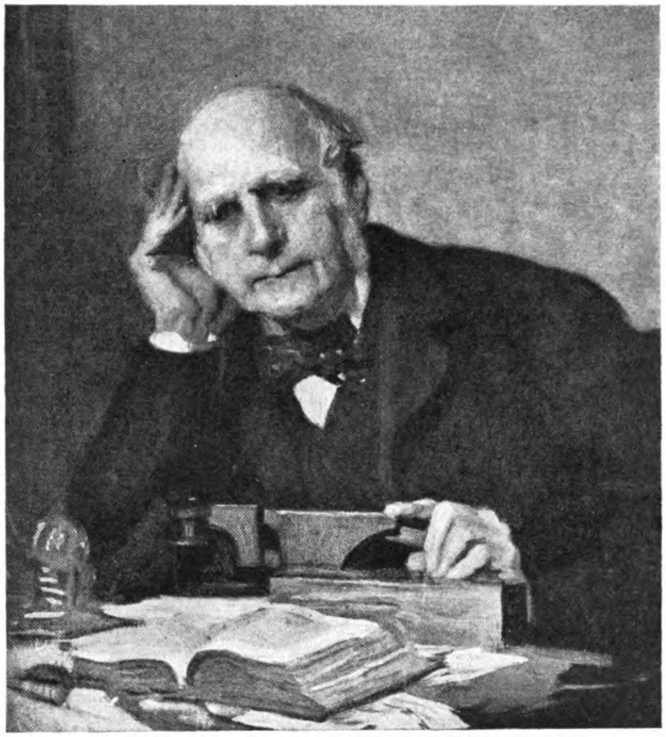


Fig. 6.—Portrait of Francis Galton (1822–1907). (For permission to publish this portrait here I am indebted to Dr. G. H. Shull, Editor of Genetics.)

fall within the realm of the calculus of probability. The foundations 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. Until he began his work, nearly all statisticians, astronomers, and physicists who had anything to do with the theory of probability, either from the standpoint of statistics or that of the theory of errors of observation, had been content to use the so-called "normal" curve of errors to describe the distribution of chance-determined events. One of the characteristics of this curve is that it is symmetric. According to it events above the mean are as likely to happen as events below the mean. Observed statistics of natural phenomena were found, as a matter of fact, to give in many cases asymmetric distributions. Indeed, some of the very examples used in the text-books to illustrate the normal curve do not accord with it when tested by an accurate measure of goodness of fit (for which

Fig. 7.—Portrait of Pierre Simon Laplace (1749–1827).

extremely valuable instrument of statistical research we are again indebted to Pearson). Starting from the sound position that the facts of nature are of more importance than any theory, even though it be one beautiful enough to excite worship, Pearson in three classic memoirs, in the series of *Mathematical Contributions to the Theory of Evolution*, developed a theory of skew frequency curves, and skew correlation, which took 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,

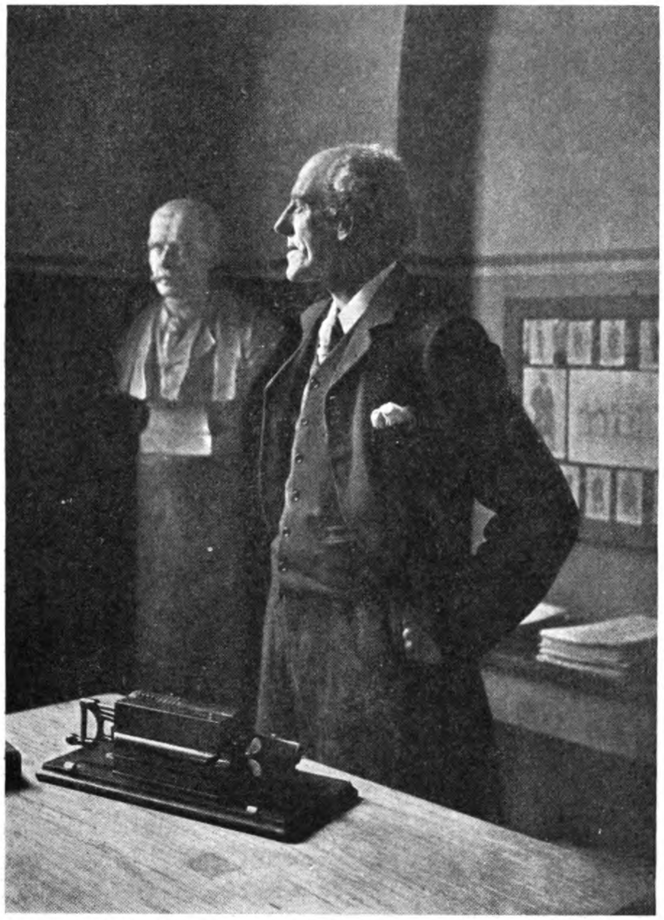


Fig. 8.—Portrait of Karl Pearson, F. R. S.

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.

In the application of biometric methods to specifically medical problems, English workers, notably Dr. Major Greenwood of the Ministry of Health, from whose work we have already quoted, and 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.

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. Rossiter, W. S.: A Century of Population Growth from the First Census of the United States to the Twelfth, 1790–1900, Washington, Govt. Printing Office, 1909.
5. The History of Statistics, Their Development and Progress in Many Countries, collected and edited by John Koren, New York (Macmillan), 1918.
6. Yule, G. U.: Introduction to the Theory of Statistics, sixth edition, London (Griffith & Company), 1922.
7. Pearl, R.: Some Landmarks in the History of Vital Statistics, Quart. Publ. Amer. Stat. Assn., June, 1920, pp. 221–223.
8. 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.)
9. Pearl, R.: The Service and Importance of Statistics to Biology, Quart. Publ. Amer. Stat. Assn., March, 1914, pp. 40–48.
10. Galton, F.: Memories of My Life, New York. (E. P. Dutton & Co.), 1909. (Every student of statistics should read this book.)
11. Greenwood, M.: Medical Statistics, Lancet, May 7, 1921.
12. 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 (Edw. Stanford), 1885. (Now a rare book, but one which every student should read.)
13. Ogle, W.: An Inquiry into the Trustworthiness of the Old Bills of Mortality, Jour. Roy. Stat. Soc., vol. 55, pp. 437–460, 1892.

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.

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 age, sex, occupation, race, etc.

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.

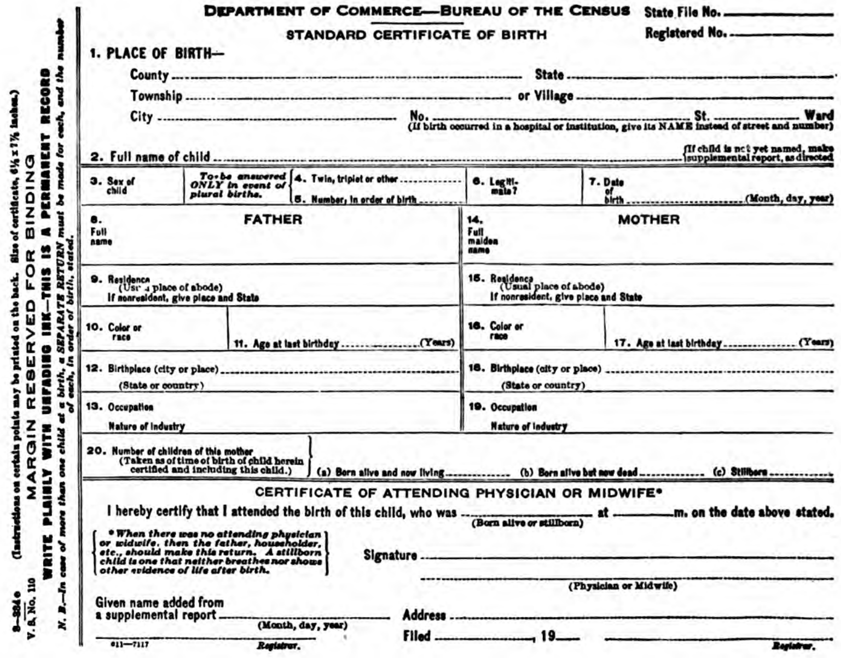
The heuristic advantages of the registration over the census method are apparent. The *course* of events can be followed. Registration gives us such knowledge as we have of births, deaths, sickness, marriages, divorces, etc., so far as concerns population aggregates.

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



8—3846
V. S. No. 110

(Instructions on certain points may be printed on the back. Size of certificate, 6½ x 7⅞ inches.)

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.

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS State File No. ____________

STANDARD CERTIFICATE OF BIRTH Registered No. ____________

1. PLACE OF BIRTH—

County ____________ State ____________

Township ____________ or Village ____________

City ____________ No. ____________ St. ____________ Ward

(If birth occurred in a hospital or institution, give its NAME instead of street and number)

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

3. Sex of child | *To be answered ONLY in event of plural births.* { 4. Twin, triplet or other ____________ 5. Number, in order of birth ____________ | 6. Legitimate? | 7. Date of birth ____________ (Month, day, year)

FATHER	MOTHER
8. Full name	14. Full maiden name
9. Residence (Usual place of abode) If nonresident, give place and State	15. Residence (Usual place of abode) If nonresident, give place and State
10. Color or race · 11. Age at last birthday ____________ (Years)	16. Color or race · 17. Age at last birthday ____________ (Years)
12. Birthplace (city or place) ____________ (State or country)	18. Birthplace (city or place) ____________ (State or country)
13. Occupation Nature of industry	19. Occupation Nature of industry

20. Number of children of this mother (Taken as of time of birth of child herein certified and including this child.) } (a) Born alive and now living ____________ (b) Born alive but now dead ____________ (c) Stillborn ____________

CERTIFICATE OF ATTENDING PHYSICIAN OR MIDWIFE*

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

{ **When there was no attending physician or midwife, then the father, householder, etc., should make this return. A stillborn child is one that neither breathes nor shows other evidence of life after birth.* }

Signature ____________

____________ (Physician or Midwife)

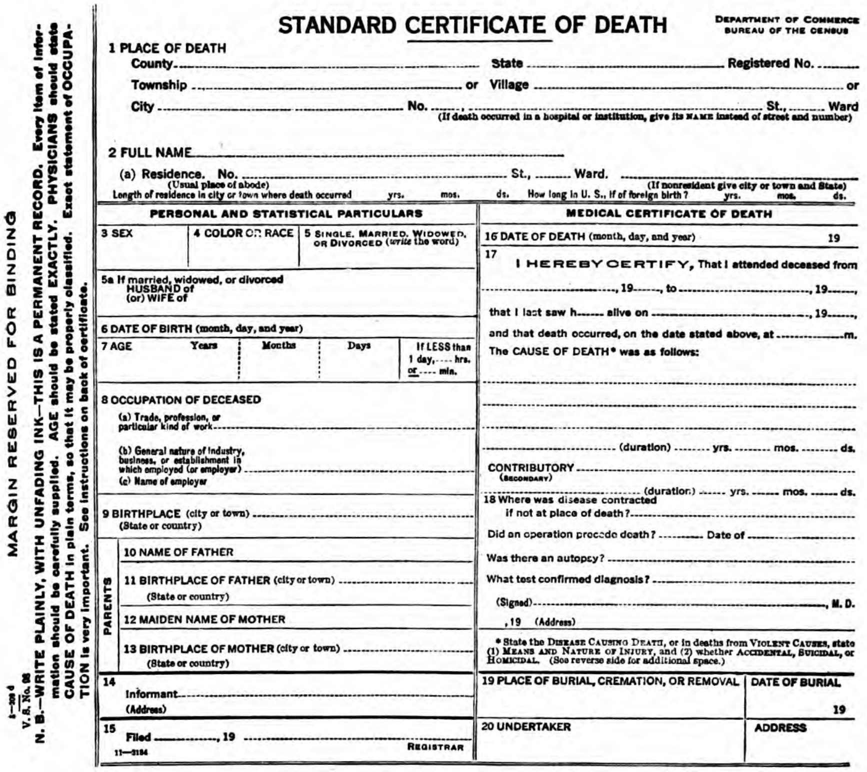
Given name added from a supplemental report ____________ (Month, day, year)

____________ *Registrar.*

Address ____________

Filed ____________, 19____ ____________ *Registrar.*

611—7117



8—208 d
V. S. No. 98

MARGIN RESERVED FOR BINDING

N. B.—WRITE PLAINLY, WITH UNFADING INK—THIS IS A PERMANENT RECORD. Every item of information should be carefully supplied. AGE should be stated EXACTLY. PHYSICIANS should state CAUSE OF DEATH in plain terms, so that it may be properly classified. Exact statement of OCCUPATION is very important. See instructions on back of certificate.

STANDARD CERTIFICATE OF DEATH

DEPARTMENT OF COMMERCE
BUREAU OF THE CENSUS

1 PLACE OF DEATH

County ____________ State ____________ Registered No. ____________

Township ____________ or Village ____________ or

City ____________ No. ____________, ____________ St., ____________ Ward

(If death occurred in a hospital or institution, give its NAME instead of street and number)

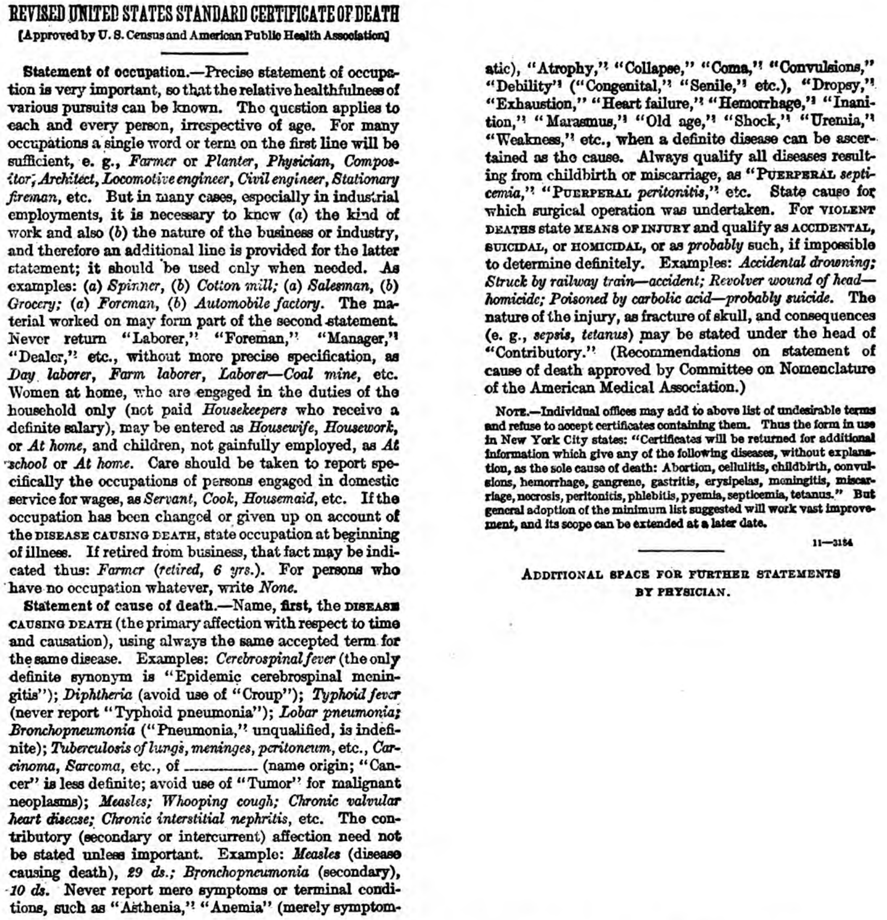
2 FULL NAME ____________

(a) Residence. No. ____________ St., ____________ Ward. ____________
(Usual place of abode) (If nonresident give city or town and State)

Length of residence in city or town where death occurred yrs. mos. ds. How long in U. S., if of foreign birth? yrs. mos. ds.

PERSONAL AND STATISTICAL PARTICULARS	MEDICAL CERTIFICATE OF DEATH
3 SEX · 4 COLOR OR RACE · 5 SINGLE, MARRIED, WIDOWED, OR DIVORCED (*write* the word)	16 DATE OF DEATH (month, day, and year) 19
5a If married, widowed, or divorced HUSBAND of (or) WIFE of	17 I HEREBY CERTIFY, That I attended deceased from ____________, 19____, to ____________, 19____, that I last saw h____ alive on ____________, 19____, and that death occurred, on the date stated above, at ____________m.
6 DATE OF BIRTH (month, day, and year)	The CAUSE OF DEATH* was as follows:
7 AGE · Years · Months · Days · If LESS than 1 day, ____ hrs. or ____ min.	____________ ____________ ____________
8 OCCUPATION OF DECEASED (a) Trade, profession, or particular kind of work ____________ (b) General nature of industry, business, or establishment in which employed (or employer) ____________ (c) Name of employer	____________ (duration) ____ yrs. ____ mos. ____ ds. CONTRIBUTORY ____________ (SECONDARY) ____________ (duration) ____ yrs. ____ mos. ____ ds.
9 BIRTHPLACE (city or town) ____________ (State or country)	18 Where was disease contracted if not at place of death? ____________ Did an operation precede death? ____ Date of ____________
PARENTS · 10 NAME OF FATHER	Was there an autopsy? ____________
PARENTS · 11 BIRTHPLACE OF FATHER (city or town) ____________ (State or country)	What test confirmed diagnosis? ____________ (Signed) ____________, M. D.
PARENTS · 12 MAIDEN NAME OF MOTHER	, 19 (Address)
PARENTS · 13 BIRTHPLACE OF MOTHER (city or town) ____________ (State or country)	* State the DISEASE CAUSING DEATH, or in deaths from VIOLENT CAUSES, state (1) MEANS AND NATURE OF INJURY, and (2) whether ACCIDENTAL, SUICIDAL, or HOMICIDAL. (See reverse side for additional space.)
14 Informant ____________ (Address)	19 PLACE OF BURIAL, CREMATION, OR REMOVAL · DATE OF BURIAL 19
15 Filed ____________, 19 ____________ REGISTRAR	20 UNDERTAKER · ADDRESS

11—3184



REVISED UNITED STATES STANDARD CERTIFICATE OF DEATH

[Approved by U. S. Census and American Public Health Association]

Statement of occupation.—Precise statement of occupation is very important, so that the relative healthfulness of various pursuits can be known. The question applies to each and every person, irrespective of age. For many occupations a single word or term on the first line will be sufficient, e. g., *Farmer* or *Planter*, *Physician*, *Compositor*, *Architect*, *Locomotive engineer*, *Civil engineer*, *Stationary fireman*, etc. But in many cases, especially in industrial employments, it is necessary to know (*a*) the kind of work and also (*b*) the nature of the business or industry, and therefore an additional line is provided for the latter statement; it should be used only when needed. As examples: (*a*) *Spinner*, (*b*) *Cotton mill;* (*a*) *Salesman*, (*b*) *Grocery;* (*a*) *Foreman*, (*b*) *Automobile factory*. The material worked on may form part of the second statement. Never return "Laborer," "Foreman," "Manager," "Dealer," etc., without more precise specification, as *Day laborer*, *Farm laborer*, *Laborer—Coal mine*, etc. Women at home, who are engaged in the duties of the household only (not paid *Housekeepers* who receive a definite salary), may be entered as *Housewife*, *Housework*, or *At home*, and children, not gainfully employed, as *At school* or *At home*. Care should be taken to report specifically the occupations of persons engaged in domestic service for wages, as *Servant*, *Cook*, *Housemaid*, etc. If the occupation has been changed or given up on account of the DISEASE CAUSING DEATH, state occupation at beginning of illness. If retired from business, that fact may be indicated thus: *Farmer* (*retired, 6 yrs.*). For persons who have no occupation whatever, write *None*.

Statement of cause of death.—Name, **first**, the DISEASE CAUSING DEATH (the primary affection with respect to time and causation), using always the same accepted term for the same disease. Examples: *Cerebrospinal fever* (the only definite synonym is "Epidemic cerebrospinal meningitis"); *Diphtheria* (avoid use of "Croup"); *Typhoid fever* (never report "Typhoid pneumonia"); *Lobar pneumonia; Bronchopneumonia* ("Pneumonia," unqualified, is indefinite); *Tuberculosis of lungs, meninges, peritoneum*, etc., *Carcinoma*, *Sarcoma*, etc., of ________ (name origin; "Cancer" is less definite; avoid use of "Tumor" for malignant neoplasms); *Measles; Whooping cough; Chronic valvular heart disease; Chronic interstitial nephritis*, etc. The contributory (secondary or intercurrent) affection need not be stated unless important. Example: *Measles* (disease causing death), *29 ds.; Bronchopneumonia* (secondary), *10 ds.* Never report mere symptoms or terminal conditions, such as "Asthenia," "Anemia" (merely symptomatic), "Atrophy," "Collapse," "Coma," "Convulsions," "Debility" ("Congenital," "Senile," etc.), "Dropsy," "Exhaustion," "Heart failure," "Hemorrhage," "Inanition," "Marasmus," "Old age," "Shock," "Uremia," "Weakness," etc., when a definite disease can be ascertained as the cause. Always qualify all diseases resulting from childbirth or miscarriage, as "PUERPERAL *septicemia*," "PUERPERAL *peritonitis*," etc. State cause for which surgical operation was undertaken. For VIOLENT DEATHS state MEANS OF INJURY and qualify as ACCIDENTAL, SUICIDAL, or HOMICIDAL, or as *probably* such, if impossible to determine definitely. Examples: *Accidental drowning; Struck by railway train—accident; Revolver wound of head—homicide; Poisoned by carbolic acid—probably suicide.* The nature of the injury, as fracture of skull, and consequences (e. g., *sepsis*, *tetanus*) may be stated under the head of "Contributory." (Recommendations on statement of cause of death approved by Committee on Nomenclature of the American Medical Association.)

NOTE.—Individual offices may add to above list of undesirable terms and refuse to accept certificates containing them. Thus the form in use in New York City states: "Certificates will be returned for additional information which give any of the following diseases, without explanation, as the sole cause of death: Abortion, cellulitis, childbirth, convulsions, hemorrhage, gangrene, gastritis, erysipelas, meningitis, miscarriage, necrosis, peritonitis, phlebitis, pyemia, septicemia, tetanus." But general adoption of the minimum list suggested will work vast improvement, and its scope can be extended at a later date.

11—3184

ADDITIONAL SPACE FOR FURTHER STATEMENTS BY PHYSICIAN.

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 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.

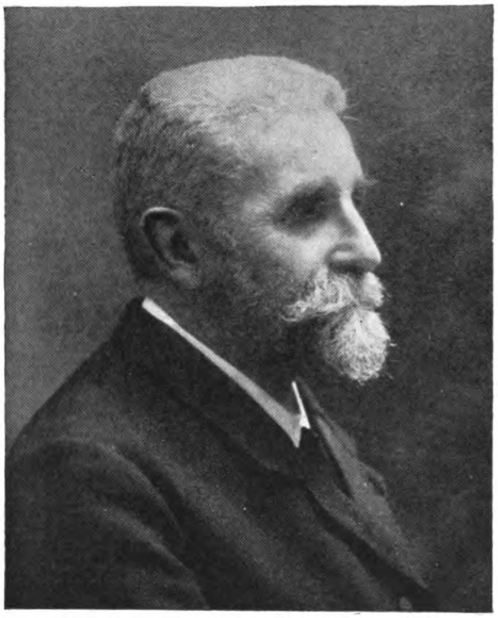


Fig. 9.—Portrait of Dr. Jacques Bertillon. (Reproduced through the kindness of Dr. Frederick L. Hoffman, to whom the original belongs, and Brig.-Gen. Robert E. Noble, Librarian of the Surgeon-General's office.)

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, and the second in 1909, and the most recent one in 1920.

The present form of the International List, after its latest revision, is as follows (kindly provided by Doctor William H. Davis, Chief for Vital Statistics, United States Census Bureau):

INTERNATIONAL LIST OF CAUSES OF DEATH

(Third Decennial Revision by the International Commission, Paris, October 11–15, 1920.)

(The lines preceded by a star indicate certain additional subdivisions which the Census Bureau intends to use to facilitate comparisons with statistics of previous years.)

I. *Epidemic, Endemic, and Infectious Diseases*

1. Typhoid and paratyphoid fever:
    (*a*) Typhoid fever.
    (*b*) Paratyphoid fever.
2. Typhus fever.
3. Relapsing fever (Spirillum obermeieri).
4. Malta fever.
5. Malaria.
6. Smallpox.
7. Measles.
8. Scarlet fever.
9. Whooping-cough.
10. Diphtheria.
11. Influenza:
    (*a*) With pulmonary complications specified.
    (*b*) Without pulmonary complications specified.
12. Miliary fever.
13. Mumps.
14. Asiatic cholera.
15. Cholera nostras.
16. Dysentery:
    (*a*) Amebic.
    (*b*) Bacillary.
    (*c*) Unspecified or due to other causes.
17. Plague:
    (*a*) Bubonic.
    (*b*) Pneumonic.
    (*c*) Septicemic.
    (*d*) Unspecified.
18. Yellow fever.
19. Spirochetal hemorrhagic jaundice.

20. Leprosy.
21. Erysipelas.
22. Acute anterior poliomyelitis.
23. Lethargic encephalitis.
24. Meningococcus meningitis.
25. Other epidemic and endemic diseases:
    *(*a*) Chickenpox.
    *(*b*) German measles.
    *(*c*) Others under this title.
26. Glanders.
27. Anthrax.
28. Rabies.
29. Tetanus.
30. Mycoses.
31. Tuberculosis of the respiratory system.
32. Tuberculosis of the meninges and central nervous system.
33. Tuberculosis of the intestines and peritoneum.
34. Tuberculosis of the vertebral column.
35. Tuberculosis of the joints.
36. Tuberculosis of other organs:
    (*a*) Tuberculosis of the skin and subcutaneous cellular tissue.
    (*b*) Tuberculosis of the bones (vertebral column excepted).
    (*c*) Tuberculosis of the lymphatic system (mesenteric and retroperitoneal glands excepted).
    (*d*) Tuberculosis of the genito-urinary system.
    (*e*) Tuberculosis of organs other than the above.
37. Disseminated tuberculosis:
    (*a*) Acute.
    (*b*) Chronic or unspecified.
38. Syphilis.
39. Soft chancre.
40. Gonococcus infection.
41. Purulent infection, septicemia.
42. Other infectious diseases.

II. *General Diseases Not Included in Class I*

43. Cancer and other malignant tumors of the buccal cavity.
44. Cancer and other malignant tumors of the stomach, liver.
45. Cancer and other malignant tumors of the peritoneum, intestines, rectum.
46. Cancer and other malignant tumors of the female genital organs.
47. Cancer and other malignant tumors of the breast.
48. Cancer and other malignant tumors of the skin.
49. Cancer and other malignant tumors of other or unspecified organs.
50. Benign tumors and tumors not returned as malignant (tumors of the female genital organs excepted).
51. Acute rheumatic fever.
52. Chronic rheumatism, osteo-arthritis, gout.
53. Scurvy.

54. Pellagra.
55. Beriberi.
56. Rickets.
57. Diabetes mellitus.
58. Anemia, chlorosis:
    (*a*) Pernicious anemia.
    (*b*) Other anemias and chlorosis.
59. Diseases of the pituitary gland.
60. Diseases of the thyroid gland:
    (*a*) Exophthalmic goiter.
    (*b*) Other diseases of the thyroid gland.
61. Diseases of the parathyroid glands.
62. Diseases of the thymus gland.
63. Diseases of the adrenals (Addison's disease).
64. Diseases of the spleen.
65. Leukemia and Hodgkin's disease:
    (*a*) Leukemia.
    (*b*) Hodgkin's disease.
66. Alcoholism (acute or chronic).
67. Chronic poisoning by mineral substances:
    *(*a*) Chronic lead-poisoning.
    *(*b*) Others under this title.
68. Chronic poisoning by organic substances.
69. Other general diseases.

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

70. Encephalitis.
71. Meningitis:
    *(*a*) Simple meningitis.
    *(*b*) Non-epidemic cerebrospinal meningitis.
72. Tabes dorsalis (locomotor ataxia).
73. Other diseases of the spinal cord.
74. Cerebral hemorrhage, apoplexy:
    (*a*) Cerebral hemorrhage.
    (*b*) Cerebral embolism and thrombosis.
75. Paralysis without specified cause:
    (*a*) Hemiplegia.
    (*b*) Others under this title.
76. General paralysis of the insane.
77. Other forms of mental alienation.
78. Epilepsy.
79. Convulsions (non-puerperal; five years and over)
80. Infantile convulsions (under five years of age).
81. Chorea.
82. Neuralgia and neuritis.
83. Softening of the brain.
84. Other diseases of the nervous system.
85. Diseases of the eye and annexa.

86. Diseases of the ear and of the mastoid process:
    *(*a*) Diseases of the ear.
    *(*b*) Diseases of the mastoid process.

IV. *Diseases of the Circulatory System*

87. Pericarditis.
88. Endocarditis and myocarditis (acute).
89. Angina pectoris.
90. Other diseases of the heart.
91. Diseases of the arteries:
    (*a*) Aneurysm.
    (*b*) Arteriosclerosis.
    (*c*) Other diseases of the arteries.
92. Embolism and thrombosis (not cerebral).
93. Diseases of the veins (varices, hemorrhoids, phlebitis, etc.).
94. Diseases of the lymphatic system (lymphangitis, etc.).
95. Hemorrhage without specified cause.
96. Other diseases of the circulatory system.

V. *Diseases of the Respiratory System*

97. Diseases of the nasal fossæ and their annexa:
    *(*a*) Diseases of the nasal fossæ.
    *(*b*) Others under this title.
98. Diseases of the larynx.
99. Bronchitis:
    (*a*) Acute.
    (*b*) Chronic.
    (*c*) Unspecified under five years of age.
    (*d*) Unspecified five years and over.
100. Bronchopneumonia:
    *(*a*) Bronchopneumonia.
    *(*b*) Capillary bronchitis.
101. Pneumonia:
    (*a*) Lobar.
    (*b*) Unspecified.
102. Pleurisy.
103. Congestion and hemorrhagic infarct of the lung.
104. Gangrene of the lung.
105. Asthma.
106. Pulmonary emphysema.
107. Other diseases of the respiratory system (tuberculosis excepted):
    (*a*) Chronic interstitial pneumonia, including occupational diseases of the respiratory system.
    (*b*) Diseases of the mediastinum.
    (*c*) Others under this title.

VI. *Diseases of the Digestive System*

108. Diseases of the mouth and annexa.
109. Diseases of the pharynx and tonsils (including adenoid vegetations):
  *(*a*) Adenoid vegetations.
  *(*b*) Others under this title.
110. Diseases of the esophagus.
111. Ulcer of the stomach and duodenum:
  (*a*) Ulcer of the stomach.
  (*b*) Ulcer of the duodenum.
112. Other diseases of the stomach (cancer excepted).
113. Diarrhea and enteritis (under two years of age).
114. Diarrhea and enteritis (two years and over).
115. Ankylostomiasis.
116. Diseases due to other intestinal parasites:
  (*a*) Cestodes (hydatids of the liver excepted).
  (*b*) Trematodes.
  (*c*) Nematodes (other than ankylostoma).
  (*d*) Coccidia.
  (*e*) Other parasites specified.
  (*f*) Parasites not specified.
117. Appendicitis and typhlitis.
118. Hernia, intestinal obstruction:
  (*a*) Hernia.
  (*b*) Intestinal obstruction.
119. Other diseases of the intestines.
120. Acute yellow atrophy of the liver.
121. Hydatid tumor of the liver.
122. Cirrhosis of the liver:
  (*a*) Specified as alcoholic.
  (*b*) Not specified as alcoholic.
123. Biliary calculi.
124. Other diseases of the liver.
125. Diseases of the pancreas.
126. Peritonitis without specified cause.
127. Other diseases of the digestive system (cancer and tuberculosis excepted).

VII. *Non-venereal Diseases of the Genito-urinary System and Annexa*

128. Acute nephritis (including unspecified under ten years of age).
129. Chronic nephritis (including unspecified ten years and over).
130. Chyluria.
131. Other diseases of the kidneys and annexa.
132. Calculi of the urinary passages.
133. Diseases of the bladder.
134. Diseases of the urethra, urinary abscess, etc.:
  (*a*) Stricture of the urethra.
  (*b*) Others under this title.
135. Diseases of the prostate.
136. Non-venereal diseases of the male genital organs.

137. Cysts and other benign tumors of the ovary.
138. Salpingitis and pelvic abscess (female).
139. Benign tumors of the uterus.
140. Non-puerperal uterine hemorrhage.
141. Other diseases of the female genital organs.
142. Non-puerperal diseases of the breast (cancer excepted).

VIII. *The Puerperal State*

143. Accidents of pregnancy:
   (*a*) Abortion.
   (*b*) Ectopic gestation.
   (*c*) Others under this title.
144. Puerperal hemorrhage.
145. Other accidents of labor:
   *(*a*) Cesarean section.
   *(*b*) Other surgical operations and instrumental delivery.
   *(*c*) Others under this title.
146. Puerperal septicemia.
147. Puerperal phlegmasia alba dolens, embolus, sudden death.
148. Puerperal albuminuria and convulsions.
149. Following child-birth (not otherwise defined).
150. Puerperal diseases of the breast.

IX. *Diseases of the Skin and of the Cellular Tissue*

151. Gangrene.
152. Furuncle.
153. Acute abscess.
154. Other diseases of the skin and annexa.

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

155. Diseases of the bones (tuberculosis excepted).
156. Diseases of the joints (tuberculosis and rheumatism excepted).
157. Amputations.
158. Other diseases of the organs of locomotion.

XI. *Malformations*

159. Congenital malformations (still-births not included):
   *(*a*) Congenital hydrocephalus.
   *(*b*) Congenital malformations of the heart.
   *(*c*) Others under this title.

XII. *Early Infancy*

160. Congenital debility, icterus, and sclerema.
161. Premature birth; injury at birth:
   *(*a*) Premature birth (not still-born).
   *(*b*) Injury at birth (not still-born).
162. Other diseases peculiar to early infancy.
163. Lack of care.

XIII. *Old Age*

164. Senility.

XIV. *External Causes*

165. Suicide by solid or liquid poisons (corrosive substances excepted).
166. Suicide by corrosive substances.
167. Suicide by poisonous gas.
168. Suicide by hanging or strangulation.
169. Suicide by drowning.
170. Suicide by firearms.
171. Suicide by cutting or piercing instruments.
172. Suicide by jumping from high places.
173. Suicide by crushing.
174. Other suicides.
175. Poisoning by food.
176. Poisoning by venomous animals.
177. Other acute accidental poisonings (gas excepted).
178. Conflagration.
179. Accidental burns (conflagration excepted).
180. Accidental mechanical suffocation.
181. Accidental absorption of irrespirable irritating or poisonous gas.
182. Accidental drowning.
183. Accidental traumatism by firearms (wounds of war excepted).
184. Accidental traumatism by cutting or piercing instruments.
185. Accidental traumatism by fall.
186. Accidental traumatism in mines and quarries:
    *(*a*) Mines.
    *(*b*) Quarries.
187. Accidental traumatism by machines.
188. Accidental traumatism by other crushing (vehicles, railways, landslides, etc.):
    *(*a*) Railroad accidents.
    *(*b*) Street-car accidents.
    *(*c*) Automobile accidents.
    *(*d*) Aëroplane and balloon accidents.
    *(*e*) Motorcycle accidents.
    *(*f*) Injuries by other vehicles.
    *(*g*) Landslide, other crushing.
189. Injuries by animals (not poisoning).
190. Wounds of war.
191. Execution of civilians by belligerent armies.
192. Starvation (deprivation of food or water).
193. Excessive cold.
194. Excessive heat.
195. Lightning.
196. Other accidental electric shocks.
197. Homicide by firearms.
198. Homicide by cutting or piercing instruments.
199. Homicide by other means.
200. Infanticide (murder of infants less than one year of age).†
201. Fracture (cause not specified).

† This title to be omitted when homicides are shown by ages under Titles 197–199.

202. Other external violence.
203. Violent deaths of unknown causation.

XV. *Ill-defined Diseases*

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

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. There have been some slight modifications in respect of details since this last manual was published in 1911. 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.*

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.*

[*] Das Budapester System der Todesursachenstatistik, 1909.

[†] 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.

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 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.

* 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.

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.

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 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.

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, if 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), rheu-

matoid 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 XIV), and also indefinite and ill-defined terms (*e. g.*, "debility," "atrophy") in Classes XI and XII that are referred, for certain ages, to Class XIV, 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.

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 make the point clear. A person has cancer of the breast, is operated upon in the hope of curing this disease, develops a postoperative pneumonia, and dies. Now if the person had not had the *cancer* and had therefore not been operated on for its relief, she would not have died when she did. 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.

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

Again, suppose she had been operated on, and had *not* developed the postoperative pneumonia. Then she might have been permanently cured of the cancer (many 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. 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 lately 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. Suicide.
3. Traumatism (Rubrics 178 to 191 inclusive and 195 and 196).
4. Homicide (Rubrics 197 to 200 inclusive).

This is neither a long nor, except in its first item, a specially important list. But personally I cannot but feel that 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 his material is fundamentally of a dubious character.

CLASSIFICATION OF THE CAUSES OF DEATH FOR RESEARCH PURPOSES

It may fairly be said that at least one of the purposes underlying the routine official collection and publication of vital statistics is the hope that from the analysis and subsequent synthesis of the mass of data so accumulated may come an increased knowledge of the fundamental biologic laws of mortality and natality. But plainly no such enhancement of knowledge is going to come if no one does anything with the statistics after they are gathered. Yet such is the inhibition engendered by the inherently official character of official statistics, on the one hand, and of any large ordered mass of figures, on the other hand, that very little has in fact been done with official vital statistics which in any way corresponds intellectually with what an ingenious boy does when he takes a machine to pieces and puts it together again in conformity with his ideas, at the moment, as to how it ought to be put together. Yet just this process of taking a nicely co-ordinated machine to pieces and putting it together again in a novel way may yield results which are not only potentially entertaining, but may be highly illuminating. It will, of course, always encounter the violent opposition of those minds which dislike to see the established disturbed, or old things looked at in new ways. Many reasons why such pernicious activity should not be indulged in can always be adduced. But the true philosopher will be undisturbed by these considerations.

Some time ago I made an essay at taking apart that piece of mechanism which is called the International List of the Causes of Death and putting it together in a new way. It seems appropriate to include some account of it here, if for no other reason than that it may stimulate some other student to embark upon the same enterprise with other and more valuable results.

In recording the statistics of death the vital statistician is confronted with the absolute necessity of putting every death record into some category or other in respect of its causation. However complex biologically may have been the train of events leading up to a particular demise, the statistician must record the terminal "cause of death" as some particular thing. The International List of the Causes of Death is a code which is the result of many years' experience and thought. Great as are its defects in certain particulars, it nevertheless has certain marked advantages, the most conspicuous of which is that by its use the vital statistics of different countries are put upon a uniform basis.

The several separate causes of death were grouped in the International List into the following general classes when this study was made:

I. General diseases.
II. Diseases of the nervous system and of the organs of special sense.
III. Diseases of the circulatory system.
IV. Diseases of the respiratory system.
V. Diseases of the digestive system.
VI. Non-venereal diseases of the genito-urinary system and annexa.
VII. The puerperal state.
VIII. Diseases of the skin and of the cellular tissue.
IX. Diseases of the bones and of the organs of locomotion.
X. Malformations.
XI. Early infancy.
XII. Old age.
XIII. External causes.
XIV. Ill-defined diseases.

It is evident enough that this is not primarily a biologic classification. The first group, for example, called "General diseases," which caused in 1916, in the Registration Area of the United States approximately one-fourth of all the deaths, is a curious biologic and clinical melange. It includes such diverse entities as measles, malaria, tetanus, tuberculosis, cancer, gonococcus infection, alcoholism, goiter, and many other equally unlike causes of death. For the purposes of the statistical registrar it perhaps has useful points to make this "General diseases" grouping, but it clearly corresponds to little that is natural in the biologic world. Again, in such part of the scheme as does have some biologic basis, the basis is different in different rubrics. Some of the rubrics have an organologic base, while others, as "Malformations," have a causational rather than an organologic base.

Altogether it is evident that, if any synthetic biologic use is to be made of mortality data, a fundamentally different scheme of classification of the causes of death will have to be worked out.

For the purposes of this study* I[1] developed an entirely different general classification of the causes of death on a reasonably consistent biologic basis. The underlying idea of this new classification is to group all causes of death under the heads of the several organ systems of the body, the functional breakdown of which is the immediate or predominant cause of the cessation of life. All except a few of the statistically recognized causes of death in the International List can be assigned to places in such a biologically grouped list. It has a sound logical foundation in the fact that, biologically considered, death results because some organ system, or group of organ systems, fails to continue its function. Practically the plan involves the reassignment of all of the several causes of death now grouped by vital statisticians under heading "I. General diseases." It also involves the re-distributing of causes of death now listed under the puerperal state, malformations, early infancy, and certain of those under external causes.

The headings finally decided upon for the new classification are as follows:

I. Circulatory system, blood, and blood-forming organs.
II. Respiratory system.
III. Primary and secondary sex organs.
IV. Kidneys and related excretory organs.
V. Skeletal and muscular systems.
VI. Alimentary tract and associated organs concerned in metabolism.
VII. Nervous system and sense organs.
VIII. Skin.
IX. Endocrinal system.
X. All other causes of death.

It should be emphasized before presenting the statistics on this new classification that the underlying idea of this rearrange-

* It should be clearly understood that this phrase "For the purposes of this study" means precisely what it says. I am not advocating a new classification of the causes of death for statistical use. I should oppose vigorously any attempt to substitute a new classification (mine or any other) for the International List now in use. Uniformity in statistical classification is essential to usable practical vital statistics. Such uniformity has now become well-established through the International List It would be most undesirable to make any radical changes in the List now.

ment of the causes of death is to put all those lethal entities together which bring about death because of the functional organic breakdown of the same general organ system. The cause of this functional breakdown may be anything whatever in the range of pathology. It may be due to bacterial infection; it may be due to trophic disturbances; it may be due to mechanical disturbances which prevent the continuation of normal function; or to any other cause whatsoever. In other words, the basis of the present classification is not that of pathologic causation, but it is rather that of organ breakdown. We are now looking at the question of death from the standpoint of the biologist, who concerns himself not with what causes a cessation of function, but rather with what part of the organism ceases to function, and therefore causes death.

The data given are in the form of death-rates per hundred thousand living at all ages from various causes of death, arranged by organ systems primarily concerned in death from the specified disease. The statistics presented are from three widely separated

TABLE 1

SHOWING THE RELATIVE IMPORTANCE OF DIFFERENT ORGAN SYSTEMS IN HUMAN MORTALITY

Group No.	Organ system.	Death rates per 100,000.			
		Registration area, U. S. A.		England and Wales.	Sao Paulo.
		1906–10.	1901–05.	1914.	1917.
II	Respiratory system	395.7	460.5	420.2	417.5
VI	Alimentary tract and associated organs	334.9	340.4	274.1	613.8
I	Circulatory system, blood	209.8	196.8	208.6	254.8
VII	Nervous system and sense organs	175.6	192.9	151.9	124.3
IV	Kidneys and related excretory organs	107.2	107.4	19.4	83.4
III	Primary and secondary sex organs	88.1	77.4	95.4	103.2
V	Skeletal and muscular system	12.6	13.7	18.2	6.8
VIII	Skin	10.1	13.3	12.0	7.9
IX	Endocrinal system	1.5	1.2	1.9	1.1
	Total death-rate classifiable on a biologic basis	1335.5	1403.6	1201.7	1612.8
X	All other causes of death	171.3	211.9	141.4	109.8

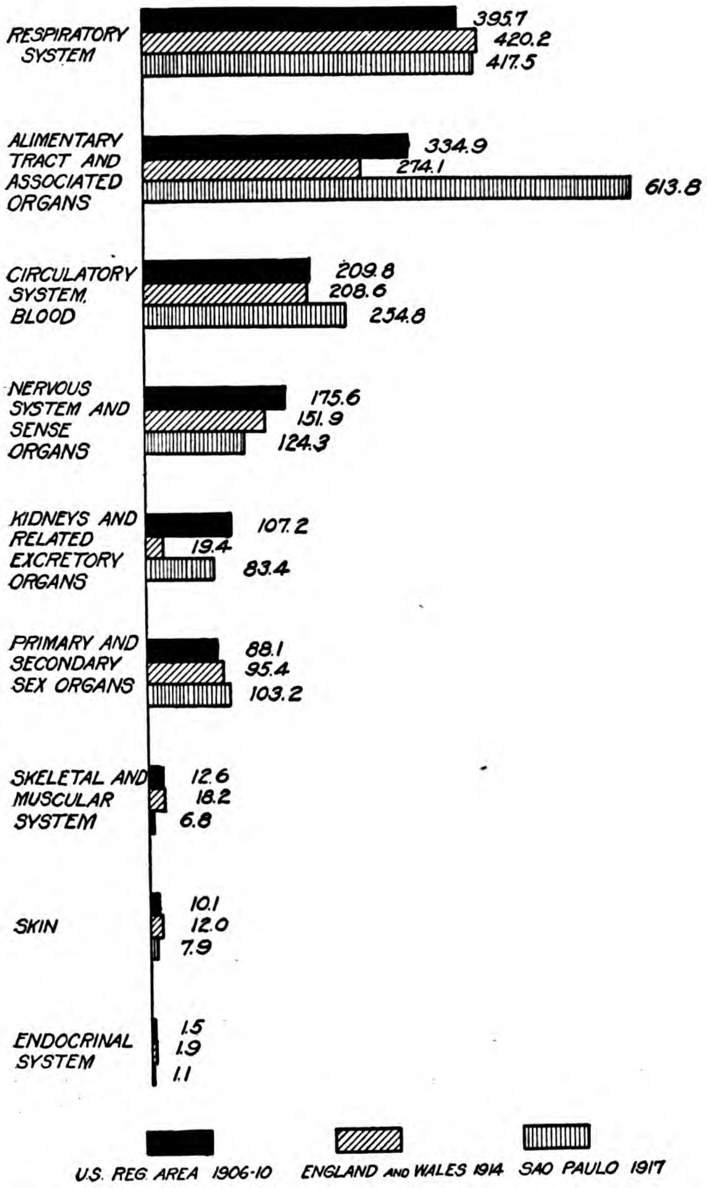


Fig. 10.—Diagram showing the relative importance of the different organ systems of the body in human mortality.

localities and times, viz.: (*a*) from the Registration Area of the United States; (*b*) from England and Wales; and (*c*) from the City of Sao Paulo, Brazil. The first two columns of each table give the death-rates, arranged in descending order of magnitude in the first column, for the Registration Area of the United States for the two periods, 1906–1910 and 1901–1905. The third column of each table gives the death-rate from the same causes of death for England and Wales in the year 1914. The fourth column gives the rates for Sao Paulo for the year 1917. The data for the United States Registration Area were extracted from the volume of Mortality Statistics for 1916, issued by the Bureau of the Census. The English data were extracted from the Report of the Registrar General of England and Wales for 1914. The Sao Paulo rates were calculated from data as to deaths and population given in the "Annuario Demographico" of Sao Paulo for 1917.

In Table 1 (p. 70) the totals are arranged in descending order of magnitude. The results are shown graphically in Fig. 10 (p. 71).

The results of this classification have been discussed in detail elsewhere and need not be gone into here. The curious reader can follow the matter up in the references at the end of this chapter.

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.: Certain Evolutionary Aspects of Human Mortality, Amer. Nat., vol. 54, pp. 5–44, 1920.
7. Pearl, R.: On the Embryological Basis of Human Mortality, Proc. Nat. Acad. Sci., vol. 5, pp. 593–598, 1919.
8. Pearl, R.: A Biological Classification of the Causes of Death, Metron., vol. 1, pp. 92–99, 1921.
9. Pearl, R.: Some Biological Aspects of Human Mortality, Proc. Pathol. Soc. Philadelphia, N. S., vol. 23, pp. 84–87, 1921.

10. Pearl, R.: The Biology of Death, Philadelphia (J. B. Lippincott Co.), 1922. (In Chapters IV and V are given age and sex specific death-rates on the basis of the organologic classification set forth in this chapter.)

11. 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.)

