THE FUNDAMENTAL PRINCIPLES OF MATHEMATICAL STATISTICS

WITH SPECIAL REFERENCE TO THE REQUIREMENTS OF ACTUARIES AND VITAL STATISTICIANS

AND

AN OUTLINE OF A COURSE IN GRADUATION

By

HUGH H. WOLFENDEN

Fellow of the Institute of Actuaries, Great Britain
Fellow of the Actuarial Society of America
Fellow of the American Institute of Actuaries
Fellow of the Royal Statistical Society

PUBLISHED FOR THE ACTUARIAL SOCIETY OF AMERICA, NEW YORK
BY
THE MACMILLAN COMPANY OF CANADA LIMITED, TORONTO
1942

Printed in Canada
at
The University of Toronto Press

FOREWORD

We appreciate the opportunity which the author has graciously given to the Actuarial Society of America to publish this book. Representing the fruits of his own intensive work in the field of Mathematical Statistics as an actuary, supplemented by his experience in lecturing to actuarial students, this volume should be of great assistance to students who are working on basic elements of theory and to actuaries who wish to acquire greater mastery in more advanced aspects of the subject. We believe that it will also be consulted by those who work in statistical fields outside the actuarial profession.

The Actuarial Society of America is making this publication available in pursuance of its policy to aid in the educational facilities for entrance into the actuarial profession and in the proficient pursuit of actuarial science. We acknowledge with deep gratefulness the author's devotion to his profession in contributing his own time and efforts to these same ends.

J. M. LAIRD,
President.

R. D. MURPHY,
Chairman of the Committee on Actuarial Studies.

THE ACTUARIAL SOCIETY OF AMERICA

"Look to the essence of a thing, whether it be a point of doctrine, of practice, or of interpretation."

Marcus Aurelius.

TABLE OF CONTENTS

THE

FUNDAMENTAL PRINCIPLES

OF

MATHEMATICAL STATISTICS

PREFACE

IT MAY possibly at first seem strange that yet another book should be prepared on Mathematical Statistics, when so many admirable text-books on various phases of the subject have been published within recent years.

Actuaries, however, and others, such as vital statisticians, who in their work find themselves concerned, sooner or later, with the foundations of the theory of probability as well as with its application to advanced and special problems concerning life contingencies and allied questions, have never yet been able to avail themselves of any single ordered and comprehensive treatment which would give them precisely those portions of the theory and practice which they require. The field is so vast, the literature so scattered and extensive, and the viewpoints from which the subject has been presented are so varied, that no present publication has seemed to the writer—in a long experience of lecturing to actuarial students—to fill the somewhat highly specialized requirements of the actuary.

These considerations form the sole reason—indeed the only justification—for this book. In it an attempt has been made to assemble and co-ordinate those portions of the theory and applications of Mathematical Statistics which are really needed, both by actuarial beginners in their studies, and by qualified actuaries in the solution of the problems which arise in practice. The selection of material, and the treatment, have thus been developed along special lines, with a particular objective. In some portions the phraseology adopted has been deliberately repetitious—for the main objective has been to clarify the mathematical foundations of the subject, without, however, bewildering the ordinary student by an intricate maze of highly condensed symbolism.

One other viewpoint should perhaps also be explained. In the teaching of these matters there is, inevitably, a cultural responsibility, as well as a merely pedagogic duty—only by an adequate presentation of the former can the latter be fully and satisfactorily achieved. It may well be desirable, amongst all the facets and complexities of modern scholarship, so to present the practical achievements in any field of learning that he who, of necessity, must run at the tempo of modern life may read with conservation of his energy; but this accomplishment must leave much of the cultural background unassimilated unless some reasonable attempt is made to picture the historical development. A recital merely of the present state of knowledge, moreover, incurs the danger—seen more than once in the scientific world—that some investigator in the future may re-discover work in ignorance of similar research undertaken years before. I therefore believe it essential to the proper understanding of any subject to absorb the history of the mental processes which have guided its development. This study, accordingly, is framed on that conviction. It is hoped, however, that the arrangement used will enable the reader to acquire the background easily, and in a manner less destructive of imaginative interest than is so often inseparable from the teaching of history *per se*.

In one major respect that arrangement is, I believe, novel. An endeavour has been made to reduce the distractions which inevitably arise when the many essential explanatory discussions and extended mathematical analyses are inserted in the main text. The body of the treatment has therefore been designed as a condensed presentation only, from which the principal ideas may be acquired in an orderly and easy manner—the subsidiary questions which naturally arise, and which must, of course, be answered, being dealt with by reference to separate portions of the book. By this means it is hoped that the student approaching the subject for the first time, or the graduate who may desire to refresh his views, will be able early in his reading to obtain a comprehensive picture of the whole, while yet possessing ample opportunities to elaborate the background or the current details as and when he may desire.

The preparation of this book has been in my thoughts for many years. On the outbreak of war in 1939 the manuscript was already nearly finished, and its early publication was intended. The dislocations of recent months, however, have caused unavoidable delay; but its completion was encouraged by the desires of many students, and by the interest of professional friends. I therefore hope that publication now may be justified, despite the war, by any assistance the book may give to those who, in the years to come, will carry the burdens of dealing, statistically or actuarially, with the problems either of a war economy or of a saner and more peaceful world.