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:

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

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

But this process can be continued indefinitely:

Every living person either	Has smallpox and	Has a fever and	Has an automobile, n_1
			or
			Has no automobile, n_2
		or	
		Has no fever and	Has an automobile, n_3
			or
			Has no automobile, n_4
	or		
	Does not have smallpox and	Has a fever and	Has an automobile, n_5
			or
			Has no automobile, n_6
		or	
		Has no fever and	Has an automobile, n_7
			or
			Has no automobile, n_8

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[2]).

TABLE 2

SHOWING THE INCIDENCE OF INFLUENZA AMONG TUBERCULOUS AND NON-TUBERCULOUS 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. An example of a statistical table based upon such a classification is seen in Table 3.

TABLE 3

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

In this table 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. may be regarded for practical purposes as alike. The correct way to state class limits in setting up a frequency table is that followed in Table 3. 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. . . .

Another model form of such a table on a linear classification is shown in Table 4, taken from a paper by Doctor Huntington Williams on "Epidemic Jaundice in New York State, 1921–1922."*

TABLE 4

AGE DISTRIBUTION OF 700 CASES OF EPIDEMIC JAUNDICE

Years.	Number.	Per cent.
0 to 4	46	6.6
5 to 14	362	51.7
15 to 24	127	18.2
25 to 34	50	7.1
35 to 44	59	8.4
45 to 54	23	3.3
55 to 64	20	2.9
65 to 74	3	0.4
75 to 84	2	0.3
85 to 94	1	0.1
Age not recorded	7	1.0
Total	700	100.0

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 5, which is an expansion of Table 2.

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 I shall not develop, but leave to the reader to work out for himself as a useful exercise in getting familiar with the reading of statistics.

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 fundamentally the most useful form of tabulation for raw, basic statistical data. Why it is the most useful is because it permits the greatest freedom and variety in the subsequent constructive and derivative use of the material.

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

TABLE 5

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			

Table 6 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 × 12 = 96 elemental cells in this table. 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.

TABLE 6

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

Tuberculous lesions			White						Colored						Totals			
			Males			Females			Males			Females						
			Under 20	20 to 49	50 and over	Under 20	20 to 49	50 and over	Under 20	20 to 49	50 and over	Under 20	20 to 49	50 and over				
Not present in lungs	Not present in heart	Not present in kidneys	1	3	..	1	..	..	12	6	1	4	3	2	33	55	59	358
		Present in kidneys	1	1	1	..	1	..	6	8	3	1	..	..	22			
	Present in heart	Not present in kidneys	..	1	..	..	2	..	..	1	..	..	..	..	4	4		
		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	269	299	
		Present in kidneys	4	17	5	7	5	1	7	28	6	12	3	..	95			
	Present in heart	Not present in kidneys	..	2	..	..	..	..	3	5	..	2	..	..	12	30		
		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	Grand Total 358			
			78			45			157			78						
			123						235									
			358															

Furthermore, the frequency of every possible *combination* of these categories is stated in Table 6.

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 tubercluous 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 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.

CLASS LIMITS

A practical question which frequently arises to vex the beginning statistician in making tables 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.[3] 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 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.

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" already referred to, 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

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

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 recorded." 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 (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.

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

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

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 evaluation 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. 18 of Chapter VI on p. 109.

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 reading, 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. Pearl, R.: Preliminary Note on the Incidence of Epidemic Influenza Among the Actively Tuberculous, Quart. Publ. Am. Stat. Ass., vol. 16, pp. 536–540, 1919.
3. Pearl, R.: A Note on the Degree of Accuracy of Biometric Constants, Amer. Nat., vol. 43, pp. 238–240, 1909.
4. Watkins, G. P.: Theory of Statistical Tabulation, Quart. Publ. Amer. Stat. Ass., vol. 14, pp. 742–757, 1915.

CHAPTER V

MEDICAL RECORDS AND THEIR MECHANICAL TABULATION*

THE COMMONEST DEFECTS IN MEDICAL RECORDS

THE fundamental and basic 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 are 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 glaringly 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 resides, far too often, in the head of the particular writer, and there only. And heads, especially of human beings, do vary so! The remedy is patent. Any investigator or administrator who desires to put his clinical records on the most scientific basis will, as a first step, draw up and have printed a

* This chapter follows closely a paper by the author entitled "Modern Methods in Handling Hospital Statistics," Johns Hopkins Hospital Bulletin, vol. 32, pp. 184–194, 1921.

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 *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, and 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 not be under necessity of having a "No information" column in their statistical tabulations relative to this point.

One realizes perfectly that any suggestion in the direction of standardizing case history writing, by the process of putting into operation methods which have been found sound and useful in other branches of science and in modern business, will at once be scornfully or even derisively received by some. It will be argued that any such process tends to cramp the individuality of great or potentially great men. This argument is perfectly valid. It will inordinately cramp such portion of their individuality as finds its expression in carelessness, inaccuracy, forgetfulness, and inattentive observation. In so far as it is desirable to foster and preserve these intellectual qualities, and embalm their results in the permanent archives of a hospital, clinicians and surgeons should be encouraged to go on writing histories in the old, more or less haphazard way.

Furthermore, the argument will be made that no other than the particular clinician or surgeon who is making the records has the competency or right to discuss at all the manner in which case histories are written. But here a little clear thinking is needed. The science and art of making accurate, comprehensive, and essentially complete records of natural phenomena is not exclusively nor even particularly a branch of the science or art of medicine. It is much broader and more basic and is, in every one of its logical principles, common to all sciences. To these principles of scientific record making many persons have devoted many years of study and thought. And it is just precisely *that* field, not medicine, that we are talking about when we are discussing the method of writing case histories.

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 fit 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.

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, there are two glaring 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 all 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.

THE ORGANIZATION OF THE ROUTINE STATISTICAL RECORDS OF A HOSPITAL

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 hospital records each history should be 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. 11*), to cards, the record on the card appearing as a series of punched holes. Then, by means of another machine, known as a "sorter" (cf. Fig. 12), the punched cards can be mechanically sorted, at a rate of about 250 cards per minute, into any desired arrangement relative to any rubric or item of information recorded upon the cards.

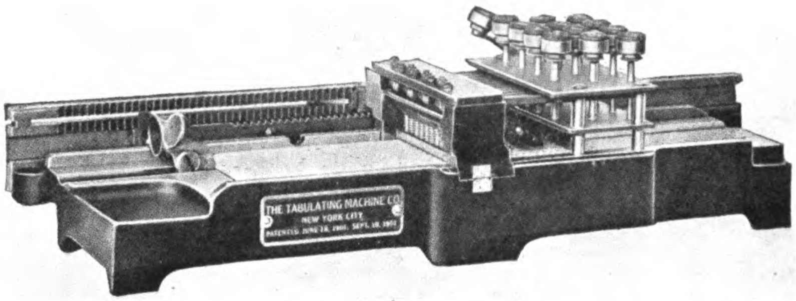

THE TABULATING MACHINE CO.
NEW YORK CITY
PATENTED JUNE 18, 1901. SEPT. 10, 1901

Fig. 11.—Key punch for transferring written records to cards to be used in mechanical tabulation and indexing.

Let us suppose, for example, that someone wishes to assemble

* 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 Co., 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.

for study all the cases of lobar pneumonia which have been treated 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

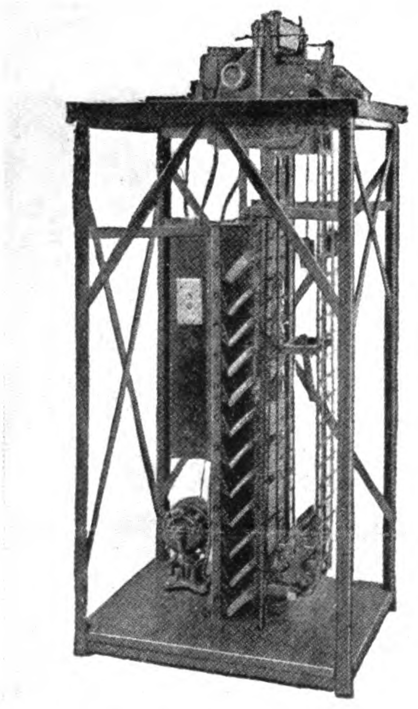


Fig. 12.—Mechanical sorter.

the lobar pneumonia cases in the hospital's records. Each card 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 250 a minute. Compare this with the laborious process that would be involved in assembling 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.

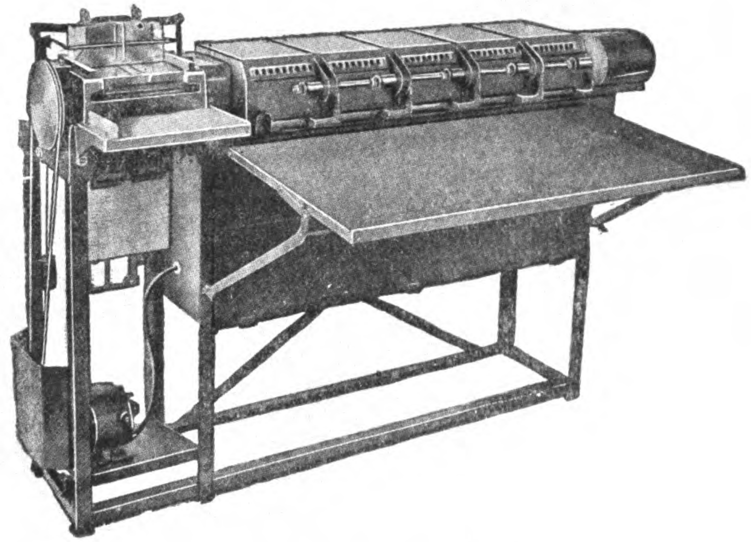


Fig. 13.—Mechanical tabulator.

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 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. 13), 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.

Examples of the usefulness of this method of handing a hospital's statistics could be multiplied indefinitely. But instead of further considering hypothetic cases, let us proceed specifically to the concrete problem of the organization of card forms for the routine statistics of a hospital.

Figures 14 and 15 (pp. 98, 99) show the necessary card forms.

A detailed explanation of these forms and the manner in which they will operate is necessary.

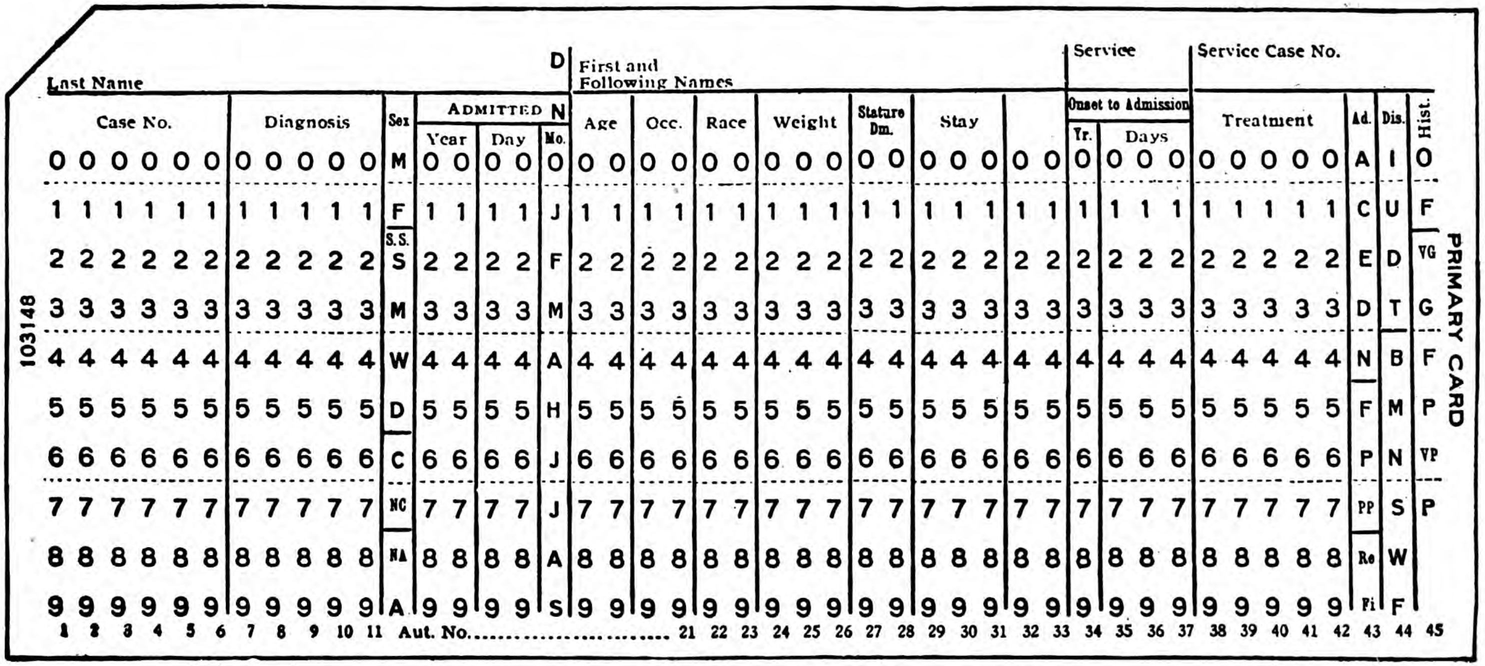
A. The Primary Card

Taking first the primary card form, it may be said that this will presumably be printed upon manila stock. Each group of numbered columns lying between vertical rules is technically known as a "field" of the card.

Across the top of the primary card is written or typewritten: (*a*) the full name of the patient, (*b*) a letter or number designating the service—whether medical, surgical, obstetric, etc.—in which the patient was admitted; and (*c*) the number of the case in that service, on the assumption that in addition to the general hospital serial number of each case there is also a special identifying service number. If a particular service does not specially number its cases, this space will be left blank.

1. The first field is a six-column one and in it is punched the general serial number of the case history. This number identifies the history and the card, and enables one to pass directly from the card to the original history. If a case, for example, is number 12,347 in the hospital's series, this field will be punched 012,347. A six-column field permits the separate serial numbering of 999,999 cases. When this number is passed, presumably the cards for the second million would be printed upon stock of another color, or else a wholly new scheme for handling records will have appeared, as much ahead of punched cards as these are in advance of clay tablets incised with cuneiform characters

2. The second field of five columns records the diagnosis of the patient's chief or primary ailment. This result is attained by the

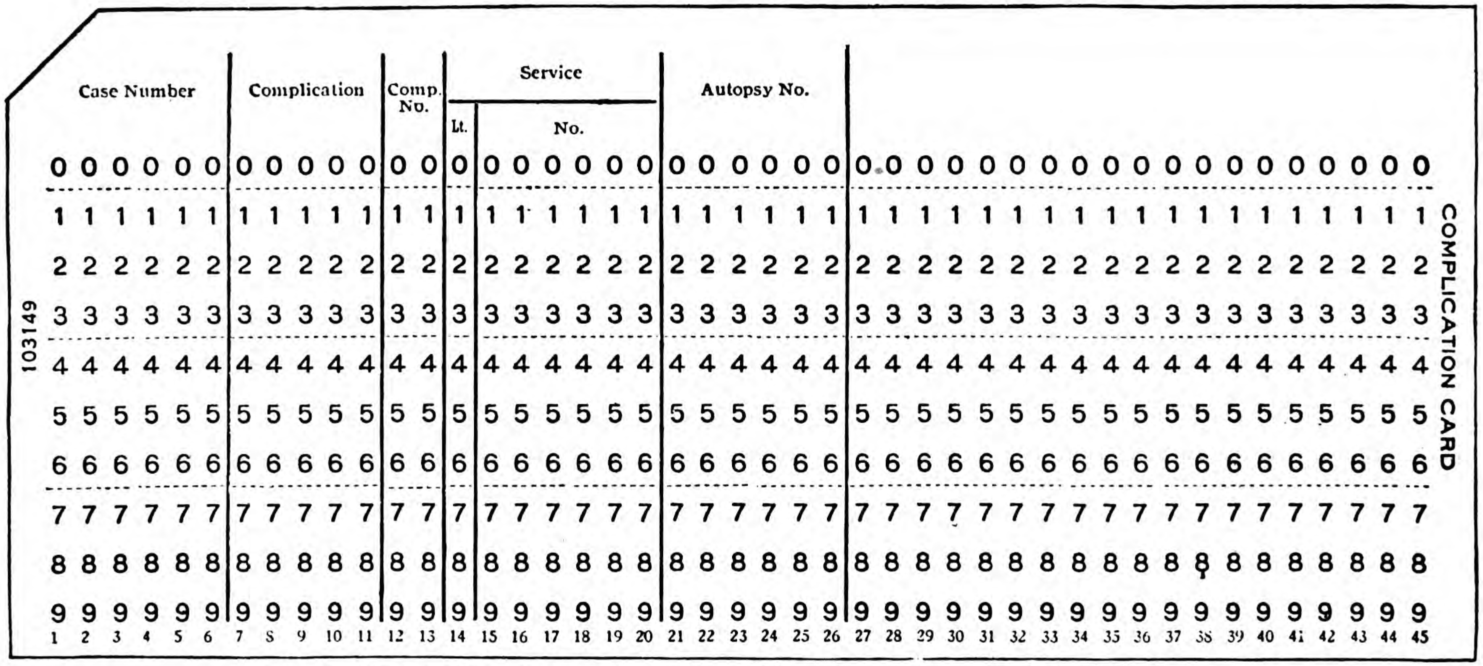


PRIMARY CARD

103148

Last Name																D	First and Following Names																Service				Service Case No.							
Case No.						Diagnosis					Sex	ADMITTED N					Age		Occ.		Race		Weight			Stature Dm.		Stay					Onset to Admission				Treatment					Ad.	Dis.	Hist.
												Year		Day		Mo.																	Yr.	Days										
0	0	0	0	0	0	0	0	0	0	0	M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	A	I	O
1	1	1	1	1	1	1	1	1	1	1	F	1	1	1	1	J	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	C	U	F
2	2	2	2	2	2	2	2	2	2	2	S.S. S	2	2	2	2	F	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	E	D	VG
3	3	3	3	3	3	3	3	3	3	3	M	3	3	3	3	M	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	D	T	G
4	4	4	4	4	4	4	4	4	4	4	W	4	4	4	4	A	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	N	B	F
5	5	5	5	5	5	5	5	5	5	5	D	5	5	5	5	H	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	F	M	P
6	6	6	6	6	6	6	6	6	6	6	C	6	6	6	6	J	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	P	N	VP
7	7	7	7	7	7	7	7	7	7	7	NC	7	7	7	7	J	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	PP	S	P
8	8	8	8	8	8	8	8	8	8	8	NA	8	8	8	8	A	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	Ro	W	
9	9	9	9	9	9	9	9	9	9	9	A	9	9	9	9	S	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	Fi	F	
1	2	3	4	5	6	7	8	9	10	11	Aut. No......................									21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45

Fig. 14.—First or face card form for mechanical tabulation and indexing of routine medical statistics.



COMPLICATION CARD

103149

Case Number						Complication					Comp. No.		Service							Autopsy No.																								
													Lt.	No.																														
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9	9
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45

Fig. 15.—Second or supplemental card form for routine medical statistics.

use of a code of diseases, each pathologic condition it is desired to distinguish being given a separate number. A five-column field permits of 100,000 different discriminatory pathologic statements. It is to be understood clearly that in this field on the primary card is recorded only what the case history states to be the primary or fundamental pathologic condition which the patient presents. The question of the recording of associated and complicating conditions is dealt with below (p. 103). In preparing the nosologic code the best advice of the clinicians, surgeons, etc., will, of course, in all cases be taken.* The field is made larger on the card than there is any present need for, to allow for development of the subject and consequent changes in viewpoint.

3. The third field of one column is a "split field," so-called, and records the following information:

(*a*) Sex of the patient, male (*M*) or female (*F*).

(*b*) Social status, whether single (*S*), married (*M*), widowed *W*), or divorced (*D*).

(*c*) Whether (*C*) or not (*N. C.*) there were complicating pathologic conditions in this case besides the primarily diagnosed condition given in the second field. The presence of this information makes it possible by a single run of the cards through the sorter to separate the uncomplicated cases of a particular disorder from the complicated ones.

(*d*) Whether (*A*) or not (*N. A.*) there was an autopsy made in this case. At the bottom of the card is written under this field, in the event that an autopsy was made, its serial number.

4. In the fourth field of five columns is punched the year, month, and day of admission of this patient to the hospital.

5. In the fifth field is punched the patient's age in years.

6. In the sixth field is punched, according to a code, the patient's occupation.

7. In the seventh field is recorded the patient's race, according to the ethnologic code used by the U. S. Bureau of Immigration, or according to some other code if preferred. This information as to race necessarily also covers color.

* A disease code for use in mechanical tabulation has been very carefully worked out in the Surgeon-General's Office by Major Albert H. Love and his associates.

8. The eighth field of three columns records the weight of the patient on admission, in kilograms (or, of course, if one prefers, in pounds). The most progressive of modern hospitals record weight on admission as a routine procedure.

9. In the ninth field of two columns the stature is punched in dekameters (as close as will ever be used in statistical groupings) or inches, if one prefers as a routine to use common rather than metric measures. The stature in centimeters may be written at the top of the field if the more exact record is desired.

10. The tenth field of three columns records the duration of the patient's stay in the hospital in days.

11. In the eleventh field of two columns is punched the pigmentation of the individual on a combined eye-color and hair-color code.

12. In the twelfth, four-column field, is recorded the duration of time in years and days as stated in the history, between the first onset or appearance of the diagnosed condition and the admission of the patient to the hospital for treatment.

13. The thirteenth field of five columns is a very important one. It records, according to a code which can be made as elaborate and detailed as is desirable, the nature of the treatment given in the hospital to this particular case. Five columns permit of 99,999 separate discriminatory items to be recorded in this field. Suppose, for example, one wishes to study the pneumonia cases in which digitalis was administered, in comparison with those in which this therapeutic measure was not employed. To pick out by hand from all the pneumonia cases the material according to this arrangement would involve an amount of labor which would deter the most enthusiastic young intern. But, mechanically, through the medium of this field, it can be very easily and quickly accomplished.

14. The fourteenth is a split single-column field. It records the following information:

(*a*) The nature of the case upon admission, whether an acute illness (*A*), a chronic (*C*), an emergency or accident case (*E*), a case admitted for purpose of diagnosis (*D*), or a normal person (*N*), as, for example, a normal pregnant woman admitted to the obstetrical service for delivery.

(*b*) The financial arrangements of the patient with the hospital, whether a free patient (*F*), paying (*P*), or partly paying (*P. P.*).*

(*c*) Whether this case represents the first admission of the patient to the hospital, or whether it is a readmission.

15. The fifteenth field is also a single column split, and records:

(*a*) The condition at discharge, whether improved (*I*), unimproved (*U*), dead (*D*), or transferred to some other service or hospital (*T*).

(*b*) The location of the patient's residence, whether in Baltimore (*B*), or in Maryland outside of Baltimore (*M*), or in the Atlantic seaboard states north of Maryland (Pennsylvania, Delaware, New Jersey, New York, and New England States) (*N*), or in the Atlantic seaboard states south of Maryland (District of Columbia, Virginia, the Carolinas, Georgia, Florida) (*S*), or in some other part of the United States not specified above (*W*), or in a foreign country (*F*).†

16. The sixteenth and last field on the card is again a single column split field. It records the following information:

(*a*) Whether or not (*F* and *O*) there is any statement in the written history as to family history of any particular disease in the patient's family. The information enables one interested in the influence of heredity on disease to pick out quickly the cases likely to be of any value to him.

(*b*) The general health of the patient prior to the present illness, as recorded in the written history. This information is punched according to the following or similar code:

Very good..................	never ill
Good.......................	minor ailments only
Fair........................	average amount of sickness
Poor........................	frequently ill
Very poor..................	an invalid throughout life

* In some hospitals, of course, this information under Item 14*b* would not be pertinent and could be omitted or replaced by something else.

† The 15*b* field is obviously drawn up solely for the Johns Hopkins Hospital and would require modification for any other hospital. Some institutions may not desire statistical information as to place of residence, in which event this portion of the card may be used for recording something else.

(*c*) Whether this card is a primary card. This is a technical point of interest only in connection with the filing of the punched cards.

This completes the description of the primary card. It records 31 different kinds or items of information.

B. The Secondary Card

The basic purpose of the secondary card shown in facsimile in Fig. 15 is to take care of the complicating diseases. An illustration from an actual case history will make the point clear:

A patient, X, was admitted under the primary diagnosis of hyperthyroidism and adenoma of the thyroid. A double lobectomy was done. A postoperative bronchopneumonia developed, and the patient died fourteen days after the operation. At autopsy besides the bronchopneumonia there were found clear evidences of (1) a cerebral embolus with softening, (2) chronic and acute verrucose mitral endocarditis, (3) multiple myomata of the uterus, (4) cystitis, and (5) fibrous pleurisy.

Now the primary card discussed in the preceding section would carry in the "Diagnosis" field only the adenoma of the thyroid. Yet clearly for any adequate statistical records there must be included some account of the other complicating conditions disclosed by the history. This is done by punching *as many of the secondary cards, shown in Fig. 15, as there are separate and distinct complications* (that is, in the present case, one secondary card for bronchopneumonia, one for cerebral embolus, one for endocarditis, one for myomata of the uterus, one for cystitis, and one for pleurisy). Each secondary card carries the same case number in the first field as the associated primary card, and will therefore automatically file with it. The "Complication" field on the secondary card registers with the "Diagnosis" field on the primary card. Therefore, when all the cards, both primary and secondary, are run through the sorting machine relative to this field, all identical diseased conditions, whether primary or complicating, will be brought together. Then by a second sorting of the cards the cases in which any particular disease, say bronchopneumonia, was the occasion of admission to the hospital, can be separated from those cases in which this disease was a secondary complication.

The remaining fields on the secondary card are used to record additional information for which there was not space on the primary card. These include: (1) the ordinal number of the complication as recorded in the history, (2) the service letter and the number of the case in that service which was written but not punched on the primary card, the purpose of the letter column being to enable the ready assembling, for any desired purpose, of all the cases in a particular service, and (3) autopsy number, which also was written but not punched on the primary card.

MECHANICAL TABULATION IN VITAL STATISTICS

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.

SUGGESTED READING

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 Co., 50 Broad St., New York, for literature regarding its application in various fields.

The following references to particular applications of mechanical tabulation, or to its development historically, will be found useful:

1. 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.)
2. 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.)
3. 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.)
4. 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.)
5. 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.)

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. I shall attempt 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. 16) 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. Generally 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.

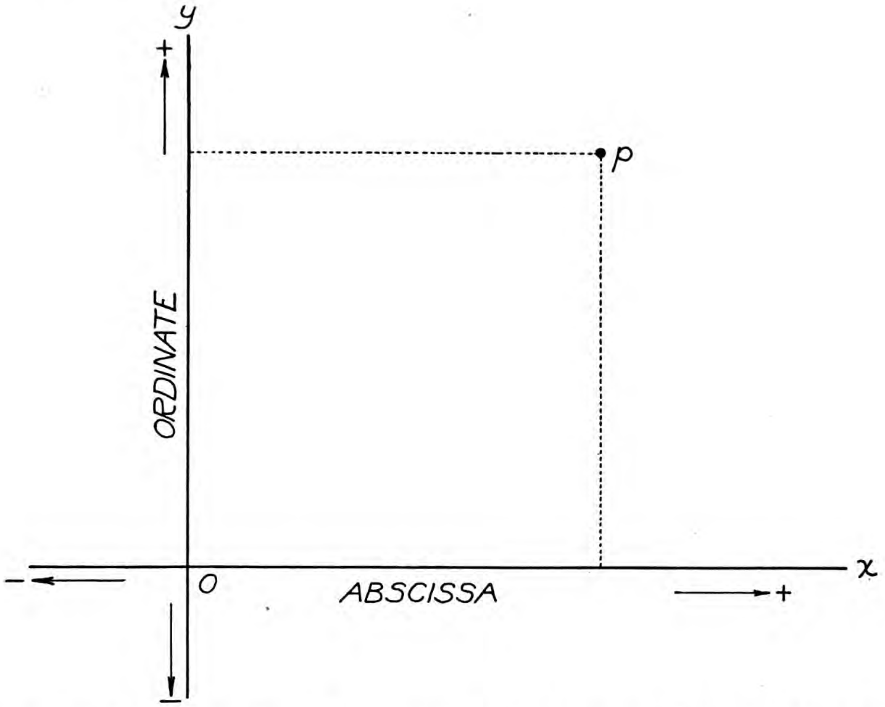


Fig. 16.—Diagram to illustrate rectangular co-ordinates. *o* is the origin. The arrows indicate the conventional directions relative to algebraic signs.

Besides rectangular co-ordinates, there are sometimes used in statistical diagrams:

(*a*) Angular co-ordinates (as in "pie" diagrams).

(*b*) Polar co-ordinates.

(*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. Things which vary discontinuously.
      Type of diagram: (*a*) Bar diagram (cf. Figs. 17 and 18).
      (*b*) "Pie" diagram (cf. Fig. 19).
   2. Things which vary continuously.
      Type of diagram: (*a*) Histogram (cf. Figs. 20–22).
      (*b*) Frequency polygon (cf. Figs. 23, 24).
      (*c*) Ogive curve (cf. Fig. 25).
      (*d*) Integral curve (cf. Figs. 26, 27).

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. 31, 32).
      (*b*) Line diagram on arithmetic grid (cf. Figs. 28–30).
   2. In Time. Cyclic.
      Type of diagram: (*a*) Line diagram on arithmetic grid (cf. Fig. 33).
      (*b*) Polar co-ordinates (cf. Fig. 34).

C. *Purpose:* To show geographic distribution of things or events.
      Type of diagram: (*a*) Spot map (cf. Fig. 35).
      (*b*) Shaded map (cf. Fig. 36).

D. *Purpose:* To facilitate or replace computation.
      Type of diagram: (*a*) Nomogram (cf. Figs. 37–39).

Bar Diagrams

The bar diagram is the simplest possible picture of a statistical situation. Figure 17 is a bar diagram* showing the proportion which each of the more important foods contributes to the total protein consumed in the United States by human beings.

From this diagram one sees at a glance the relative significance of the great staple foods in furnishing protein for human consumption. Wheat stands first. Beef contributes roughly one-half as

* From R. Pearl, The Nation's Food, Philadelphia, 1920, p. 237. Data on which diagram is based are there given.

much protein to the national dietary as wheat, and poultry and eggs about half as much as beef, etc. The whole story of the

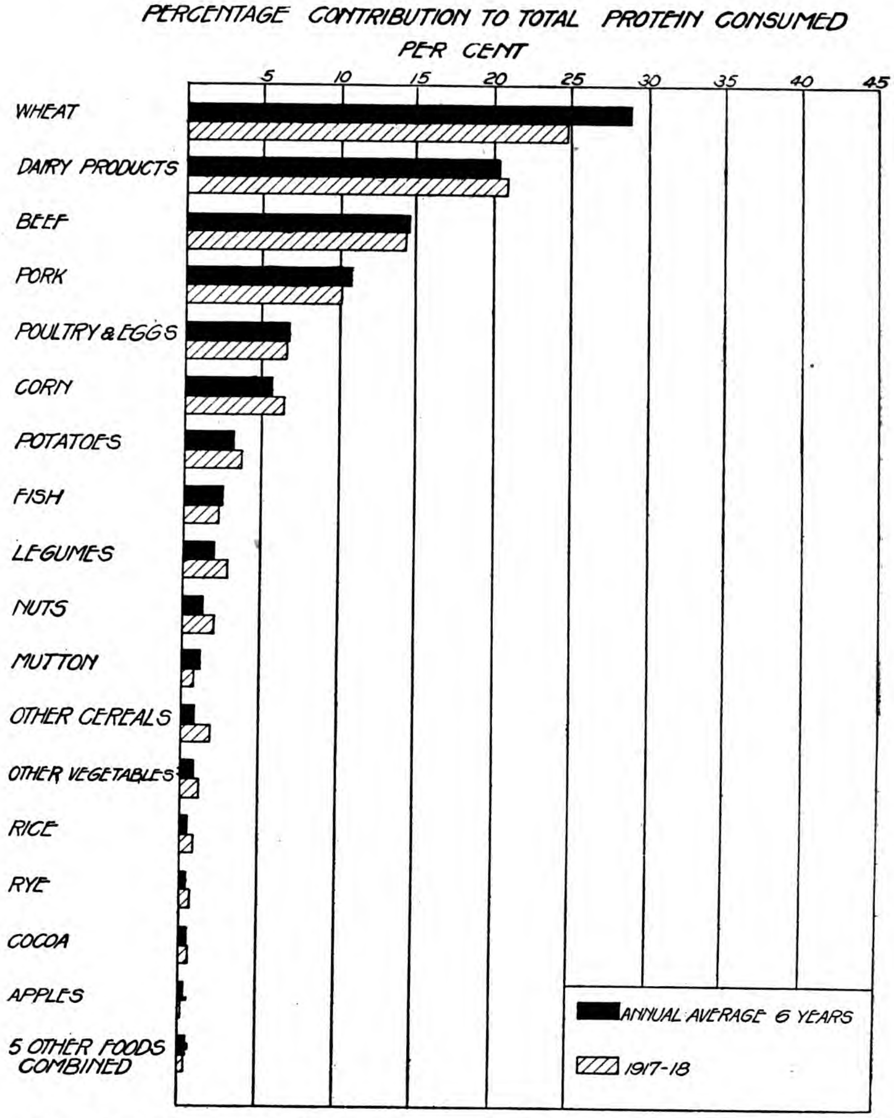


Fig. 17.—Diagram showing the percentage of the total protein consumed in the United States contributed by each of 23 commodities. The solid bars denote the average consumption in the six years preceding our entry into the war. The cross-hatched bars denote the consumption in 1917 and 1918.

sources of the protein we, as a people, consume is accurately visualized.

The percentage columns of Table 8 in Chapter IV make the bar diagram shown in Fig. 18. This is a slightly different form of bar diagram from that shown in Fig. 17.

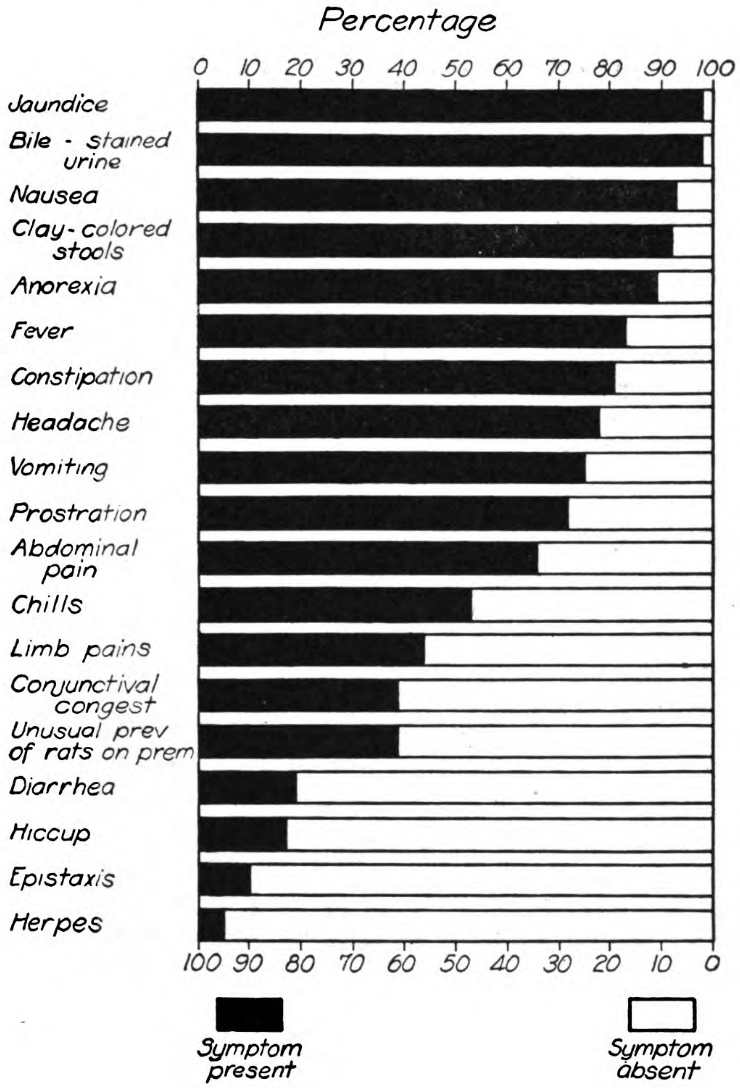


Fig. 18.—Bar diagram based upon data of Table 8, Chapter IV, showing the relative frequency of different symptoms in epidemic jaundice.

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. 19.

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, and Integral Frequency Diagrams

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 9 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).

* Tocher, J. F.: Anthropometric Survey of the Inmates of Asylums in Scotland, Henderson Trust Reports, vol. i, Edinburgh, 1905.

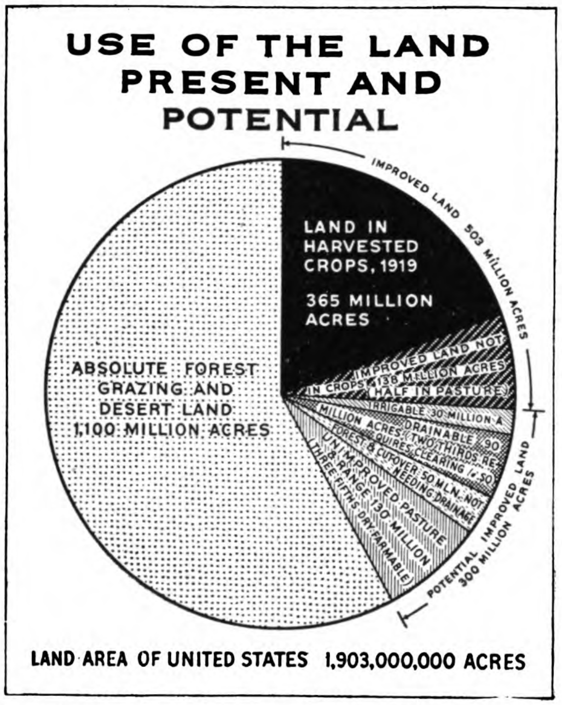


Fig. 19.—Example of diagram to angular co-ordinates. (Reproduced by permission of Dr. O. E. Baker and the editor of the Geographical Review from an article by Dr. Baker entitled "Land Utilization in the United States: Geographical Aspects of the Problem," published in the Geographical Review, vol. 13, January, 1923.)

The data of Table 9 (p. 112) 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 10 (p. 113).

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.

"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 9

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 10. The result will be that shown in Fig. 20.

Now it is at once evident that Fig. 20 is an inadequate and misleading graphical representation of the important facts about variation in head height in this group of people. It is a long, flat

TABLE 10

FREQUENCY DISTRIBUTION OF HEAD HEIGHTS FROM TABLE 9

Head heights, mm.	Ungrouped frequencies.	Frequencies grouped in 5 mm. classes.	Class limits for group frequencies, mm.
119	1	1	115–119
120		....	120–124
121			
122			
123			
124			
125	3	10	125–129
126	2		
127			
128	1		
129	4		
130	2	15	130–134
131	5		
132	1		
133	2		
134	5		
135		17	135–139
136	1		
137	6		
138	5		
139	5		
140	2	16	140–144
141	3		
142	2		
143	6		
144	3		
145	3	5	145–149
146	1		
147			
148			
149	1		
150	2	3	150–154
151			
152	1		
153			
154			
155	1	1	155–159
Totals	68	68	—

thing with many gaps and only roughly indicates what general sorts of head heights occur most frequently. The grouping, in 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. 21, which is a histogram plotted from the grouped data of the latter half of Table 10.

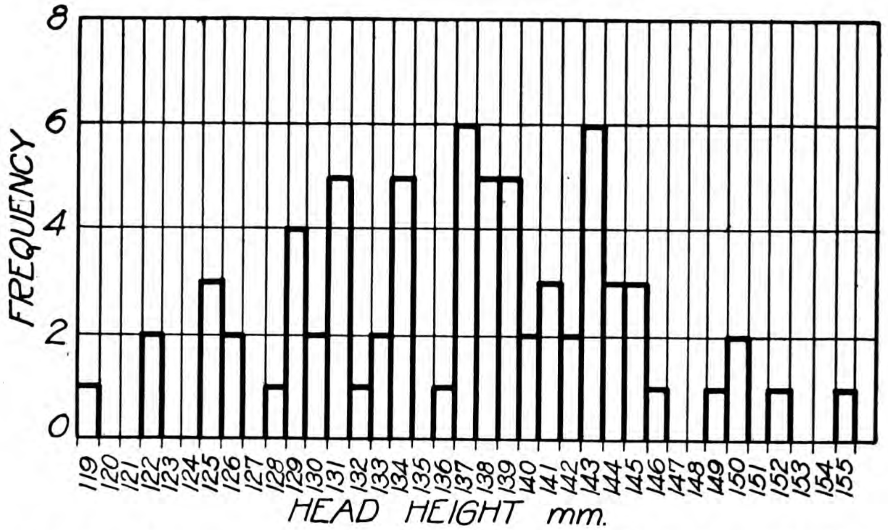


Fig. 20.—Histogram of ungrouped frequencies of head height from Table 10.

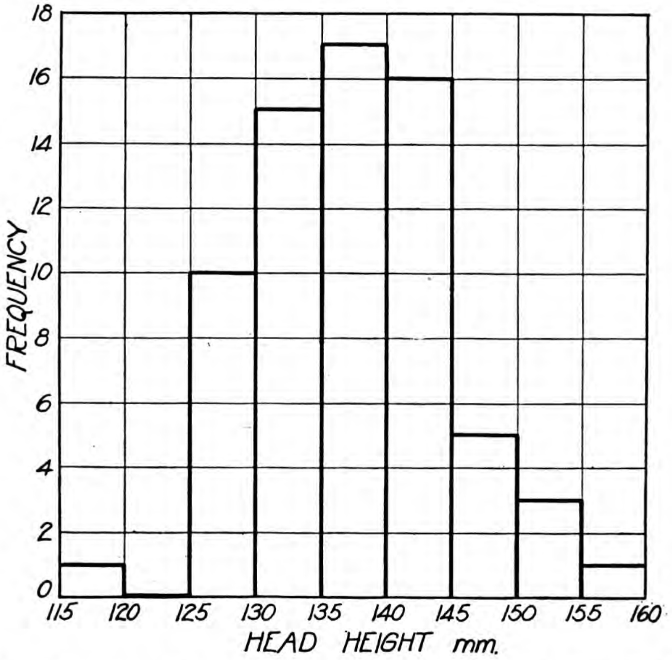


Fig. 21.—Histogram of grouped frequencies of head height from Table 10.

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. Other examples of histograms are seen in Figs. 60, 63 *infra*. A method of drawing a histogram which is preferred by some statisticians is that shown in Fig. 22. 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 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.

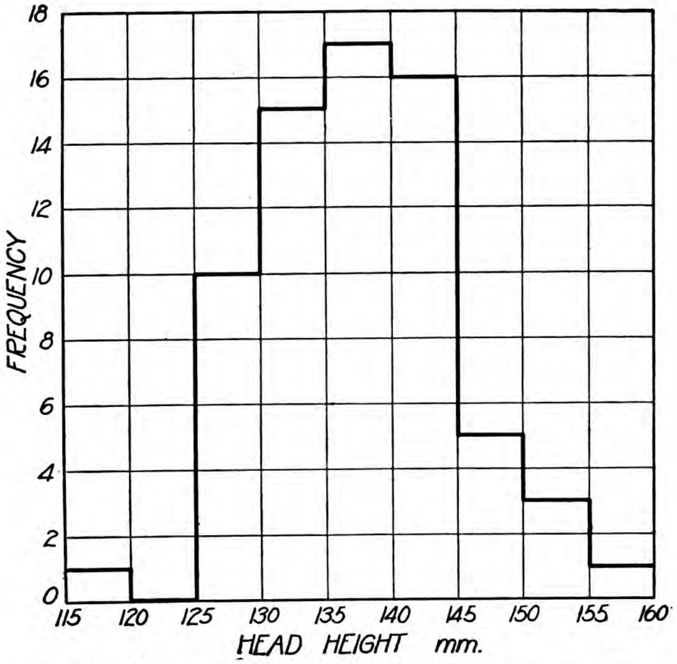


Fig. 22.—Alternative form of histogram shown in Fig. 21.

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. 21 is shown plotted as a frequency polygon in Fig. 23.

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 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. 24.

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 10 are plotted as an ogive curve in Fig. 25.

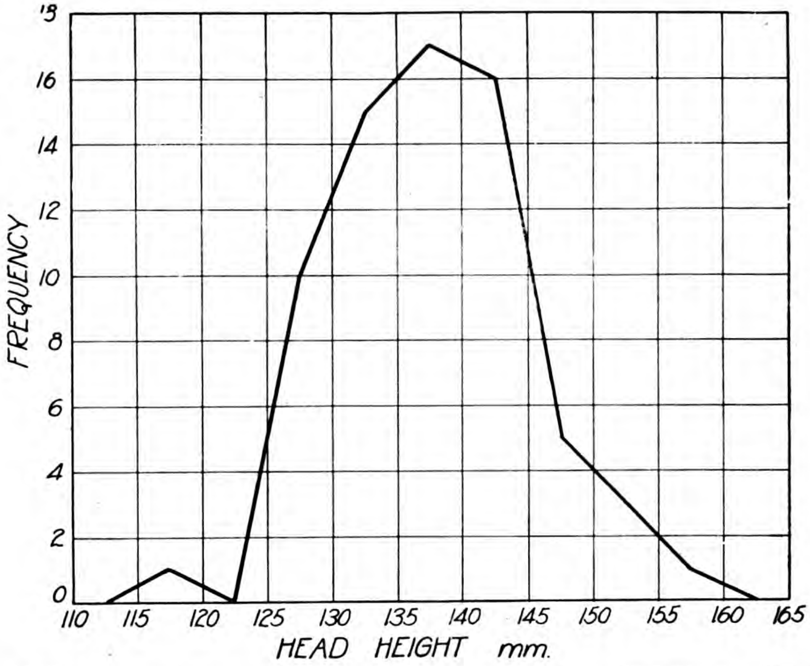


Fig. 23.—Frequency polygon of grouped frequencies of head heights from Table 10.

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. 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.

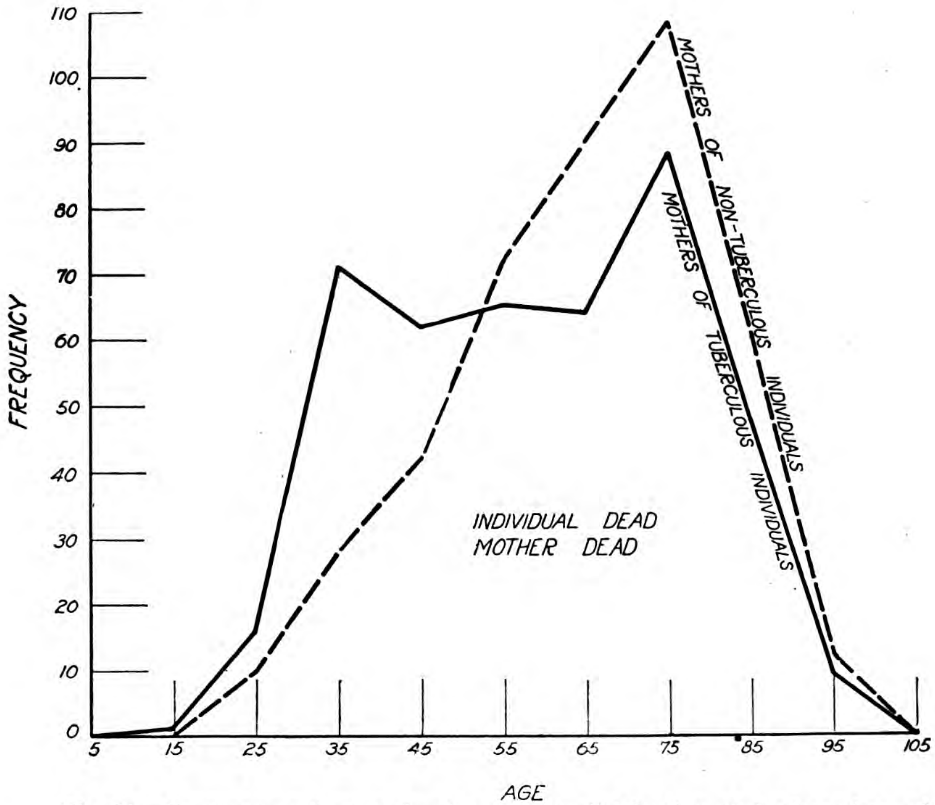


Fig. 24.—Frequency polygons showing the age distribution of dead mothers of dead (*a*) tuberculous (solid line) and (*b*) non-tuberculous (broken line) individuals. (Reproduced from Pearl, R., "The Age at Death of the Parents of the Tuberculous and the Cancerous," Amer. Jour. Hygiene, vol. 3, pp. 71–89, 1923.)

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 end on. The integral curve for the data of Table 10 is shown in Fig. 26.

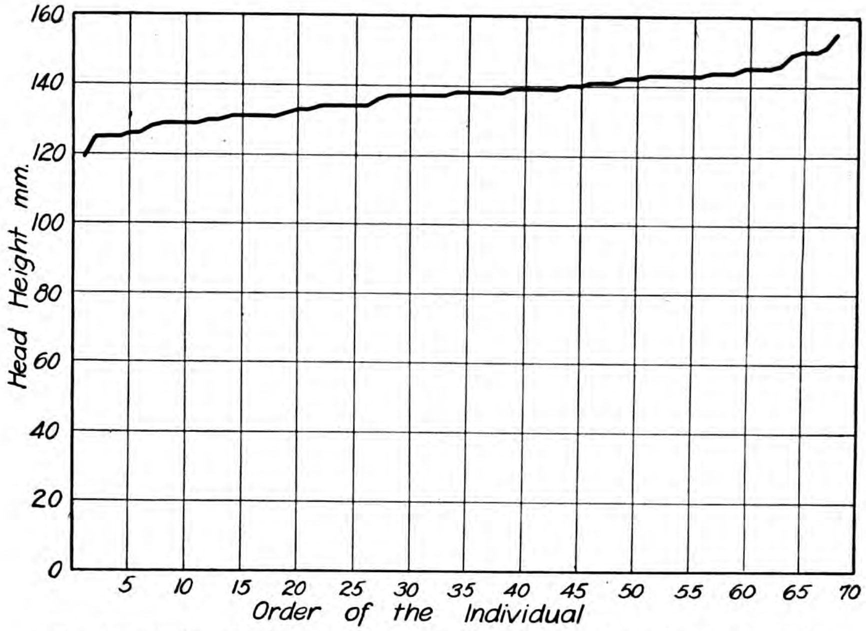


Fig. 25.—Ogive of ungrouped frequencies of head height, from Table 10.

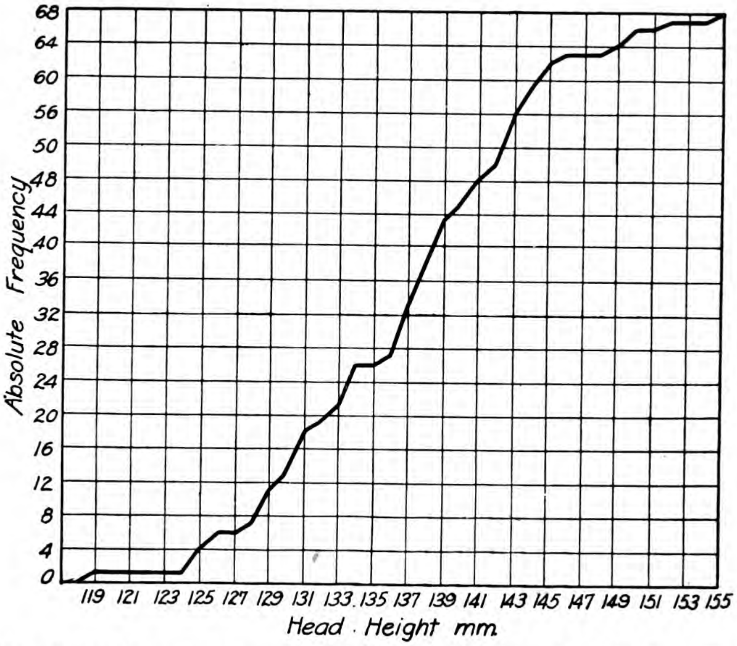


Fig. 26.—Integral curve of ungrouped frequencies of head height from Table 10.

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. 26, 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 margin still remains an absolute scale for frequencies in the same group. The resulting diagram is shown as Fig. 27.

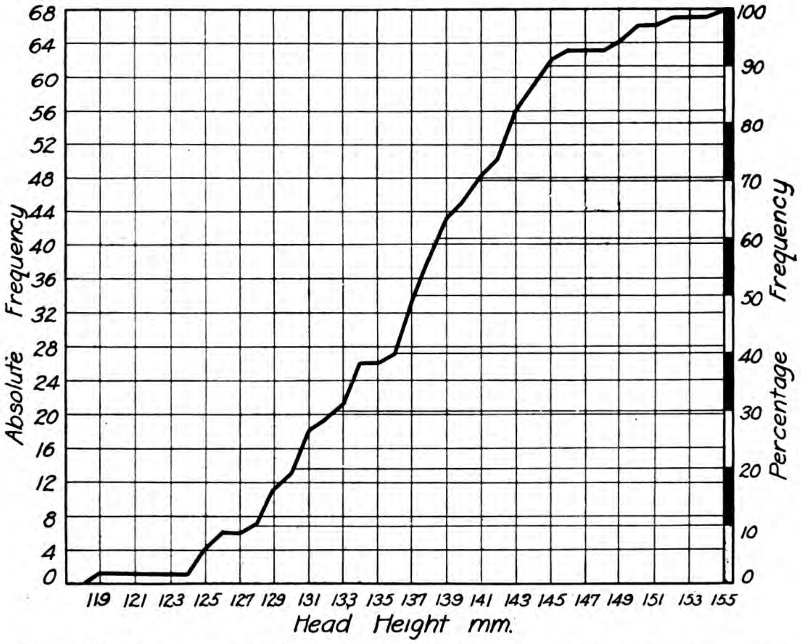


Fig. 27.—Like Fig. 26, but with added scale of relative or percentage frequencies.

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 143 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 difference is that in the case of the ogive frequencies are plotted along the abscissal axis, and in the integral along the y axis as usual.

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 frequency of occurrence of the event in question as ordinate. Thus suppose it is desired to show the decline in the death-rate in Baltimore from 1889 to 1919 inclusive. A diagram like that shown in Fig. 28 may be prepared.

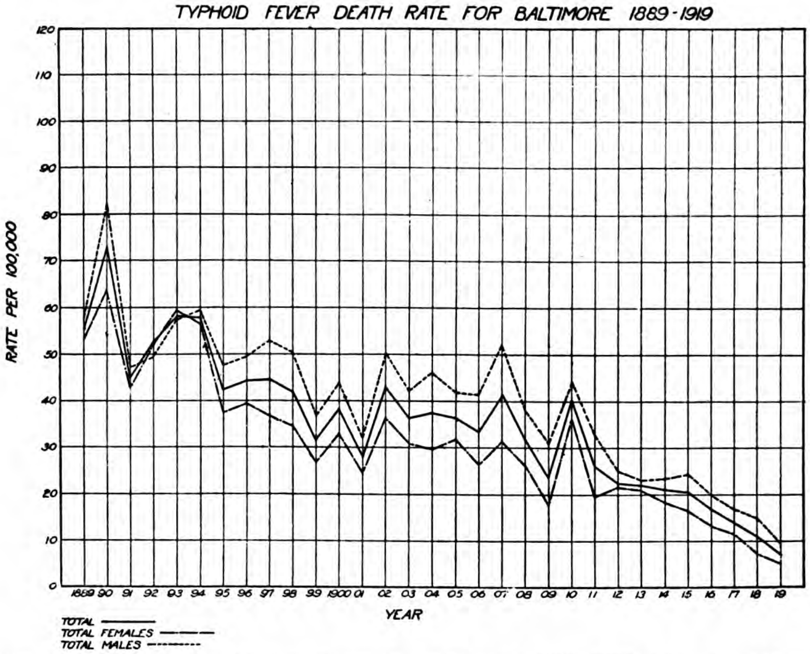


Fig. 28.—Death-rate from typhoid in Baltimore 1889–1919 inclusive for males, females, and total population. (From Howard, W. T., "The Natural History of Typhoid Fever in Baltimore, 1851–1919," Johns Hopkins Hospital Bulletin, vol. 31, pp. 276–286, 319–334, 1920.)

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 11 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 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.

TABLE 11

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

Using the same graphic methods as in Fig. 28 and the data from Table 11 we get the result shown in Fig. 29.

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 slopes downward much more steeply.

But is this conclusion correct? The diagram presented in Fig. 29 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 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

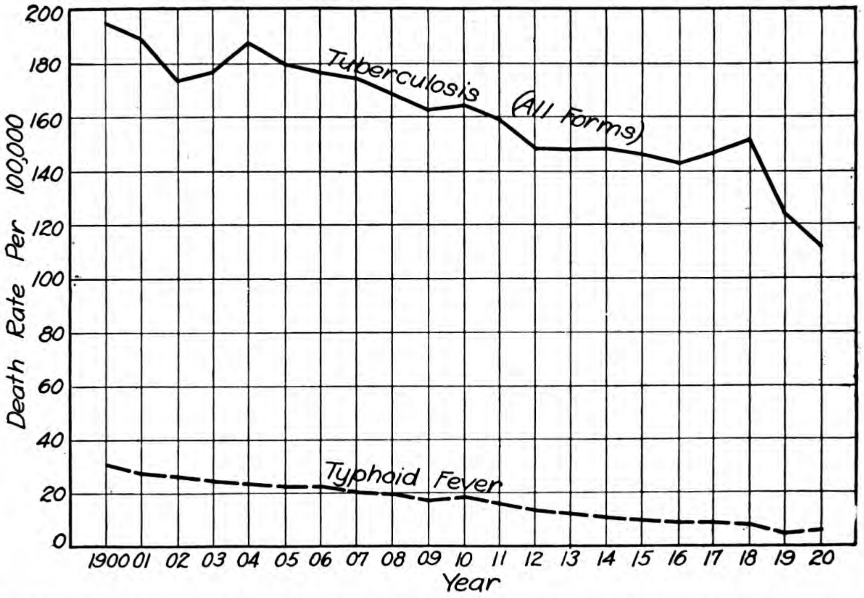


Fig. 29.—Death-rates from (*a*) tuberculosis (all forms) and (*b*) typhoid fever in the Registration Area, 1900–1920 inclusive. Arithmetic grid.

on an arithmetic grid as is Fig. 29. The result will be as shown in Fig. 30.

Anyone looking at this diagram would surely conclude that the decline in mortality had been much more rapid in the first community 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.

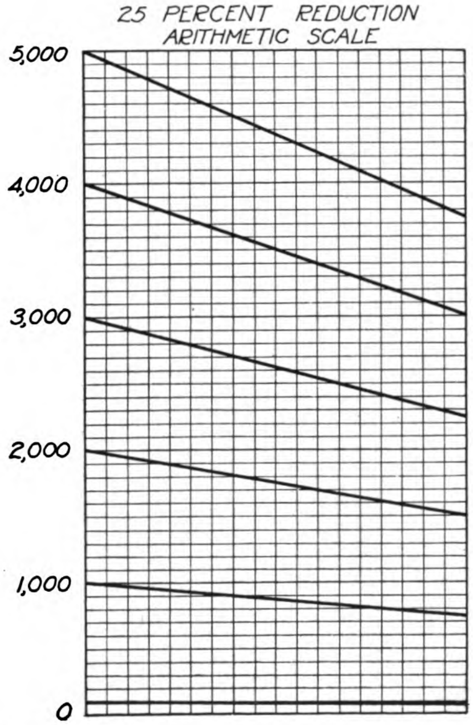


Fig. 30.—Diagram on arithmetic grid to show result of 25 per cent. reduction in mortality in each of six places of different size. Hypothetic case.

But plainly such a result cannot be attained by plotting these data on 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 progression. Such a ruling is called an *arithlog* grid. The result is shown in Fig. 31.

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.

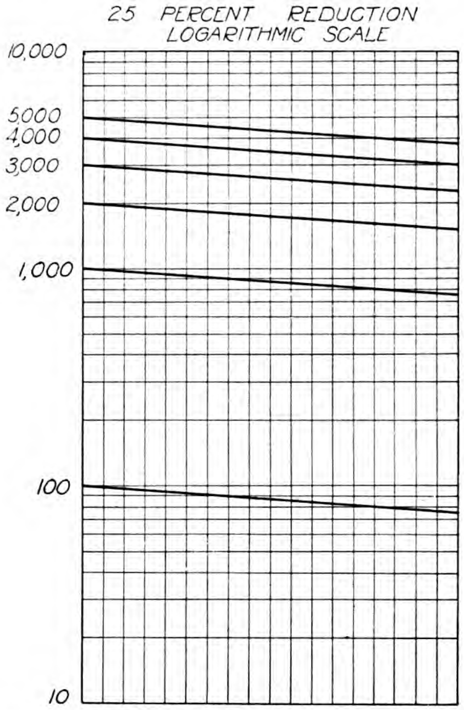


Fig. 31.—Showing the result of plotting the data of Fig. 30 on an arithlog grid. Compare with Fig. 30.

It will be well now to go back and replot the data of Fig. 29 on an arithlog grid. The result is that shown in Fig. 32.

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:

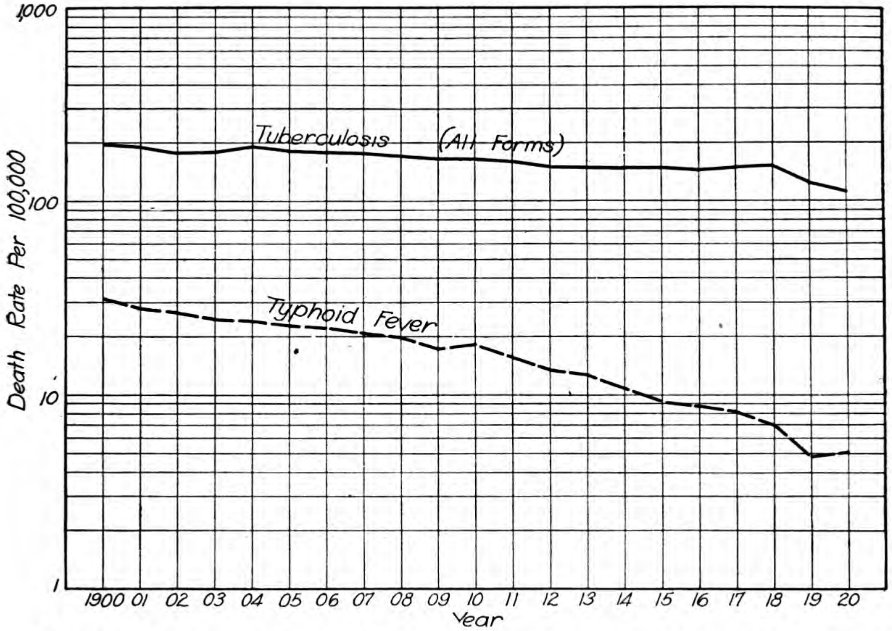


Fig. 32.—Death-rates from (*a*) tuberculosis (all forms) and (*b*) typhoid fever in the original registration states, 1900–1920 inclusive. Arithlog grid. Compare with Fig. 29.

"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.

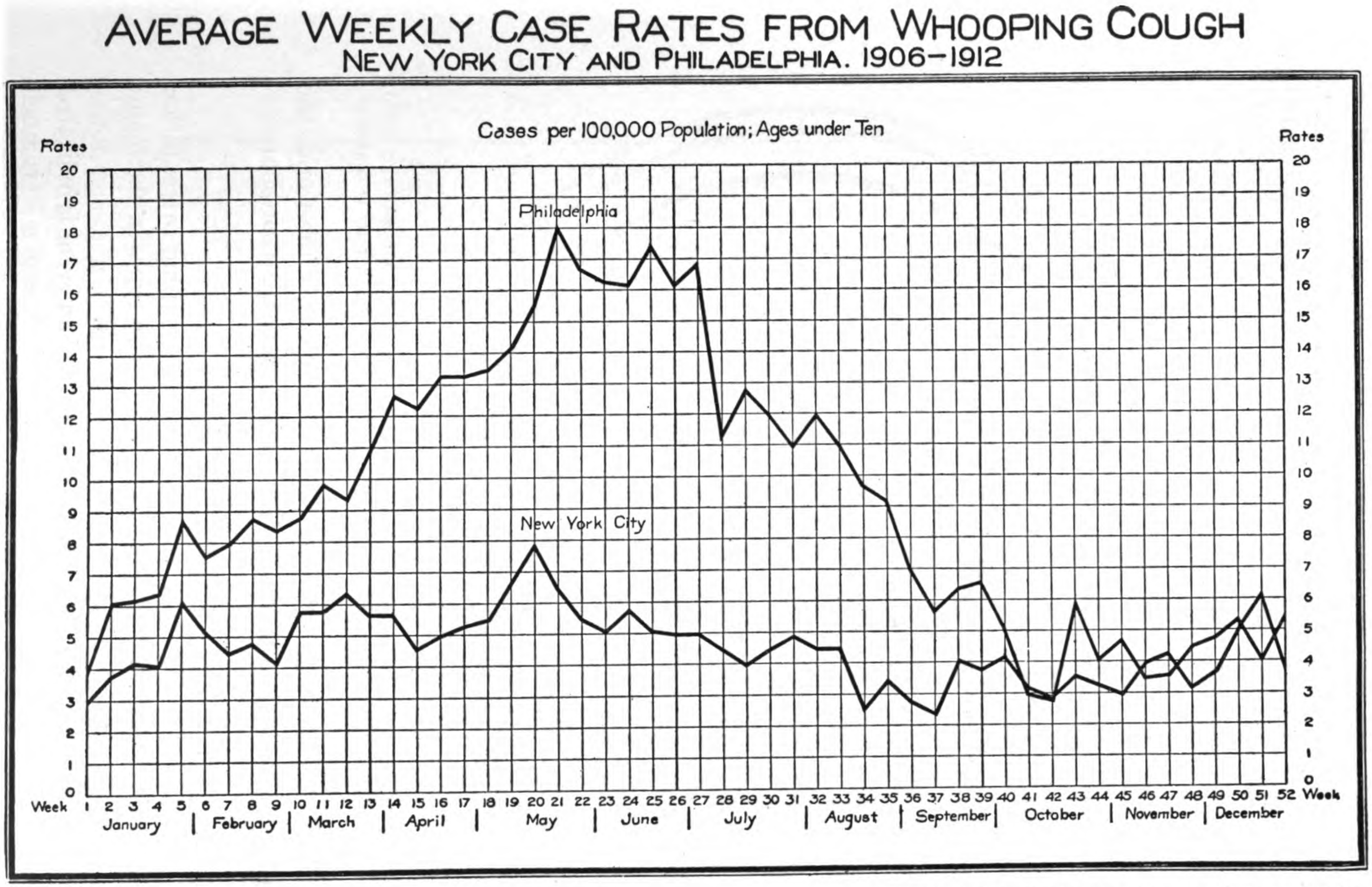
Cyclic Time Trend Diagrams

A cyclic event is one whose frequency of occurrence varies in an orderly recurring manner. An example is found in the seasonal incidence of various diseases, as shown in Fig. 33 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 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. 33 these are perhaps to be preferred in ordinary statistical work to polar co-ordinate diagrams for cyclic events.

An interesting and useful method of showing graphically the



STATISTICIAN'S DEPT., THE PRUDENTIAL INSURANCE COMPANY OF AMERICA

Fig. 33.—Average weekly case incidence rates from whooping-cough in two cities. (Reproduced by courtesy of the Statistician's Department of the Prudential Insurance Company.)

time relations of certain kinds of cyclic phenomena is presented in Fig. 34. This diagram, taken from the Annual Report for 1922 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.

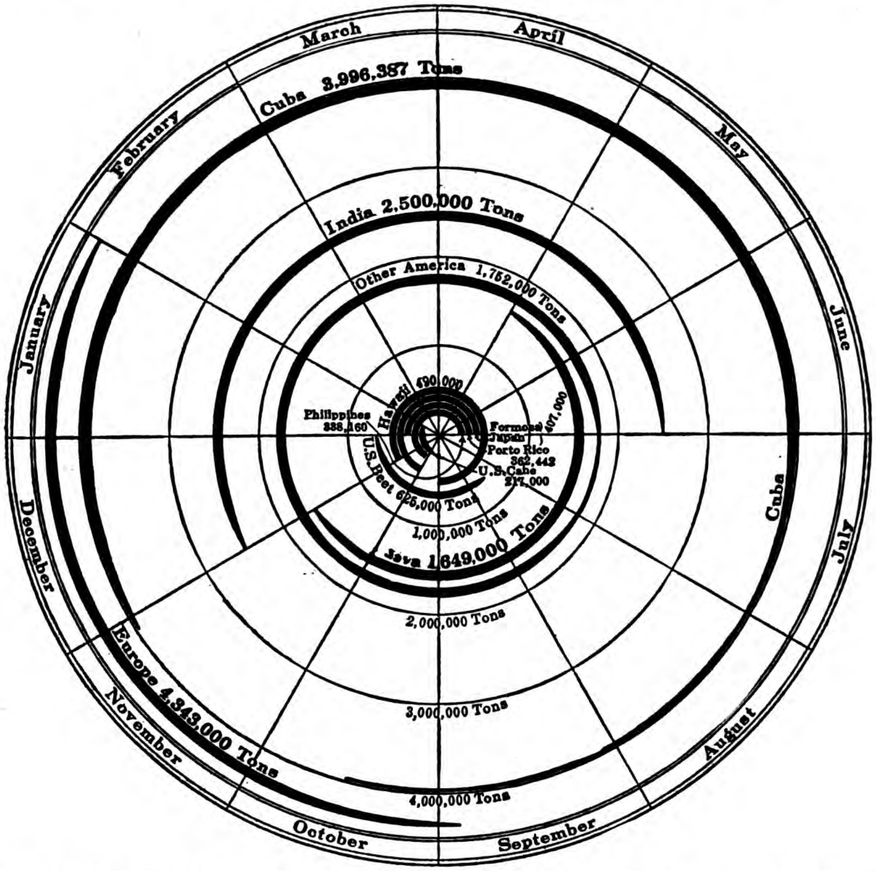


Fig. 34.—Diagram showing time of harvesting of principal sugar crops of the world. (Reproduced from source indicated in text, by permission of Mr. Earl D. Babst.)

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

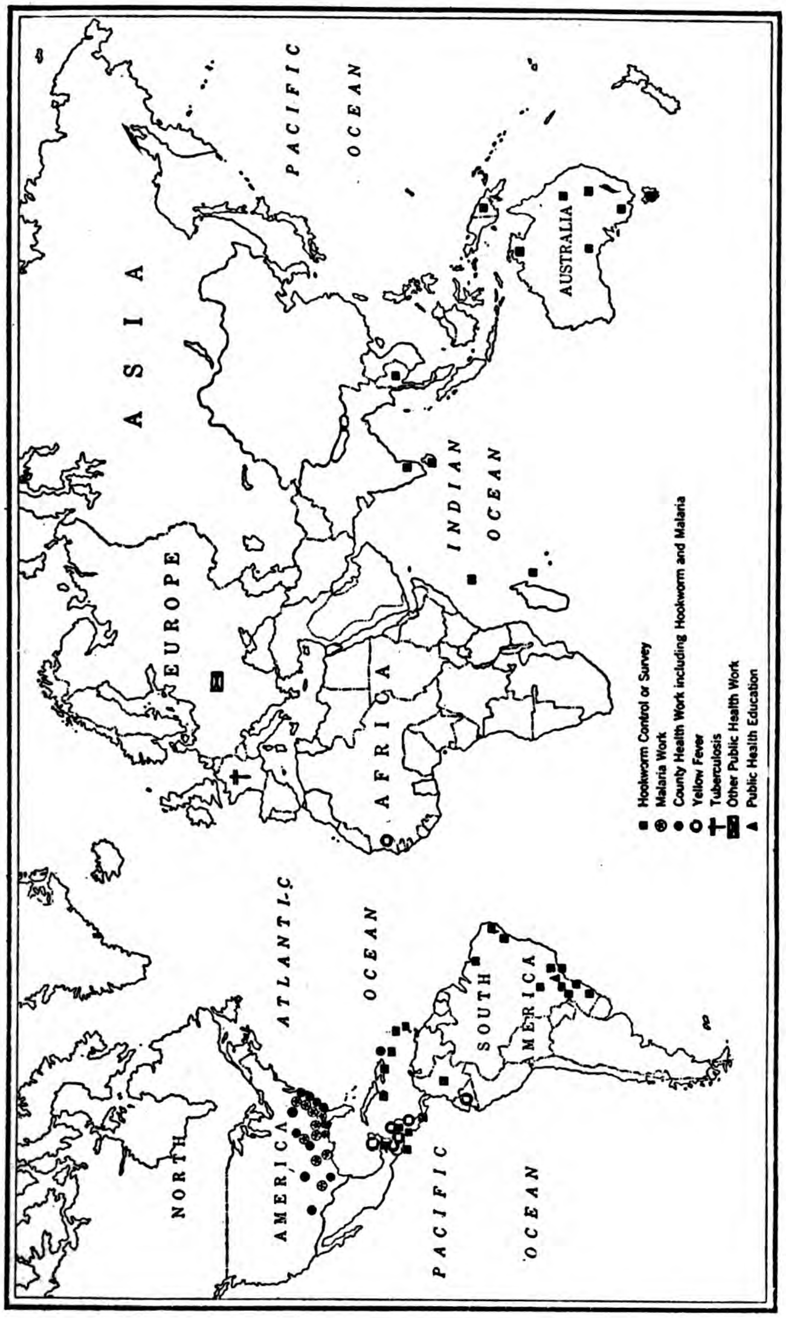


Fig. 35.—World map of activities of International Health Board during 1920. (Reproduced by permission of Mr. Wickliffe Rose from Seventh Ann. Rept. International Health Board, 1921.)

used in epidemiologic work. An example of such a map is given in Fig. 35, showing the distribution of the different sorts of activities of the International Health Board in 1920.

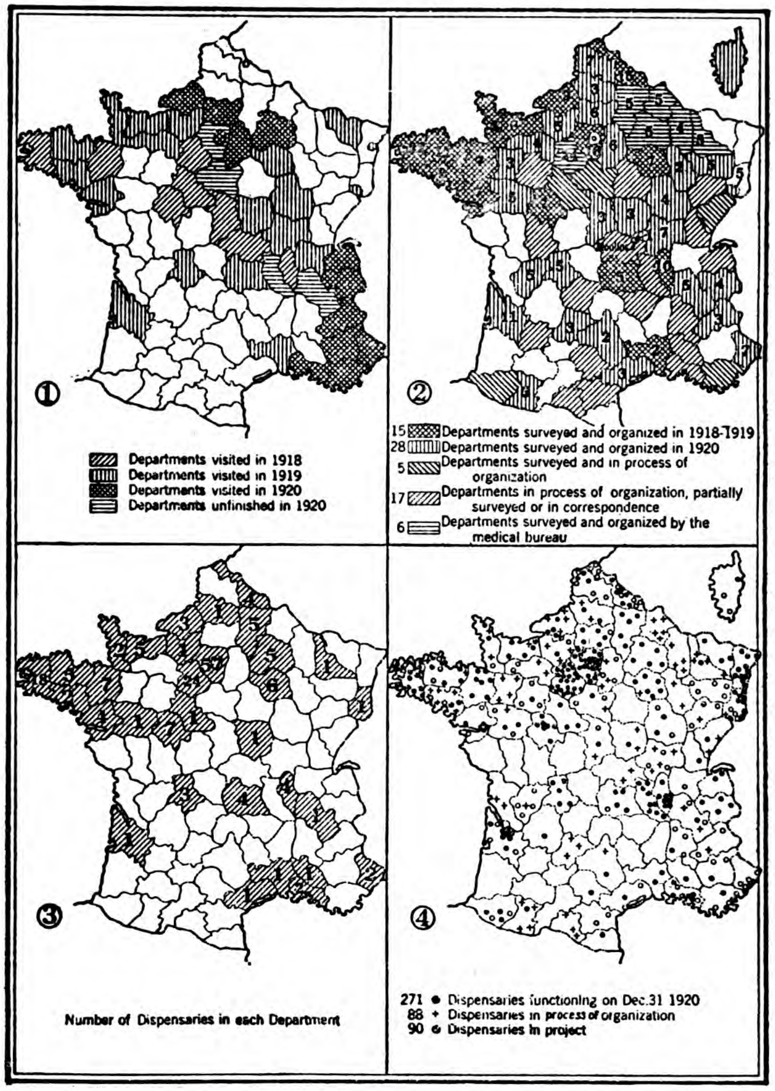

①
▨ Departments visited in 1918
▥ Departments visited in 1919
▩ Departments visited in 1920
▤ Departments unfinished in 1920
②
15 ▩ Departments surveyed and organized in 1918-1919
28 ▥ Departments surveyed and organized in 1920
5 ▧ Departments surveyed and in process of organization
17 ▨ Departments in process of organization, partially surveyed or in correspondence
6 ▤ Departments surveyed and organized by the medical bureau
③
Number of Dispensaries in each Department
④
271 ● Dispensaries functioning on Dec. 31 1920
88 + Dispensaries in process of organization
90 ⊙ Dispensaries in project

Fig. 36.—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.)

Figure 35 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 36 gives examples of such maps.

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. D'Ocagne's[9] own writings are, of course, the final authority, but not particularly adapted to the medical man with a meager equipment of mathematics.

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

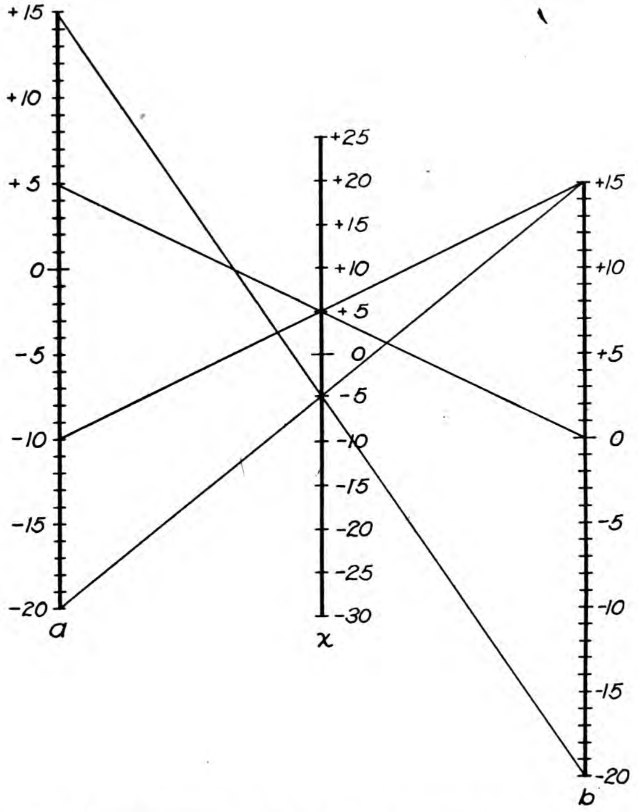


Fig. 37.—Construction of addition nomogram. See text.

between the a and b lines of Fig. 37 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. 37.

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 recently published a nomogram of the DuBois equation

$$S = 71.84\ W^{0.425}\ H^{0.725}$$

This is reproduced as Fig. 38.

The second example is Lawrence J. Henderson's† nomogram relating six variables in the physiology of the blood. It is shown in Fig. 39.

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, [H_2CO_3] and [$BHCO_3$]; 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 upon the oxygen dissociation curve of blood, and of Christiansen,

* 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.

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.

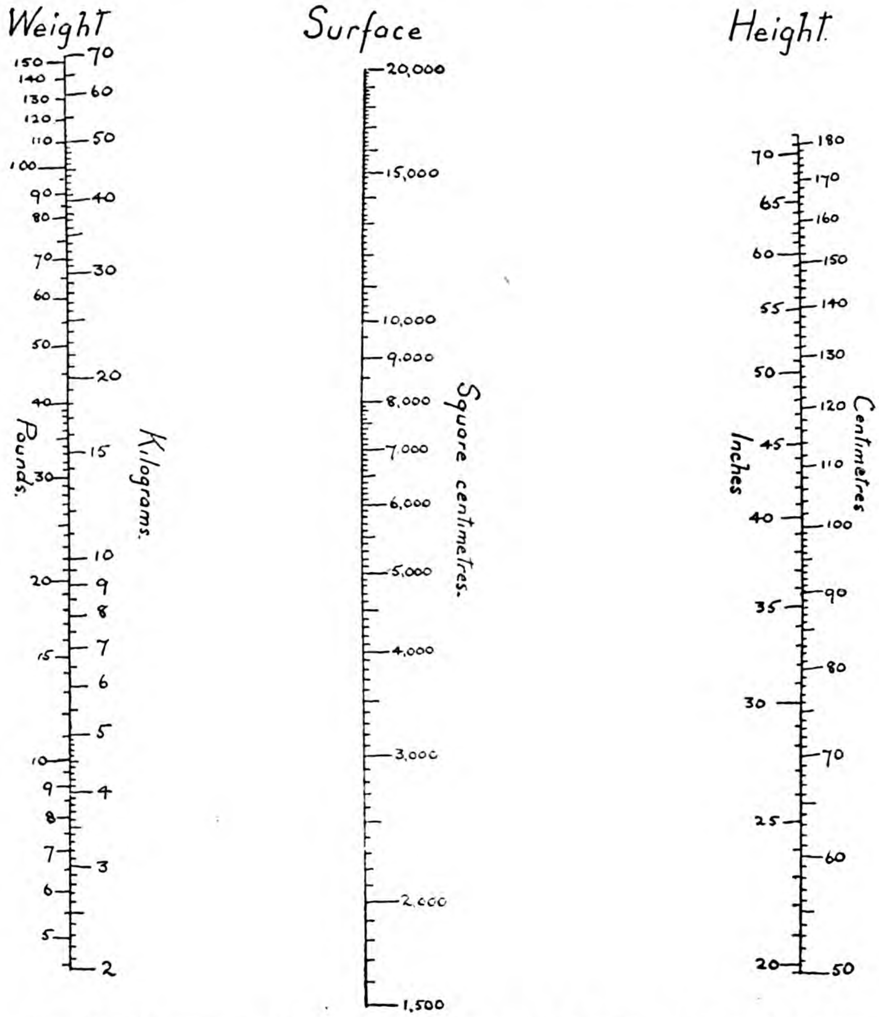


Fig. 38.—Nomogram for $S = 71.84\ W^{0.425}\ H^{0.725}$, where S = surface in sq. cm., W = weight in kg., and H = height in cm. A straight line joining given values of W and H cuts the middle scale at the correct value of S. Thus a line joining the points 24 on the weight scale, with the point 110 on 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.)

Regarding the nomogram Henderson says (loc. cit.):

"The significance of the nomogram is most easily appreciated by considering particular points, for example, A_1 and V_1, which may be taken to represent the cases of arterial and venous blood respectively. The co-ordinates of these points are as follows:

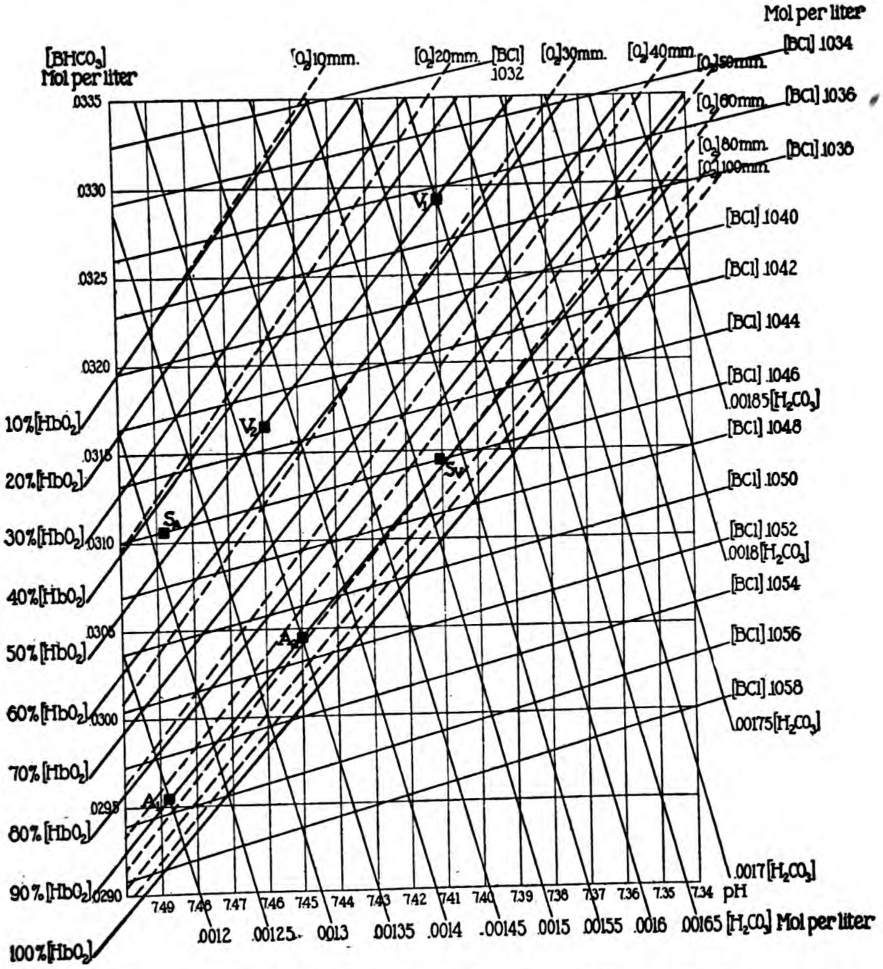


Fig. 39.—Nomogram for certain physicochemical relations of the blood. (From L. J. Henderson.)

	pH.	[O_2] mm.	[HbO_2] per cent.	[H_2CO_3].	[$BHCO_3$].	[BCl].
A_1.....	7.488	56	90	12 × 10^{-4} N	295.4 × 10^{-4} N	1,055.5 × 10^{-4} N
V_1.....	7.411	26	50	16 × 10^{-4} N	329.3 × 10^{-4} N	1,036.5 × 10^{-4} N

Evidently these co-ordinates give a fair representation of the difference between arterial and venous blood.

"The difference between arterial and venous serum bicarbonate concentrations is 33.9×10^{-4} N. This may be contrasted with the difference between the bicarbonate concentration corresponding to two points, S_A and S_V, which both fall upon the co-ordinate $1,046 \times 10^{-4}$ N for [BCl], and which have the co-ordinates for pH, 7.488 and 7.411 respectively, of the points A_1 and V_1. Since S_A and S_V have the same co-ordinate for [BCl], it follows that in passing from one to the other, and in general in passing from any point to any other point with the same value of [BCl], there can be no exchange of electrolyte between serum and corpuscles. Therefore, the increase of [$BHCO_3$] in passing from S_A to S_V is due to the reaction of the constitutents of the serum under the influence of increasing concentration of carbonic acid. In short, for such a case, the serum behaves as if isolated from the corpuscles.

"The co-ordinates of [$BHCO_3$] for S_A and S_V are 310.6×10^{-4} N and 314.4×10^{-4} N respectively. The difference between these concentrations is 3.8×10^{-4} N, in contrast with 33.9×10^{-4} N for the difference between A_1 and V_1. Thus it is evident that the escape of carbonic acid in the lung and its absorption in the tissues must depend chiefly upon the corpuscles. Even in the serum, and presumably therefore in the plasma, much the greater part of the variation in bicarbonate concentration is the result of a heterogeneous reaction with the corpuscles, and only a small amount, if the present estimate is quantitatively correct, approximately 10 per cent., of the loading and unloading of carbonic acid in the serum depends upon a reaction exclusively within the serum.

"It remains to say a word regarding the general significance of the nomogram. Previous investigations have led to the proof of a relationship, in certain cases, between three of the six variables in question. This relationship always has the character of an ordinary algebraic equation in three unknowns, corresponding to a contour line chart where all three variables are determined if definite values are assigned to any two. Such is the case for [$\overset{+}{H}$],

$[H_2CO_3]$, and $[BHCO_3]$; for $[O_2]$, $[HbO_2]$, and $[H_2CO_3]$; or for $[H_2CO_3]$, $[BHCO_3]$, and $[HbO_2]$; and only lack of adequate experimental data has heretofore prevented the establishment of such a relationship between $[BCl]$, $[H_2CO_3]$, and any of the other four variables.

"All these relationships are expressed by the nomogram. But, among six variables, taking three at a time, there are twenty different combinations. Hence, in addition to the three familiar cases above mentioned, the nomogram expresses seventeen other similar relationships."

ELEMENTARY STANDARDS IN GRAPHIC WORK

In 1915 a widely representative joint committee of engineering statistical, economic, biologic, and other societies, interested in the promotion of sound methods of graphic presentation of

1 The general arrangement of a diagram should proceed from left to right.

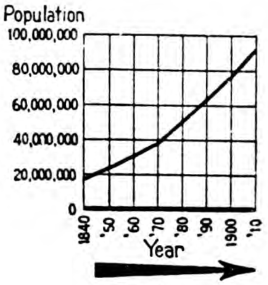


Fig. 1

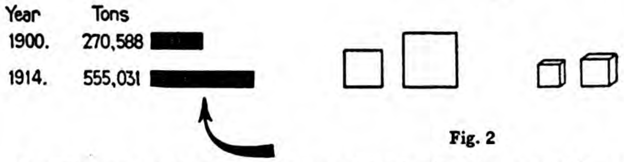


Fig. 2

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.

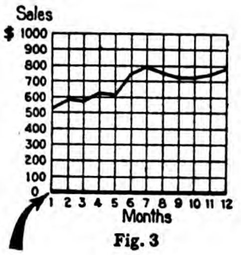


Fig. 3

data, published[10] a preliminary report on standards. This report 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.

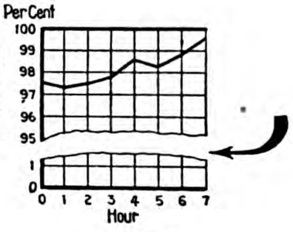


Fig. 4

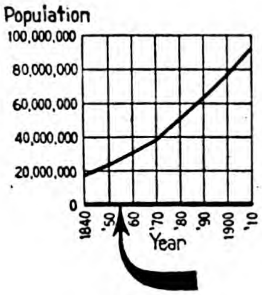


Fig. 5A

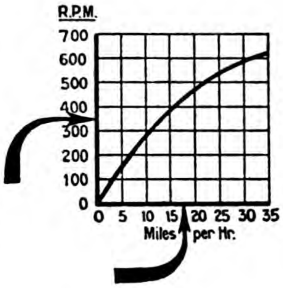


Fig. 5B

5 The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

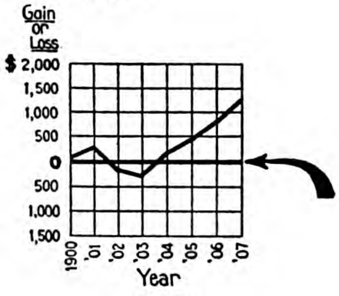


Fig. 5C

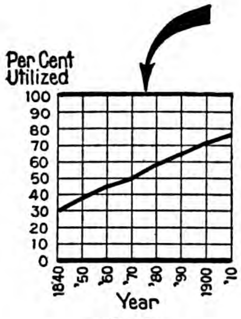


Fig. 6A

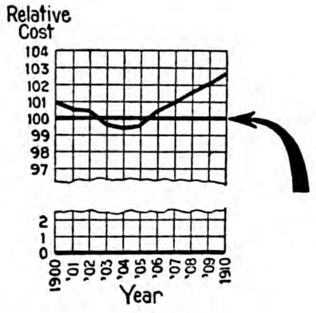


Fig. 6B

6 For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

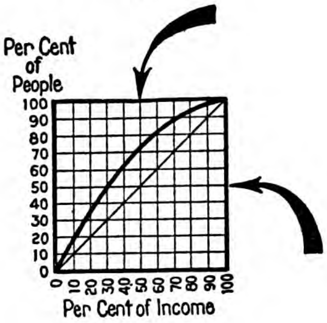


Fig. 6C

7 When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

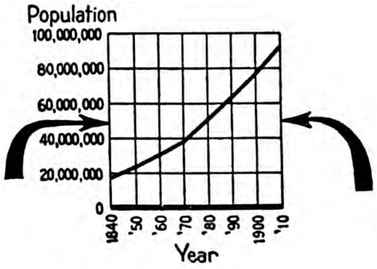


Fig. 7

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.

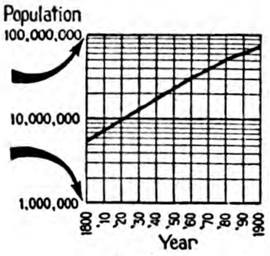


Fig. 8

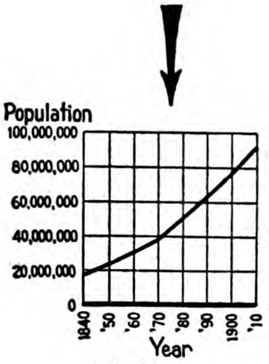


Fig. 9A

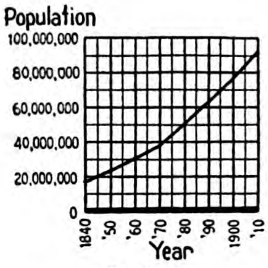


Fig. 9B

9 It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10 The curve lines of a diagram should be sharply distinguished from the ruling.

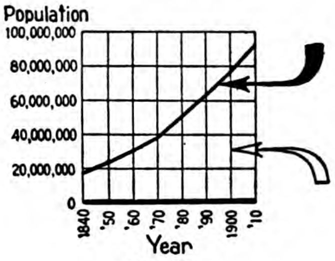


Fig. 10

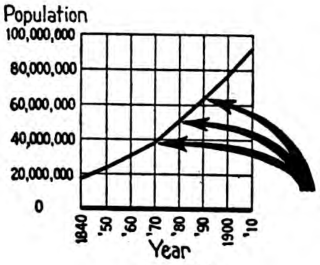


Fig. 11A

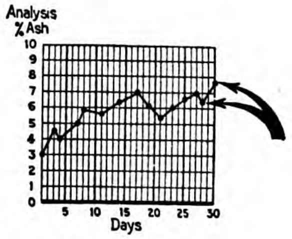


Fig. 11B

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.

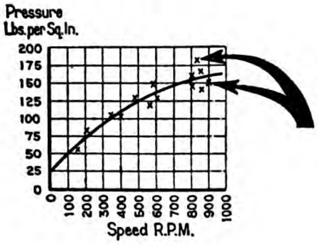


Fig. 11C

12 The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

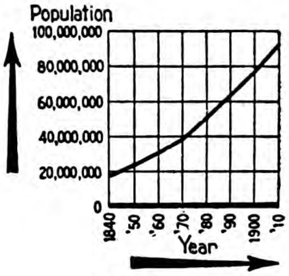


Fig. 12

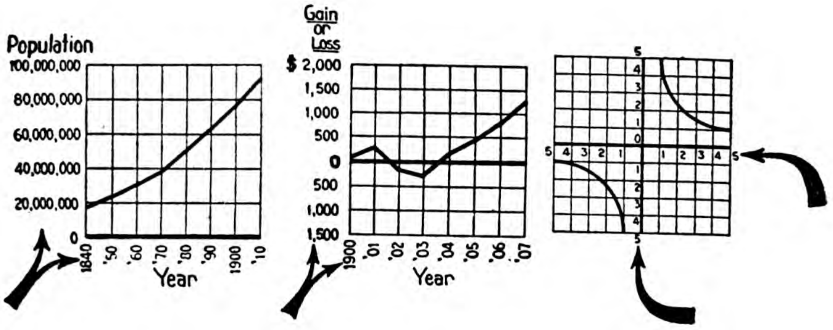


Fig. 13A Fig. 13B Fig. 13C

13 Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

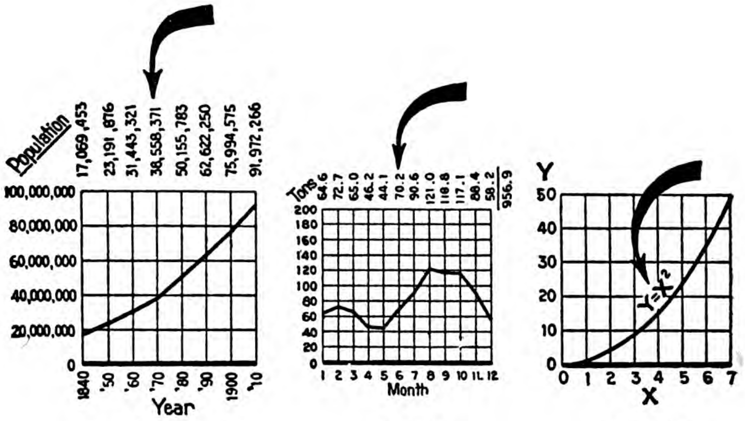
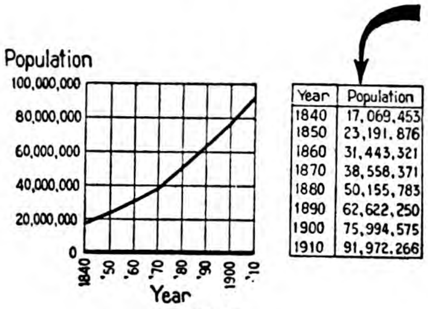


Fig. 14A Fig. 14B Fig. 14C

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.

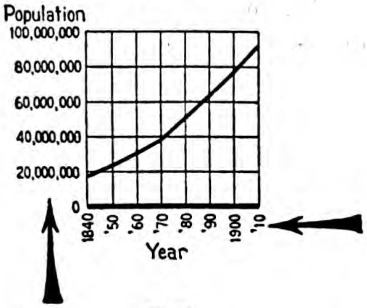
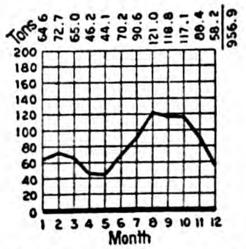


Fig 16

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.

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 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 ÷ 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 ÷ 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 lately 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. A recent example may be cited.* Litchfield and Hardman report excellent results in the treatment of laryngeal diphtheria by suction to remove the membrane. They present a table, here reproduced as Table 12, 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 12

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. of the two modes of treatment, in 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:

$$R = K\left(\frac{a}{a+b}\right), \qquad \text{(i)}$$

which, expressed in words, means

$$\text{Rate} = \text{a constant } (K) \times \left\{ \frac{\text{The number of times the event actually occurs.}}{\text{The whole number of exposures to risk of its occurrence, } i.\ e.\text{, the number of times it actually occurs} + \text{the number of times it might occur, but does not.}} \right\}$$

The part of the right-hand number of the rate equation which is in brackets limits the universe of discourse to which the rate applies *in space.*

A rate is also always 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.

The constant K is generally taken as some power of 10: either 10^2 or 10^3 or 10^4 or 10^5 or 10^6. There is no reason for this except convention. When $K = 10^2$ the rate is per centum; when $K = 10^3$ the rate is per thousand, etc.

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.

$$R_o = K\left(\frac{a}{c+d}\right) \qquad \text{(ii)}$$

where

R_o = a ratio,
K = a constant,
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 and ratios 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 K in rate formulas 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 or ratio measures.

CRUDE DEATH-RATES

Here the fundamental equation (i) becomes

$$R_c = K\left(\frac{D}{P}\right)$$

where

R_c = crude death-rate,
D = deaths from all causes,
P = total population $= D + (P - D) = P$.

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 distribution, 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 while 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

$$R'_c = K\left(\frac{D'}{P}\right)$$

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 the figures of Table 13, which shows the mean age of the living population of Amsterdam, at nine consecutive census periods (1829 to 1909 inclusive), together with the probable errors of these means (the meaning of probable errors will be explained in a later chapter).

It is at once apparent that in this long period the age constitution of the population of Amsterdam has changed but slightly.

TABLE 13

Mean Age of Living Population of Amsterdam at Each of Nine Consecutive Censuses, 1829–1909.

Census years.	Mean age.		
	Male, years.	Female, years.	Both, years.
1829	27.820 ± .045	30.521 ± .041	29.318 ± .030
1839	27.120 ± .042	29.874 ± .040	28.637 ± .029
1849	27.352 ± .040	30.301 ± .038	28.963 ± .028
1859	27.469 ± .039	30.180 ± .037	28.944 ± .027
1869	27.891 ± .038	30.444 ± .037	29.268 ± .027
1879	27.445 ± .035	29.754 ± .034	28.674 ± .024
1889	26.783 ± .030	28.901 ± .030	27.905 ± .021
1899	26.709 ± .027	28.682 ± .026	27.755 ± .019
1909	27.772 ± .025	29.639 ± .025	28.750 ± .018

It has been shown analytically by Lotka[2] that, under certain conditons 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 distribution, a condition which is usually not carried out.

SPECIFIC DEATH-RATES

Here the fundamental equation becomes

$$R_s = K\left(\frac{D_e}{E}\right),$$

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.

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 $R_{as} = K\left(\frac{10}{100}\right)$, which means that the annual death-rate, specific for age and sex (R_{as}) in this community was 0.1 K for males between twelve and thirteen years of age.

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 14 and Fig. 40, in which the rates are specific for quinquennial age groups.

TABLE 14

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

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.

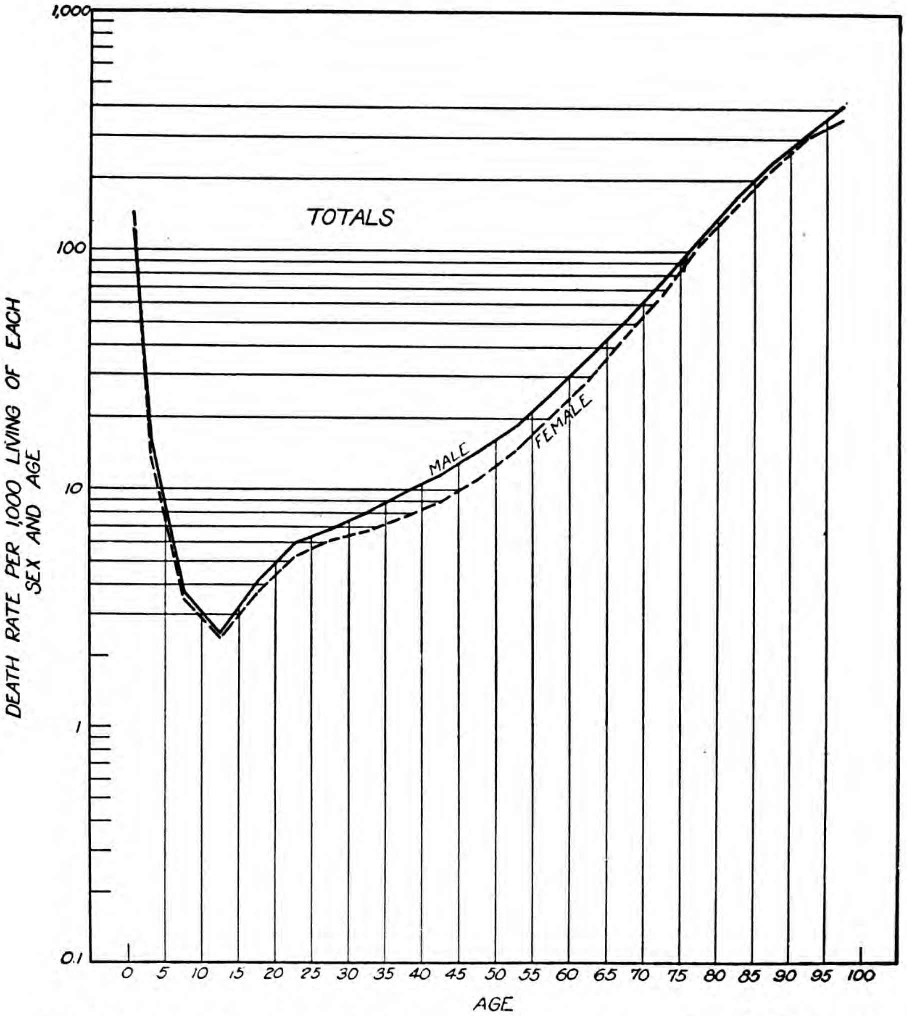


Fig. 40.—Age and sex specific death-rates from all causes for the U. S. Registration Area (exclusive of North Carolina) in 1910. Plotted from data of Table 14, on an arithlog grid.

From that point on it rises steadily, though not entirely evenly till the end of the life span. The specific death-rates are lower 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. In Appendix I there are given age and sex specific death-rates (on a quinquennial age grouping), for each statistically recognized cause of death, in the United States Registration Area (exclusive of North Carolina) in 1910. These tables the reader should study in order to get a general understanding of mortality. 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.
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

$$R_i = K\left(\frac{D_i}{B}\right),$$

where

R_i = infant mortality rate,
D_i = deaths of infants under one year of age,
B = births.

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 15.

It will be noted from this table that there is great variation among the different cities in the rate of infant mortality. This

TABLE 15

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

variation I have 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 country to rural country, 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, and (2) from the racial composition of the several populations, and (3) from fundamentally uncontrollable environmental differences, such as climate, soil, etc. Now with the actually existing condition of variation between different communities in respect of infant mortality, it is obvious that there must be particulate and presumably in large degree determinable reasons for each 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. Operating on a basis largely of empiricism and *a priori* reasoning, efforts to reduce infant mortality have in the past been attempted with considerable success. Also, 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 16

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 presently must needs 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 difficulties 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 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 true. But Greenwood and Brown,[4] in what may fairly be regarded 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 we must 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 we must get an adequate conception of the magnitude and character of the differences themselves.

TABLE 17

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	—	3	1	—	—	—	—	—	—	—	—	—	—	—	1
20– 39	—	1	—	—	—	2	1	—	2	9	33	32	—	—	—	—	1	9	—	—	—	—	2	1
40– 59	2	1	5	1	11	4	12	6	49	45	152	174	1	—	—	1	30	33	—	—	—	1	8	5
60– 79	16	18	22	17	25	27	49	37	130	125	396	342	4	1	6	3	71	63	—	1	1	—	26	22
80– 99	27	24	50	43	44	42	76	65	99	107	316	298	14	13	8	5	66	64	—	1	1	1	29	19
100–119	29	34	45	40	35	39	61	48	52	57	140	165	7	8	8	6	46	38	—	1	3	1	36	42
120–139	13	14	13	27	20	23	24	38	17	21	59	64	1	5	3	4	13	16	3	1	—	4	40	37
140–159	9	5	7	13	11	7	13	15	6	15	18	31	—	—	1	5	3	8	1	2	3	1	28	30
160–179	1	2	1	1	3	5	5	15	1	2	4	11	—	—	—	2	1	2	6	2	1	2	21	19
180–199	1	—	1	2	3	6	4	4	—	—	4	8	—	—	—	—	1	—	5	2	—	—	18	17
200–219	—	—	—	—	—	—	—	6	1	—	1	1	—	—	—	—	—	1	6	5	2	2	5	10
220–239	—	—	—	—	—	—	1	—	—	—	1	—	—	—	—	—	—	—	3	6	5	4	8	12
240–259	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	2	2	3	1	4	8
260–279	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	1	1	4
280–299	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	2	4	2	2
300–319	—	—	—	—	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1	1	2	—	1
320–339	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	1	1	—
340–359	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—
360–379	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
380–399	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	—
400–419	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
420–439	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—
440–459	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
460–479	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1
480–499	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
500–519	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1	1	—	—	—	—
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.

The distribution of variation in infant mortality in cities and rural areas in the United States is shown in Table 17, taken from the paper cited.

The infant mortality rates of various countries are given in Table 18.

TABLE 18

INFANT MORTALITY RATES (DEATHS UNDER ONE YEAR PER 1000 BIRTHS) FOR VARIOUS COUNTRIES. (Rearrangement of Data from Birth Statistics, 1920, p. 40.)

Country and year.	Male.	Female.
Hungary (1915)	281.9	244.6
Russia (1909)	264.9	236.9
Chile (1918)	260.9	248.2
Ceylon (1919)	227.8	217.3
Austria (1913)	204.2	174.6
Japan (1917)	181.8	164.2
German Empire (1914)	177.1	149.2
Prussia (1914)	177.1	150.2
Italy (1916)	174.5	157.7
Jamaica (1919)	167.7	155.4
Bulgaria (1911)	166.1	145.7
Spain (1917)	163.5	146.1
Serbia (1910)	144.7	132.4
Belgium (1912)	132.1	107.2
Uruguay (1920)	124.7	109.5
France (1913)	122.7	101.7
Finland (1918)	122.6	107.5
Scotland (1919)	112.9	89.6
Denmark (1919)	101.3	81.2
United Kingdom (1919)	101.3	79.0
England and Wales (1919)	100.0	77.6
Ireland (1919)	97.3	77.5
Switzerland (1918)	96.9	79.1
United States (registration area, 1920)	95.1	76.1
Australian Commonwealth (1920)	76.7	61.1
Sweden (1916)	76.6	62.5
Norway (1917)	70.6	57.0
The Netherlands (1919)	55.2	43.9
New Zealand (1918)	53.6	43.0

CASE FATALITY RATES

Here the fundamental equation becomes

$$R_F = K\left(\frac{D_c}{C}\right),$$

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.

This is, provided age, sex, race, occupation, and locality of dwelling are taken into account, the most refined form of a 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 example, 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, either favorably or unfavorably, 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

$$R_B = K\left(\frac{B}{P}\right),$$

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.

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 ten and sixty.* 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 19, taken from Knibbs.[5]

* The limits usually taken are 15 and 45, 50 or 55. Actually, however, there are every year 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 small 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 19

CRUDE BIRTH-RATES FOR VARIOUS COUNTRIES—1860–1914—PER 10,000 OF THE POPULATION. (From Knibbs.)

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	227	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	247
1913	282	239	255	228	190	..	..	..	252	231	256	281	..	..	..	246
Mean	354	335	338	243	235	366	357	284	296	287	304	335	296	374	411	

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 in this age group as the numerator of the rate fraction. Such figures measure the *fertility* of women of the specified class. Matthews Duncan 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 20, adapted from Knibbs.[5]

TABLE 20

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

It is to be understood that the figures in Table 20 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 high as compared with 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:

* Births per 1000 married women of indicated age.

† 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.

"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.

MORBIDITY RATES

The fundamental equation for a crude morbidity rate is as follows:

$$R_M = K\left(\frac{M}{P}\right)$$

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.

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.

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

$$Rt_D = K\left(\frac{D'}{D}\right),$$

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.

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 birth-rate, 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 employed in researches on tuberculosis by Greenwood and Tebb,[6] and by Arne Fisher* 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* unhealthy. 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 be from a social standpoint. It is curious, in view of the obvious significance of this constant, the vital index of a population, that so little attention is paid to it by demographers. After much study of it I am convinced that no single figure gives so sensitive a measure of the vitality of a

* Fisher, A.: On the Construction of Mortality Tables by Means of Compound Frequency Curves, Scandinavian Insurance Magazine, 1920, *passim*.

nation or any subgroup of people as this does. There appears to have been no adequate general discussion of it since that of Wernicke[*] in 1889, and even he does 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. Recently Pell[§] has dealt with the idea implicit in the birth/death ratio, but in a most inadequate manner.

In Table 21 are shown four vital indices for urbàn, rural, and 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.

$$\text{Vital index } D = \frac{100 \text{ (births of whites)}}{\text{Deaths of whites}}$$

[*] 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 Förenhingens 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.

TABLE 21

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	86.0	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	103.6	70.6	150.0	116.2	72.3	65.6	114.7
	Rural	*	*	*	*	*	*	*	*	223.7	58.1	74.5	208.2	190.4	50.5	67.4	177.9
	Total	*	*	*	*	*	*	*	*	207.3	68.6	72.2	195.3	171.9	50.1	66.4	162.2
Kentucky	Cities	*	*	*	*	*	*	*	*	135.8	26.6	47.4	123.1	105.6	25.1	34.7	98.5
	Rural	*	*	*	*	*	*	*	*	241.4	35.9	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	*35.0*	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	*150.0*	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	*20.0*	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	128.8	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

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.

TABLE 21—*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	160.2	66.1	39.3
	Rural	*	*	*	*	*	*	*	*	157.5	113.3	73.8	156.3	138.2	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*	186.9
	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.

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 no 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 21. 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 novel points of interest. We may first 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 first we note that for each grouping and each year Index B is much larger than Index A. Except for the rural population B is more than twice as large as A. Generally speaking the foreign population produces in this country approximately two babies for every death. The native population (as defined in Vital Index A) produces only a small fraction over one baby for each death. In other words, the native population, even when so broadly defined as by Index A, is in about the same state as France before the war, and not in as vigorous a state as the French population is now.

The vital indices of Table 21 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 reproduction. 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 (and cannot be because of lack of published data). But it would be an age-specific vital index for the females as reproductive units.

TABLE 22

AGE SPECIFIC VITAL INDICES FOR NATIVE-BORN AND FOREIGN-BORN WOMEN IN B. R. A. 1919

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 results of exactly such an experiment for the women of the Birth Registration Area in the year 1919 are shown in Table 22.

The figures in Table 22 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. 41.

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

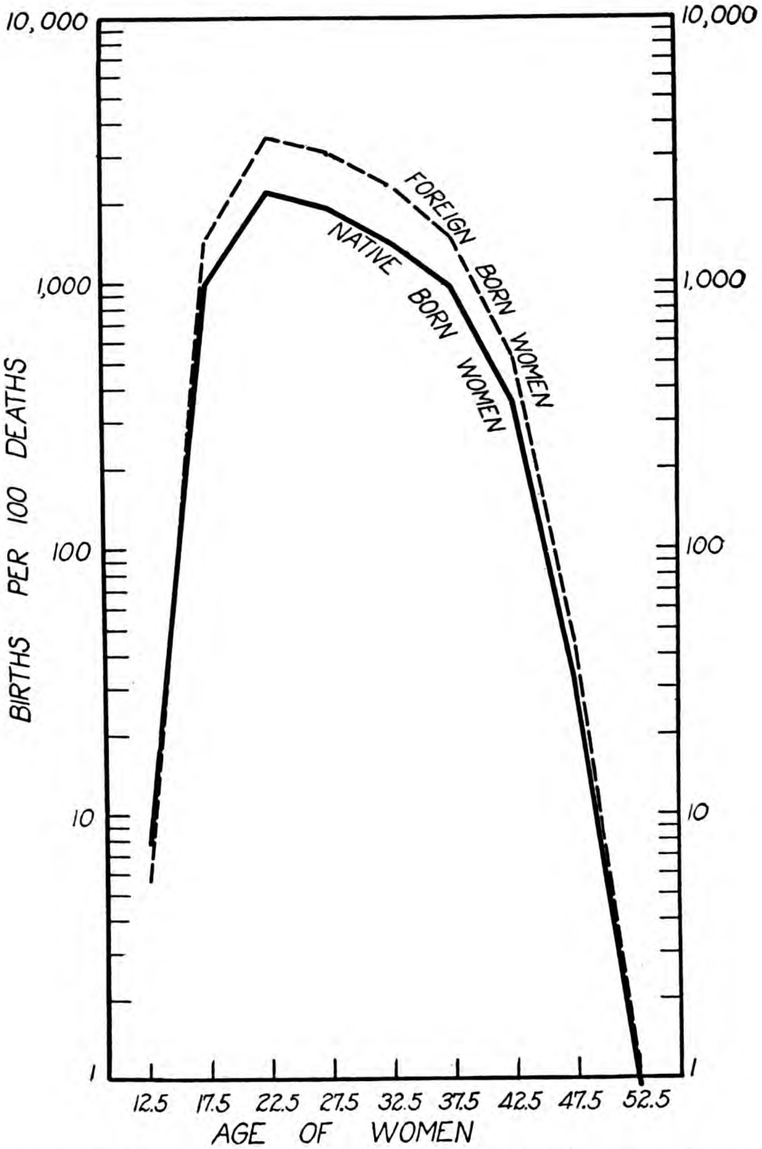


Fig. 41.—Showing the differences in specific vital indices for native-born and foreign-born women in 1919. Solid line, native-born women; dash line, foreign-born women.

for the race, "best" being taken here in the sense of biologic efficiency and economy.

So far as I am aware no attempt had been made before this work[7] to calculate age-specific vital indices. They picture, as exhibited in Table 22 and Fig. 41, an extremely interesting biologic fact. If we had such 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 22 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 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 22 would be about one-half as large absolutely as they really are in that table, but the general *form* of the curve of Fig. 41 would be unchanged.

For further discussion of vital indices see Pearl and Burger,[8] Pearl* and Miner.†

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 Registration 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 suggestive ideas and results. The present writer disagrees fundamentally with

* 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.

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.)

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. I do not intend 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,*

$$q_1 = K \frac{d_1}{100{,}000}.$$

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,

$$q_2 = K \frac{d_2}{(100{,}000 - d_1)}$$

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		100,000	
1–2			
etc.			

* 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 I am trying to do is 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 $\overset{\circ}{e}_x$ 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 approximately 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 concrete realization of what a life table looks like, Table 23 and Figs. 42, 43, and 44 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 23

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	81,116
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 23 (*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 23 (*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	$1000q_x$	$\mathring{e}_x$	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 mortaity 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. 42) shows the form of the life table

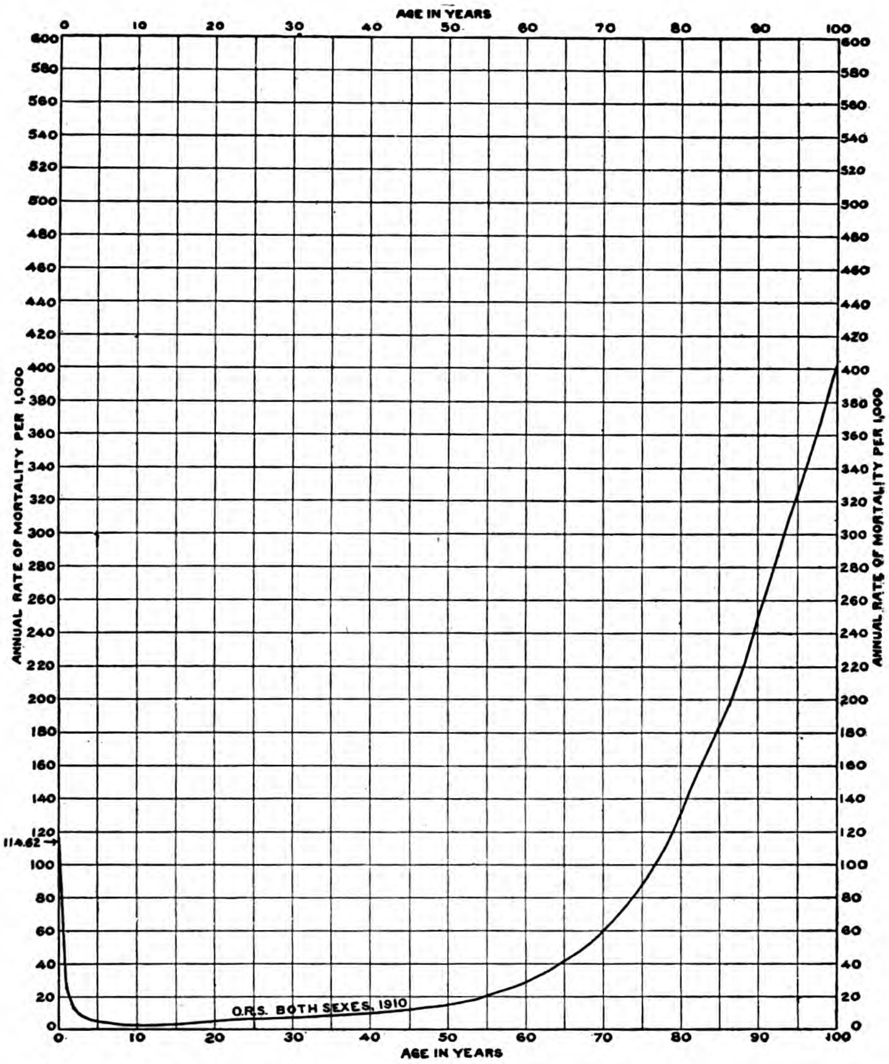


Fig. 42.—Annual mortality rate per thousand. The original registration states, both sexes, 1910 (from Glover,[1] p. 243).

specific death-rate curve (q_x), being the plot of this column of Table 23 above.

The next diagram (Fig. 43) 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.

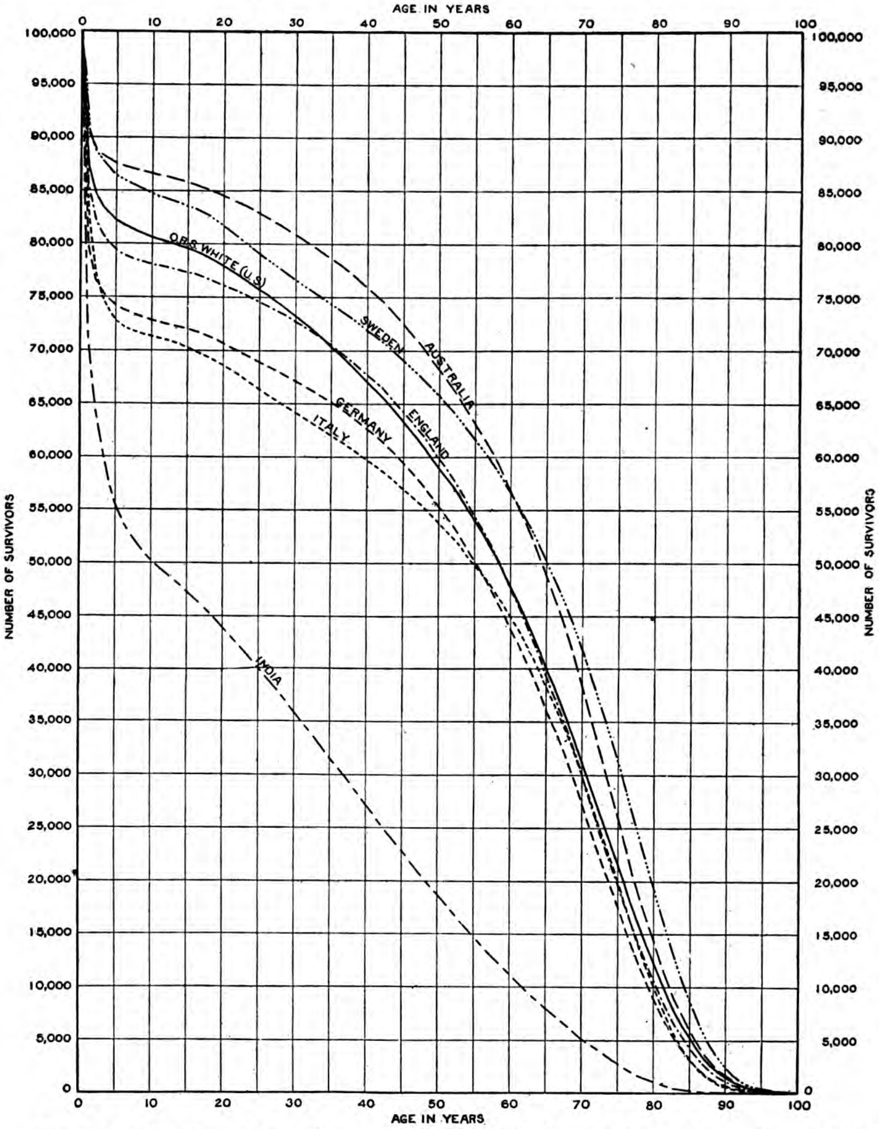


Fig. 43.—Number of survivors out of 100,000 born alive. Australia, England, Germany, India, Italy, Sweden, and whites in the original registration states. Males, 1901–10 (from Glover,[1] p. 260).

The next diagram (Fig. 44) shows the form of the d_x curve, and again the life tables of several countries are drawn upon for comparison.

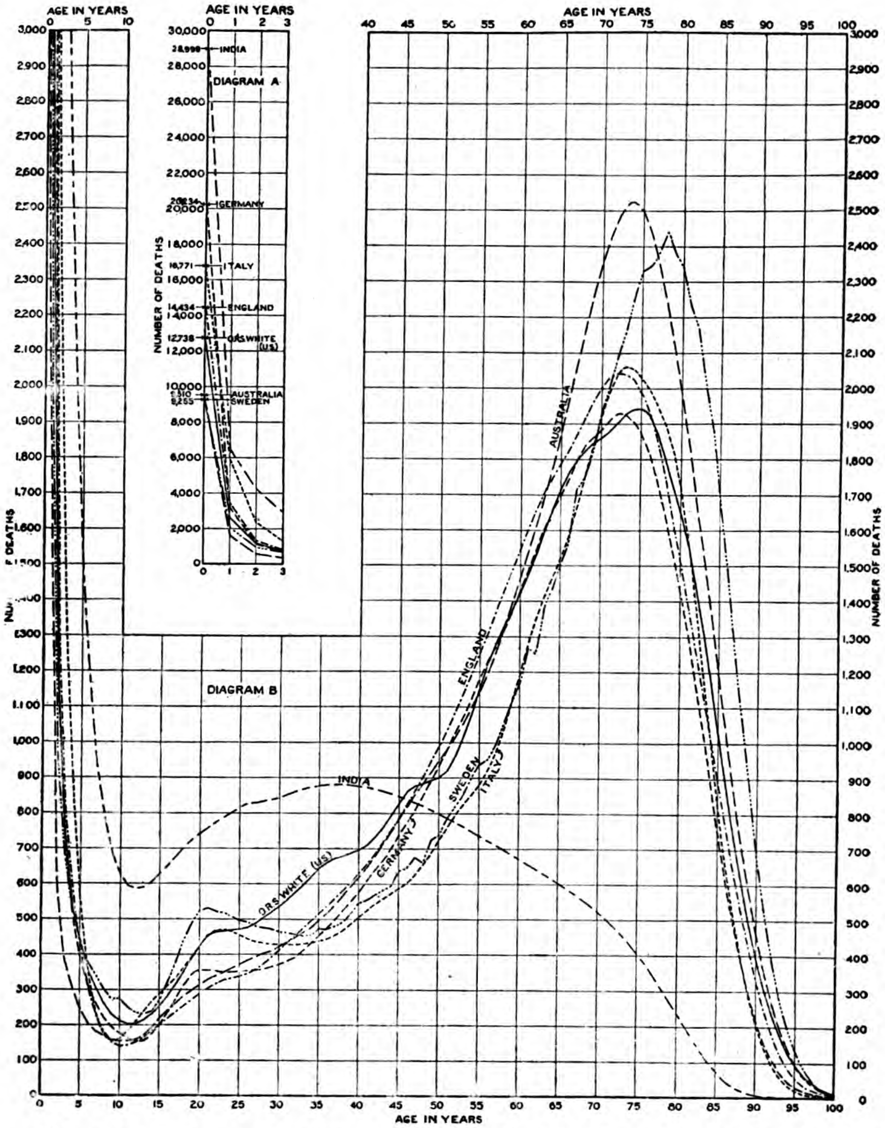


Fig. 44.—Number of deaths out of 100,000 born alive. Australia, Germany, England, India, Italy, Sweden, and whites in the original registration states. Males, 1901–10 (from Glover,[1] p. 270).

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

TABLE 24

LIFE TABLE FOR DROSOPHILA—LONG-WINGED MALES

Age in Days	Observed		Calculated			Age in Days	Observed		Calculated		
	d_x	l_x	l_x	q_x	e_x		d_x	l_x	l_x	q_x	e_x
1........	5	1,000	1,000	9.6	41.0	50........	15	363	368	45.8	14.2
2........	12	995	990	9.7	40.4	51........	20	348	351	47.7	13.8
3........	6	983	981	9.7	39.7	52........	19	328	334	49.6	13.5
4........	13	977	971	9.7	39.1	53........	22	309	318	51.6	13.1
5........	10	964	962	9.9	38.5	54........	16	287	301	53.7	12.8
6........	10	954	952	10.0	37.9	55........	13	271	285	55.7	12.4
7........	15	944	943	10.1	37.2	56........	19	258	269	57.9	12.1
8........	9	929	933	10.3	36.6	57........	12	239	254	60.2	11.8
9........	9	920	924	10.4	36.0	58........	19	227	238	62.5	11.5
10........	9	911	914	10.6	35.4	59........	13	208	224	64.8	11.2
11........	12	902	904	10.8	34.7	60........	12	195	209	67.3	10.9
12........	8	890	895	11.0	34.1	61........	18	183	195	69.8	10.6
13........	8	882	885	11.3	33.5	62........	8	165	181	72.4	10.3
14........	11	874	875	11.6	32.8	63........	13	157	168	75.2	10.1
15........	8	863	865	12.0	32.2	64........	12	144	156	77.9	9.8
16........	14	855	854	12.2	31.6	65........	13	132	143	80.8	9.5
17........	8	841	844	12.6	31.0	66........	14	119	132	83.6	9.3
18........	13	833	833	13.0	30.3	67........	7	105	121	86.7	9.0
19........	10	820	822	13.4	29.7	68........	8	98	110	89.8	8.8
20........	11	810	811	13.9	29.1	69........	5	90	100	92.9	8.6
21........	16	799	800	14.4	28.5	70........	8	85	91	96.1	8.4
22........	6	783	789	14.9	27.9	71........	5	77	82	99.6	8.1
23........	13	777	777	15.4	27.3	72........	7	72	74	102.9	7.9
24........	11	764	765	16.0	26.7	73........	8	65	67	106.4	7.7
25........	11	753	753	16.6	26.2	74........	3	57	59	110.0	7.5
26........	10	742	740	17.3	25.6	75........	9	54	53	113.8	7.3
27........	10	732	727	17.9	25.0	76........	2	45	47	117.3	7.1
28........	14	722	714	18.7	24.4	77........	8	43	41	121.5	6.9
29........	11	708	701	19.4	23.9	78........	4	35	36	125.4	6.8
30........	15	697	687	20.2	23.3	79........	4	31	32	129.2	6.6
31........	13	682	673	21.1	22.8	80........	3	27	28	133.6	6.4
32........	11	669	659	21.9	22.3	81........	3	24	24	137.5	6.3
33........	15	658	645	22.9	21.8	82........	4	21	21	142.0	6.1
34........	7	643	630	23.8	21.2	83........	2	17	18	146.4	5.9
35........	18	636	615	24.8	20.7	84........	1	15	15	151.0	5.8
36........	15	618	600	25.8	20.2	85........	2	14	13	156.2	5.7
37........	19	603	584	26.9	19.7	86........	2	12	11	160.2	5.5
38........	13	584	569	28.1	19.3	87........	1	10	9	164.5	5.4
39........	22	571	553	29.3	18.8	88........	2	9	8	170.6	5.2
40........	15	549	536	30.5	18.3	89........	2	7	6	175.6	5.1
41........	13	534	520	31.8	17.9	90........	0	5	5	180.4	5.0
42........	23	521	503	33.2	17.4	91........	1	5	4	185.0	4.9
43........	19	498	487	34.5	17.0	92........	1	4	3	189.7	4.8
44........	22	479	470	36.0	16.6	93........	1	3	3	196.3	4.6
45........	18	457	453	37.5	16.1	94........	0	2	2	201.8	4.5
46........	22	439	436	39.0	15.7	95........	1	2	2	207.5	4.4
47........	19	417	419	40.7	15.3	96........	0	1	1	212.8	4.3
48........	15	398	402	42.3	14.9	97........	0	1	1	218.3	4.2
49........	20	383	385	44.0	14.6						

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 24 and 25, and Figs. 45 and 46. These life tables were worked out in the author's laboratory.[4] The l_x curves in the diagrams 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.

TABLE 25

LIFE TABLE FOR DROSOPHILA—SHORT-WINGED MALES

Age in Days	Observed		Calculated			Age in Days	Observed		Calculated		
	d_x	l_x	l_x	q_x	e_x		d_x	l_x	'l_x	q_x	e_x
1........	6	1,000	1,000	25.6	14.2	24........	16	151	155	107.5	8.0
2........	27	994	974	31.5	13.6	25........	13	135	139	107.3	7.8
3........	30	967	944	37.5	13.0	26........	11	122	124	107.1	7.7
4........	34	937	908	43.3	12.4	27........	6	111	111	107.2	7.5
5........	38	903	869	48.9	12.0	28........	20	105	99	107.3	7.2
6........	36	865	826	54.6	11.5	29........	8	85	88	107.9	7.0
7........	85	829	781	60.0	11.1	30........	7	77	79	109.2	6.7
8........	66	744	734	65.5	10.8	31........	13	70	70	111.1	6.4
9........	55	678	686	70.4	10.5	32........	7	57	62	114.5	6.1
10........	52	623	638	75.2	10.2	33........	9	50	55	119.3	5.8
11........	44	571	590	79.8	9.9	34........	5	41	49	126.2	5.4
12........	48	527	543	84.0	9.7	35........	4	36	42	135.4	5.0
13........	21	479	497	87.9	9.5	36........	6	32	37	147.4	4.7
14........	49	458	454	91.5	9.3	37........	4	26	31	162.7	4.3
15........	53	409	412	94.8	9.1	38........	1	22	26	182.2	3.9
16........	43	356	373	97.7	9.0	39........	6	21	21	206.2	3.6
17........	24	313	337	100.2	8.9	40........	2	15	17	234.7	3.3
18........	28	289	303	102.1	8.7	41........	4	13	13	268.6	3.0
19........	22	261	272	104.1	8.6	42........	1	9	10	308.0	2.7
20........	19	239	244	105.3	8.5	43........	2	8	7	352.9	2.4
21........	24	220	218	106.3	8.4	44........	5	6	4	403.0	2.2
22........	17	196	195	106.8	8.3	45........	0	1	3	457.9	2.0
23........	28	179	174	107.2	8.1	46........	0	1	1	516.0	1.8

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?

In a recent paper Pearl[5] has suggested what appears to be a valid method of dealing with this difficulty, in making a comparison of the mortality of *Drosophila* with that of man. The nature of the solution is indicated in the following quotation from that paper:

"Upon what basis shall any life table function, say l_x, of the *Drosophila* life table be compared with that of man? The life span of one of these organisms is best measured in years; that of the other in days. This fact, however, offers no insuperable difficulty to the comparison. What is needed is to superimpose the two curves so that at least two *biologically equivalent points* coincide. The best two points would be the beginning and the end of the life span. But in the case of *Drosophila* our life tables start with the beginning of *imaginal* life only. The larval and pupal durations are omitted.

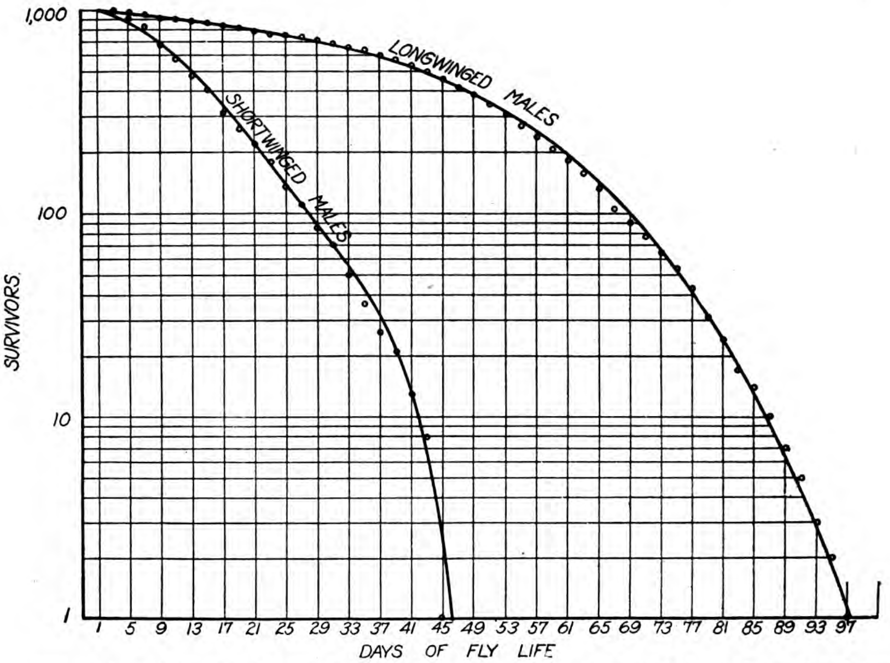


Fig. 45.—Diagram showing the observed and graduated l_x points for long-winged (normal wings) and short-winged (vestigial wings) males of *Drosophila*. The small circles are the observations and the smooth lines the fitted curves. In order not to overcrowd the diagram only every second observation is shown.

"I think we can get at this starting point . . . by putting the human and *Drosophila* l_x curves together as a starting point at the age for each organism *where the instantaneous death-rate* q_x *is a minimum.* In the case of *Drosophila* I think we are safe in concluding, on the basis of the work of Loeb and Northrop as well as from our own observations, that this point is at or very near the beginning of imaginal life. We shall accordingly take *Drosophila* age one day as this point. Our life tables show that certainly *after* this time q_x never again has so low a value.

"For the other end of the life span we may conveniently take the age at which there is left but *one* survivor out of 1000 starting at age one day for *Drosophila* and age twelve years for white males."

When the above was written we were aware of the existence of complete life tables for only the two organisms, *Drosophila* and man. Since then Dr. Carl R. Doering and the present writer have published[6] a life table for a third form, the rotifer *Proales decipiens*, on the basis of data as to its mortality recently furnished by Dr. Bessie Noyes.*

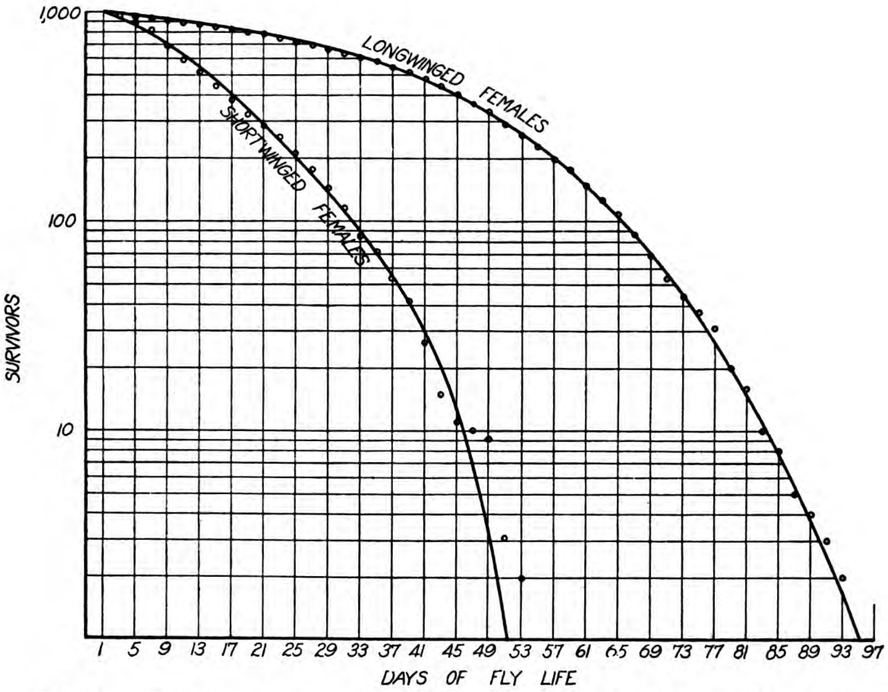


Fig. 46.—Diagram showing the observed and graduated l_x points for long-winged (normal wings) and short-winged (vestigial wings) females of *Drosophila*. The small circles are the observations and the smooth lines the fitted curves. In order not to overcrowd the diagram only every second observation is shown.

Miss Noyes provides in her paper, in two different but apparently homogeneous series, data on the life history of 1454 individuals. The observations were taken only once in twenty-four hours, an interval far too long to give a smooth curve for an animal having a maximum total life span of only about eight days. This fact makes the construction of a life table more difficult and much less accurate than if the observations had been more closely spaced.

* Noyes, B.: Experimental Studies on the Life History of a Rotifer Reproducing Parthenogenetically (*Proales decipiens*), Jour. Exp. Zoöl., vol. 35, pp. 225–255, 1922.

It is as though one tried to construct a life table for man from data as to age at death recorded only to the nearest decade.

Taking the data as they stand, however, the central death-rates were computed and graduated with the results shown in Table 26.

TABLE 26

OBSERVED AND CALCULATED q_x VALUES FROM NOYES' DATA ON PROALES

Days of life.	Observed death-rate (per 1000) within interval.	Calculated q_x.	Calculated l_x (number living at beginning of each interval).
0–0.9	0	.06	1000.0
1.0–1.9	1.4	1.39	999.9
2.0–2.9	9.6	9.99	998.5
3.0–3.9	47.3	44.98	988.5
4.0–4.9	136.5	144.60	944.0
5.0–5.9	393.9	349.90	807.5
6.0–6.9	575.9	653.50	525.0
7.0–7.9	1000.0	956.10	181.9
8.0–8.9	......	.....	8.0
9.0–9.9	......	.....	0

The next step was to calculate a *Proales* life table in terms of centiles of the life span rather than in absolute age. This was done.

In order that it may be seen how the forces of mortality operate in *Proales* as compared with man and *Drosophila*, the diagram shown as Fig. 47 is presented.

Comparing the three curves, we note the following points:

1. The *Proales* curve lies above the other two at all parts of the comparable life span. This means that out of 1000 individuals starting together at biologically equivalent points in the life span (*i. e.*, at the age when q_x is a minimum for each organism) at any subsequent age centile there will be more surviving rotifers than men, and more surviving men than flies.

2. The median durations of life, or, put in another way, the ages prior to which just 500 of the 1000 individuals starting together will have died, are approximately:

For *Proales*,	74.0 per cent. of the equivalent life span.
For Man,	62.0 per cent. of the equivalent life span.
For *Drosophila*,	42.5 per cent. of the equivalent life span.

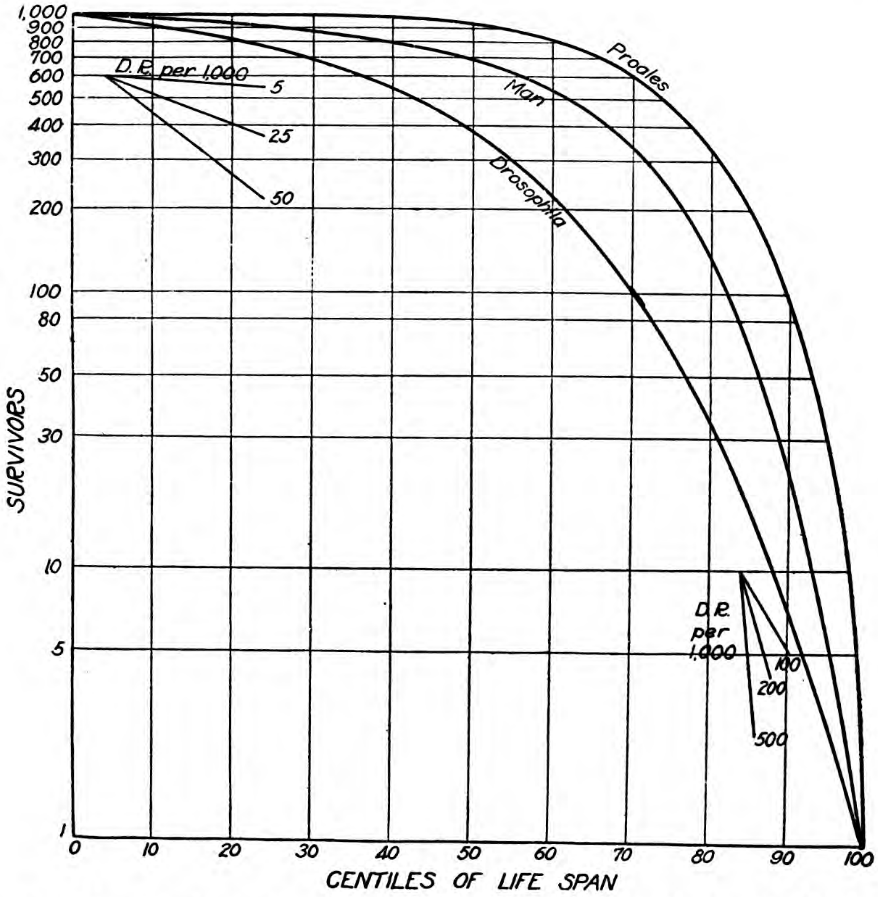


Fig. 47.—Showing survivorship distributions for (*a*) the rotifer *Proales decipiens*, (*b*) man (males in original registration states, 1910), and (*c*) *Drosophila melanogaster* (wild type males). The death-rates corresponding to given slopes of the l_x line are also given by the groups of fine lines at the two ends of the diagram. Age is measured in each of the three organisms in terms of centiles of the equivalent life span.

3. The comparison the other way about indicates that when 50 per cent. of the equivalent life spans have been passed there are still surviving:

In *Proales*,	93.0 per cent. of the individuals starting.
In Man,	68.5 per cent. of the individuals starting.
In *Drosophila*,	38.0 per cent. of the individuals starting.

The outstanding thing about the life curve for *Proales* is that it approaches nearer to the theoretically possible right-angled form in which all the individuals live to a given age *x* and then all die at once, than any other that has yet been observed. Whether this is the result of (*a*) the greater uniformity of environment, on the average, for the *Proales* under the experimental conditions than for the other forms, or (*b*) the greater uniformity of the population in genetic constitution, consequent upon the fact that *Proales* reproduces parthenogenetically and that all of the cultures were descended from at most not over six different individuals, or (*c*) a combination of both, cannot be definitely stated. Both of the factors mentioned undoubtedly do in some degree operate to produce the form of life curve exhibited.

There is need for data regarding the mortality of other organisms. It is an interesting commentary on the development of biology that the distribution of mortality in respect of age is known for only three species of animal life with sufficient accuracy to permit the formation of age-specific death-rates, and hence of a life table. Into every discussion of the problem of evolution, and into every attempt to determine its causes, there must necessarily enter the question of the mortality of the forms being dealt with. There seems no good reason for indefinitely continuing to handle the matter by the current methods, which are either to make large *a priori* guesses about the distribution of mortality in the particular case, or to assume that it is the same as that of man. In the nearly universal neglect of the problem of mortality and duration of life, biologists have missed an interesting and obviously important field.

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 persons living. The necessary computations have been done for three sizes of age classes and the results are presented in Tables 27, 28, and 29, on the basis of the L_x data of Table 23 above. This then is the population derived from the life table for the original registration states in 1910, both sexes together.

TABLE 27

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
33–34	14,367	68–69	7,139	103–104	1
34–35	14,259	69–70	6,760	104–105	1

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 30 (p. 196). Table 30 exactly corresponds to Table 28 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.

TABLE 28

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 29

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

Figure 48 compares the life table standard million (from Table 28) with the standard million of the actual population.

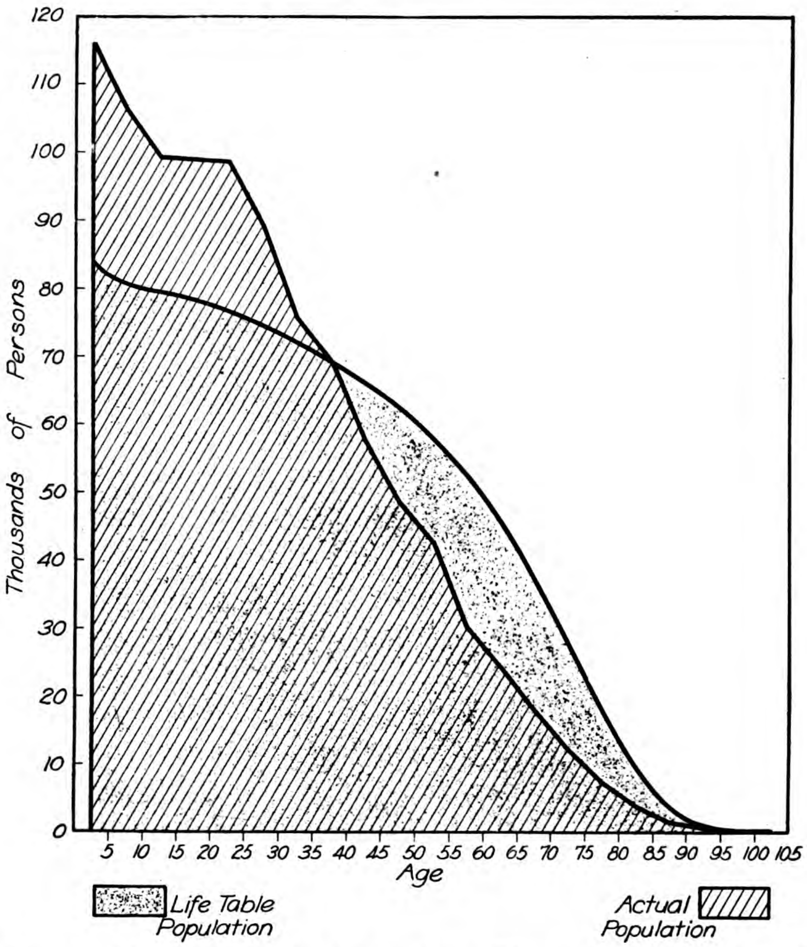


Fig. 48.—Diagram comparing the standard million of (*a*) the life table stationary population (stippled area), and (*b*) the actual population (cross-hatched area); both for the year 1910, and for both sexes together. (Data of Tables 28 and 30.)

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 30

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

I have already emphasized the fact that I do not intend 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 are discussed by Hayward.[3]

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. Like other really valuable government documents it will probably soon be allowed to go out of print.)
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 Parker, S. L.: Experimental Studies on the Duration of Life. I. Introductory Discussion of the Duration of Life in *Drosophila*, Amer. Nat., vol. 55, pp. 481–509, 1921.
5. Pearl, R.: Experimental Studies on the Duration of Life. VI. A Comparison of the Laws of Mortality in *Drosophila* and in Man, Amer. Nat., vol. 56, pp. 398–405, 1922.
6. 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.

CHAPTER IX

STANDARDIZED AND CORRECTED DEATH-RATES

It has been seen in Chapter VII (Table 14 and Fig. 40) 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:

$$R_{St} = K \frac{\Sigma\,(P_x \times q_x)}{\Sigma\,P_x}$$

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.

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 31.

TABLE 31

LIFE TABLE DEATH-RATES, FROM TABLE 23 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 32.

TABLE 32

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 33.

TABLE 33

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 = $(P_x \times q_x)$	2951.31 = $(P_x \times q_x)$

Hence

$$\text{For Providence } R_{St} = 1000\left(\frac{3247.96}{224.050}\right) = 14.50$$

$$\text{For Seattle } R_{St} = 1000\left(\frac{2951.31}{234.719}\right) = 12.57$$

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 31 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,

$17.66 \times 1.339 = 23.65$ = 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

$10.05 \times 1.545 = 15.53$ = 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,

$$R_{Co} = K \frac{\Sigma\ (L_x \times R_{sx})}{\Sigma\ (L_x)}$$

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.

Coming back to the Providence-Seattle example we have already had given in Table 32 the populations of these two cities at ages. Table 34 gives the deaths at ages in columns (1) and (2). By dividing each figure in column (1) of Table 34 by the corresponding population figure of Table 32, we shall get the specific death-rates of Providence set down in column (3), and similarly for Seattle in column (4).

TABLE 34

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	$\frac{1175}{21.814} = 53.86$	$\frac{453}{17.043} = 26.58$
5– 9	74	50	$\frac{74}{18.707} = 3.96$	$\frac{50}{15.123} = 3.31$
10–19	144	107	$\frac{144}{38.315} = 3.76$	$\frac{107}{32.666} = 3.28$
20–39	596	623	$\frac{596}{83.563} = 7.13$	$\frac{623}{109.340} = 5.70$
40–59	854	625	$\frac{854}{46.482} = 18.37$	$\frac{625}{49.817} = 12.55$
60–79	954	447	$\frac{954}{14.111} = 67.61$	$\frac{447}{10.140} = 44.08$
80 and over	182	103	$\frac{182}{1.058} = 172.02$	$\frac{103}{.590} = 174.58$
Totals	3979	2408		

The next step is to multiply the appropriate standard population figures derived from Tables 27, 28, and 29 of the preceding chapter by the specific death-rates of Table 34 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 35.

TABLE 35

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:

$$\text{For Providence: } R_{C_0} = 1000\,\frac{23{,}100}{1{,}000{,}000} = 23.10$$

$$\text{For Seattle: } R_{C_0} = 1000\,\frac{15{,}850}{1{,}000{,}000} = 15.85$$

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 35 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 moieties 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 he can, 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 36 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 35 above.

TABLE 36

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 36 as a standard. The result is that shown in Table 37.

TABLE 37

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 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, as (*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, he will not 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.

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.)

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 the fact particularly 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 38, 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 38

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 I take five names in order as they are given and determine the average age at which these five persons married, and then take the next five names in order and do 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 39 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 39. 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, I decided to take a second fifty names from another part of the alphabet. So I opened the book 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 39.

TABLE 39

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. 49, 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 39 and Fig. 49 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 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.

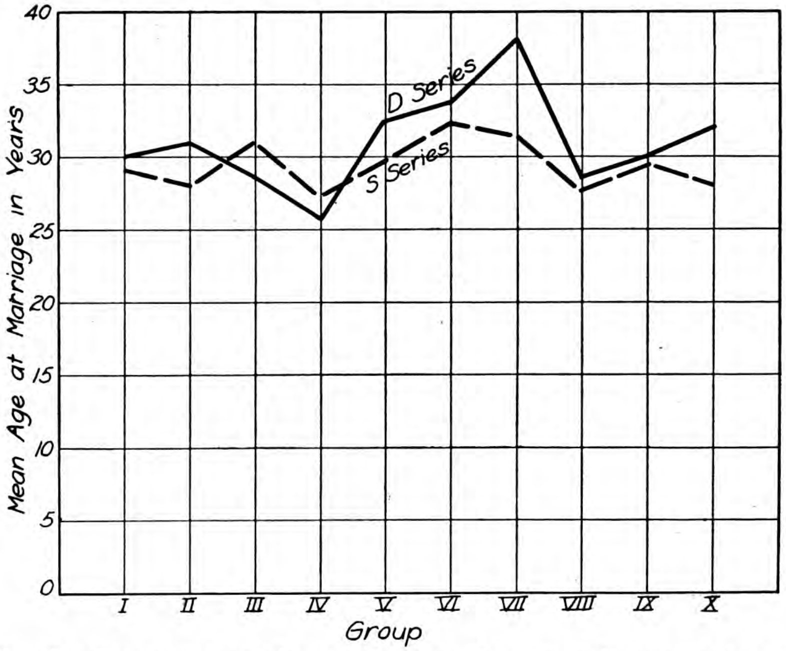


Fig. 49.—Group averages of age at marriage of persons taken at random. (Data from Table 39 above.) The Roman numerals indicate the order of the groups from the starting-points indicated in the text. Solid line = data from upper half of table. Broken line = data from lower half of table.

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 approach to parallelism in the courses of the lines of means for the D and S series. I think that most 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. 49 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 ten possible cases (I–II, II–III, VI–VII, and IX–X) the D and S lines have opposite slopes, against six 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.

Suppose, for example, 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 ± .20 beats per minute, while the normal condition in the same group was 79.68 ± .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:

$$\begin{aligned} &\text{Difference} &&= 81.12 - 79.68 = 1.44, \\ &(.20)^2 + (.15)^2 &&= .0400 + .0225 = .0625, \\ &\sqrt{.0625} &&= .25 \end{aligned}$$

Or we see that the difference in the two cases is 1.44 ± .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.

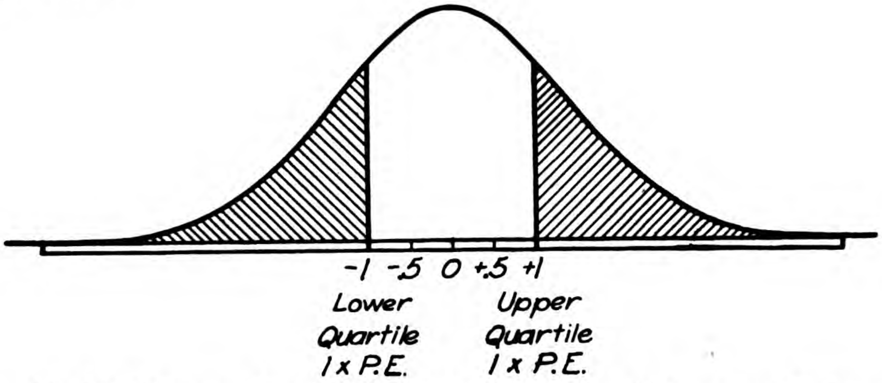


Fig. 50.—The area of a normal curve inside (blank) and the area outside (cross-hatched) the lower and upper quartiles. The quartiles are the points on the abscissa where perpendiculars to the base cut off just one-quarter of the total area of the curve at each end. By definition of the probable error given above, it is seen that the quartile distance on the x axis is 1 × P. E. The sum of the two cross-hatched areas is exactly equal to the blank area in the center.

In order that a more adequate conception may be had of just what the probable error, and various multiples of it, mean, Figs. 50 to 53 are inserted here. They show the areas of the normal curve inside and outside certain limits.

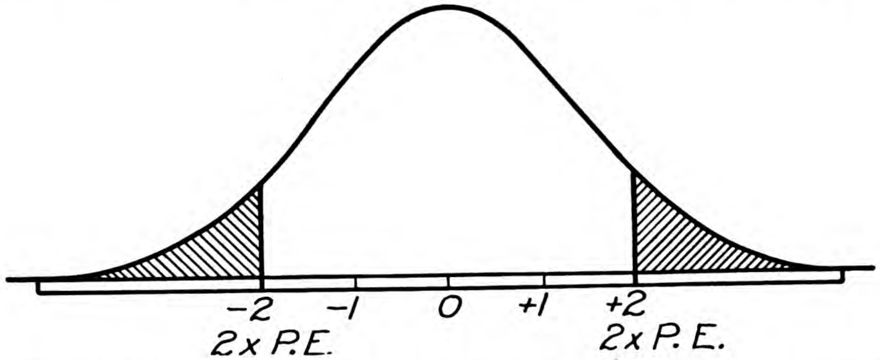


Fig. 51.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *twice the probable error*.

From these diagrams one may perceive exactly what he means 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. 52 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. 53, that a conclusion based upon a difference four or more times its probable error is practically safe.

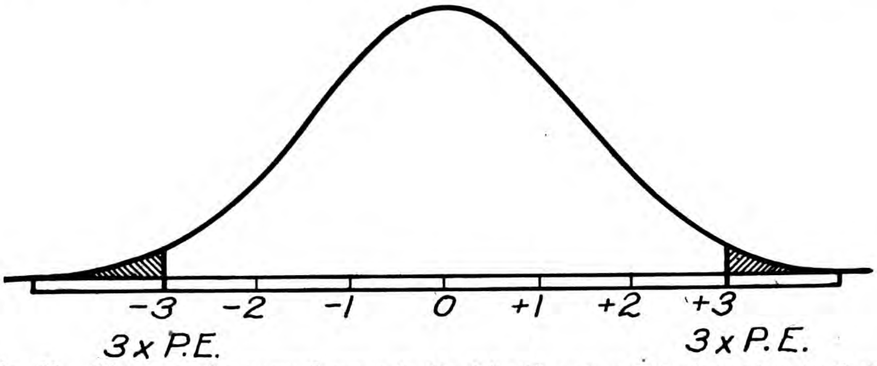


Fig. 52.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *three times the probable error*.

The following table (Table 40) 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

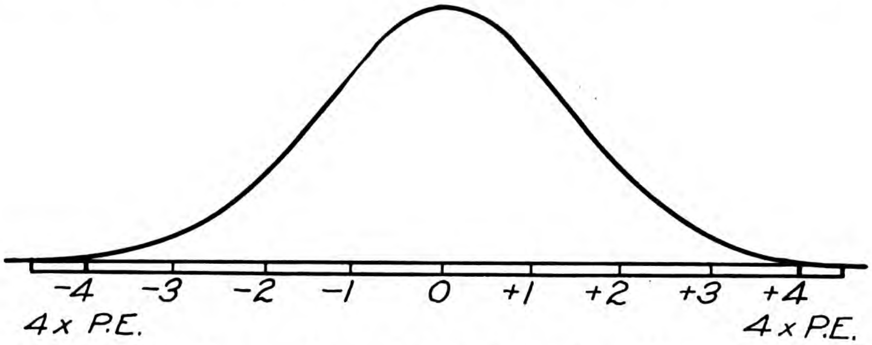


Fig. 53.—The area of a normal curve inside (blank) and outside (cross-hatched) the limits set by *four times the probable error*.

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 comparable series of measurements. These means, which may be called A and B, differ by a certain amount. The difference is found 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 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 case of each of the 100 pairs, there would 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 has significant meaning.

TABLE 40

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.
1.0....	50.00	1.00 to 1
1.1....	45.81	1.18 to 1
1.2....	41.83	1.39 to 1
1.3....	38.06	1.63 to 1
1.4....	34.50	1.90 to 1
1.5....	31.17	2.21 to 1
1.6....	28.05	2.57 to 1
1.7....	25.15	2.98 to 1
1.8....	22.47	3.45 to 1
1.9....	20.00	4.00 to 1
2.0....	17.73	4.64 to 1
2.1....	15.67	5.38 to 1
2.2....	13.78	6.25 to 1
2.3....	12.08	7.28 to 1
2.4....	10.55	8.48 to 1
2.5...	9.18	9.90 to 1
2.6....	7.95	11.58 to 1
2.7....	6.86	13.58 to 1
2.8....	5.89	15.96 to 1
2.9....	5.05	18.82 to 1
3.0....	4.30	22.24 to 1
3.1....	3.65	26.37 to 1
3.2....	3.09	31.36 to 1

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.
3.3...	2.60	37.42 to 1
3.4...	2.18	44.80 to 1
3.5...	1.82	53.82 to 1
3.6...	1.52	64.89 to 1
3.7...	1.26	78.53 to 1
3.8...	1.04	95.38 to 1
3.9...	.853	116.3 to 1
4.0...	.698	142.3 to 1
4.1...	.569	174.9 to 1
4.2...	.461	215.8 to 1
4.3...	.373	267.2 to 1
4.4...	.300	332.4 to 1
4.5...	.240	415.0 to 1
4.6...	.192	520.4 to 1
4.7...	.152	655.3 to 1
4.8...	.121	828.3 to 1
4.9...	.0950	1,052. to 1
5.0...	.0745	1,341. to 1
6.0...	.0052	19,300. to 1
7.0...	.00023	427,000. to 1
8.0...	.0000068	14,700,000. to 1
9.0...	.00000013	730,000,000. to 1
10.0...	.0000000015	65,000,000,000. to 1

It is hoped that this chapter will have given the reader a general idea of what a probable error is, and something of its purpose and significance. Throughout the remainder of the book examples will be given of probable errors, and the methods used in their calculation described.

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.

3. Wilson, E. B.: The Statistical Significance of Experimental Data, Science, vol. 58, pp. 93–100, 1923.

(To the reader with sufficient penetration to perceive and discount the writer's ironical method of presentation this will prove a valuable discussion. By some it will be taken as a brilliant and irrefragable proof that all statistical procedures are idle and futile, a result which the author probably did not foresee or desire.)

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 very 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, and conclude that that experience gives us no warrant for believing that the slight differences in relief at the two ends of such a cylinder as a penny is, do in fact sensibly influence or determine which of the two possible positions of equilibrium shall in a particular case eventuate.

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. 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 is of three sorts:

A. Experience of machines or structures like pennies, namely, cylinders excessively short in proportion to their diameter.

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.

C. Experience of tossing pennies many times.

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,

* The expert in the theory of probability will have perceived by this time that my position is as far as possible antipodal to that of Keynes in his recent book, "A Treatise on Probability." This is intentional and deliberate. I am only interested, and I think the audience for which this book has been written is also only interested, in the theory of probability as a working tool of science. From this point of view a more unsafe and unreliable guide than Keynes I have never chanced upon, whatever metaphysical merits (?) his lucubrations may have.

$$p = \frac{a}{a + b}$$

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 $\frac{1}{2}$, 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?

$$\begin{aligned} \text{In 1919 male births} &= 705{,}593 = a \\ \text{In 1919 total births} &= 1{,}373{,}438 = a + b \end{aligned}$$

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

$$p = \frac{a}{a + b + c + \ldots\ldots}$$

$$q = \frac{b}{a + b + c + \ldots\ldots}$$

$$r = \frac{c}{a + b + c + \ldots\ldots}$$

etc.

$$p + q + r + \ldots\ldots = 1$$

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 no doubt forgotten.

PERMUTATIONS AND COMBINATIONS

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

$$\begin{matrix} a\ b\ c \\ a\ c\ b \\ b\ a\ c \\ b\ c\ a \\ c\ a\ b \\ c\ b\ a \end{matrix}$$

These six different arrangements are the *permutations* of three things taken three at a time.

Generally we may write

$${}_nP_r = n\,(n-1)\,(n-2)\,(n-3)\,.....\,(n-r+1) = \frac{\lfloor\underline{n}}{\lfloor\underline{(n-r)}}$$

which means that the number of permutations of n things taken r at a time (${}_nP_r$) is equal to factorial n, $\left(\lfloor\underline{n}\right)$ divided by factorial n minus r, $\left(\lfloor\underline{(n-r)}\right)$.

From this it will be perceived that

$${}_nP_n = \lfloor\underline{n}$$

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 $e\,\underline{|n}$ 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 41.

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 merely be different, not in respect of the order of the letters, but of the letters themselves. Thus only one of the combinations

TABLE 41

VALUES OF PERMUTATIONS

Permutations of

	$r \diagdown n$	10	9	8	7	6	5	4	3	2	1
At a time.	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	1,630	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

abc and *cab* can be used, because each contains the same letters, *a*, *b*, and *c*.

Writing down the possibilities we get

$$\begin{matrix} a\ b\ c \\ a\ b\ d \\ a\ c\ d \\ b\ c\ d \end{matrix}$$

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

$$ {}_nC_r = \frac{\underline{|n}}{\underline{|r}\ \underline{|(n-r)}}$$

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

$$ {}_nC_n = \frac{\underline{|n}}{\underline{|n}} = 1$$

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

$$\Sigma\, {}_nC_r = n + \frac{n.\,(n-1)}{1.2} + \frac{n.\,(n-1)\,(n-2)}{1.2.3} + \ldots\ldots + n + 1$$

But the right-hand side of the equation is plainly

$$(1+1)^n - 1.$$

Hence

$$\Sigma\, {}_nC_r = 2^n - 1.$$

Again, for the sake of convenience, a brief table of combinations is inserted as Table 42.

TABLE 42

VALUES OF COMBINATIONS

Combinations of

	n r	10	9	8	7	6	5	4	3	2	1
At a time.	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

It will be noted from this table that in each column the values rise to a maximum and then decline.

$$\text{The maximum } {}_nC_r = \frac{\underline{|n}}{\left(\underline{\left|\frac{n}{2}\right.}\right)^2} \text{ when } n \text{ is even.}$$

$$\text{The maximum } {}_nC_r = \frac{\underline{|n}}{\underline{\left|\frac{n+1}{2}\right.}\;\underline{\left|\frac{n-1}{2}\right.}} \text{ when } n \text{ is odd.}$$

Approximations to $\underline{|n}$

In all practical work with probability it is useful to have an easily computed approximation to the value of $\underline{|n}$ in cases when

n is large. In Pearson's "Tables for Statisticians and Biometricians" a table of log $\underline{|n}$ 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:

$$\underline{|n} = \sqrt{2\,\pi\,n} \times n^n\,e^{-n} \times \left\{1 + \frac{1}{12n} + \frac{1}{288n^2} + \ldots\ldots\right\}$$

Forsyth's:

$$\underline{|n} = \sqrt{2\,\pi}\;\left\{\frac{\sqrt{n^2 + n + \frac{1}{2}}}{e}\right\}^{n + \frac{1}{2}}$$

This is accurate to $\frac{1}{240n^3}$.

To indicate the closeness of these approximations we may calculate $\underline{|n}$ for $n = 2$.

The result is

$$\begin{aligned} \underline{|n} &= 1.999479 \text{ (Forsyth)} \\ \text{Actual error} &= \ .000521 \\ \frac{1}{240n^3} &= \ .00052083 \end{aligned}$$

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, $\frac{1}{2}$. 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 $\frac{1}{2}\ n$ AH, B –. But now the B penny, which is associated with each of these AH pennies in the $\frac{1}{2}\ n$ throws, will be subject to the same influences as though it were tossed alone. Consequently we shall have in these $\frac{1}{2}\ n$ 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 $\frac{1}{4}\ n$ AH, BT, and $\frac{1}{4}\ n$ 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,

$$\tfrac{1}{4}\, n \text{ AH, BH} = 2 \text{ heads,}$$
$$\tfrac{1}{2}\, n \left\{\begin{matrix}\text{AH, BT}\\ \text{or}\\ \text{AT, BH}\end{matrix}\right\} = 1 \text{ head, } 1 \text{ tail,}$$
$$\tfrac{1}{4}\, n \text{ AT, BT} = 2 \text{ tails.}$$

Whence we arrive at the rule:

If the separate probabilities of each of several independent events are respectively p_1, p_2, p_3 . . . . ., 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 $\frac{1}{4}$ 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, $\frac{1}{4}$, 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

$$\begin{aligned} p_1 &= .0628 \text{ as before} \\ p'_2 &= \frac{372}{23{,}464} = .0159 \\ P &= p_1 \times p'_2 = .000999 \end{aligned}$$

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

$$p \times p \times p \times \ldots\ldots \times q = p^{n-1}.q$$

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:

H H H T
H H T H
H T H H
T H H H

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

$$p \times p \times \ldots\ldots \times q \times q = p^{n-2}.q^2$$

But again this may happen in

$\dfrac{n\ (n-1)}{1.2} = {}_nC_r$ (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:

H H T T
H T H T
H T T H
T H H T
T H T H
T T H H

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. 54, 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 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

$q = \frac{5}{6}$ = 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

$$(\tfrac{1}{6} + \tfrac{5}{6})^4$$

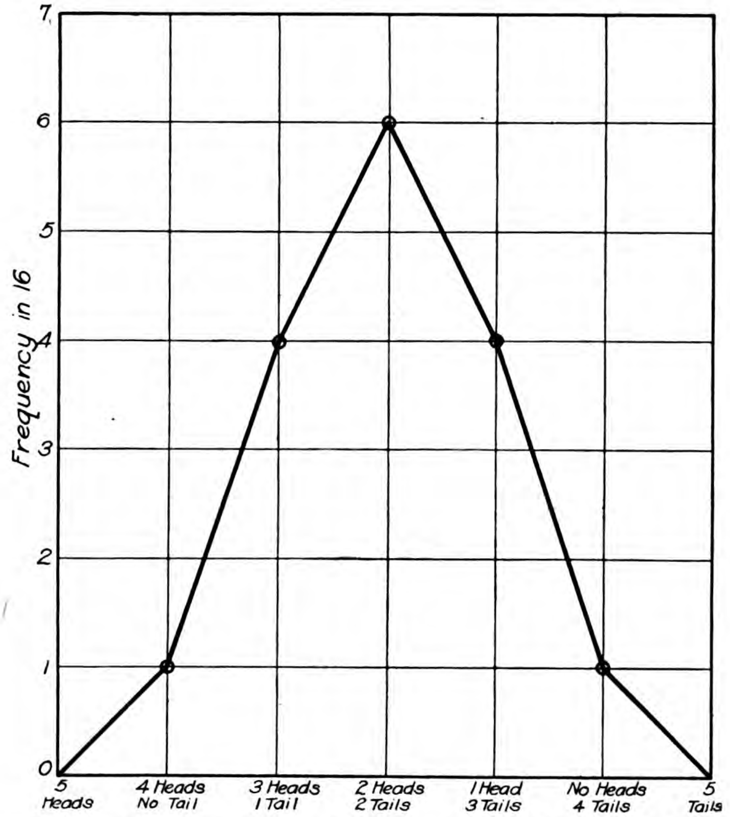


Fig. 54.—The results of tossing four pennies together at random, as given by the binomial $(\frac{1}{2} + \frac{1}{2})^4$.

These are:

$$p^n = \frac{1}{1296} = \text{probability that all 4 dice will fall with the 6 face up.}$$

$$np^{n-1}q = \frac{20}{1296} = \text{probability that 3 dice will fall 6's and 1 die something other than 6.}$$

$$\frac{n\,(n-1)}{1.2}\,p^{n-2}q^2 = \frac{150}{1296} = \text{probability that 2 dice will fall 6's, and the other 2 something other than 6.}$$

$$\frac{n\,(n-1)\,(n-2)}{1.2.3}\,p^{n-3}q^3 = \frac{500}{1296} = \text{probability that 1 die will fall 6 and the other three something else.}$$

$$\frac{n\,(n-1)\,(n-2)\,(n-3)}{1.2.3.4}\,p^{n-4}q^4 = \frac{625}{1296} = \text{probability that no die will fall 6.}$$

This distribution is shown graphically in Fig. 55, 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

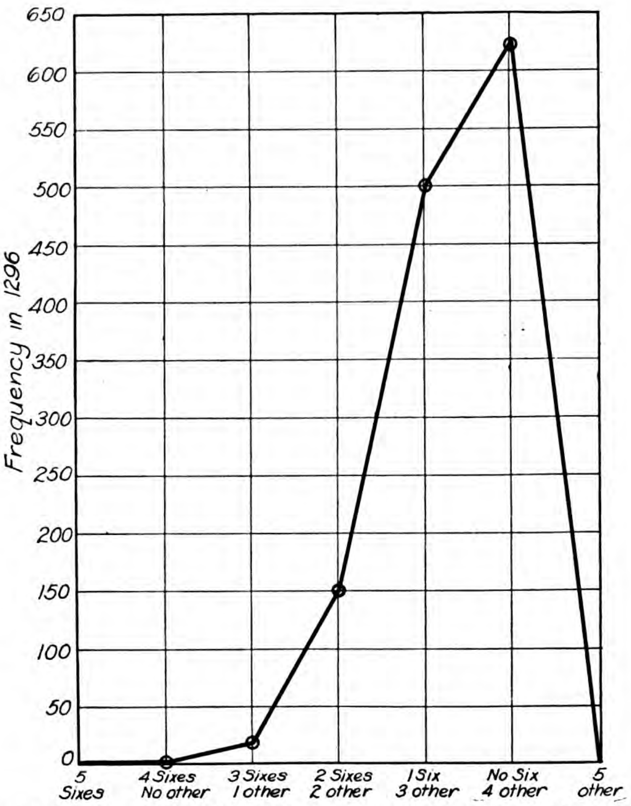


Fig. 55.—The probability of getting different numbers of 6's in the throws of 4 dice together, as given by $(\frac{1}{6} + \frac{5}{6})^4$.

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. 55 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 43.

TABLE 43

SUCCESSIVE TERMS OF $(\frac{1}{2} + \frac{1}{2})^{10}$

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	

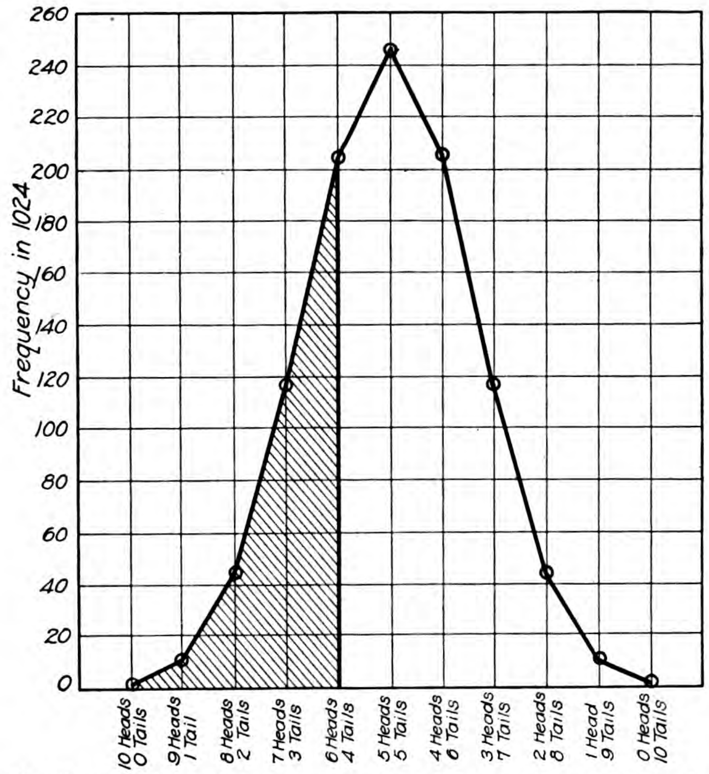


Fig. 56.—The binomial $(\frac{1}{2} + \frac{1}{2})^{10}$. The meaning of the cross-hatched area is explained in the text.

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 43 are plotted in Fig. 56.

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

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 cocked hat or a sugar loaf,* 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,

* Both somewhat mythical objects which scarcely any living American can ever have seen, but of such hoary antiquity in the literature of probability as terms for the description of the shape of curves such as we are talking about, that I feel compelled to use them. One should not lightly break with ancient traditions!

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 I only wish to point 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 43 in another form, as shown in Table 44.

TABLE 44

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

Normally, of course, one would never carry so many places of decimals in x/σ. But this example will indicate that the position 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 57 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 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

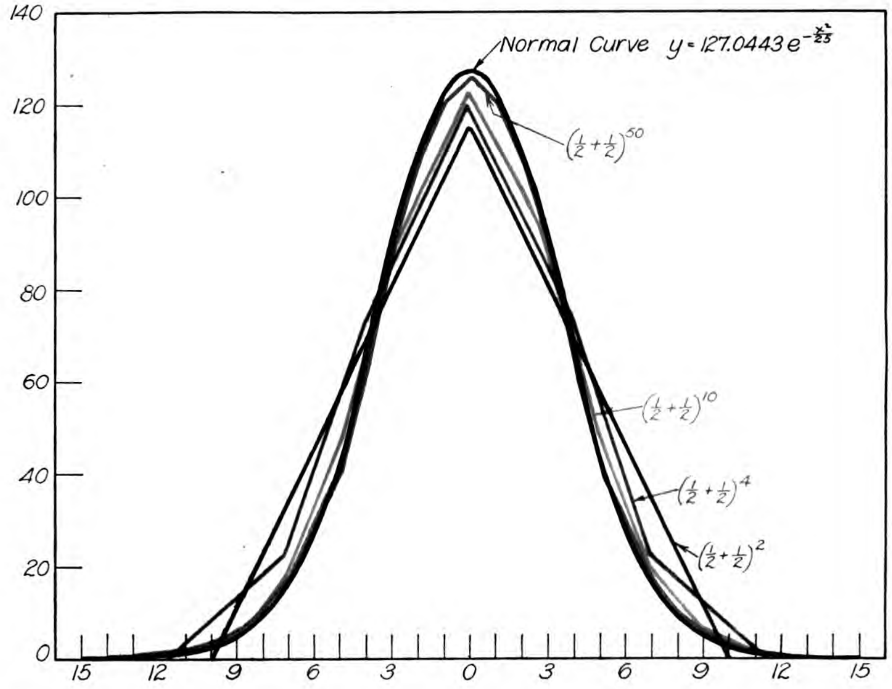


Fig. 57.—Point binomials for several values of n, and a superimposed normal curve.

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 we know the equation to the smooth curve we can integrate it 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 our derived probabilities read off from the table of these integrations, without any work at all. Now it is apparent from Fig. 57 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 ~~XI~~ IX 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 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 treatment 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. 58.

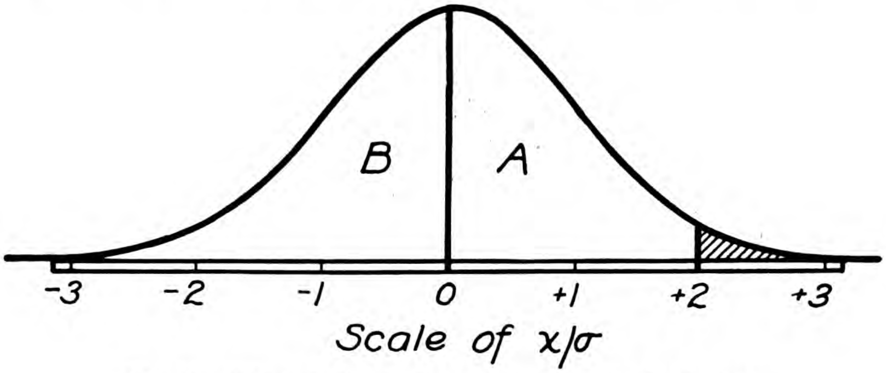


Fig. 58.—Diagram of probability example given in the text.

Now in the table in Appendix IM 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 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:

$$\text{P. E.} = .6744898 \ldots \sigma.$$

Because there is frequent use for this knowledge Table 45 is presented, giving the relations between certain multiples of σ and the probable error.

TABLE 45

MULTIPLES OF σ AND THE PROBABLE ERROR

σ times any number in Column A is the same as the probable error times the corresponding number in Column B.

A.	B.	A.	B.
.5	.741	2.3	3.410
.6	.890	2.4	3.558
.7	1.038	2.5	3.707
.8	1.186	2.6	3.855
.9	1.334	2.7	4.003
1.0	1.483	2.8	4.151
1.1	1.631	2.9	4.300
1.2	1.779	3.0	4.448
1.3	1.927	3.1	4.596
1.4	2.076	3.2	4.744
1.5	2.224	3.3	4.893
1.6	2.372	3.4	5.041
1.7	2.520	3.5	5.189
1.8	2.669	3.6	5.337
1.9	2.817	3.7	5.486
2.0	2.965	3.8	5.634
2.1	3.113	3.9	5.782
2.2	3.262	4.0	5.930

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.)

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.

PAST EXPERIENCE AND FUTURE EXPECTATION

A theorem in probability which is likely to be most useful to medical men is one enabling us to measure the reliability of predictions of future results on the basis of past experience. This theorem is due to Pearson.[1] It should be understood that this section is merely a *refinement* of the methods dealt with in Chapter XI. It has significance, as distinct from those methods only when meticulous accuracy is desired.

Let it be supposed that a first sample of $n = p + q$ be drawn from the population, p denoting the number of times the event dealt with occurs in the n trials, and q the number of times it fails.

Write

$$\bar{p} = \frac{p}{n}, \bar{q} = \frac{q}{n}$$

whence, of course,

$$\bar{p} + \bar{q} = 1.$$

We then have for the chief constants of the error distribution for a second sample, of magnitude m, drawn from the same population the following values:

$$\text{Mean} = m\bar{p} + \frac{m}{n+2}(q - \bar{p}) \text{.......................(i)}$$

$$\text{Mode} = \text{the integral portion of } m\bar{p} + p$$

Standard deviation

$$= \left\{ m\left(p + \frac{\bar{q} - p}{n+2}\right)\left(q - \frac{q - p}{n+2}\right)\left(1 + \frac{m-1}{n+3}\right)\right\}^{\frac{1}{2}} \text{...........(ii)}$$

The last mentioned constant, the standard deviation, is needed because the probable error of the number of occurrences is related to it in the following way:

$$\text{P. E.} = \pm\ .67449 \text{ Standard Deviation}$$

The writer* has applied this theorem to a problem in the treatment of pneumonia. In a certain camp 966 acute pneumonia patients were treated on the open ward plan. Of these so treated, 135 died. The treatment was then changed to the closed ward plan, and 435 were so treated, with a mortality of 14. From these figures we may state the problem in the following terms: If, under the open ward treatment, out of a sample of 966 patients with acute pneumonia 135 died, what would be the probable number to die in a second example of 435 acute pneumonia patients from the same population, given the same treatment?

We have here, in our mathematical notation,

$$\begin{aligned} n &= 966 \\ m &= 435 \\ p &= 135 \\ q &= 831 \\ \bar{p} &= \frac{135}{966} = .1397 \quad \bar{q} = \frac{831}{966} = .8603 \end{aligned}$$

Whence, from the equation given earlier, we readily deduce

Mean deaths expected in second sample = 61.12 ± 5.88.
Modal, or most probable number of deaths in second sample = 60.
Standard deviation = 8.72.

But the actually observed number of deaths in the second sample, under the close ward treatment, was 14 instead of 60. Hence we may safely conclude that under the close ward treatment significantly fewer persons died than would have been expected to die on the basis of chance, if the same force of mortality had prevailed in the latter period as did in the former.

But can it be assumed that the same force of mortality was impinging, and its results were mitigated only by the new method of treatment? In order to settle this question we must use data from a comparable outbreak of post-influenzal pneumonia where

* Pearl, R.: A Statistical Discussion of the Relative Efficacy of Different Methods of Treating Pneumonia, Arch. Int. Med., vol. 24, pp. 398–403, 1919.

the same treatment was followed throughout the outbreak. Would there be in such a case a falling off of the case fatality rate in the latter part of the epidemic corresponding entirely or in some degree to that observed above? Such data were obtained, and may be put in the following way:

In the same notation as before

Cases of pneumonia through October 6 = n = 1000.
Deaths from pneumonia throughout October 14 = p = 364, whence q = 636, and $\bar{p}$ = .364 and $\bar{q}$ = .636.
Cases of pneumonia beginning October 7 and going to end of the epidemic = m = 499.

The problem, then, may be stated in this way: If, with no change of treatment, 364 patients died out of a sample of 1000, what is the probable number of deaths in a second sample of 499 cases? By the same method as before the data give

Mean deaths expected in second sample = 181.77 ± 8.87.
Modal, or most probable number of deaths in second sample = 182.
Standard deviation = 13.15.

Now, the actual number of deaths in the last 499 cases of this second epidemic was only 77 instead of the expected 182. But there had been no change in method of treatment. Hence, it is clear that fewer deaths in proportion to cases occur in the later as compared with the earlier portion of these epidemic outbreaks, quite without change of treatment.

This result obviously tends to cast reasonable doubt on the efficacy of the closed ward as compared with the open ward treatment of these epidemic pneumonias. It is necessary, however, to make a further quantitative comparison before drawing any final conclusion. In the second sample (latter part) from the second epidemic the actual deaths formed 42 per cent. of the deaths expected on the basis of chance from the results shown in the first sample from the same epidemic

$$\frac{77 \times 100}{182} = 42 \text{ per cent.}$$

In the first epidemic the actual deaths under the close ward treatment formed only 23 per cent. of the deaths expected on the basis of chance from the results shown in the first sample from the same epidemic, under the open ward treatment

$$\frac{14 \times 100}{60} = 23 \text{ per cent.}$$

This result gives the significant comparison. The whole matter may be summarized in this way. While it is true that the case fatality rate tends under a constant form of treatment to be markedly lower in the later portions of epidemic outbreaks of pneumonia, nevertheless the data show, when given proper mathematic analysis, that under the closed ward treatment only about half (23 versus 42 per cent.) as many deaths occurred relatively in the latter part of the particular epidemic studied, as would probably have occurred if the open ward method of treatment had been used in this epidemic, and worked the same way that it did in the one to which it was applied throughout, after making allowance for the normally diminishing case fatality rate of later portions of epidemics.

An interesting special case of the theorem just discussed sometimes has application in practical problems, notably in Mendelian experiments.

The case is this: Suppose $p = 0$. This means that the event did not occur at all in the first sample. Ordinarily, on long current views of the theory of probability one concludes, either that the population from which the sample is drawn does not contain that which would make the event capable of happening, which makes the probability of its occurrence in second or other samples always equal to zero, or else, falling back upon a supposed applicability of Bayes' theorem, it is concluded that, having no knowledge of the population, the event at the next trial is as likely as not to occur, whence the probability equals $\frac{1}{2}$. Obviously neither of these conclusions is true. The first, because neither (i) nor (ii) vanishes when $p = 0$, the second, because the first sample *does*, in fact, give us some knowledge of the constitution of the population.

Putting $p = 0$, $q = 1$, we have, for the expectation *of occurrence of the event*, in the second sample

$$\text{Mean} = m \frac{1}{n + 2} \qquad \text{(iii)}$$

and

$$\sigma^2 = m\left(\frac{1}{n+2}\right)\left(1 - \frac{1}{n+2}\right)\left(1 + \frac{m-1}{n+3}\right) \,\ldots\ldots\ldots\ldots \text{(iv)}$$

Equation (iii) shows that the expectation of the occurrence of the event in a second sample m if it did not happen at all in the first sample, n, varies from just over zero if m is very small as compared with n, through 1 if $m = n + 2$ to any number whatever when $m > n + 2$.

To take an example: What is the probability that the sun will *fail* to rise tomorrow? Here n = 36,500,000,000, on Lord Kelvin's estimate of one hundred million years for the age of the earth, and on the assumption that the rising of the sun has been observed (or could have been) on each of the days in this period, and $m = 1$. Hence expected mean number of occurrences tomorrow of failure of the sun to rise = $\frac{1}{36{,}500{,}000{,}002}$.

This is admittedly a small number, but not precisely equal to zero. Which means that we cannot be *quite* sure that the sun will rise tomorrow, on the basis of our past experience with that useful body.

THE CHI-SQUARE TEST

Another theorem in probability 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 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

$$m_1,\ m_2,\ m_3,\ m_4 \ldots\ldots\ m_p,\ m_q \ldots\ldots\ m_s$$

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	...	f_p	f_q	...	f_s	N
Second sample......	f'_1	f'_2	f'_3	...	f'_p	f'_q	...	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 46

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. 59. The numerical work is set forth in Table 47.

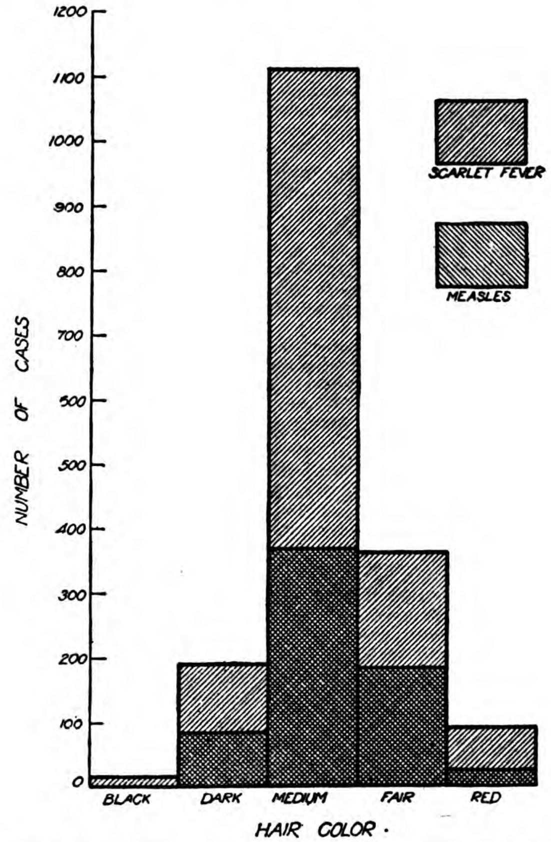


Fig. 59.—Distribution of scarlet fever and measles in respect of hair color of those attacked.

Therefore

$$\chi^2 = NN' \times .000,0211 = 1864 \times 661 \times .000,0211 = 26.00$$

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 indifferently all individuals whatever their hair pigmentation; or, that scarlet fever and measles are differential in their selection.

TABLE 47

NUMERICAL WORK TO CALCULATE PROBABILITY THAT THE MEASLES AND SCARLET FEVER DISTRIBUTIONS OF TABLE 46 ARE RANDOM SAMPLES OF THE SAME POPULATION

		Black.	Dark.	Medium.	Fair.	Red.		Totals.
Scarlet fever....................	(i)	12	289	1109	360	94	f	1864
Measles........................	(ii)	0	85	367	184	25	f'	661
(i) + (ii).......................	(iii)	12	374	1476	544	119	$f+f'$	2525
(i)/1864.......................	(iv)	.0064	.1551	.5950	.1931	.0504	f/N	1.0000
(ii)/661........................	(v)	.0000	.1286	.5552	.2784	.0378	f'/N'	1.0000
(iv) − (v)......................	(vi)	+.0064	+.0265	+.0398	−.0853	+.0126	$f/N-f'/N'$	
Square of (vi)...................	(vii)	.000,041	.000,702	.001,584	.007,276	.000,159	$(f/N-f'/N')^2$	
(vii) ÷ (iii)....................	(viii)	.000,0034	.000,0019	.000,0011	.000,0134	.000,0013	$\frac{(f/N-f'/N')^2}{f+f'}$	.000,0211

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 differentiated 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 contains 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 representative of the universe, must contain some individuals from each of the differentiated classes. Thus, to take a representative sample from the population of a given locality relative to our hookworm problem, it would theoretically 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 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 I 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, etc.; but supposing I had accurate data to do all this, the resulting probability would have only a very academic interest, because I already know before I begin that elephants 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 48), 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 48

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 48 was calculated needs some discussion. First, for each value of *m* and of the percentages of infestation 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 48. Wishing to see just how much there was in this objection, and also desiring to give the reader of this book a concrete idea of the behavior of

TABLE 49

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 50

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	.144559	.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.

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 49 and 50.

Consider the most unfavorable case in Table 48 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 49, we have $\sum_{0}^{6} = 0.00$. 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 48 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 50.

Again, let us test the worst case. Table 48 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 50 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 48 is a safe guide. The bearing of these results on the first section of this chapter should not be overlooked.

The refinement of the hypergeometric series over the binomial will ordinarily not be required in practical work.

The practical uses of Table 48 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 Influence of Past Experience on Future Expectation, Phil. Mag., 1907, pp. 365–378.
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 Certain Properties of the Hypergeometrical Series, and On the Fitting of Such Series to Observation Polygons in the Theory of Chance, Phil. Mag., 1899, pp. 236–246.

4. 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.

5. 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.

6. Pearl, R.: A Statistical Note on Epidemic Encephalitis, Johns Hopkins Hospital Bulletin, vol. 32, pp. 221–225, 1921.

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. An example of such a distribution is given in Table 51 and is exhibited graphically in Fig. 60.

TABLE 51

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

* The treatment of the subject in this chapter follows essentially the account given in an article by the present writer in the Nelson Loose-leaf Medicine, vol. 7, entitled The Significance of Biometry and Vital Statistics to the Science of Medicine.

† Whiting, M. H.: A Study of Criminal Anthropometry, Biometrika, vol. 11, pp. 1–37, 1915.

A word should be said about the designation of the class limits in the first column of Table 51. The pulse rates, as actually *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 51.

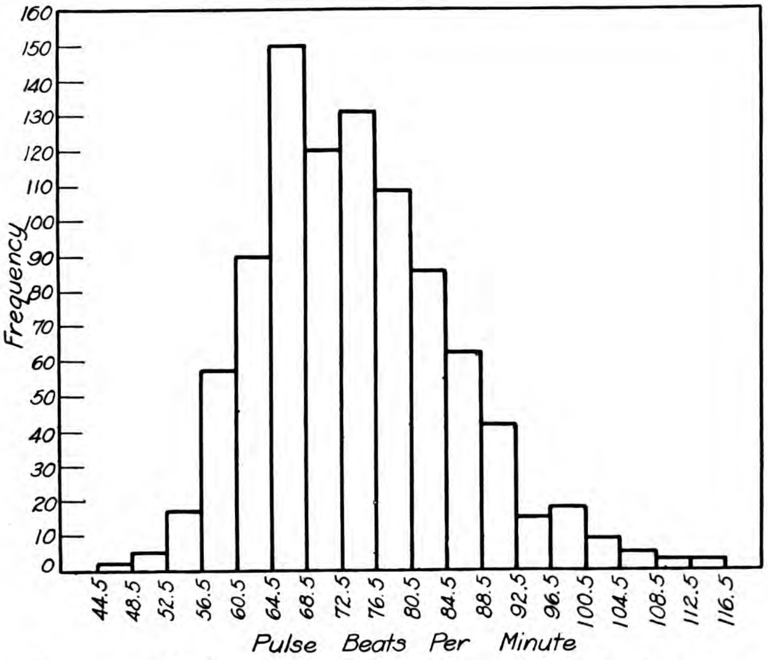


Fig. 60.—Histogram showing frequency distribution of variation in pulse beats per minute in English convicts. (Data of Table 51.)

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 requency becomes smaller and smaller till it disappears. The

f

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 51 the results shown in Table 52.

TABLE 52
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	± 7.7
48.4........	2	922	± 0.95	88.4......	829	95	± 6.2
52.4........	7	917	± 1.8	92.4......	871	53	± 4.8
56.4........	24	900	± 3.3	96.4......	886	38	± 4.1
60.4........	81	843	± 5.8	100.4......	904	20	± 3.0
64.4........	171	753	± 8.0	104.4......	913	11	± 2.2
68.4........	321	603	± 9.8	108.4......	918	6	± 1.6
72.4........	441	483	± 10.2	112.4......	921	3	± 1.2
76.4........	572	352	± 10.0	116.4......	924	0	....
80.4........	681	243	± 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 × 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 precise 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 53. 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 53.

TABLE 53
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, we have

$$\begin{aligned}\pi_1 &= 0 \text{ (by definition of the mean)}\\ \pi_2 &= \nu_2 - \nu_1^2\\ \pi_3 &= \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3\\ \pi_4 &= \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4\end{aligned}$$

For the pulse-rate example we have:

$$\begin{aligned}\pi_2 &= 55.698052 - 47.960126 = 7.737926\\ \pi_3 &= 505.964286 - 1157.181336 + 664.278920 = 13.061870\\ \pi_4 &= 5095.827922 - 14015.868476 + 16027.713552 - 6900.521118 = 207.151880\end{aligned}$$

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:

$$\begin{aligned}\mu_1 &= 0\\ \mu_2 &= \pi_2 - \tfrac{1}{12}. \quad (\tfrac{1}{12} = .083333)\\ \mu_3 &= \pi_3\\ \mu_4 &= \pi_4 - \tfrac{1}{2}\pi_2 + \tfrac{7}{240}. \quad (\tfrac{7}{240} = .029167)\end{aligned}$$

We then have, from the pulse-rate example

$$\begin{aligned}\mu_2 &= 7.654593\\ \mu_3 &= 13.061870\\ \mu_4 &= 203.312084\end{aligned}$$

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. 60 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

$$\text{P. E. Mean} = \chi_1 \pm \sigma.$$

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 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 above 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 somewhat 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. 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 and σ are constants which will be explained below. Here they may be taken as given. It should, however, be expressly noted that this χ 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

$$\text{P. E.}_{d\ (\text{normal curve})} = \pm\ .67449 \sqrt{\frac{3}{2\,N}}\,\sigma.$$

Consequently, unless d amounts to three or four times its 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 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. 61 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.

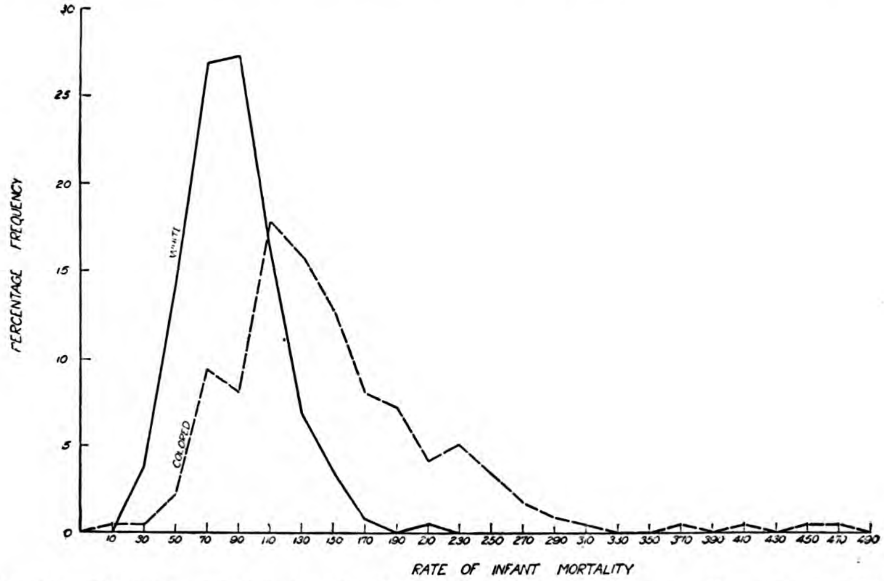


Fig. 61.—Frequency polygons showing variation in infant mortality rate in 1918 of (*a*) the white population and (*b*) the colored population of rural counties.

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. 58, 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

$$\text{P. E.}_{\sigma} = \pm\ \chi_2\sigma,$$

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

$$\text{P. E.}_{\text{c. v.}} = \pm\ .67449 \frac{V}{\sqrt{2\ N}} \left\{ 1 + 2 \left(\frac{\text{V}}{100}\right)^2 \right\}^{\frac{1}{2}} = \chi_2 \times \psi,$$

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 a chief 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 54 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, as compiled in an earlier paper by the writer.*

TABLE 54

Coefficients of Variation for Man

	♂	♀
Weight of spleen (General Hospital population)[1]	50.58	
Weight of spleen (healthy)[2]	38.21	
Dermal sensitivity[3]	35.70	45.70
Weight of heart (General Hospital population)[1]	32.39	
Keenness of sight[3]	28.68	32.21
Weight of kidneys (General Hospital population)[1]	24.63	
Weight of body (Bavarians)	21.32	24.715
Weight of liver (General Hospital population)[1]	21.12	
Swiftness of blow[3]	19.4	17.1
Weight of heart (healthy)[2]	17.71	
Weight of kidneys (healthy)[2]	16.80	
Breathing capacity[3]	16.6	20.4
Strength of pull[3]	15.0	19.3
Weight of liver (healthy)[2]	14.80	
Height of mandible (English, both sexes)[4]	11.73	11.73
Weight of body (English)[3]	10.37	13.37
Skull capacity (Etruscan)[5]	9.58	8.54
Brain weight (French)[3]	9.16	9.14
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)	8.118	8.340
Brain weight (Hessian)	8.096	8.125
Brain weight (Bohemian)	7.809	7.382
Skull capacity (modern German)[5]	7.74	8.19
Skull capacity (Naqada)[5]	7.72	6.92
Brain weight (Swedish)	7.592	8.043
Skull capacity (Parisian, French)[5]	7.36	7.10
Skull capacity (Aino)[5]	7.07	6.90
Mandible, distance between foramina mentalia (English, both sexes)[4]	6.23	6.23
Length of forearm[8]	5.24	5.21
Length of femur (French)[3]	5.05	5.04
Length of tibia (French)[3]	4.975	5.365
Length of humerus (French)[3]	4.89	5.61
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
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
Skull, length of base (English)[6]	4.07	4.11
Skull, cephalic index for greatest length (English)[6]	3.95	4.03
Stature (English)[8]	3.99	3.83
Skull, ratio of height to greatest length (English)[6]	3.80	4.21
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
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

[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.

* Pearl, R.: Biometrical Studies in Man. I. Variation and Correlation in Brain-weight, Biometrika, vol. 4, pp. 13–104, 1905.

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, in which this rule fails, and the sign of the skewness may be checked by the relations between mean and median. 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 the 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. 62 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 Pearson's "Tables for Statisticians and Biometricians," or in more detail in Elderton's "Frequency Curves and Correlation." Here space is lacking to go further into this phase of the matter.

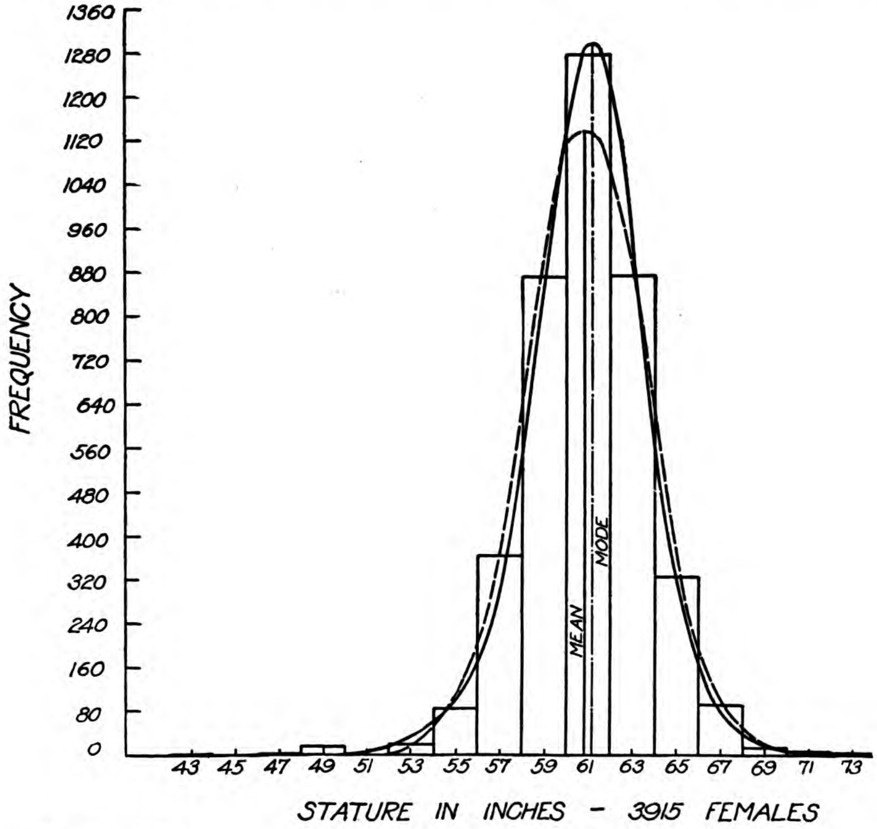


Fig. 62.—Histogram and fitted curves for variation in stature of 3915 Scottish females (insane). The solid curve is the skew curve appropriate to the distribution. The broken curve is the corresponding normal or Gaussian curve. The skew curve is leptokurtic. (Plotted from data of Tocher, Biometrika, 5, pp. 298–350.)

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. 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 $z = \frac{x_1}{x_2}$ 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 54 *bis* 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.

* Cf. Pearl, R.: Biometrika, vol. 6, pp. 437, 438, 1909; Reed, L. J.: Jour. Washington Acad. Sci., vol. 11, pp. 449–455, 1921.

† Proc. Roy. Soc., vol. 60, pp. 489–498, 1897.

σ_1, σ_2, and σ_3 the standard deviations of x_1, x_2, and x_3.

$v_1 = \frac{\sigma_1}{m_1}$, $v_2 = \frac{\sigma_2}{m_2}$, $v_3 = \frac{\sigma_3}{m_3}$. (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.

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 54 *bis*

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	$\sqrt{(\sigma_1^2+\sigma_2^2-2r_{12}\sigma_1\sigma_2)}$
$z=x_1 \,.\, x_2$	$m_1 m_2+r_{12}\sigma_1\sigma_2$	$m_1m_2[v_1^2+v_2^2+2r_{12}v_1v_2+v_1^2v_2^2(1+r_{12})]\sigma$,* or approximately $m_1m_2[v_1^2+v_2^2+2r_{12}v_1v_2]^{\frac{1}{2}}$
$z=x_1 \,.\, x_2 \,.\, x_3$	$m_1m_2m_3[1+r_{12}v_1v_2+r_{13}v_1v_3+r_{23}v_2v_3]$	$m_1m_2m_3[v_1^2+v_2^2+v_3^2+2r_{12}v_1v_2+2r_{13}v_1v_3+2r_{23}v_2v_3]^{\frac{1}{2}}$ approximately
$z=\frac{x_1}{x_2}$	$\frac{m_1}{m_2}(1+v_2^2-r_{12}v_1v_2)$	$\frac{m_1}{m_2}\sqrt{(v_1^2+v_2^2-2r_{12}v_1v_2)}$

* 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.

AN EXAMPLE OF THE APPLICATION OF BIOMETRIC METHODS MEASURING VARIATION TO A PARTICULAR PUBLIC HEALTH PROBLEM

By way of concrete illustration of the principles set forth in this chapter it seems desirable to give the detailed results of their application in a specific case, that of variation in infant mortality in different places (cf. Pearl, R[4]).

TABLE 55

CONSTANTS OF VARIATION IN RATE OF INFANT MORTALITY (DEATHS UNDER 1 PER 1000 BIRTHS) COMPUTED FROM DATA OF TABLE 17, CHAPTER VII

Group.	Mean.†	Mode.†	Median.†	Standard† deviation.	Coefficient of variation (per cent.).	Skewness.	Kurtosis.
Cities over 25,000,* Total, 1915	104.49± 1.78	96.26	102.76	26.14±1.26	25.02±1.28	+.3148±.0937	.145± .491
" " " " 1916	102.53± 1.67	104.47	103.24	24.69±1.18	24.09±1.22	−.0786±.0848	.224± .511
" " " " 1917	99.58± 1.32	93.83	98.00	23.45± .93	23.55± .99	+.2455±.0858	1.127±1.558
" " " " 1918	107.78± 1.41	99.66	105.50	25.07±1.00	23.26± .96	+.3237±.0800	.271± .471
Cities under 25,000,* Total, 1915	100.98± 1.68	92.87	97.95	30.81±1.18	30.51±1.28	+.1934±.0657	.419± .603
" " " " 1916	104.23± 1.75	97.05	101.03	32.38±1.24	31.07±1.30	+.2217±.0678	.725± .916
" " " " 1917	99.24± 1.32	84.74	94.74	29.94± .93	30.17±1.02	+.4840±.1197	1.624±1.197
" " " " 1918	111.61± 1.66	90.36	104.17	37.78±1.17	33.85±1.13	+.5625±.2647	3.212±2.945
Rural counties, Total, 1915	83.07± .85	75.40	79.54	23.95± .60	28.83± .79	+.3204±.1454	2.200±2.690
" " " 1916	85.28± .90	76.10	82.15	29.94± .63	30.42± .81	+.3536±.0509	.362± .319
" " " 1917	82.01± .52	74.73	78.96	25.71± .37	31.35± .49	+.2833±.1157	2.392±2.379
" " " 1918	84.43± .57	72.14	80.97	28.40± .40	33.64± .48	+.4328±.0409	1.281± .482
Cities over 25,000,* White, 1917	92.22± 2.02	—	92.14	15.60±1.43	16.91±1.60	—	—
" " " " 1918	102.59± 2.00	—	99.23	15.42±1.42	15.03±1.41	—	—
Cities under 25,000,* White, 1917	98.46± 2.75	—	97.50	20.82±1.95	21.15±2.06	—	—
" " " " 1918	114.62± 4.17	—	113.33	31.49±2.95	27.47±2.80	—	—
Rural counties, White, 1917	86.21± 1.07	81.66	84.24	24.15± .76	28.02± .99	+.1799	2.847
" " " 1918	85.90± 1.27	77.80	83.75	28.90± .90	33.65±1.05	+.2802±.0650	.999±1.055
Cities over 25,000,* Colored, 1917	202.59± 8.88	—	194.00	68.43±6.28	33.78±3.44	—	—
" " " " 1918	216.67±11.15	—	214.00	75.87±7.88	39.63±3.65	—	—
Cities under 25,000,* Colored, 1917	213.08± 9.92	—	228.00	74.96±7.01	35.18±3.68	—	—
" " " " 1918	217.69±11.46	—	225.00	86.65±8.10	39.80±4.27	—	—
Rural counties, Colored, 1917	134.76± 2.55	106.17	127.25	57.37±1.80	42.57±1.56	+.4984±.5222	4.036±6.160
" " " 1918	147.26± 2.92	108.83	134.59	66.15±2.06	44.92±1.66	+.5819±.4154	4.239±4.998

* In 1910.

† In concrete units, i.e., rate of deaths under 1 per 1000 births.

The original frequency distributions are given in Table 17, Chapter VII. The simple biometric constants derived from these distributions are presented in Table 55.

From the data presented in Table 55 the following points are to be noted:

1. There is no certainly significant decline in the mean value of the rate of infant mortality during the four years covered by these statistics in any of the demographic units considered. In the cities of over 25,000 the mean rate declined during the years 1915, 1916, and 1917, but the total amount of this drop cannot be regarded as statistically significant, having regard to the probable errors involved. In other words, the change from a mean rate of 104 in 1915 to a mean rate of 100 (99.58) in 1917 may easily have been simply the result of chance.

2. In 1918 there was a general tendency toward an increase in the mean rate of mortality over that which obtained in 1917. This increase is unquestionably to be attributed to the influenza epidemic of the autumn and winter of 1918. A careful examination of the rates by months will convince one that the mortality of infants increased very materially during the period of the epidemic. Whether this increased number of deaths was truly to be charged to influenza does not concern us here. The important fact is that the rate of infant mortality markedly increased coincidently with the existence of the epidemic. In a number of cases the increase in the mean rate of 1918 over 1917 is probably statistically significant having regard to the probable errors involved. Thus we have the following large differences:

Cities under 25,000,	Total, 1918 mean to 1917 mean	= 12.37 ± 2.12
Cities over 25,000,	White, 1918 mean to 1917 mean	= 10.37 ± 2.84
Cities under 25,000,	White, 1918 mean to 1917 mean	= 16.16 ± 5.00

It is noteworthy that this increase in the infant mortality rate in 1918 is practically confined entirely to the cities. The rural counties, whether for white or colored or total population, show little or no change in 1918 as compared with 1917.

3. There is no unequivocal difference in the mean rates of infant mortality in the larger as compared with the smaller cities. Con-

sidering the largest differences in mean rates for total populations in cities of 25,000 and over, as compared with cities of under 25,000, there is no difference which is as much as even three times its probable error. This result, that there is no marked or striking difference in the mean rate of infant mortality in large as compared with small cities, is somewhat surprising. It suggests by inference that when the matter is adequately investigated it will probably be found that there is no definite or significant correlation between the rate of infant mortality and the density of population in American cities. It should be understood, however, that this is here suggested only as a probable inference. Positive statements on the matter cannot be made until the point has been carefully investigated by the method of multiple correlation.

4. The mean rates of infant mortality are notably smaller in the rural than in the urban areas. The fact has, of course, long been well known. The first writer on vital statistics, in the sense in which we now understand that subject, Captain John Graunt (1662), more than two hundred and fifty years ago pointed out that rural communities exhibited generally a lower rate of mortality than urban communities. We are still nearly as far as he was from a scientific understanding of *why* this is so. There has long been current a certain glib patter of explanation for the superiority of rural communities over urban in rate of mortality, but the subject still awaits careful analytic quantitative investigation, which will measure the relative influence of each one of the considerable number of factors which may be obviously directly concerned in producing this difference. The difference between urban and rural rates of infant mortality is reflected just as clearly in the high absolute rates of the colored population as it is in the lower rates of the white population.

5. The mean rates of infant mortality are, roughly speaking, something like twice as high for the colored population as for the white population in each of the demographic units considered, and at all times. This, again, is a fact in general well known, but here we have precise figures on the point, with probable errors, which show definitely how tremendously poorer the negro baby's chances of surviving the first year of life are than the white baby's.

6. The mode is seen in every case but one (1916, cities over 25,000, total) to be smaller than the mean; that is, to lie to the left of the mean in the distribution. The differences between mean and mode are fairly considerable for most of the distributions. They are largest in the case of the rural colored distributions.

7. The cities of over 25,000 exhibit distinctly less variation in respect of infant mortality than do either the smaller cities (under 25,000) or the rural counties. This is true, however the variation is measured, whether absolutely, in terms of standard deviation, or relatively, in percentage terms. The smaller cities and the rural counties exhibit about the same degree of variation relative to their means, but absolutely, in terms of standard deviation, the rural counties show less variability than the cities under 25,000. It is probable that in this case the coefficient of variation represents more truly the real biologic fact than the standard deviation. The colored distributions exhibit a much higher degree of variation in respect of infant mortality, however measured, whether absolute or relative, than do the white populations. Probably part of this greater variation in the colored population arises from the fact that these populations are absolutely much smaller in size in every case than the white populations of the same communities, and therefore less likely to give steady and characteristic rates. How much of the actually observed variation, however, is to be explained in this way is at present undeterminable. In general, it may fairly be assumed that the greater the variation exhibited by a given class of the community in respect of infant mortality, the greater the chance of effective control and reduction of the average infant mortality by administrative measures. There can be no question that there is no field which offers so great opportunities in this direction as the colored population.

8. The skewness is seen to be positive in sign in every case but one. In that case (1916, cities over 25,000, total) the skewness is not significant in comparison with its probable error. With this exception the curves tend to tail off more gradually and farther toward the right end than toward the left end of the range, and in consequence, as we have already seen, the mode lies to the left of the mean. In many instances (notably in the distributions for

the colored population of rural counties) the skewness values rise to considerable magnitudes and may be regarded as significantly different from zero, having regard to their probable errors. In other words, the rate of infant mortality in these different American demographic units tends generally to distribute itself in a substantially unsymmetric fashion about the mean, extremely high rates occurring more frequently than correspondingly low rates. This fact might perhaps be taken to indicate that the task confronting the administrative control of infant mortality in certain communities of the United States and yet to be accomplished is even greater than what has already been accomplished in the past, great and worthy of commendation as that is.

9. The kurtosis is seen to be positive in sign and relatively large in amount in most cases. This confirms analytically the conclusion already reached from mere inspection, namely, that the curves of variation in infant mortality are, with great uniformity, leptokurtic, that is, more sharply peaked than the corresponding normal curve would be.

10. A noteworthy point is the remarkable similarity, evident both from inspection and from the analytic constants, in all of these frequency distributions. Evidently infant mortality variation curves are of a quite definitely characteristic and uniform type, at least in this country. It will be interesting to examine similar curves for other countries.

For the sake of further analytic work the moments and certain derived constants from the longer distributions are given in Table 56.

It has seemed desirable, in the case of certain of the distributions, which may be fairly considered as typical representatives of all, to go on and fit the appropriate skew frequency curves to the observations. The results are shown in Figs. 63–67.

It is evident that the curves in general fit closely the observed facts. The greatest discrepancy between theory and observation is found in Fig. 67, the negro rates for rural counties. Here the observations are obviously rough, and probably the curve given is as satisfactory a result as could be obtained with such material.

TABLE 56

Analytic Constants for Infant Mortality Curves

Group.	N	ν_1†	μ_2†	μ_3†	μ_4†	β_1	$\sqrt{\beta_1}$	β_2
Cities over 25,000,* Total, 1915	98	2.7245	1.7081	1.1649	9.1772	.2723	.5218	3.1454
" " " " 1916	99	3.6263	1.5245	.3296	7.4924	.0307	.1751	3.2440
" " " " 1917	144	2.4792	1.3746	1.0234	7.7984	.4033	.6350	4.1274
" " " " 1918	144	2.8889	1.5710	1.1084	8.0719	.3168	.5629	3.2706
Cities under 25,000,* Total, 1915	153	4.5490	2.2734	1.5528	19.2617	.1804	.4247	3.4194
" " " " 1916	156	4.7115	2.6219	2.2304	25.6044	.2760	.5254	3.7246
" " " " 1917	236	3.4619	2.2415	3.3963	23.2314	1.0243	1.0121	4.6218
" " " " 1918	236	3.0805	3.5670	9.2407	79.0312	1.8302	1.3529	6.2116
Rural counties, Total, 1915	358	3.6536	1.4336	1.5898	10.6856	.8579	.9262	5.1995
" " " 1916	381	2.7638	1.6824	1.3432	9.5167	.3789	.6155	3.3623
" " " 1917	1127	3.6007	1.6525	1.8928	14.7237	.7939	.8910	5.3916
" " " 1918	1127	3.7214	2.0165	2.4618	17.4078	.7391	.8597	4.2810
Rural counties, White, 1917	232	2.8103	1.4583	1.2374	9.2685	.4937	.7027	5.8472
" " " 1918	234	2.7949	2.0883	2.0117	17.4371	.4444	.6666	3.9986
Rural counties, Colored, 1917	231‡	5.2381	8.2279	33.3269	476.3210	1.9940	1.4121	7.0359
" " " 1918	234	6.8632	10.9407	54.8500	866.4468	2.2973	1.5157	7.2386

* In 1910.

† It should be expressly noted that the values of the moments ν_1 (about arbitrary origin at the lower range end), and μ_2, μ_3 and μ_4 (about the mean) as here given, are in *units of grouping*, and not in terms of death rates. The unit of grouping, as can be seen in Table 17 is 20 deaths per 1000 births.

‡ These are the constants after dropping the single aberrant observation in the death rate class 600–619 deaths per 1000 births.

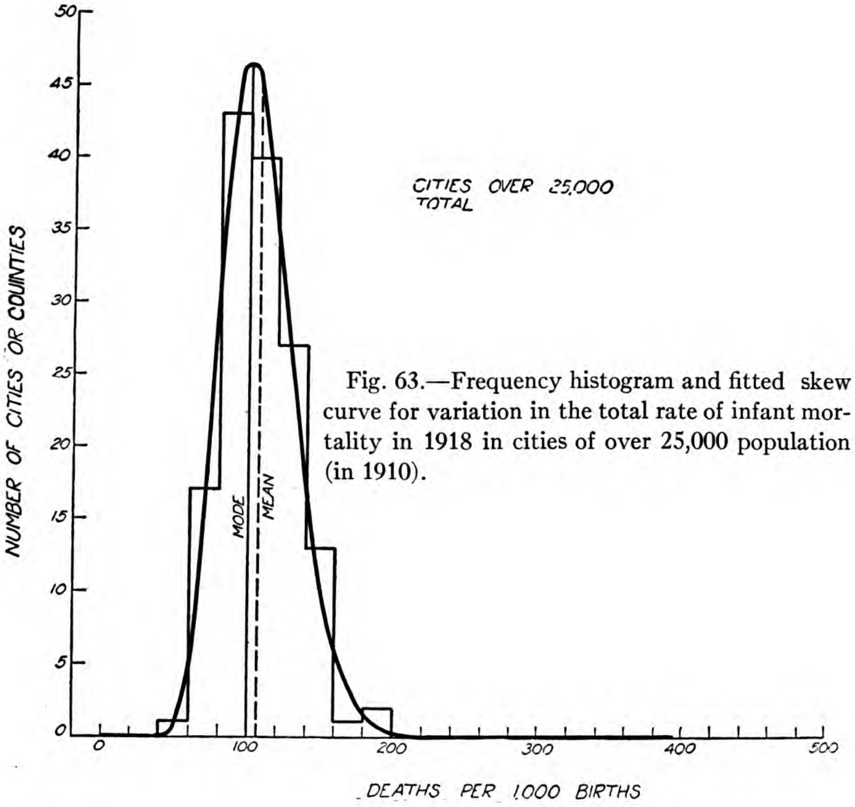


Fig. 63.—Frequency histogram and fitted skew curve for variation in the total rate of infant mortality in 1918 in cities of over 25,000 population (in 1910).

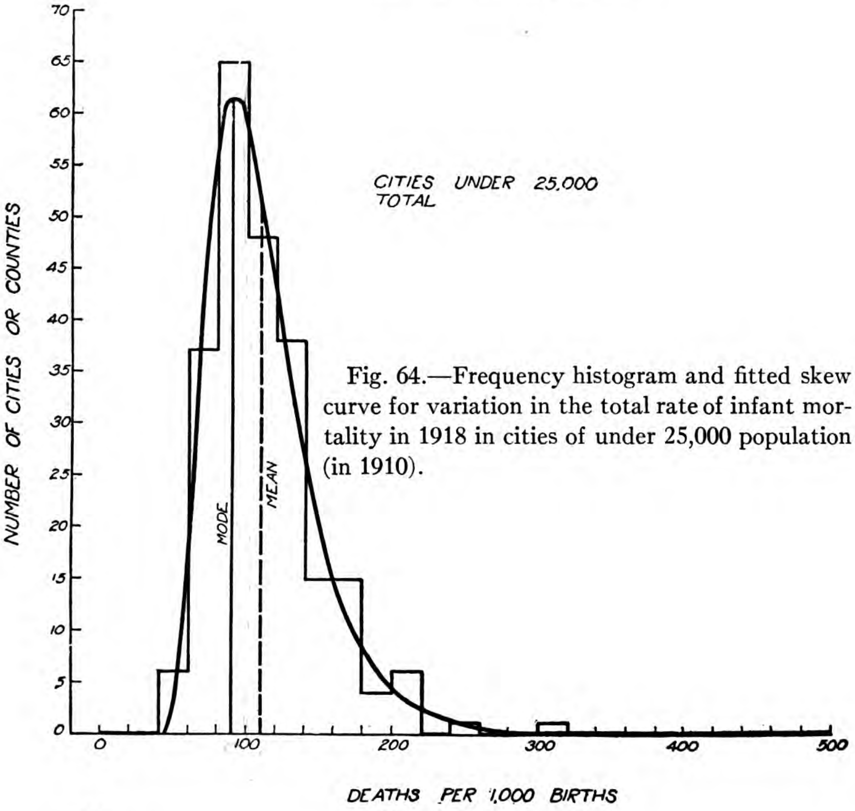


Fig. 64.—Frequency histogram and fitted skew curve for variation in the total rate of infant mortality in 1918 in cities of under 25,000 population (in 1910).

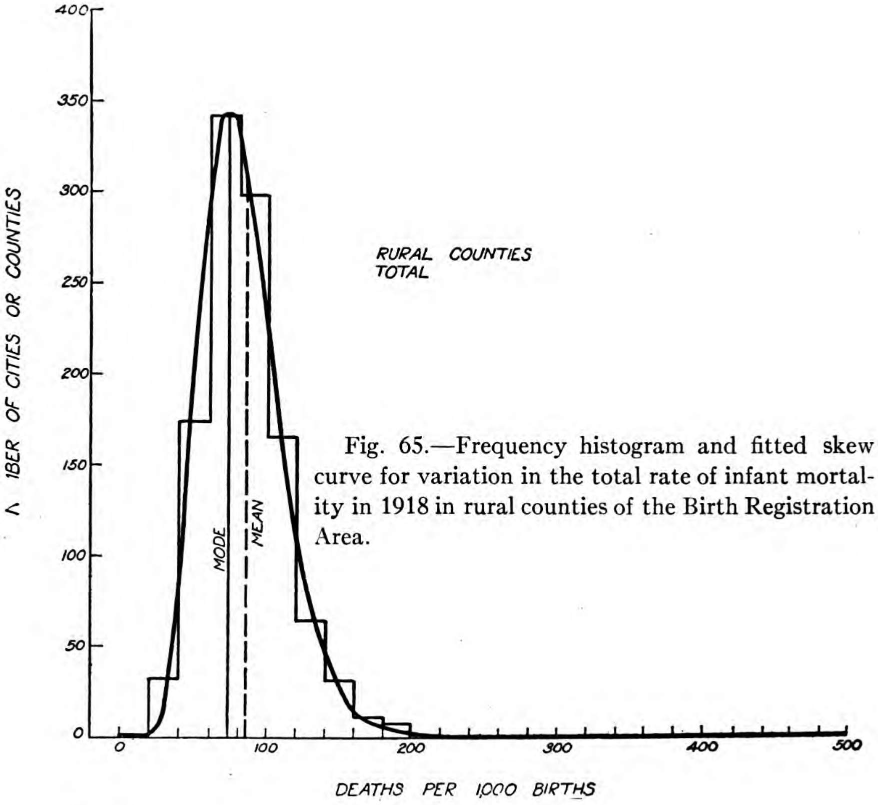


Fig. 65.—Frequency histogram and fitted skew curve for variation in the total rate of infant mortality in 1918 in rural counties of the Birth Registration Area.

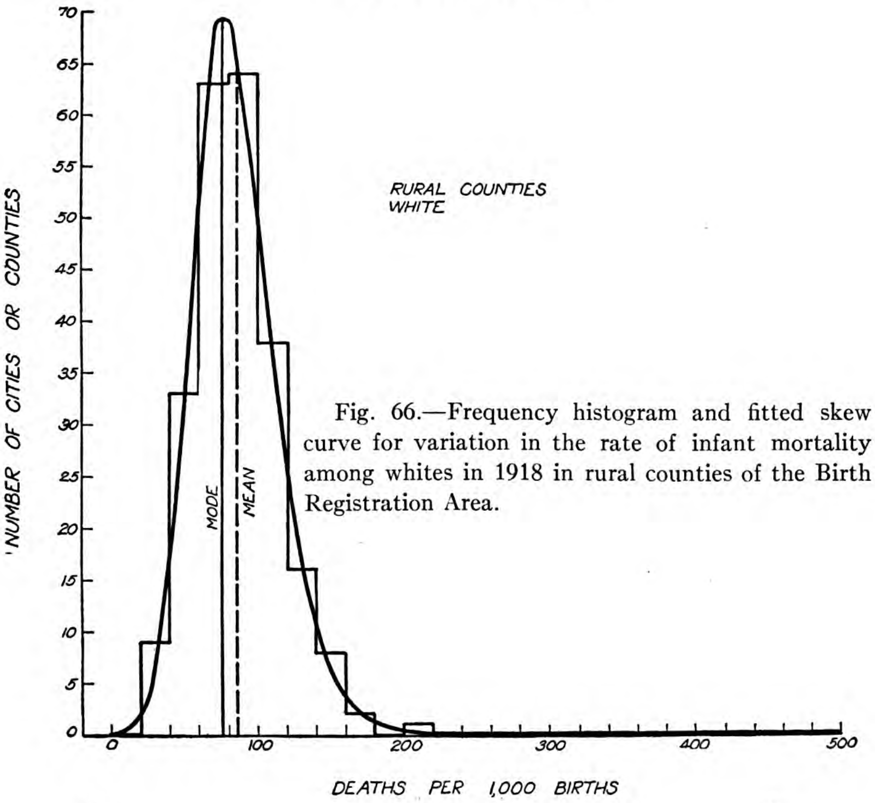


Fig. 66.—Frequency histogram and fitted skew curve for variation in the rate of infant mortality among whites in 1918 in rural counties of the Birth Registration Area.

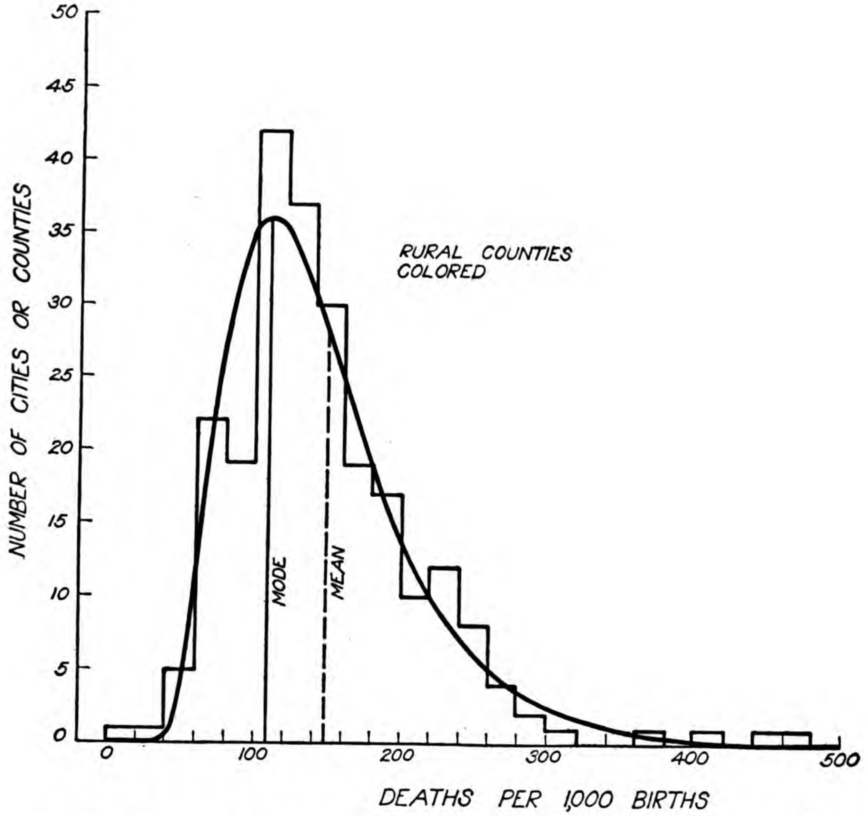


Fig. 67.—Frequency histogram and fitted skew curve for variation in the rate of infant mortality among negroes in 1918 in rural counties of the Birth Registration Area.

The equations for the curves are as follows:

Cities over 25,000, total:

$$\text{(Type I)} \quad y = 46.5502 \left(1 + \frac{x}{2.8946}\right)^{4.4296} \left(1 - \frac{x}{14.5155}\right)^{22.2126}$$

Cities under 25,000, total:

$$\text{(Type VI)} \quad y = 8.8238 \times 10^{35} (x - 18.4653)^{2.9924} x^{-26.7505}$$

Rural counties, total:

$$\text{(Type VI)} \quad y = 9.2560 \times 10^{70} (x - 17.7766)^{8.9630} x^{-55.2026}$$

Rural counties, white:

$$\text{(Type IV)} \quad y = 0.005609 \left[1 + \frac{x^2}{(4.5095)^2}\right]^{-12.5407} e^{25.9862 \tan^{-1} \frac{x}{4.5095}}$$

Rural counties, colored:

$$\text{(Type VI)} \quad y = 4.2887 \times 10^{27} (x - 20.9629)^{3.0627} x^{-19.9804}$$

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. 169–212, 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 matter 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.

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 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 go back and 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 57 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 like that exhibited in Table 57.*

* 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.

TABLE 57

RELATION BETWEEN THE NUMBER OF HEADS FALLING IN SUCCESSIVE RANDOM TOSSES OF 12 PENNIES 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				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

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 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 58.

TABLE 58

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

It is at once evident that this Table 58 is not quite like Table 57. The frequencies are tending, very slightly but still evidently, to concentrate along a diagonal from the upper left to the lower right corners of the table.

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 59 shows the result of marking 2 pennies and leaving them down; Table 60, of marking 3 pennies and leaving them down, and so on up to all 12 pennies.

TABLE 59

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.

Heads in second 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

TABLE 60

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.

Heads in second 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	25	22	9	3	1			92
6		1		10	16	17	25	28	22	5	1			125
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

TABLE 61

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.

Heads in second 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	4	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

TABLE 62

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.

Heads in second 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

TABLE 63

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.

Heads in second 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

TABLE 64

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.

Heads in second 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

TABLE 65

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

TABLE 66

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

TABLE 67

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

TABLE 68

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

TABLE 69

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 second toss.	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

In this series of tables is seen the genesis of correlation. In Table 57 the results of the first toss have no influence on the results of the second. There is no correlation between them. In Table 69 the results of the first toss completely determine, *or cause*, the results of the second. This gives perfect correlation—or causation —between the two.

In all the tables the diagonal lines cut off the cells in which events cannot possibly happen.

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 70. This table is taken from a paper by the present writer,* the original data having been collected by Matiegka.†

TABLE 70

Correlation Between Brain-weight and Skull Length. Bohemian Males, Twenty to Fifty-nine Years of Age

		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			
Skull length (mm.)	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			

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

* 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.

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 like 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 70. 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. 68.

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. 68 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 would be no increase in brain-weight associated with an increase in

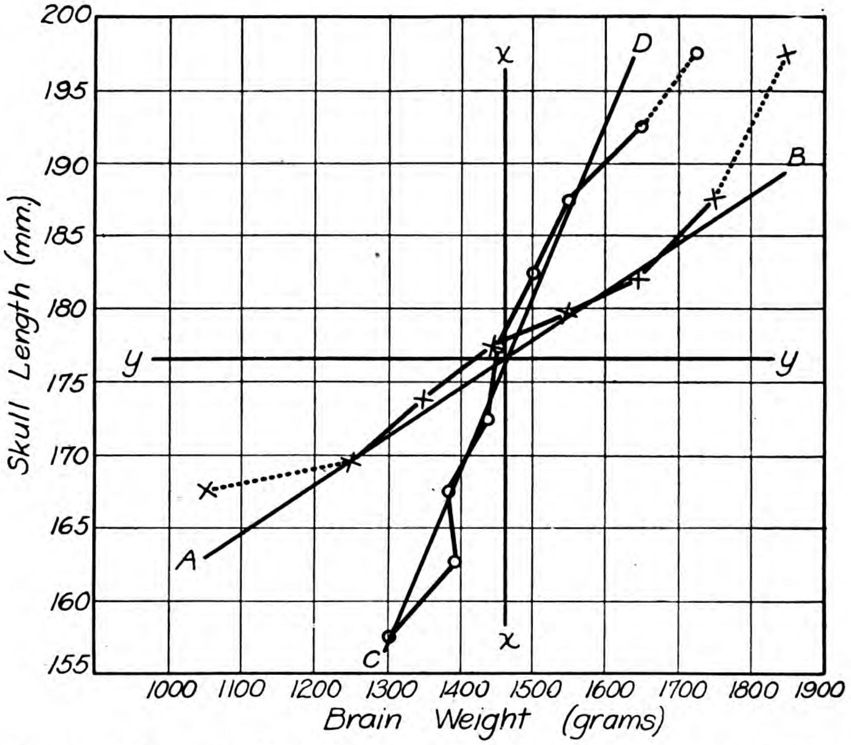


Fig. 68.—Observed and calculated regressions for brain-weight and skull length from Table 70. The crosses are the means of the observed skull length arrays (observed regression of skull length on brain-weight). *AB* is the calculated regression line of skull length on brain-weight. The circles are the 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.

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

$$\text{P. E.}_{\cdot r} = .67449 \frac{1 - r^2}{\sqrt{N}}.$$

For very small numbers ($N < 25$) special caution must be used in estimating the reliability of a correlation coefficient.

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.[1]

As an example we may take Table 70 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 71.

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 71

Showing the Steps in the Calculation of a Correlation Coefficient

Skull length (mm.)	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						
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 deviations for the two variables. Proceeding just as in Chapter XIII, we have:

$$\nu_{1x} = \frac{S\ (Z_x x)}{S\ (Z_x)} = \frac{41}{299} = .137124$$

$$\nu_{2x} = \frac{S\ (Z_x x^2)}{S\ (Z_x)} = \frac{429}{299} = 1.434783$$

Omitting Sheppard's corrections for the sake of simplicity, we then have

$$\pi_{2x} = 1.434783 - (.137124)^2 = 1.415980,$$

whence

$$\sigma_x = \sqrt{\pi_{2x}} = 1.189950 \text{ in class units.}$$

We then have

$$\text{Mean brain-weight} = 1450 + (100 \times .1371) = 1463.71 \pm 4.64 \text{ grams.}$$

$$\begin{array}{l}\text{Standard deviation}\\ \text{(in brain-weight)}\end{array} = 100 \times 1.18995 = 119.00 \pm 3.28 \text{ grams.}$$

Similarly for skull length we have:

$$\nu_{1y} = \frac{S\ (Z_y y)}{S\ (Z_y)} = \frac{242}{299} = .809365$$

$$\nu_{2y} = \frac{S\ (Z_y y^2)}{S\ (Z_y)} = \frac{870}{299} = 2.909699$$

$$\pi_{2y} = 2.909699 - (.809365)^2 = 2.254627$$

$$\sigma_y = \sqrt{2.254627} = 1.501542 \text{ in class units.}$$

$$\text{Mean skull length} = 172.5 + (.809365 \times 5) = 176.55 \pm .29 \text{ mm.}$$

$$\begin{array}{c}\text{Standard deviation}\\ \text{(in skull length)}\end{array} = 5 \times 1.501542 = 7.51 \pm .21 \text{ mm.}$$

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

$$\begin{aligned} 1 \times (-2) &= -2 \\ 1 \times (-1) &= \underline{-1} \\ \text{Sum} &= -3 \end{aligned}$$

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:

$$\begin{aligned} 2 \times (-2) &= -4 \\ 6 \times (-1) &= -6 \\ 4 \times 0 \quad &= \quad 0 \\ 2 \times (+1) &= \quad \underline{2} \\ \text{Sum} &= -8 \end{aligned}$$

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_x) 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

$$\begin{array}{l} (-3) \times (-\ 3) = +\ 9 \\ (-2) \times (-\ 8) = +16 \\ (-1) \times (-27) = +27 \\ 0 \quad\ \times (-\ 7) = \quad 0 \\ (+1) \times (+\ 4) = +\ 4 \end{array}$$

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

$$\begin{aligned} A = \frac{S\ (z_{xy}xy)}{N} - \nu_{1x}\nu_{1y} &= \frac{+312}{299} - (.137124 \times .809365) \\ &= 1.043478 - .110983 \\ &= +\ .932495 \end{aligned}$$

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

$$r_{12} = \frac{+ .932495}{1.786760} = +.522 \pm .028.$$

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. 68 *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 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,

$$\bar{x} = (\bar{X} - 1463.7)$$

$$y = (Y - 176.5)$$

So then,

$$\overline{X} - 1463.7 = 8.272\,(Y - 176.5).$$

Simplifying, we get

$$\text{Brain-weight (in grams)} = 3.7 + 8.272 \text{ skull length (in mm.).}$$

This is the equation of the regression line *CD* of Fig. 68. 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

$$b_2 = r_{12}\frac{\sigma_2}{\sigma_1} = .521892\,\frac{7.508}{118.995} = .033.$$

$$\overline{y} = .033\,x$$

$$\overline{Y} - 176.5 = .033\,(X - 1463.7)$$

$$\text{Skull length (in mm.)} = 128.2 + .033 \text{ brain-weight (in grams).}$$

This is the equation of the line *AB* in Fig. 68.

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 72 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 72

Correlation in Man

Correlated Characters.	Coefficient of correlation.
Age (adults) and temperature (oral) [1]	−.150±.022
Age (adults) and pulse rate [1]	+.121±.022
Age (adults) and respiration rate [1]	+.077±.022
Age (adults) and body weight [1]	+.136±.030
Temperature (oral) and pulse rate [1]	+.288±.020
Temperature (oral) and respiration rate [1]	+.142±.022
Temperature (oral) and height [1]	+.003±.022
Temperature (oral) and body weight [1]	+.043±.022
Pulse rate and respiration rate [1]	+.060±.022
Pulse rate and height [1]	−.078±.022
Pulse rate and body weight [1]	+.114±.022
Respiration rate and height [1]	−.144±.022
Respiration rate and body weight [1]	−.089±.022
Corrected death rates from (*a*) cancer of the liver, and (*b*) cancer of the stomach (Switzerland) [2]	+.161±.140
Corrected death rates from (*a*) cancer of the stomach, and (*b*) cancer of the rectum and intestines (Switzerland) [2]	+.263±.134
Occupation and cancer mortality (occupied and retired males, 1900–2, weighted) [3]	+.40 ±.06
Weight and length of infants at birth (males) [4]	+.644±.012
Body weight and height (adult males) [4]	+.486±.016
Strength of pull and height (adult males) [4]	+.303±.019
Strength of pull and body weight (adult males) [4]	+.545±.015
Length of first joint of forefinger in (*a*) right hand, and (*b*) left hand [5]	+.925±.004
Stature in (*a*) brother and (*b*) sister [6]	+.375±.017
Cephalic index in (*a*) brother, and (*b*) sister [6]	+.340±.050
Birth rate and infant death rate (London, 1901) [7]	+.51 ±.10
Birth rate and poverty rate [8]	+.420±.047
Infant mortality and artificial feeding rate [8]	+.760±.029
Heart weight and body weight [9]	+.65 ±.04
Heart weight and kidney weight [9]	+.56 ±.05
Heart weight and liver weight[9]	+.52 ±.06
Heart weight and brain weight [9]	+.08 ±.08
Obstetric conjugate and inter-crests diameters of pelvis [10]	+.17 ±.04
Obstetric conjugate and inter-spines diameters of pelvis [10]	+.13 +.05
Obstetric conjugate and transverse diameters of pelvis [10]	+.07 ±.05
Obstetric conjugate and diagonal conjugate diameters of pelvis [10]	+.91 ±.01
Obstetric conjugate and antero-posterior diameters of pelvis [10]	+.30 ±.04
Duration of life of (*a*) father, and (*b*) adult son [11]	+.135±.021
Duration of life of (*a*) father and (*b*) minor son [11]	+.087±.022
Duration of life of (*a*) father, and (*b*) adult daughter [11]	+.130±.020
Duration of life of (*a*) mother, and (*b*) adult son [11]	+.131±.019
Duration of life of (*a*) mother, and (*b*) adult daughter [11]	+.149±.020
Duration of life of (*a*) adult brother and (*b*) adult brother [11]	+.285±.020
Duration of life of (*a*) adult sister and (*b*) adult sister [11]	+.332±.019
Vaccination and recovery from smallpox [12]	+.656±.009
Lung capacity and body weight (age 19, males) [13]	+.62 ±.02
Number of decayed teeth and use of tooth-brush (boys) [14]	+.074±.030
Mean age at death of (*a*) husband, and (*b*) wife [15]	+.224±.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 CORRELATION 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 approached even. 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 called these cases of non-linear regression *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, proposes 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 70 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 70 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 plainly is the ratio of reduction of average variability of an array below the variability of the sample as a whole. Now one can see by studying again Table 57 to 69 *supra* that when the correlation or association between the two variables is *high* $\sigma_{ax}{}^2$ 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*, $\sigma_{ax}{}^2$ 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

$$\sigma_{mx} = \sigma_x - \sigma_{ax}$$

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 70. The work is shown in Table 73.

TABLE 73

Calculation of Correlation Ratio from Data of Table 70

Skull length classes.	Means of the x arrays (brain-weight).	x	x^2	Z_x	$Z_x x^2$
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

$$\sigma_{mx} = \sqrt{\frac{1272513}{299}} = \sqrt{4255.896} = 65.237$$

$$\sigma_x = 118.995 \text{ (from p. 305 } \textit{supra}\text{)}$$

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x} = \frac{65.237}{118.995} = .548$$

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 74.

TABLE 74

Alternative Calculation of Correlation Ratio from Data of Table 70

Brain-weight classes.	Means of the y arrays.	y	y^2	Z_y	$Z_y y^2$
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 ± .028. Thus it appears from analytic, as well as visual evidence, that the regressions of Table 70 are strictly linear.

Let us take another example where the regression is more evidently non-linear. Such a case is furnished in Table 75, 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 75

CORRELATION BETWEEN WEIGHT AND SITTING HEIGHT OF EMBRYOS BELOW 400 GRAMS IN WEIGHT

(Weight in grams).

Sitting height in mm.	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	Totals
30- 44	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5
45- 59	38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38
60- 74	40	22	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	62
75- 89	-	32	17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	49
90- 104	-	-	23	27	8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	58
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	83	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 76.

TABLE 76

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_x	$Z_x \times (m_x - M_x)^2$
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...............	128.2464 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

$$\begin{aligned} \eta &= .9804 \\ r &= \underline{.9440} \\ \text{Difference} &= \underline{.0364} \end{aligned}$$

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,

$$\text{P. E.}_{\cdot\zeta} = 2\,\chi_1 \sqrt{\zeta}\,\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1},$$

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:

$$\text{P. E.}_{\cdot\zeta} = 2 \times .03166 \times .2648\sqrt{.0388^2 - .1089^2 + 1}$$

$$= .01677\sqrt{.9896} = .01677 \times .9948 = .017$$

$$\zeta = \eta^2 - r^2 = .961 - .891 = .070 \pm .017$$

ζ is 4.1 times its probable error, and quite certainly significant. We may then conclude that the regression of sitting height on weight is certainly 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 $(\kappa - 1)/N$, 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] lately has returned to a consideration of the subject and 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 all 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 precisely the same sort and meaning as 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 constant conditions or values of the other variables in the complex, while any other selected one varies.

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.*, $b_{12.3\ldots n}$ and $b_{21.3\ldots n}$ 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 $r_{12.34\ldots n}$ is defined by the relation

$$r_{12\cdot 34\ldots n} = (b_{12\cdot 34\ldots n} \cdot b_{21\cdot 34\ldots n})^{\frac{1}{2}}. \tag{2}$$

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

$$r_{12\cdot 34\ldots n}$$

the variables represented by the secondary subscripts $34\ldots.n$ *are held constant*, 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

$$r_{12\cdot 3} = -\ .8955.$$

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 table then constructed, *for these selected eggs*, between length and

* 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.

breadth. There were 42 such eggs and the table is shown as Table 77.

TABLE 77

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 78 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 78

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

value. Table 79 gives the length-breadth correlation for 13 eggs each having a weight between 62 and 62.9 grams, that is, heavy eggs.

TABLE 79

Correlation Between Egg Length and Breadth for Eggs Weighing 62 to 62.9 Grams

Egg breadth (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

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:

$$\begin{array}{rcrcr}
-.9117 & \times & 42 & = & -38.2914 \\
-.8911 & \times & 46 & = & -40.9906 \\
-.8739 & \times & \underline{13} & = & \underline{-11.3607} \\
\text{Totals} & & 101 & & -90.6427
\end{array}$$

Whence

$$\begin{array}{rcr}
\text{Mean } r_{12} & = & -.8975 \\
\text{(By partial correlation) } r_{12\cdot3} & = & \underline{--.8955} \\
\text{Difference} & = & .0020
\end{array}$$

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 $r_{12.3}$.

The only difference between the perfectly simple biologic procedure, which anyone can understand, of selecting individuals alike in respect of n variable 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,

$$r_{12\cdot34\ldots\ldots n} = \frac{r_{12\cdot34\ldots\ldots(n-1)} - r_{1n\cdot34\ldots\ldots(n-1)} \, . \, r_{2n\cdot34\ldots\ldots(n-1)}}{(1 - r^2_{1n\cdot34\ldots\ldots(n-1)})^{\frac{1}{2}}(1 - r^2_{2n\cdot34\ldots\ldots(n-1)})^{\frac{1}{2}}}, \qquad (3)$$

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,

$$r_{12\cdot 3} = \frac{r_{12} - r_{13}\cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}}\,(1 - r^2_{23})^{\frac{1}{2}}} \qquad (5)$$

The zero order correlations r_{12}, r_{13}, and r_{23} will be calculated from the observed correlation tables like Table 70. 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: $r_{12.4}$, $r_{12.5}$, $r_{13.2}$, $r_{13.4}$, $r_{13.5}$, $r_{14.2}$, $r_{14.3}$, $r_{14.5}$, $r_{15.2}$, $r_{15.3}$, $r_{15.4}$, $r_{23.1}$, $r_{23.4}$, $r_{23.5}$, $r_{24.1}$, $r_{24.3}$, $r_{24.5}$, $r_{25.1}$, $r_{25.3}$, $r_{25.4}$, $r_{34.1}$, $r_{34.2}$, $r_{34.5}$, $r_{35.1}$, $r_{35.2}$, $r_{35.4}$, $r_{45.1}$, $r_{45.2}$, $r_{45.3}$. Each one of these can be determined from the zero order coefficient just as $r_{12.3}$ 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 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 my 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 80, which include 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 80. 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 80

PARTIAL CORRELATIONS. INFLUENZA. ZERO AND FIRST ORDER COEFFICIENTS

r 0 Order		$(1 - r^2)^{\frac{1}{2}}$.	Product term of numerator.	Whole numerator.	Denominator.	*r* First order.	
Subscript.	Coefficient.					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 81, which is of exactly the same form as Table 80, 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 82 gives the third order coefficient in which we are interested, again calculated in two ways as a check.

TABLE 81

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....... 25.4....... 35.4.......	+.5965 −.0269 −.1220	..... .9996 .9925	+.0033	+.5932	.9921	23.45	+.5979
23.5....... 24.5....... 34.5.......	+.4896 +.0205 +.6022	..... .9998 .7983	+.0123	+.4773	.7981	23.45	+.5980
23.4....... 26.4....... 36.4.......	+.5965 +.1198 −.0220	..... .9928 .9998	−.0026	+.5991	.9926	23.46	+.6036
23.6....... 24.6....... 34.6.......	+.5148 +.0514 +.5960	..... .9986 .8030	+.0306	+.4842	.8019	23.46	.+6038
23.5....... 26.5....... 36.5.......	+.4896 +.1114 −.1592	..... .9938 .9872	−.0177	+.5073	.9811	23.56	+.5171
23.6....... 25.6....... 35.6.......	+.5148 −.0314 −.1679	..... .9995 .9858	+.0053	+.5095	.9853	23.56	+.5171
25.4....... 26.4....... 56.4.......	−.0269 +.1198 −.0126	.9996 .9928 .9999	−.0015 +.0003	−.0254 +.1195	.9927 .9995	25.46 26.45	−.0256 +.1196
24.5....... 26.5....... 46.5.......	+.0205 +.1114 −.2337	.9998 ..... .9723	−.0048	+.1162	.9721	26.45	+.1195
24.6....... 25.6....... 45.6.......	−.0514 −.0314 −.1181	.9986 ..... .9930	−.0061	−.0253	.9916	.25.46	−.0255
35.4....... 36.4....... 56.4.......	−.1220 −.0220 −.0126	.9925 .9998 .9999	+.0003 +.0015	−.1223 −.0235	.9997 .9924	35.46 36.45	−.1223 −.0237
34.5....... 36.5....... 46.5.......	+.6022 −.1592 −.2337	.7983 ..... .9723	−.1407	−.0185	.7762	36.45	−.0238
34.6....... 35.6....... 45.6.......	+.5960 −.1679 −.1181	.8030 ..... .9930	−.0704	−.0975	.7974	35.46	−.1223

TABLE 82

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...... 26.45...... 36.45......	+.5979 +.1195 −.0237	..... .9928 .9997	−.0028	+.6007	.9925	23.456	+.6052
23.46...... 25.46...... 35.46......	+.6037 −.0255 −.1223	..... .9997 .9925	+.0031	+.6006	.9922	23.456	+.6053

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 sort of 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.

The only method which science offers in the premises is that which Newton followed in discovering the law of gravitation, namely, to fit a curve to the observations, and use the equation of the curve as the expression of the law. Newton studied the observed positions of the planets, relative to the sun and to each other, and found that these observed positions could be fitted with great exactness by a family of curves based upon the assumption that between each of the heavenly bodies there is an attraction, having at any moment a force inversely proportional to the squares of their distances from each other, and directly proportioned to their masses. He called this relationship the law of gravitation.

It is doubtless too much to hope that this chapter will make Newtons out of all its readers, but 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; whereas actually if he does what has been described he emphasizes the fluctuations and 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 I have said, only on his 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. I propose 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 I shall use 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 mine, 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.

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

$$\begin{aligned} S\,(y) \;- n\,a \;\;\; - b\,S\,(x) &= 0 \\ S\,(xy) - a\,S\,x - b\,S\,(x^2) &= 0 \end{aligned}$$

Transposing terms in form for computation these become

$$\begin{aligned} n\,a + b\,S\,(x) &= S\,(y) \\ a\,S\,(x) + b\,S\,(x^2) &= S\,(xy), \end{aligned}$$

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 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 delightfully simple this all is will be to work out an example. This is done in Table 83. The data are drawn from Table 75 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.

TABLE 83

MEAN SITTING HEIGHTS OF EMBRYO. CURVE FITTING

Weight of embryo in grams.	Mean sitting height in mm. y	x	xy	xy^2	$y \log x$.	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	2,580.3	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		

From this table, and a table of the sums of the powers of the natural numbers, we have,

$$
\begin{aligned}
n &= 20 \\
S(x) &= 210 \\
S(x^2) &= 2870 \\
S(y) &= 2673.7 \\
S(xy) &= 31640.5
\end{aligned}
$$

Whence the equations are

$$
\begin{aligned}
20\,a + 210\,b &= 2673.7 \\
210\,a + 2870\,b &= 31640.5
\end{aligned}
$$

Solving, we get

$$
\begin{aligned}
a &= 77.37 \\
b &= 5.36 \\
y &= 77.37 + 5.36\,x.
\end{aligned}
$$

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. 69.

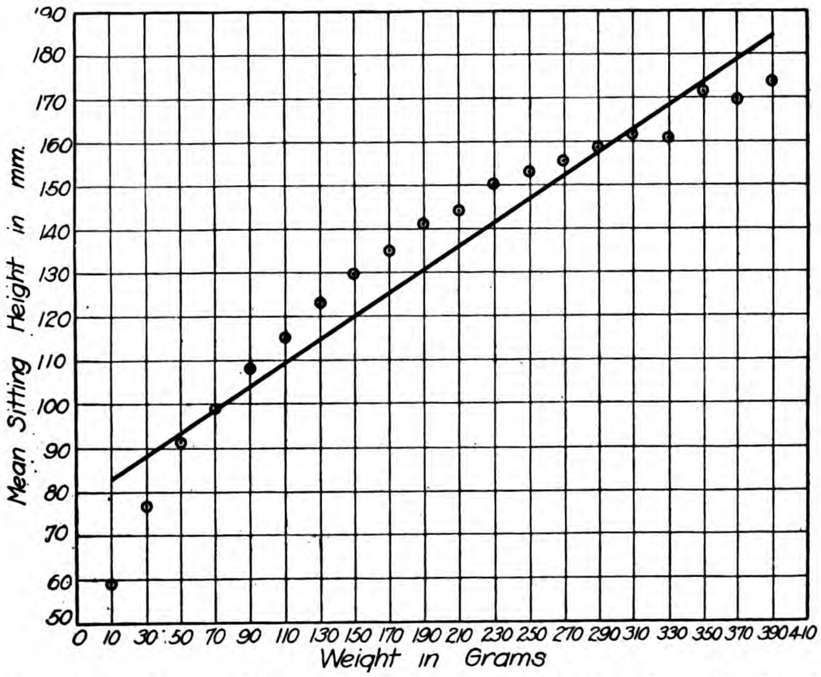


Fig. 69.—Observed mean sitting heights of embryos (circles) and straight line fitted by least squares.

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. 317).

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 83 these become

$$\begin{aligned} 20\,a + 210\,b + 2870\,c &= 2673.7 \\ 210\,a + 2870\,b + 44{,}100\,c &= 31{,}640.5 \\ 2870\,a + 44{,}100\,b + 722{,}666\,c &= 453{,}700.3 \end{aligned}$$

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 83. 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. 70.

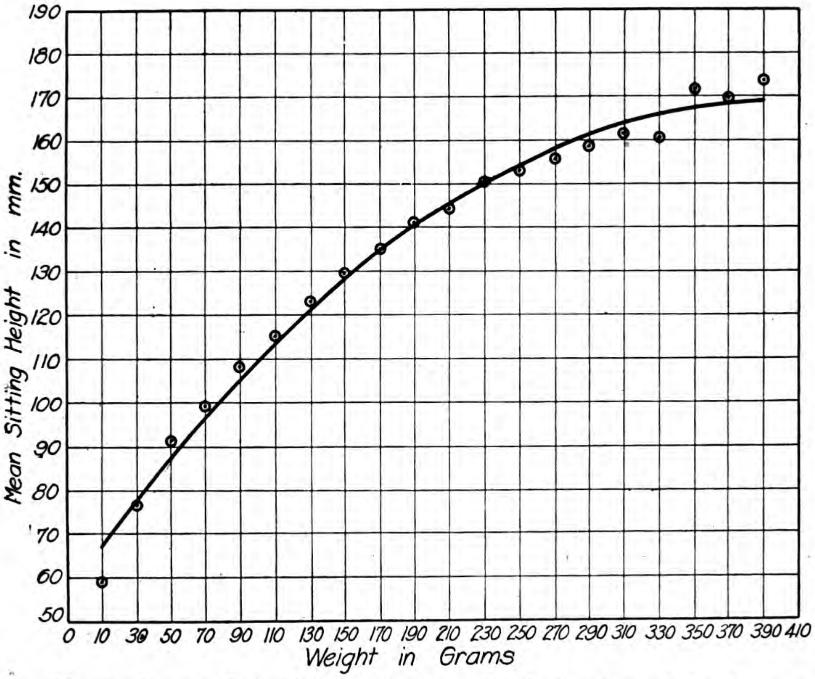


Fig. 70.—Observed mean sitting height of embryo (circles) and parabola of the second order fitted by least squares.

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 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:

$$\begin{aligned} n\,a + b\,S\,(x) \quad & + c\,S\,(\log x) &= S\,(y) \\ a\,S\,(x) + b\,S\,(x^2) \quad & + c\,S\,(x \log x) &= S\,(xy) \\ a\,S\,(\log x) + b\,S\,(x \log x) & + c\,S\,(\log x)^2 &= S\,(y \log x) \end{aligned}$$

The numerical values here are again drawn from Table 83 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

$$\begin{aligned} 20\,a + 210\,b \quad & + \;\; 18.3861246\,c = 2673.7 \\ 210\,a + 2870\,b \quad & + 230.0033043\,c = 31640.5 \\ 18.3861246\,a + 230.0033043\,b & + \;\; 19.2694686\,c = 2677.6433 \end{aligned}$$

Solving, we have

$$= 54.347 + 1.555\,x + 68.549 \log x$$

Substituting successive values of x as before and solving for y gives the values in the last column of Table 83, which are shown graphically in comparison with the observations in Fig. 71.

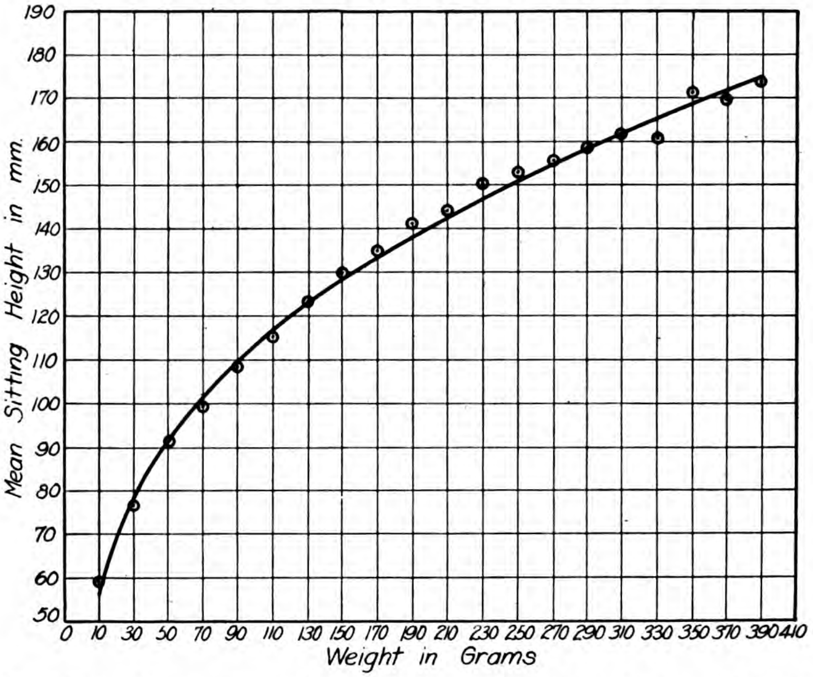


Fig. 71.—Observed mean sitting heights of embryo (circles), and a logarithmic curve fitted by least squares.

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 roundabout process of trial and error to get a proper value of a. It cannot be determined directly by the least square method.

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.

APPENDIX I

AGE AND SEX SPECIFIC DEATH-RATES FOR THE UNITED STATES REGISTRATION AREA (EXCLUSIVE OF NORTH CAROLINA) IN 1910

As a matter of reference, and because they are nowhere available for the United States Registration Area, it is thought desirable to present age and sex specific death-rates for each of the principal causes of death in the International List, as it existed in 1910. These rates have been for some years used in manuscript form in this laboratory.

All the rates are per 1000 living in the designated class, and are tabled to four places of decimals. The rates are, in general, not significant to this degree, but it seems desirable to have the extra figures, in case one wishes to make derivative use of the rates involving computations, so that a merely arithmetic error may not accumulate in the last place which is significant. Furthermore, in many of the causes of death where the rates are inherently low, their definite and orderly trend is better shown with the four place figures than it would be with figures corrected to the last significant place.

In order that the reader may judge, in every case, the extent to which the rate is significant, the following preliminary table is introduced, which shows *how large a rate has to be in order to be as much as three times the probable error different from zero.*

The relationship between p and n, when p is three times its probable error is

$$p = 3\,(.67449)\sqrt{\frac{pq}{n}} = 2.02347\sqrt{\frac{pq}{n}}$$

$$= \frac{4.09443}{n + 4.09443}$$

Population of Registration Area (Exclusive of North Carolina) 1910

Age.	Males.	Probably significant death-rate per 1000.	Females.	Probably significant death-rate per 1000.
All ages..........	27,340,093	.0001	25,920,337	.0002
Under 1..........	578,096	.0071	563,247	.0073
1- 4...........	2,167,541	.0019	2,123,575	.0019
5- 9...........	2,492,138	.0016	2,455,673	.0017
10–14...........	2,384,739	.0017	2,362,477	.0017
15–19...........	2,476,198	.0017	2,504,352	.0016
20–24...........	2,700,695	.0015	2,622,154	.0016
25–29...........	2,623,398	.0016	2,386,018	.0017
30–34...........	2,295,174	.0018	2,063,546	.0020
35–39...........	2,108,397	.0019	1,906,943	.0021
40–44...........	1,802,044	.0023	1,602,854	.0026
45–49...........	1,523,143	.0027	1,360,843	.0030
50–54...........	1,289,199	.0032	1,146,788	.0036
55–59...........	902,617	.0045	829,995	.0049
60–64...........	709,372	.0058	687,427	.0060
65–69...........	525,942	.0078	525,080	.0078
70–74...........	347,919	.0118	360,859	.0113
75–79...........	204,211	.0200	221,533	.0185
80–84...........	95,295	.0430	110,206	.0372
85–89...........	34,681	.1180	43,540	.0940
90–94...........	8,319	.4919	11,423	.3583
95–99...........	1,414	2.8875	2,266	1.8037
100 or over.....	259	15.5682	494	8.2217

It is to be understood that whenever in the following tables a rate is larger than those given in the third and fifth columns of the above table, it is more than three times its probable error different from zero.

Specific Death-rates per 1000 Population. Registration Area, 1910, Exclusive of North Carolina

Diseases.	Males.																					
	Under 1.	1–4.	5–9.	10–14.	15–19.	20–24.	25–29.	30–34.	35–39.	40–44.	45–49.	50–54.	55–59.	60–64.	65–69.	70–74.	75–79.	80–84.	85–89.	90–94.	95–99.	100 or over.
All causes	147.9927	15.0913	3.7109	2.4988	4.1394	5.9685	6.7725	8.0007	9.7463	11.6079	14.5390	18.5123	25.6620	36.0728	51.4410	75.0979	112.2369	168.1410	237.9112	313.0184	410.1839	494.2085
1 Typhoid fever	.0623	.1250	.1324	.1493	.3679	.4854	.4460	.3647	.3073	.2536	.2593	.2529	.2515	.2002	.1939	.1696	.1812	.1049	.0577	.3606		
4 Malaria	.0484	.0226	.0116	.0034	.0141	.0156	.0126	.0157	.0152	.0150	.0197	.0256	.0465	.0578	.0551	.0504	.1273	.1259	.2307	.2404	1.4144	
5 Smallpox	.0311	.0060	.0012	.0017	.0048	.0030	.0046	.0052	.0028	.0055	.0053	.0023	....	.0028	.0038	.0029	.0049	.0105				
6 Measles	1.5032	.9093	.1196	.0243	.0222	.0130	.0053	.0083	.0066	.0061	.0072	.0039	.0066	.0042	.0095	.0115	.0049	.0315				
7 Scarlet fever	.2889	.7299	.3330	.0805	.0424	.0252	.0126	.0113	.0066	.0039	.0020	.0016	.0033									
8 Whooping-cough	2.7902	.4683	.0365	.0042	.0016	.0015	.0008	....	....	.0006	.0007	....	....	....	....	.0029	.0098					
9 Diphtheria and croup	.8805	1.5418	.5710	.1342	.0505	.0237	.0160	.0118	.0100	.0089	.0079	.0140	.0100	.0085	.0019	.0029	.0098					
10 Influenza	.5068	.0900	.0225	.0143	.0262	.0244	.0282	.0344	.0484	.0605	.0893	.1241	.2094	.3158	.5856	1.1324	1.9734	3.4524	5.7668	8.2943	9.9010	15.4440
13 Cholera nostras	.1090	.0138	.0036	.0017	.0008	.0007	.0008	.0017	.0028	.0039	.0066	.0093	.0122	.0324	.0209	.0891	.1518	.1994	.4613	.3606	.7072	
14 Dysentery	.6798	.1449	.0092	.0029	.0028	.0041	.0053	.0083	.0138	.0266	.0230	.0411	.0631	.0789	.1863	.3219	.6709	1.4271	1.8742	2.7648	6.3649	3.8610
17 Leprosy	....	....	....	....	....	.0004	.0004	....	....	....	.0013	.0016										
18 Erysipelas	.5899	.0129	.0012	.0042	.0069	.0081	.0145	.0231	.0351	.0499	.0643	.0745	.0997	.1212	.2091	.2386	.3575	.5142	.9227	1.0819		
19 Other epidemic diseases	.0692	.0161	.0040	.0038	.0012	.0019	.0004	.0013	.0014	....	.0007	.0023										
20 Purulent infection and septicemia	.2180	.0217	.0185	.0172	.0182	.0148	.0198	.0266	.0356	.0455	.0532	.0752	.0931	.0874	.1312	.1667	.1714	.2623	.3748	.1202		
21 Glanders	....	....	....	....	....	.0004	.0008	....	.0014	....	....	....	.0011									
22 Anthrax	.0017	....	....	....	.0008	.0007	....	.0013	.0005	....	.0013	.0039	....	.0014	....	.0029	.0049					
23 Rabies	....	.0028	.0036	.0025	.0012	.0004	.0004	.0004	.0019	.0017	.0020	.0016	.0033	.0028	.0038							
24 Tetanus	.3961	.0180	.2490	.0503	.0299	.0185	.0183	.0209	.0228	.0166	.0243	.0186	.0388	.0211	.0190	.0144	.0294					
25 Mycoses	....	....	....	.0004	....	.0007	.0008	....	....	.0011	.0033	....	....	.0056	.0038	.0086	.0049					
26 Pellagra	....	.0005	.0004	.0013	.0012	.0019	.0011	.0035	.0033	.0039	.0059	.0054	.0066	.0070	.0038	.0057	.0098					
27 Beriberi	....	....	....	....	....	.0007	.0008	.0004	....	.0006	.0007	.0008										
28 Tuberculosis of the lungs	.8026	.2648	.0823	.1308	.8832	1.7055	2.0111	2.2870	2.4274	2.4616	2.3760	2.2634	2.3986	2.2738	2.2588	1.9746	1.8559	1.6475	1.0669	1.6829	2.1216	
29 Acute miliary tuberculosis	.0830	.0323	.0148	.0105	.0363	.0559	.0545	.0558	.0569	.0405	.0361	.0411	.0454	.0338	.0323	.0201	.0049					
30 Tuberculous meningitis	.9012	.4498	.1099	.0411	.0444	.0307	.0255	.0301	.0213	.0211	.0236	.0147	.0177	.0127	.0152	.0086	.0147	.0210				
31 Abdominal tuberculosis	.2076	.0770	.0197	.0197	.0323	.0396	.0412	.0449	.0550	.0527	.0617	.0527	.0875	.0775	.0989	.0977	.1077	.0630	.0577	.1202		
32 Pott's disease	.0138	.0212	.0205	.0101	.0157	.0152	.0156	.0100	.0138	.0150	.0138	.0209	.0188	.0155	.0209	.0201	.0245	.0210				
33 White swellings	.0086	.0078	.0060	.0117	.0093	.0104	.0069	.0087	.0095	.0083	.0131	.0124	.0144	.0113	.0209	.0287	.0098	.0210	.0288			
34 Tuberculosis of other organs	.1072	.0249	.0088	.0109	.0117	.0156	.0225	.0209	.0209	.0211	.0295	.0303	.0366	.0381	.0399	.0661	.0490	.0315	....	.1202		
35 Disseminated tuberculosis	.0813	.0138	.0072	.0075	.0182	.0215	.0194	.0213	.0237	.0222	.0184	.0248	.0222	.0296	.0247	.0259	.0098	.0210	.0288			
36 Rickets	.2318	.0383	.0024	.0013	.0004	.0004	.0008	.0004	....	....	.0007	.0008	.0022	....	....	....	.0049	....	.0288			
37 Syphilis	1.5862	.0323	.0052	.0038	.0065	.0259	.0374	.0471	.0593	.0583	.0696	.0698	.0842	.0775	.0551	.0460	.0490	.0315	.0288			
38 Gonococcus infection	.0796	.0009	....	....	.0008	.0011	.0008	.0022	.0014	.0017	.0013	.0023	.0011	.0028	.0038	.0029	.0049	.0105				
39 Cancer of the buccal cavity	....	.0009	.0016	.0008	.0012	.0007	.0030	.0057	.0095	.0277	.0670	.1334	.1595	.2749	.3822	.4628	.4848	.9444	1.3840	1.0819	2.1216	
40 Cancer of the stomach, liver	.0052	.0051	.0040	.0004	.0048	.0056	.0244	.0457	.1006	.2370	.4261	.7268	1.2109	1.8650	2.3235	2.8599	3.1046	3.1061	2.3932	1.5627	1.4144	
41 Cancer of the peritoneum, etc.	....	.0042	.0020	.0038	.0032	.0052	.0160	.0222	.0398	.0733	.1274	.1838	.3113	.4229	.5932	.7128	.7296	.7451	.7785	.2404	.7072	
44 Cancer of the skin	....	.0014	.0012	.0017	.0016	.0015	.0030	.0044	.0071	.0139	.0322	.0458	.0864	.1494	.2225	.4139	.6660	1.0809	1.6436	2.4041	2.8289	
45 Cancer of other organs, etc.	.0242	.0254	.0092	.0105	.0206	.0322	.0217	.0466	.0607	.1043	.1871	.2816	.4620	.6555	.9393	1.2273	1.6062	1.5741	1.9896	1.2021	1.4144	
46 Other tumors	.0173	.0032	.0008	.0013	....	.0007	.0023	.0013	.0028	.0050	.0053	.0085	.0144	.0127	.0304	.0575	.0490	.0420	.0288	.1202		

47 Acute articular rheumatism...	.0294	.0291	.0646	.0709	0533	.0293	.0225	.0314	0413	.0511	.0558	.0760	.0908	.1311	.2225	.2501	.3330	.2938	.6055	1.2021		
48 Chronic rheumatism and gout.	....	....	....	....	.0020	.0019	.0019	.0030	.0038	.0067	.0072	.0124	.0299	.0507	.0513	.0977	.1812	.2309	.2307	.1202	.7072	
50 Diabetes....................	.0121	.0281	.0277	.0428	.0517	.0422	.0438	.0571	.0702	.0971	.1733	.2971	.4709	.6710	.9184	1.0347	1.0528	1.0599	.7497	.8414		
51 Exophthalmic goiter..........	.0017	....	....	....	.0008	.0015	.0023	.0030	.0024	.0055	.0065	.0047	.0111	.0014	....	.0057	.0049					
52 Addison's disease............	....	....	....	.0004	.0024	.0015	.0015	.0035	.0052	.0072	.0059	.0124	.0111	.0127	.0076	.0029	.0196					
53 Leukemia....................	0311	.0194	.0100	.0113	.0101	.0122	.0156	.0161	.0204	.0239	.0249	.0334	.0388	.0395	.0551	.0604	.0539	.0105	.0288			
54 Anemia, chlorosis............	.1730	.9226	.0080	.0071	.0081	.0081	.0088	.0161	.0232	.0400	.0578	.0861	.1274	.1762	.2301	.2271	.2351	.2414	.2595	.4818		
55 Other general diseases.........	.3771	.0600	.0156	.0075	.0093	.0100	.0069	.0070	.0095	.0117	.0151	.0155	.0188	.0310	.0418	.0489	.0735	.0735	.1153	.3606		
56 Alcoholism (acute or chronic)...	....	....	.0008	....	.0028	.0167	.0587	.1237	.1836	.2231	.2357	.2226	.2360	.2284	.2187	.2012	.1518	.1574	.0577			
57 Chronic lead-poisoning.......	....	....	....	.0004	.0004	.0011	.0019	.0052	.0081	.0105	.0118	.0163	.0122	.0099	.0114	.0086	.0098					
58 Other chronic occupation poisonings................	....	....	....	....	.0004	.0004	....	.0004	....	.0011	.0007											
59 Other chronic poisonings......	....	....	....	....	....	.0015	.0030	.0039	.0085	.0055	.0079	.0078	.0166	.0169	.0266	.0287	.0196	.0630				
60 Encephalitis.................	.1263	.0235	.0088	0084	.0105	.0118	.0095	.0135	.0133	.0139	.0138	.0186	.0377	.0296	.0171	.0287	.0392	.0630	.0577	.1202		
61 Simple meningitis............	1.6624	.4055	.0815	.0449	0367	.0307	.0294	.0296	.0308	.0366	.0440	.0489	.0465	.0578	.0627	.0747	.0588	.0839	.2883	.1202		
Cerebrospinal meningitis......	.5259	.1582	.0437	.0250	.0287	.0189	.0099	.0109	.0095	.0078	.0131	.0116	.0133	.0085	.0076	.0086	.0049	.0105				
Cerebrospinal fever...........	.0450	.0161	.0096	.0034	.0036	.0022	.0011	.0022	.0014	.0017	.0013	.0016	....	.0014								
62 Locomotor ataxia............	....	.0005	....	.0004	....	0004	.0050	.0131	.0270	.0549	.0893	.1272	.1828	.2044	.2643	.2529	.2351	.1994	.0577			
63 Acute anterior poliomyelitis...	.2456	.1629	.0494	.0222	.0198	.0118	.0069	.0035	.0009	.0039	.0033	.0023	.0011	.0028	.0038	.0029	.0049					
Other diseases of spinal cord..	.0329	.0106	.0044	.0101	0137	.0118	.0141	0257	.0261	.0422	.0735	.1032	.1418	.2143	.3365	.4886	.6660	.6821	.7497	.3606		
64 Cerebral hemorrhage, apoplexy.	5276	.0346	.0096	.0101	.0206	.0330	0606	.1159	.1836	.3296	.6073	1.0619	2.0485	3.4693	5.6622	8.9245	13.4175	18.5739	23.4999	23.4403	22.6308	30.8880
65 Softening of the brain.........	....	....	....	....	....	.0004	.0023	.0044	.0052	.0061	.0151	.0240	.0355	.0761	.1445	.2386	.5680	.7346	.8515	.7212		
66 Paralysis without specified cause.....................	.0467	.0101	.0052	.0063	0044	0070	.0084	.0179	.0365	.0549	.1031	.1598	.2770	.5047	1.0000	1.5693	2.8794	4.2290	6.0552	6.9720	6.3649	3.8610
67 General paralysis of the insane.	....	..	....	....	.0008	.0033	.0133	.0697	.1598	.2164	.2258	.2536	.2160	.1706	.1977	.2041	.3771	.3568	.4325			
68 Other forms of mental alienation.....................	.0017	....	.0004	....	.0057	.0107	.0107	.0244	.0223	.0322	.0374	.0582	.0787	.1085	.1521	.2529	.3526	.7136	.5767	.4808		
69 Epilepsy....................	.0882	.0217	.0189	.0294	.0404	.0415	.0435	.0444	.0574	.0572	.0617	.0683	.0654	.0930	.0875	.1121	.1861	.2414	.0577	.1202	.7072	
70 Convulsions (non-puerperal)...	....	....	.0104	.0021	.0016	.0011	.0015	.0009	.0019	.0039	.0007	....	....	....	.0019	.0057	.0049	.0315	....	.1202		
71 Convulsions of infants........	3.9924	.1772																				
72 Chorea.....................	.0069	.0028	.0012	.0021	.0065	.0011	.0008	....	.0005	....	.0020	....	.0011	.0014	.0038	....	....	.0105				
73 Neuralgia and neuritis........	....	....	.0012	.0004	.0004	.0007	.0015	.0030	.0057	.0061	.0131	.0140	.0144	.0127	.0494	.0345	.0539	.0420	.0577	.1202		
74 Other diseases of the nervous system..................	.1263	.0300	.0156	.0113	.0125	.0170	.0156	.0274	.0294	.0427	.0565	.0721	.0576	.0789	.0837	.1610	.2399	.3148	.2595	.3606	.7072	
75 Diseases of the eyes and annexa	.0086	.0014	....	....	....	.0004	....	.0009	.0009	.0006	....	....	....	.0014	....	....	....	.0105	....	.1202		
76 Diseases of the ears..........	.2422	.0434	.0197	.0143	.0125	.0111	.0114	.0091	.0085	.0133	.0158	.0155	.0166	.0197	.0304	.0086	.0098	.0315				
77 Pericarditis.................	.0173	.0032	.0048	.0067	.0028	.0041	.0050	.0087	.0128	.0100	.0184	.0178	.0366	.0381	.0589	.0776	.1371	.1574	.2307	.1202		
78 Acute endocarditis............	.1453	.0281	.0353	.0369	.0396	.0385	.0469	.0693	.0901	.1154	.1740	.2203	.3590	.1410	.1787	.2501	.3379	.4197	.6632	.7212	.7072	
79 Organic diseases of the heart...	.9116	.0983	.1316	.1874	.2290	.2240	.2855	.4313	.6446	.9894	1.4030	2.1859	3.5209	6.1054	9.9289	15.3829	22.7657	30.9670	38.7532	39.9086	43.1400	54.0541
80 Angina pectoris..............	.0035	.0005	.0012	.0021	.0032	.0041	.0069	.0174	.0223	.0433	.0893	.1536	.3002	.5159	.7396	.9600	1.3026	1.3747	1.5282	1.5627	1.4144	
81 Diseases of the arteries, atheroma, etc..................	.0052	.0009	....	.0013	.0020	.0056	.0103	.0266	.0413	.0671	.1385	.2288	.4343	.8486	1.6523	3.2766	6.2142	10.5567	16.6085	20.0745	25.4597	27.0270
82 Embolism and thrombosis.....	.0519	.0042	.0040	.0038	.0040	.0093	.0118	.0152	.0190	.0327	.0387	.0520	.0798	.1226	.1825	.2817	.4456	.4932	.3172	.8414	.7072	
83 Diseases of the veins.........	.0069	.0009	.0004	.0004	.0008	.0007	.0008	.0035	.0028	.0100	.0085	.0093	.0089	.0197	.0285	.0316	.0441	.0315	.1153	.1202		
84 Diseases of the lymphatic system.......................	.1194	.0115	.0040	.0025	.0024	.0007	.0015	.0009	.0019	.0017	.0013	.0023	.0033	.0014	.0095	.0115	.0245	.0210	.0288			
85 Hemorrhage, etc............	.1401	.0032	.0012	.0004	.0012	.0011	.0004	.0057	.0043	.0033	.0085	.0085	.0066	.0141	.0133	.0144	.0539	.0525	.0577	.1202		

Specific Death-rates Per 1000 Population. Registration Area, 1910, Exclusive of North Carolina—*Continued*

Diseases.	Males.																					
	Under 1.	1–4.	5–9.	10–14.	15–19.	20–24.	25–29.	30–34.	35–39.	40–44.	45–49.	50–54.	55–59.	60–64.	65–69.	70–74.	75–79.	80–84.	85–89.	90–94.	95–99.	100 or over.
86 Diseases of the nasal fossæ	.0571	.0042	.0020	.0013	.0012	.0004	....	.0004	.0005	....	....	.0008	.0011	....	....	.0029	.0049	.0420	....	.3606		
87 Diseases of the larynx	.1816	.0775	.0197	.0046	.0016	.0015	.0034	.0044	0062	.0050	.0092	.0070	.0078	.0070	.0152	.0201	.0196	.0839	.0288	.1202		
88 Diseases of the thyroid body	.0104	.0005	....	.0017	.0004	....	.0004	.0004	.0009	.0011	.0020	.0016	.0022	.0056	.0095	.0115	....	.0105				
89 Acute bronchitis	4.0132	.3017	.0152	.0042	.0044	.0048	.0050	.0035	.0095	.0061	.0098	.0171	.0310	.0663	.1046	.2156	.4505	1.1333	2.0472	5.6497	4.2433	
90 Chronic bronchitis	.1055	.0245	.0128	.0042	.0069	.0096	.0091	.0192	.0223	.0300	.0506	.0861	.1429	.2763	.4810	.9571	1.9147	3.3265	5.8533	8.5347	5.6577	7.7220
91 Bronchopneumonia	10.3616	1.5806	.1031	.0264	.0307	.0304	.0431	.0444	.0863	.1004	.1156	.1675	.2482	.4441	.6674	1.1842	1.9441	3.0432	4.4405	5.8901	11.3154	19.3050
92 Lobar pneumonia	2.3024	.5744	.0991	.0621	.1490	.2318	.2638	.3886	.5103	.6315	.7892	.9463	1.1444	1.4858	1.7207	2.0062	2.6492	3.1061	3.2294	3.9668	8.4866	7.7220
Pneumonia undefined	5.7793	.8936	.1352	.0797	.1345	.1537	.1757	.2244	.3206	.4034	.5016	.6686	.9018	1.1982	1.6808	2.5092	3.6433	6.0129	8.3908	10.3378	16.9731	11.5830
93 Pleurisy	.0744	.0655	.0165	.0067	.0234	.0281	.0274	.0292	.0346	.0516	.0538	.0761	.0831	.1156	.1217	.2271	.2889	.2833	.3748	.4808	2.1216	
94 Pulmonary congestion, pulmonary apoplexy	.5812	.0424	.0068	.0021	.0036	.0078	.0091	.0118	.0138	.0172	.0177	.0295	.0432	.1043	.1198	.2357	.5338	1.2068	2.3356	3.4860	7.0721	
95 Gangrene of the lung	....	.0009	.0004	....	....	.0015	.0019	.0022	.0028	.0044	.0066	.0101	.0155	.0169	.0190	.0144	.0098	.0105	.0865			
96 Asthma	.0208	.0046	.0016	.0004	.0008	.0007	.0038	.0057	.0090	.0189	.0295	.0520	.0798	.1297	.1654	.2932	.4750	.6716	.6920	.6010	.7072	
97 Pulmonary emphysema	.0121	.0009	.0012	.0004	.0004	.0007	.0011	.0009	.0014	.0033	.0066	.0047	.0078	.0127	.0171	.0315	.0392	.0735	.1153			
98 Other diseases of the respiratory system	.0917	.0125	.0040	.0034	.0061	.0096	.0183	.0213	.0247	.0316	.0440	.0590	.0787	0973	.1198	.1207	.1077	.1364	.2307	.2404	1.4144	
99 Diseases of the mouth and annexa	.2595	.0120	.0028	....	.0012	.0004	.0019	.0013	.0014	.0011	.0020	.0008	.0044	.0028	.0057	.0115	.0049	.0315	.0288	.1202		
100 Diseases of the pharynx	.1245	.0415	.0269	.0101	.0101	.0100	.0084	.0100	.0076	.0122	.0072	.0132	.0199	.0113	.0304	.0201	.0490	.1049	.1442			
101 Diseases of the esophagus	.0035	.0023	....	....	.0008	.0007	.0004	.0004	.0005	.0017	.0026	.0039	.0100	.0099	.0095	.0201	.0294	.0525	.0865	....	.7072	
102 Ulcer of the stomach	.0156	.0037	.0028	.0029	.0052	.0196	.0252	.0383	.0522	.0666	.0939	.1055	.1473	.1635	.1730	.2184	.2546	.2623	.3172	.3606		
103 Other diseases of the stomach	2.7158	.1269	.0233	.0067	.0125	.0126	.0221	.0344	.0470	.0666	.0959	.1249	.1839	.3073	.4335	.7473	1.1704	2.2142	3.6908	5.4093	4.2433	7.7220
104 Diarrhea and enteritis (under two years)	42.3476	2.3196																				
105 Diarrhea and enteritis (two years and over)	....	.6879	.0943	.0264	.0170	.0159	.0248	.0288	.0413	.0494	.0663	.1063	.1451	.2693	.4468	.7789	1.5082	2.8753	4.3251	7.6932	4.9505	3.8610
106 Ankylostomiasis	....	....	....	.0004	.0004																	
107 Intestinal parasites	.0017	.0088	.0028	.0004	.0004	.0011	.0008	.0013	....	.0006												
108 Appendicitis and typhlitis	.0242	.0494	.1188	.1673	.1741	.1492	.1254	.1311	.1361	.1249	.1287	.1559	.1185	.1508	.1559	.1178	.1028	.0525	.0577			
109 Hernia	.2110	.0088	.0016	.0008	.0057	.0093	.0118	.0190	.0232	.0305	.0414	.0745	.0820	.1198	.2130	.3564	.5827	.7555	.8939	1.2021	.7072	
Intestinal obstruction	.7905	.0803	.0281	.0210	.0279	.0285	.0259	.0309	.0460	.0483	.0657	.0931	.1429	.1804	.2852	.3737	.5778	.7241	.7497	1.3223	1.4144	
110 Other diseases of the intestines	.2975	.0157	.0060	.0038	.0032	.0085	.0095	.0161	.0161	.0144	.0328	.0341	.0510	.0663	.0780	.1236	.2302	.4722	.4325	.7212	2.1216	
111 Acute yellow atrophy of the liver	.0086	.0014	....	.0004	.0012	.0022	.0027	.0035	.0024	.0017	.0046	.0054	.0022	.0099	.0095	.0057	.0490	.0315				
112 Hydatid tumor of the liver	.0035	.0005	....	....	....	.0004	.0011	.0004	.0009	.0017	.0013	....	.0022	.0014	.0019	.0029						
113 Cirrhosis of the liver	.0035	.0028	.0036	.0034	.0040	.0118	.0313	.0845	.1660	.2708	.3992	.5251	.7046	.9050	1.0648	1.0951	1.1312	.9759	.8650	.6010	2.1216	
114 Biliary calculi	.0017	.0005	....	.0004	....	.0004	.0046	.0039	.0100	.0133	.0256	.0496	.0709	.0761	.0856	.1552	.2155	.2309	.1153			
115 Other diseases of the liver	.1401	.0203	.0092	.0063	.0085	.0156	.0255	.0344	.0427	.0583	.0637	.1094	.1318	.2044	.2244	.2903	.4358	.5876	.7497	.8414	.7072	
116 Diseases of the spleen	.0052	....	....	.0008	....	.0004	.0015	.0030	.0038	.0022	.0020	.0023	.0033	.0070	.0076	.0057	.0196					
117 Simple peritonitis (non-puerperal)	.1470	.0374	.0237	.0247	.0258	.0252	.0194	.0283	.0275	.0311	.0289	.0403	.0399	.0789	.0913	.1207	.1861	.1259	.2018			

118 Other diseases of the digestive system	.0086	.0028	.0004	.0013	.0020	.0033	.0027	.0052	.0085	.0089	.0105	.0194	.0144	.0141	.0133	.0144	.0294	.0315	.0288			
119 Acute nephritis	.4013	.0904	.0546	.0289	.0355	.0485	.0694	.0771	.0996	.1199	.1405	.1784	.2260	.2876	.3023	.4254	.4848	.7555	1.0380	1.0819		
120 Bright's disease	.3217	.0830	.0457	.0453	0816	.1144	.2104	.3233	.5217	.8196	1.2697	1.8872	2.9171	4.4377	6.2859	9.0969	12.6144	16.0764	20.9625	18.7522	19.8020	30.8880
121 Chyluria	....	....	....	....	....	....	....	....	....	....	....	....	....	....	....	.0029						
122 Other diseases of the kidneys and annexa	.1470	.0055	.0032	.0008	.0020	.0067	.0103	.0139	.0128	.0233	.0453	.0403	.0499	.1085	.1312	.2328	.3379	.5667	.8074	1.2021	.7072	3.8610
123 Calculi of the urinary passages	.0104	.0009	.0004	.0004	.0008	.0022	.0034	.0039	.0076	.0039	.0118	.0140	.0288	.0310	.0589	.0948	.1077	.0839	. 0577			
124 Diseases of the bladder	.0467	.0028	.0008	.0004	.0020	.0019	.0038	.0039	.0095	.0133	.0171	.0233	.0543	.1128	.2681	.5547	1.3271	2.2666	3.8926	5.1689	5.6577	3.8610
125 Diseases of the urethra, etc.	.0086	....	....	....	.0004	.0015	.0038	.0039	.0081	0161	.0164	.0217	.0255	.0183	.0380	.0546	.0441	.0630	....	.1202		
126 Diseases of the prostate	....	....	....	....	....	.0004	.0008	.0017	.0024	.0044	.0118	.0225	.0798	.2171	.6274	1.2043	2.1155	3.5993	4.3540	4.2072	8.4866	
127 Non-venereal diseases of male genital organs	.0571	.0014	.0004	....	....	.0004	.0004	.0022	.0005	.0022	.0013	.0008	.0011	.0042	.0019	.0259	.0294	.0105	....	.1202	.7072	
142 Gangrene	.0311	.0051	.0024	.0013	.0024	.0015	.0019	.0035	.0062	.0117	0151	.0240	.0465	.0761	.1996	.3966	.7884	1.7105	2.5951	4.6881	6.3649	
143 Furuncle	.0346	.0005	.0004	....	.0032	.0030	.0019	.0052	.0033	.0061	.0072	.0140	.0244	.0197	.0418	.0460	.0294	.0525	.0865	....	.7072	
144 Acute abscess	.1540	.0120	.0020	.0008	.0036	.0022	.0034	.0061	.0066	.0089	.0092	.0140	.0155	.0268	.0437	.0460	.0490	.0839	.0577			
145 Other diseases of the skin and annexa	.2041	.0055	....	....	....	.0019	....	.0017	....	.0033	.0053	.0093	.0022	.0113	.0228	.0402	.0539	.0735	.2018	.1202	.7072	
146 Diseases of the bones (tuberculosis excepted)	.1090	.0401	.0173	.0226	.0218	.0130	.0126	.0161	.0223	.0222	.0243	.0287	.0321	.0282	.0589	.0661	.0686	.1889	.0865	.1202	2.1216	
147 Diseases of the joints (tuberculosis, etc., excepted)	.0086	.0009	.0012	.0013	.0012	.0004	.0023	.0004	.0047	.0044	.0026	.0047	.0022	.0014	.0038	.0086	.0294	....	.0288	.2404		
148 Amputations	....	....	....	....	....	....	....	....	....	.0006	....	....	.0011	.0014	....	.0057						
149 Other diseases of the organs of locomotion	.0017	.0005	.0004	.0004	.0012	.0004	....	.0009	.0005	.0011	.0020	....	.0011	.0028	.0019	.0057	.0147	.0315	.0577			
150 Hydrocephalus	.4722	.0355	.0080	.0034	.0008	.0011	.0004	....	.0005	.0006	....	.0008										
Congenital malformations of the heart	4.6705	.0304	.0052	.0063	.0028	.0011	....	....	.0005													
Other congenital malformations	2.2436	.0189	.0032	....	.0008	.0004	.0004															
151 Premature birth	19.6507																					
Congenital debility, "atrophy," "marasmus"	11.7852	.0028	.0008	....	.0004	.0004	.0008															
152 Injuries at birth	3.9578																					
Other causes peculiar to early infancy	3.1413																					
153 Lack of care	.1228																					
154 Senility	....	....	....	....	....	....	....	....	....	....	....	....	....	.1085	.4126	1.7964	5.3131	17.8288	38.3207	77.5334	123.0552	227.7992
155 163 Suicide	....	....	.0004	.0075	.0594	.1996	.2501	.2802	.3349	.3624	.4747	.5088	.6404	.6583	.6236	.5950	.7052	.6716	.6055	.7212	.7072	3.8610
164 186 Accidental or undefined	1.3042	.9375	.5778	.5485	.8691	1.3500	1.4089	1.3916	1.5514	1.5566	1.6564	1.6359	1.7039	1.8848	2.0934	2.5379	3.2173	4.8691	7.3816	12.0207	26.1669	19.3050
182 184 Homicide	.1245	.0097	.0072	.0126	.0448	.1592	.1719	.1534	.1622	.1182	.1156	.0838	.0775	.0648	.0494	.0431	.0441	.0525	.0577			
187 189 Ill-defined diseases	6.2152	.2630	.0152	.0088	.0141	.0193	.0313	.0292	.0536	.0522	.0775	.1342	.2050	.3017	.4355	.6553	1.1606	2.4660	4.2675	7.6932	8.4866	11.5830

Specific Death-rates Per 1000 Population. Registration Area, 1910, Exclusive of North Carolina

Diseases.	Females.																					
	Under 1.	1–4.	5–9.	10–14.	15–19.	20–24.	25–29.	30–34.	35–39.	40–44.	45–49.	50–54.	55–59.	60–64.	65–69.	70–74.	75–79.	80–84.	85–89.	90–94.	95–99.	100 or over.
All causes	119.8213	13.8441	3.4891	2.3916	3.6884	5.2148	6.1169	6.7941	7.7685	8.9347	11.0321	14.6304	20.6230	29.2671	44.2904	66.8156	100.8969	155.8717	222.7377	309.7260	368.9320	471.6599
1 Typhoid fever	.0604	.1224	.1425	.2023	.2911	.2727	.2154	.1740	.1652	.1635	.1609	.1526	.1627	.1280	.1333	.1884	.1129	.1633	.0919			
4 Malaria	.0586	.0212	.0086	.0097	.0104	.0141	.0163	.0155	.0126	.0125	.0198	.0244	.0229	.0320	.0571	.0499	.0451	.1361	.1608	.2626	.4413	
5 Smallpox	.0107	.0047	.0012	.0008	.0020	.0053	.0042	.0019	.0037	.0044	.0037	.0026	.0012	....	.0038							
6 Measles	1.1576	.8344	.1144	.0385	.0220	.0160	.0163	.0233	.0184	.0175	.0162	.0105	.0157	.0102	.0114	.0222	.0181	.0091				
7 Scarlet fever	.2432	.7412	.3661	.1054	.0507	.0336	.0251	.0136	.0105	.0037	.0073	.0052	.0048	.0029	....	....	.0045					
8 Whooping-cough	3 0111	.6932	.0558	.0030	.0024	.0011	.0004	....	.0005	.0012	....	.0009	.0012	.0029	.0038	.0083	.0045	....	.0230	.1751		
9 Diphtheria and croup	.6782	1.3661	.6137	.1600	.0403	.0305	.0193	.0121	.0126	.0150	.0096	.0096	.0060	.0029	.0095	.0194	.0090	.0091	....	....	.4413	
10 Influenza	.3977	.0838	.0269	.0165	.0212	.0267	.0377	.0344	.0503	.0649	.0816	.1291	.2735	.4582	.8170	1.4632	2.3653	4.4008	6.3160	9.1920	9.7087	10.1215
13 Cholera nostras	.0604	.0118	.0020	.0017	.0020	.0011	.0038	.0015	.0037	.0025	.0059	.0044	.0145	.0175	.0419	.0665	.1580	.2178	.2986	.5253	1.3239	
14 Dysentery	.5699	.1121	.0098	.0030	.0024	.0042	.0101	.0107	.0173	.0168	.0272	.0314	.0590	.1178	.2552	.4766	.7629	1.5335	1.9063	2.9765	2.6478	4.0486
17 Leprosy	....	....	....	....	....	.0004	....	.0005	....	.0006	.0007											
18 Erysipelas	.6285	.0226	.0020	.0017	.0072	.0095	.0109	.0160	.0157	.0237	.0316	.0375	.0542	.0742	.1333	.2023	.2347	.3720	.4134	.3502	.8826	2.0243
19 Other epidemic diseases	.0497	.0151	.0049	.0008	....	....	.0004	.0005	....	....	.0007	....	....	....	.0019	.0055	.0045					
20 Purulent infection and septicemia	.1687	.0179	.0110	.0089	.0152	.0229	.0235	.0199	.0236	.0237	.0250	.0401	.0373	.0436	.0686	.0748	.0813	.1089	.2986	.1751	....	2.0243
22 Anthrax	....	....	....	....	....	....	....	....	....	....	.0007	....	....	.0015	.0019							
23 Rabies	.0018	.0019	.0016	.0008	.0012	....	.0004	....	....	....	.0007	.0009										
24 Tetanus	.3160	.0067	.0155	.0114	.0052	.0042	.0071	.0029	.0131	.0056	.0073	.0078	.0145	....	.0095	.0166	.0226	.0454				
25 Mycoses	....	....	....	....	.0008	....	.0004	....	.0010	.0006	.0015	....	....	.0015	....	....	....	.0091				
26 Pellagra	.0018	.0009	.0008	....	.0020	.0050	.0109	.0136	.0121	.0112	.0081	.0742	.0193	.0087	.0133	.0166	.0045					
27 Beriberi	....	....	....	....	....	.0008																
28 Tuberculosis of the lungs	.6480	.2274	.1112	.3043	1.1536	1.8546	2.0704	1.9554	1.7284	1.5017	1.2500	1.1938	1.2253	1.2481	1.4074	1.6267	1.6431	1.3066	.8268	.7003	.4413	
29 Acute miliary tuberculosis	.0639	.0301	.0163	.0233	.0491	.0614	.0486	.0465	.0330	.0231	.0213	.0201	.0120	.0087	.0229	.0194	.0226	.0091	.0230			
30 Tuberculous meningitis	.8788	.4346	.1157	.0508	.0515	.0294	.0256	.0218	.0184	.0175	.0132	.0113	.0133	.0102	.0076	.0055	.0045	.0181				
31 Abdominal tuberculosis	.2361	.0593	.0224	.0246	.0587	.0698	.0708	.0746	.0792	.0661	.0603	.0759	.0627	.0946	.1181	.0970	.0993	.1543	.1608			
32 Pott's disease	.0231	.0217	.0175	.0089	.0060	.0076	.0134	.0131	.0058	.0069	.0110	.0105	.0241	.0145	.0229	.0333	.0316					
33 White swellings	.0142	.0066	.0037	.0076	.0064	.0038	.0034	.0039	.0052	.0031	.0051	.0035	.0108	.0058	.0095	.0139	.0135	.0363				
34 Tuberculosis of other organs	.0852	.0193	.0057	.0102	.0148	.0160	.0142	.0102	.0205	.0168	.0228	.0183	.0301	.0247	.0343	.0471	.0542	.0635	.0919	.1751		
35 Disseminated tuberculosis	.0479	.0170	.0069	.0068	.0164	.0202	.0222	.0141	.0215	.0175	.0184	.0140	.0193	.0160	.0267	.0221	.0181	.0272				
36 Rickets	.1864	.0311	.0029	.0021	.0012	.0004	.0008	.0005	.0016	.0012	.0015	....	....	.0015	.0057	.0055	.0135	.0181				
37 Syphilis	1.3049	.0344	.0041	.0008	.0076	.0172	.0260	.0291	.0409	.0349	.0228	.0296	.0265	.0247	.0209	.0249	.0135	.0181	.0230			
38 Gonococcus infection	.1012	.0014	....	.0004	.0056	.0057	.0021	.0024	.0025	.0019	.0037	....	.0024	....	.0019	.0028						
39 Cancer of the buccal cavity	.0036	.0005	.0016	.0004	.0012	.0008	.0021	.0015	.0021	.0044	.0088	.0096	.0337	.0335	.0667	.0970	.1490	.1543	.2986	.5253	.8826	
40 Cancer of the stomach, liver	.0053	.0042	.0020	.0017	.0024	.0065	.0235	.0635	.1301	.2539	.4968	.7979	1.3675	1.6845	2.4377	2.9319	3.1553	3.1759	2.8939	1.9259		
41 Cancer of the peritoneum, intestines, etc.	.0018	.0019	.0020	.0017	.0028	.0057	.0189	.0363	.0729	.1185	.1940	.3183	.4193	.6139	.7675	.9533	1.2323	1.0798	.8957	.4377		
42 Cancer of the female genital organs	....	.0005	.0004	.0017	.0056	.0114	.0407	.1265	.2292	.4529	.6783	.7761	.9458	1.0110	.9827	.9948	1.0518	.8348	.6431	.7003	.4413	4.0486

43 Cancer of the breast..........	.0036	.0009	....	....	.0008	.0042	.0113	.0465	.0991	.2059	.3601	.4412	.5422	.5441	.7732	.8507	1.1691	1.3974	1.9293	1.9259	.8826	2.0243
44 Cancer of the skin............	.0053	.0009	.0004	....	.0004	.0011	.0004	.0024	.0047	.0075	.0162	.0218	.0229	.0655	.1047	.2051	.3476	.7985	1.0106	.4882	.8826	2.0243
45 Cancer of other organs or of organs not specified........	.0178	.0202	.0086	.0089	.0160	.0118	.0239	.0426	.0724	.1204	.1499	.2581	.3108	.4568	.5237	.6568	.8983	.9528	.9646	1.3131	1.3239	
46 Other tumors (tumors of the female genital organs excepted)....................	.0142	.0024	.0024	.0004	.0024	.0042	.0038	.0087	.0142	.0187	.0316	.0349	.0398	.0320	.0667	.0942	.1625	.2268	.2297	....	.8826	
47 Acute articular rheumatism....	.0213	.0330	.0672	.0796	.0511	.0362	.0293	.0334	.0467	.0487	.0522	.0802	.1036	.1120	.1619	.2771	.4830	.7078	.5053	.6128		
48 Chronic rheumatism and gout..	.0018	.0009	....	.0008	.0012	.0008	.0029	.0019	.0058	.0062	.0169	.0253	.0458	.0524	.1219	.1968	.2302	.3539	.2297	.5253	.4413	
50 Diabetes....................	.0107	.0179	.0305	.0440	.0511	.0351	.0474	.0567	.0666	.1004	.1764	.3453	.6542	.8961	1.1598	1.1168	1.2549	1.0979	.9187	.6128	.4413	
51 Exophthalmic goiter..........	....	....	....	.0013	.0124	.0206	.0231	.0267	.0372	.0505	.0500	.0506	.0651	.0466	.0457	.0527	.0406	.0272				
52 Addison's disease............	....	....	....	.0004	.0012	.0015	.0029	.0082	.0058	.0069	.0073	.0087	.0169	.0145	.0190	.0166	.0181					
53 Leukemia....................	.0107	.0108	.0077	.0034	.0056	.0092	.0113	.0111	.0105	.0175	.0125	.0235	.0277	.0262	.0267	.0333	.0406	.0454	.0230			
54 Anemia, chlorosis............	.1793	.0188	.0077	.0097	.0196	.0206	.0243	.0286	.0472	.0631	.0874	.1221	.1627	.1993	.2438	.2771	.2483	.3539	.3445	.2626		
55 Other general diseases.........	.3373	.0570	.0171	.0072	.0084	.0107	.0096	.0102	.0105	.0137	.0103	.0122	.0133	.0262	.0438	.0804	.0903	.1724	.1378	.0875	....	2.0243
56 Alcoholism (acute or chronic)..	....	....	....	.0004	.0004	.0034	.0122	.0262	.0404	.0381	0345	.0279	.0205	.0175	.0133	.0055	.0045	....	.0230			
57 Chronic lead-poisoning........	....	.0005	....	....	....	.0004	....	....	....	.0012	.0007	....	....	.0015	.0038	.0055						
59 Other chronic poisonings......	....	....	....	....	.0008	.0015	.0025	.0039	.0063	.0075	.0081	.0113	.0072	.0175	.0286	.0277	.0316	.0363	.0230			
60 Encephalitis.................	.0959	.0188	.0049	.0072	.0052	.0061	.0059	.0078	.0084	.0087	.0118	.0052	.0145	.0145	.0152	.0222	.0361	.0272	.0230			
61 Simple meningitis	1.3742	.3367	.0900	.0457	.0327	.0217	.0243	.0228	.0220	.0287	.0367	.0340	.0349	.0305	.0476	.0748	.0632	.0726	.1378	.2626		
Cerebrospinal meningitis......	.4687	.1573	.0448	.0288	.0148	.0118	.0105	.0048	.0105	.0075	.0066	.0096	.0024	.0058	.0114	.0055	.0090	.0272	.0230	....	.4413	
Cerebrospinal fever...........	.0373	.0118	.0045	.0038	.0012	.0034	.0004	.0015	.0021	....	.0015	.0009	....	.0015	....	.0028						
62 Locomotor ataxia............	....	....	.0004	....	.0008	.0004	.0034	.0068	.0089	.0131	.0257	.0358	.0482	.0669	.0838	.0998	.0587	.0272				
63 Acute anterior poliomyelitis...	.2184	.1304	.0472	.0241	.0108	.0057	.0059	.0034	.0016	.0019	.0015	.0017	....	.0073	.0019	.0028	....	....	.0230			
Other diseases of the spinal cord.....................	.0355	.0122	.0057	.0051	.0076	.0080	.0113	.0184	.0210	.0324	.0558	.0959	.1277	.2095	.2952	.3769	.4017	.4446	.4364	.4377	....	2.0243
64 Cerebral hemorrhage, apoplexy	.3835	.0316	.0086	.0089	.0204	.0301	.0469	.0838	.1673	.3294	.6349	1.1414	1.9048	3.0854	5.1763	7.9532	12.9146	18.5199	21.4745	25.2998	25.5958	26.3158
65 Softening of the brain.........	....	.0005	....	....	.0016	.0004	.0013	.0039	.0026	.0050	.0184	.0209	.0373	.0611	.1162	.2217	.3295	.6533	.8039	1.3131	1.3239	
66 Paralysis without specified cause....................	.0284	.0099	.0057	.0025	.0068	.0053	.0113	.0233	.0315	.0449	.0735	.2049	.2916	.5659	.9618	1.6627	2.9747	4.5551	7.0280	7.6162	7.0609	8.0972
67 General paralysis of the insane.....................	....	....	....	.0004	....	.0023	.0096	.0228	.0378	.0580	.0536	.0602	.0590	.0742	.1047	.1413	.2663	.3357	.4364	.4377	.8826	
68 Other forms of mental alienation.....................	....	....	....	.0008	.0048	.0103	.0214	.0281	.0346	.0393	.0625	.0759	.0771	.1047	.1390	.1968	.2889	.5081	.6661	1.1381	1.3239	
69 Epilepsy....................	.0586	.0137	.0130	.0202	.0280	.0328	.0386	.0426	.0414	.0455	.0397	.0602	.0386	.0495	.0648	.0721	.1399	.1724	.1608	.0875	.4413	2.0243
70 Convulsions (non-puerperal)...	....	....	.0114	.0017	.0056	.0092	.0088	.0073	.0037	.0037	.0029	.0035	....	.0029	.0057	.0028	.0090					
71 Convulsions of infants........	3.1159	.1526																				
72 Chorea.....................	.0036	.0019	.0041	.0051	.0096	.0027	....	.0010	.0005	....	.0007	.0009	.0012	.0029	.0076	.0028	.0045					
73 Neuralgia and neuritis........	.0018	.0005	....	.0004	.0024	.0027	.0038	.0082	.0063	.0087	.0125	.0166	.0133	.0116	.0438	.0471	.0497	.1180	.2756	.1751		
74 Other diseases of the nervous system..................	.1261	.0193	.0126	.0131	.0120	.0172	.0218	.0354	.0399	.0468	.0507	.0794	.0759	.0873	.1181	.1552	.2438	.2541	.4593	.9630	.4413	
75 Diseases of the eyes and annexa....................	.0053	.0014	....	.0004	....	.0008	....	....	....	.0006	....	.0099	.0012	.0029	.0038	.0055	....	.0091	.0230			
76 Diseases of the ears..........	.1474	.0386	.0175	.0127	.0056	.0065	.0071	.0073	.0047	.0056	.0132	.0131	.0120	.0087	.0209	.0111	.0226	.0363	.0689			
77 Pericarditis.................	.0124	.0042	.0081	.0063	.0028	.0042	.0050	.0058	.0068	.0094	.0110	.0113	.0193	.0189	.0571	.0582	.0903	.1996	.2297	.0875	.4413	
78 Acute endocarditis............	.1314	.0424	.0464	.0580	.0387	.0503	.0499	.0654	.0084	.0998	.1330	.2058	.3036	.1236	.2133	.2162	.2618	.3357	.3904	.5253	.8826	
79 Organic diseases of the heart...	.6906	.0914	.1576	.2362	.2324	.2437	.3210	.4420	.6303	.9502	1.4278	1.9829	3.1024	5.2558	8.2102	13.2268	18.8279	27.0130	33.4864	32.3032	41.0415	38.4615

SPECIFIC DEATH-RATES PER 1000 POPULATION. REGISTRATION AREA, 1910, EXCLUSIVE OF NORTH CAROLINA—*Continued*

Diseases.	Females.																					
	Under 1.	1-4.	5-9.	10-14.	15-19.	20-24.	25-29.	30-34.	35-39.	40-44.	45-49.	50-54.	55-59.	60-64.	65 69.	70-74.	75-79.	80-84.	85-89.	90-94.	95-99.	100 or over.
80 Angina pectoris	.0036	.0005	.0016	.0030	.0036	.0050	.0117	.0155	.0210	.0406	.0470	.0863	.1554	.2793	.4323	.6152	.7719	1.0072	1.0795	1.0505	.4413	
81 Diseases of the arteries, etc	.0052	.0009	....	.0013	.0024	.0031	.0042	.0131	.0194	.0275	.0544	.1055	.2205	.5310	1.0056	2.0562	3.8459	7.9850	12.4943	19.9597	24.7132	24.2915
82 Embolism and thrombosis	.0320	.0028	.0041	.0042	.0068	.0149	.0168	.0276	.0362	.0412	.0522	.0654	.1036	.1397	.2076	.2660	.3972	.7894	.6431	1.2256		
83 Diseases of the veins	.0053	....	....	....	.0004	.0008	.0038	.0097	.0094	.0069	.0154	.0087	.0217	.0291	.0362	.0582	.0767	.1633	.0459	.0875	.4413	
84 Diseases of the lymphatic system	.0905	.0075	.0012	.0013	.0012	....	....	.0010	....	.0012	.0007	.0009	.0048	.0015	.0038	.0055	.0090	....	.0230	.0875		
85 Hemorrhage, etc.	.1101	.0014	.0012	.0008	.0012	.0031	.0013	.0029	.0047	.0044	.0037	.0052	.0084	.0102	.0095	.0083	.0406	.0635	.0689	.1751		
86 Diseases of the nasal fossæ.	.0426	.0019	.9016	.0008	.0008	....	.0004	....	.0005	....	....	.0017	.0012	....	.0076	.0028	.0045	.0544	.1148	.2626	.4413	
87 Diseases of the larynx	.1083	.0523	.0167	.0008	.0030	.0031	.0029	.0039	.0052	.0012	.0051	.0035	.0024	.0102	.0114	.0083	.0181	.0181	.0459	.0875		
88 Diseases of the thyroid body	.0053	.0009	.0001	.0013	.0040	.0053	.0046	.0053	.0105	.0087	.0125	.0131	.0169	.0175	.0305	.0222	.0361	.0181	.0230	.1751		
89 Acute bronchitis	3.1958	.3014	.0212	.0047	.0040	.0027	.0071	.0039	.0094	.0119	.0103	.0227	.0398	.0931	.1676	.3963	.7042	1.7150	3.0547	6.1280	3.5305	4.0486
90 Chronic bronchitis	.0799	.0235	.0118	.0072	.0068	.0126	.0147	.0174	.0236	.0200	.0353	.0732	.1434	.2764	.5599	1.2165	2.1848	4.0016	7.2118	9.6297	12.3566	12.1457
91 Bronchopneumonia.	8.1261	1.5168	.1063	.0360	.0311	.0404	.0310	.0456	.0519	.0643	.0926	.1482	.2819	.4859	.9560	1.5602	2.7490	4.7910	6.4309	9.2795	11.4740	16.1943
92 {Lobar pneumonia	1.7541	.5015	.1002	.0525	.0843	.1102	.1706	.2244	.2475	.3001	.3836	.5188	.7277	1.0925	1.5350	2.1283	2.8032	3.4390	4.0652	4.2896	2.6478	6.0729
{Pneumonia undefined	4.7386	.7916	.1197	.0800	.0803	.1014	.1354	.1580	.2119	.2496	.3285	.4194	.6277	1.0765	1.6607	2.6769	4.3470	6.4969	8.6817	2.6937	12.7979	16.1943
93 Pleurisy	.1047	.0593	.0102	.0068	.0096	.0137	.0142	.0218	.0257	.0243	.0265	.0323	.0566	.0887	.1295	.1690	.2392	.2994	.4364	.4377	....	2.0243
94 Pulmonary congestion, pulmonary apoplexy	.4563	.0311	.0045	.0047	.0032	.0069	.0067	.0082	.0131	.0100	.0191	.0305	.0458	.0727	.1695	.3187	.5417	1.2159	2.2508	4.2896	5.2957	4.0486
95 Gangrene of the lung	.0018	.0009	.0004	.0004	.0004	....	.0021	.0019	.0010	.0012	.0029	.0026	.0024	....	.0076	.0055	.0045					
96 Asthma	.0142	.0047	.0016	0013	.0012	.0023	.0038	.0073	.0121	.0175	.0265	.0262	.0554	.0975	.1924	.2383	.4063	.4628	1.0106	.6128	.8826	2.0243
97 Pulmonary emphysema	.0071	.0005	....	....	....	.0008	.0017	....	....	.0006	.0029	.0061	.0048	.0029	.0152	.0249	.0226	.0817	.0689	....	.4413	
98 Other diseases of the respiratory system	.0941	.0071	.0024	.0021	.0048	.0034	.0101	.0087	.0126	.0150	.0088	.0201	.0217	.0378	.0628	.0721	.1219	.2178	.1378	.3502	.4413	
99 Diseases of the mouth and annexa	.2131	.0146	.0016	.0017	.0012	.0011	.0004	....	.0037	.0012	.0015	.0035	.0012	.0029	.0038	.0028	.0090	.0091	.0459	.1751	.4413	
100 Diseases of the pharynx	.0675	.0320	.0228	.0106	.0060	.0061	.0046	.0063	.0079	.0037	.0059	.0105	.0120	.0102	.0190	.0249	.0497	.0726	.0919	.0875		
101 Diseases of the esophagus	.0053	.0005	.0004	.0004	.0004	....	....	.0010	.0016	.0006	.0022	.0009	.0024	.0087	.0114	.0222	.0135	.0454	.0230			
102 Ulcer of the stomach	.0178	.0024	.0024	.0047	.0136	.0244	.0243	.0281	.0409	.0399	.0536	.0732	.0904	.0989	.1276	.2272	.2483	.1633	.2067	.3502		
103 Other diseases of the stomach	2.2246	.1342	.0232	.0174	.0160	.0263	.0298	.0422	.0566	.0774	.0794	.1308	.1759	.3084	.4952	.8258	1.4355	2.5225	3.7207	5.2526	4.4131	4.0486
104 Diarrhea and enteritis (under two years)	35.3096	2.0579																				
105 Diarrhea and enteritis (two years and over)	....	.6216	.0933	.0284	.0192	.0324	.0402	.0456	.0503	.0792	.0948	.1352	.2325	.3142	.6456	1.1667	1.9230	3.3846	5.2595	7.0910	9.7087	8.0972
106 Ankylostomiasis	....	.0014	....	....	....	....	....	....	....	.0006												
107 Intestinal parasites	.0071	.0085	.0020	.0004	.0008	.0015	....	....	....	....	....	....	....	....	.0019	....	.0045					
108 Appendicitis and typhlitis	.0284	.0301	.1104	.1342	.1270	.1072	.0964	.0911	.1028	.0961	.0933	.0837	.1096	.0931	.0857	.0942	.0767	.1180	.1148	.0875		
109 {Hernia	.0462	.0028	.0012	.0013	.0028	.0029	.0038	.0078	.0147	.0418	.0830	.0985	.1422	1789	.2628	.2854	.3882	.4991	.3675	.7003	.4413	
{Intestinal obstruction	.5646	.0560	.0228	.0157	.0180	.0320	.0457	.0572	.0556	.0686	.0933	.1125	.1542	.2182	.3428	.4406	.6636	.7804	.8268	.9630	.4413	
110 Other diseases of the intestines.	.2592	.0198	.0041	.0047	.0052	.0107	.0138	.0170	.0246	.0293	.0309	.0375	.0313	.0509	.0724	.1469	.3024	.2994	.4593	1.0505	.4413	

111 Acute yellow atrophy of the liver	.0160	.0014	.0004	.0004	.0032	.0053	.0092	.0039	.0084	.0050	.0044	.0009	.0108	.0116	.0114	.0166	.0316	.0181	.0230			
112 Hydatid tumor of the liver	....	....	....	....	....	.0004	....	....	....	....	....	....	.0012	.0015	.0019							
113 Cirrhosis of the liver	.0089	.0061	.0024	.0034	.0060	.0092	.0264	.0470	.0781	.1260	.1852	.2468	.2964	.4059	.4780	.6180	.7087	.7804	.8039	.5253	.4413	
114 Biliary calculi	....	....	....	.0004	.0008	.0072	.0080	.0141	.0294	.0437	.0838	.1107	.1530	.1833	.2247	.2743	.3431	.2722	.3215	.4377		
115 Other diseases of the liver	.1332	.0188	.0053	.0085	.0084	.0137	.0147	.0267	.0472	.0599	.0801	.1142	.1518	.2080	.3104	.4406	.5733	.7894	1.3780	.5253	.4413	2.0243
116 Diseases of the spleen	.0053	.0005	....	.0004	....	.0004	....	.0019	.0010	.0037	.0007	.0044	.0060	.0102	.0057	....	.0090	.0091	....	.0875		
117 Simple peritonitis (non-puerperal)	.1474	.0325	.0297	.0207	.0379	.0576	.0608	.0693	.0750	.0549	.0529	.0532	.0542	.0669	.0743	.1025	.1309	.1815	.3215	2875		
118 Other diseases of the digestive system	.0071	.0005	.0033	.0004	.0008	.0027	.0063	.0053	.0084	.0081	.0088	.0131	.0181	.0175	.0133	.0222	.0271	.0181	.0230	....	.4413	
119 Acute nephritis	.2521	.1022	.0464	.0394	.0439	.0545	.0687	.0843	.0949	.1142	.1205	.1291	.1723	.2400	.2971	.3353	.4153	.4718	.6661	1.3131	1.3239	
120 Bright's disease	.2503	.0650	.0436	.0652	.0938	.1419	.2443	.3518	.5129	.7549	1.0648	1.5556	2.2639	3.0752	4.5269	6.6702	8.9061	11.5239	14.0790	13.8317	20.7414	8.0972
122 Other diseases of the kidneys and annexa	.1119	.0080	.0049	.0017	.0044	.0080	.0172	.0121	.0147	.0281	.0184	.0340	.0470	.0582	.0552	.1025	.1896	.2087	.2756	.1751	.4413	
123 Calculi of the urinary passages	.0018	.0005	.0004	.0008	.0004	.0031	.0013	.0029	.0016	.0050	.0059	.0061	.0133	.0058	.0114	.0222	.0271	.0272				
124 Diseases of the bladder	.0107	.0014	.0004	.0004	....	.0008	.0034	.0005	.0037	.0031	.0059	.0061	.0096	.0218	.0533	.1081	.1399	.2813	.7579	.7003	.8826	2.0243
125 Diseases of the urethra, etc	.0018	....	....	....	.0004	....	....	.0005	.0005	.0006	.0015	.0009	....	.0029	.0019	.0028	.0090					
128 Uterine hemorrhage	....	....	....	.0004	.0016	.0046	.0038	.0048	.0058	.0062	.0081	.0035	....	.0029	.0038							
129 Uterine tumor (non-cancerous)	.0018	.0005	....	....	.0004	.0027	.0092	.0363	.0614	.1148	.1389	.1012	.0566	.0742	.0743	.0804	.1219	.1543	.0459	....	.4413	
130 Other diseases of the uterus	.0036	.0009	.0004	.0021	.0116	.0301	.0503	.0499	.0561	.0661	.0551	.0410	.0265	.0233	.0552	.0360	.0316	.0635	.0459	.0875		
131 Cysts and other tumors of the ovary	....	....	.0008	....	.0016	.0088	.0189	.0218	.0278	.0343	.0382	.0384	.0325	.0553	.0762	.0804	.1219	.0544	.1148	.1751		
132 Salpingitis and other diseases of the female genital organs.	.0018	....	.0004	.0008	.0299	.0801	.1136	.1129	.1232	.0805	.0639	.0201	.0108	.0116	.0095	.0055	.0045	.0091	.0230			
133 Non-puerperal diseases of the breast	.0036	.0005	....	....	.0004	.0004	.0004	....	.0021	.0025	.0015	.0035	.0012	.0015	.0038	....	.0090	.0272				
134 Accidents of pregnancy	....	....	....	.0004	.0140	.0503	.0985	.1013	.0939	.0406	.0051	.0009										
135 Puerperal hemorrhage	....	....	....	....	.0068	.0385	.0738	.0858	.1002	.0437	.0088											
136 Other accidents of labor	....	....	....	....	.0144	.0564	.0729	.0741	.0965	.0618	.0073											
137 Puerperal septicemia	....	....	....	.0017	.1294	.3596	.4053	.3925	.2858	.1522	.0184	.0009										
138 Puerperal albuminuria and convulsions	....	....	....	.0021	.0703	.1586	.1639	.1657	.1489	.0973	.0088	.0017										
139 Puerperal phlegmasia alba dolens, embolus, sudden death	....	....	....	....	.0040	.0149	.0231	.0267	.0231	.0143	.0022											
140 Following childbirth (not otherwise defined)	....	....	....	....	.0020	.0038	.0059	.0058	.0037	....	.0007											
141 Puerperal diseases of the breast	....	....	....	....	.0004	....	.0008	.0005	.0005													
142 Gangrene	.0320	.0047	.0016	.0017	.0008	.0011	.0008	.0024	.0063	.0062	.0147	.0166	.0433	.0815	.1295	.3215	.6590	1.0344	2.0441	2.5387	4.8544	4.0486
143 Furuncle	.0337	.0019	....	....	.0012	.0015	.0013	.0019	....	.0012	.0015	.0070	.0084	.0073	.0057	.0222	.0181	.0181	.0459	....	.4413	
144 Acute abscess	.1119	.0075	.0029	.0013	....	.0031	.0025	.0039	.0031	.0019	.0044	.0070	.0072	.0145	.0114	.0333	.0497	.0544	.0230	.0875	.4413	
145 Other diseases of the skin and annexa	.1775	.0042	.0012	.0008	.0012	.0019	.0017	.0019	.0047	.0019	.0096	.0096	.0096	.0175	.0209	.0388	.0813	.1270	.1378	.0875	....	2.0243
146 Diseases of the bones (tuberculosis excepted)	.0799	.0264	.0200	.0152	.0096	.0053	.0122	.0073	.0079	.0112	.0125	.0183	.0229	.0218	.0305	.0360	.0677	.0544	.0919	.0875		

SPECIFIC DEATH-RATES PER 1000 POPULATION. REGISTRATION AREA, 1910, EXCLUSIVE OF NORTH CAROLINA—*Concluded*

Diseases.		Females.																					
		Under 1.	1–4.	5–9.	10–14.	15–19.	20–24.	25–29.	30–34.	35–39.	40–44.	45–49.	50–54.	55–59.	60–64.	65–69.	70–74.	75–79.	80–84.	85–89.	90–94.	95–99.	100 or over.
147	Diseases of the joints (tuberculosis and rheumatism excepted)	.0071	.0005	.0012	.0004	.0008	.0011	.0013	.0015	.0005	.0012	.0007	.0026	.0012	.0058	.0095	.0055	.0090	.0454	.0230	.0875		
148	Amputations	....	....	....	....	....	....	....	....	....	.0006	....	....	....	.0015								
149	Other diseases of the organs of locomotion	.0018	....	....	....	....	....	.0008	....	....	.0012	....	.0009	.0024	.0015	.0019	.0028	....	.0363	.0230			
150	Hydrocephalus	.4030	.0226	.0041	.0013	.0008	.0004	.0004	.0005														
	Congenital malformations of the heart	3.3502	.0325	.0077	.0042	.0024	.0011	.0008	.0010	.0010	....	....	.0009										
	Other congenital malformations	1.9405	.0141	.0020	....	.0004	.0008	....	....	.0005													
151	Premature birth	15.5900																					
	Congenital debility, "atrophy," "marasmus"	9.1097	.0033	.0004	....	....	.0004	.0004	.0005	....	....	.0007	....	....	....	....	....	....	.0091				
152	Injuries at birth	2.5371																					
	Other causes peculiar to early infancy	2.3826																					
153	Lack of care	.1083																					
154	Senility	....	....	....	....	....	....	....	....	....	....	....	....	....	.1615	.5656	2.0479	6.1706	18.0843	40.6752	81.7649	116.9462	224.6964
155 163	Suicide	....	....	....	.0055	.0711	.1148	.0972	.1071	.1211	.1067	.1161	.1186	.1217	.1047	.1295	.1219	.1309	.0817	.0919	.1751		
164 186	Accidental or undefined	1.0724	.7502	.2883	.1198	.1426	.1449	.1408	.1420	.1883	.1872	.2234	.3183	.3771	.5411	.9199	1.6073	3.0244	5.5986	9.9908	18.2964	21.6240	16.1943
182 184	Homicide	.0870	.0094	.0053	.0068	.0240	.0416	.0398	.0412	.0351	.0306	.0250	.0140	.0169	.0044	.0152	.0166	.0181	....	....	.0875		
187 189	Ill-defined diseases	5.5375	.2501	.0126	.0089	.0120	.0137	.0163	.0233	.0351	.0449	.0625	.0759	.1157	.2066	.3618	.6180	1.0879	1.7422	3.8585	6.3906	5.7370	8.0972

APPENDIX II

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 III

MATHEMATICAL FORMULÆ AND CONSTANTS

MULTIPLICATION

$$(1)\quad 1a = a;\ \ 3a = a + a + a$$

$$(2)\quad (a + b)\, c = ac + bc$$

$$(3)\quad (a - b)\, c = ac - bc$$

$$\begin{aligned}(4)\quad (a + b) \cdot (c + d) &= (a + b)\, c + (a + b)\, d \\ &= ac + ad + bc + bd\end{aligned}$$

$$\begin{aligned}(5)\quad (a - b) \cdot (c + d) &= (a - b)\, c + (a - b)\, d \\ &= ac + ad - bc - bd\end{aligned}$$

$$\begin{aligned}(6)\quad (a + b)\, (c - d) &= (a + b)\, c - (a + b)\, d \\ &= ac - ad + bc - bd\end{aligned}$$

$$\begin{aligned}(7)\quad (a - b)\, (c - d) &= (a - b)\, c - (a - b)\, d \\ &= ac - ad - bc + bd\end{aligned}$$

(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) $\dfrac{a}{b} + 1 = \dfrac{a + b}{b}$

$$(12)\quad \frac{a}{b} - 1 = \frac{a-b}{b}$$

$$(13)\quad \frac{1}{a} + \frac{1}{b} = \frac{a+b}{ab}$$

$$(14)\quad \frac{1}{a} - \frac{1}{b} = \frac{b-a}{ab} = -\frac{a-b}{ab}$$

$$(15)\quad \frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}$$

$$(16)\quad \frac{a}{b} - \frac{c}{d} = \frac{ad-bc}{bd}$$

$$(17)\quad \frac{a}{a+b} + 1 = \frac{2a+b}{a+b}$$

$$(18)\quad \frac{a+b}{2} + \frac{a-b}{2} = a$$

$$(19)\quad \frac{a+b}{2} - \frac{a-b}{2} = b$$

$$(20)\quad \frac{\frac{1}{a} + \frac{1}{b}}{\frac{1}{a} - \frac{1}{b}} = \frac{b+a}{b-a}$$

$$(21)\quad \frac{o}{a} = o$$

$$(22)\quad \frac{a}{o} = \infty$$

$$(23)\quad \frac{+a}{+b} \text{ or } \frac{-a}{-b} = +\frac{a}{b}$$

$$(24)\quad \frac{+a}{-b} \text{ or } \frac{-a}{+b} = -\frac{a}{b}$$

Powers

$$a^4 = aaaa;\ a^1 = a;\ 1^a = 1$$

(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) $(a : b)^m = \left(\frac{a}{b}\right)^m = \frac{a^m}{b^m}$

(6) $a^m \cdot a^n = a^{m+n}$

(7) $a^m : a^n = \frac{a^m}{a^n} = a^{m-n}$

(8) $a^{n+1} : a^n = a$

(9) $a^n : a^{n-1} = a$

(10) $\frac{1}{a^n} = \left(\frac{1}{a}\right)^n$

(11) $(a^m)^n = a^{mn}$

(12) $3a^0 = 3; (3a)^0 = 1$

(13) $a^{-1} = \frac{1}{a}$

(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) $\frac{a^x}{a^{-y}} = a^{x+y}$

(19) $\frac{a^{-x}}{a^y} = \frac{1}{a^{x+y}}$

(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) $\sqrt[3]{1 - x} = 1 - \frac{1}{3}x - \frac{1}{9}x^2 - \frac{5}{81}x^3 - \frac{10}{243}x^4 - \frac{22}{729}x^5 - \ldots$

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) $\left(\frac{a}{b}\right)^{\frac{m}{n}} = a^{\frac{m}{n}} : b^{\frac{m}{n}}$

(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) $\log \frac{M}{N} = \log M - \log N$.

(4) $\log (M)^n = n \log M$.

(5) $\log \sqrt[n]{M} = \frac{1}{n} \log M$.

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) $a = \frac{bc}{d}$; $d = \frac{bc}{a}$; $b = \frac{ad}{c}$; $c = \frac{ad}{b}$

(4) $ma : mb = c : d$

$ma : b = mc : d$, etc.

(5) $\frac{a}{n} : \frac{b}{n} = c : d$

$\frac{a}{n} : b = \frac{c}{n} : d$, 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) $y = x^n, \frac{dy}{dx} = nx^{n-1}$

$y = ax^n, \frac{dy}{dx} = nax^{n-1}$

(2) $y = \frac{a}{x^n}, \frac{dy}{dx} = -\frac{na}{x^{n+1}}$

$= ax^{-n}, \frac{dy}{dx} = -nax^{-n-1}$

(3) $y = a\sqrt[n]{x}, \frac{dy}{dx} = \frac{a}{n}\sqrt[n]{x^{1-n}}$

$= ax^{\frac{1}{n}}, \frac{dy}{dx} = \frac{1}{n}ax^{\frac{1}{n}-1}$

(4) $y = a\sqrt[n]{x^m}, \frac{dy}{dx} = \frac{m}{n}a\sqrt[n]{x^{m-n}}$

$= ax^{\frac{m}{n}}, \frac{dy}{dx} = \frac{m}{n}ax^{\frac{m}{n}-1}$

(5) $y = \sqrt{x}, \frac{dy}{dx} = \frac{1}{2\sqrt{x}}$

$= x^{\frac{1}{2}}, \frac{dy}{dx} = \frac{1}{2}x^{-\frac{1}{2}}$

(6) $y = e^x, \frac{dy}{dx} = e^x$

(7) $y = a^x, \frac{dy}{dx} = a^x \log_e a$

(8) $y = \log_e x, \dfrac{dy}{dx} = \dfrac{1}{x}$

$y = \log x, \dfrac{dy}{dx} = M \cdot \dfrac{1}{x}$, where $M = 0.43429$

(9) $y = \sin x, \dfrac{dy}{dx} = \cos x$

(10) $y = \cos x, \dfrac{dy}{dx} = - \sin x$

(11) $y = \text{tg}\, x, \dfrac{dy}{dx} = \dfrac{1}{\cos^2 x}$

(12) $y = \text{ctg}\, x, \dfrac{dy}{dx} = - \dfrac{1}{\sin^2 x}$

(13) $y = \text{arc} \sin x, \dfrac{dy}{dx} = \dfrac{1}{\sqrt{1 - x^2}}$

(14) $y = \text{arc} \cos x, \dfrac{dy}{dx} = - \dfrac{1}{\sqrt{1 - x^2}}$

(15) $y = \text{arc tg}\, x, \dfrac{dy}{dx} = \dfrac{1}{1 + x^2}$

(16) $y = \text{arc ctg}\, x, \dfrac{dy}{dx} = - \dfrac{1}{1 + x^2}$

Simple Integrals

(1) $\int a\, dx = ax + C$

(2) $\int ax^n\, dx = \dfrac{ax^{n+1}}{n + 1} + C$

(3) $\int e^x\, dx = e^x + C$

(4) $\int \dfrac{1}{x}\, dx = \log_e x + C$

(5) $\int a^x\, dx = \dfrac{a^x}{\log_e a} + C$

(6) $\int \sin x\, dx = - \cos x + C$

(7) $\int \cos x\, dx = \sin x + C$

$$(8)\quad \int a \cos x \, dx = a \cdot \sin x + C$$

$$(9)\quad \int \frac{1}{\cos^2 x} \, dx = \operatorname{tg} x + C$$

$$(10)\quad \int \frac{1}{\sin^2 x} \, dx = -\operatorname{ctg} x + C$$

$$(11)\quad \int \frac{1}{\sqrt{1 - x^2}} \, dx = \operatorname{arc} \sin x + C$$

$$= -\operatorname{arc} \cos x + C'$$

$$(12)\quad \int \frac{1}{1 + x^2} \, dx = \operatorname{arc} \operatorname{tg} x + C$$

$$= -\operatorname{arc} \operatorname{ctg} x + C'$$

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
	$\frac{1}{\pi} =$	0.3183099	9.5028501 − 10
	$\pi^2 =$	9.8696044	0.9942997
	$\sqrt{\pi} =$	1.7724539	0.2485749
	$\sqrt[3]{\frac{\pi}{6}}$		9.9063329 − 10

APPENDIX IV

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