The final stages before publication have been facilitated notably by correspondence and conversations with Dr. W. Edwards Deming on certain points in the history of statistical theory, and the "Student" and χ^2 tests. Mr. Donald D. Cody has generously devoted many hours to a critical reading of almost the whole manuscript, and has contributed valuable suggestions resulting in clarifications and some enlargement of the text. The diagrams (except Figure 20, which was evolved by Dr. Deming during a discussion of the "Student" theory) have been prepared by Mr. John B. McKinnon, who also has assisted greatly in the arrangement of the bibliography. Mr. Ray D. Murphy has given most encouraging co-operation as Chairman of the Committee on Actuarial Studies of the Actuarial Society of America. It is a real pleasure to record my thanks to these four friends.

HUGH H. WOLFENDEN

Toronto.
February, 1942.

I. INTRODUCTION

CERTAIN aspects of the study of Mathematical Statistics have always seemed to present special difficulties to students. This is particularly true in those portions which involve the classical **Theory of Errors** and the **Method of Least Squares,** and their relation to some of the more advanced developments of the modern **Theory of Sampling,** systems of **Frequency Curves,** and the **Method of Moments.**

In the case of actuarial students these difficulties have been due only slightly to any lack of basic training in the earlier mathematical requirements—for they come to the subject with a sound practical knowledge of the elements of the **Theory of Probability,** and with an adequate facility in most of the necessary fundamentals of the differential and integral calculus. They are attributable rather to a hiatus which exists in the usual courses of study. Notwithstanding some improvement within recent years, students are still generally plunged almost directly from those simple elements of probability and calculus, as they are taught in the text-books, into all the complex ramifications of Mathematical Statistics, without any really sufficient preparation in the underlying theories. It is consequently little wonder, when there, almost for the first time, they meet constant references to an enormous mass of historical literature, and encounter the classical disagreements and still prevailing conflicts engendered by metaphysical speculations on the nature of probability and the theories flowing from it, that even the best students sometimes feel bewildered, and have sought a presentation of the subject which would lead by easy stages through the apparent maze.

It is the purpose of this volume to attempt that task. Only the elements of probability and calculus will be assumed as known. From that point the development will follow largely the classical discussions, with emphasis upon the essential place of the Theory of Errors and Least Squares. It has always seemed to me impossible to approach the modern theories with any hope of success if we leap over all those concepts—"errors", the "normal curve", "mean square of error", "probable error", "weights", "normal equations", and so forth—which impelled so many of the controversies and yet have formed the stepping-stones of history towards more recent methods. The ultimate result of neglecting any of these fundamental matters can only be confusion of the student's mind.

In thus treating the subject from its elements and through its classical discussions an endeavour will be made, however, to free the body of the text, as much as possible, from historical descriptions, elaborate demonstrations, and instances of application. These three types of information will therefore be found in three appended sections. Their relegation in that manner is not, of course, intended to imply their unimportance—for, as a principle of pedagogy, nothing is more essential in the presentation of any subject than a picture of the background (given in Section A on History), an understanding of the technical analysis (provided by Section B on Mathematics and Interpretations), and an appreciation of the special practical utilities and applications (shown in Section C on Applications). It is therefore hoped that references to these portions of the study will not be disregarded. They will be given in brackets, in the form (p. 163; A; 7), for example, meaning that the supplementary material will be found at p. 163 of this volume, in Section A, Part 7, in that case.

Historically important and currently useful publications will be brought together in a Bibliography, and identified by italic numerals in two lists, with page reference where possible, and the letter H or P to indicate "historical" or "present" value—as H:*33*:329, meaning the "H" list of works of historical significance, publication number *33*, p. 329 thereof. An attempt has been made to give complete (though not redundant) documentation, for two reasons—firstly, because it seems desirable, in matters so inseparably concerned with the logical dilemmas of philosophical probability, to refer wherever possible to those publications which may properly be considered as authoritative in their respective fields; and secondly, in order that the student, if he should feel dissatisfied at any stage with the development or views set out herein, may have readily available the sources of original deductions and supplementary discussions on each topic.

II. THE NATURE OF THE PROBLEMS

THE actuarial student is ordinarily introduced to the elementary notion of Probability in his text-books on algebra. Following instruction in permutations and combinations, he is led to a "definition" of the concept, is taken through the "addition" and "multiplication" rules, and is quietly permitted to assimilate—in his own fashion, and generally without question at that stage—the ideas of "mutually exclusive", "independent", "dependent", and "equally likely" events. He is then exercised actively in the combination and unravelling of the probabilities of complex sets of occurrences, in preparation for the further drilling which he must receive in his manipulation of such probabilities in dealing with life contingencies.

It is well that, in this curriculum, he is not brought deliberately into touch with those subtleties of thought and metaphysical speculations which he now must meet, and in some way resolve, in attempting to apply his text-book training in the wider field of Mathematical Statistics. For in this wider field he will encounter early certain philosophic doubts as to the precision of the text-book "definitions"; he will wonder whether the hitherto accepted axioms may not perhaps, from another point of view, be theorems; he will discover intriguingly conflicting attempts to interpret "equally likely" events; and he will find, all through the history, and still today, disputes concerning the validity of "inverse" ("inductive" or "*a posteriori*") probability (the "probability of causes" in the past), much argument expended on its relationship to "direct" ("deductive" or "*a priori*") probability (the probability of a specified event in the future), and disagreement yet about those "paradoxes" to which so much analysis has been devoted notwithstanding often a fundamental inability to settle precisely the nature of the problems.*

It is accordingly in an atmosphere of controversy that we must make a start. In order to avoid the danger of confusing the student at this stage, we may omit from the body of this treatment any discussion of the extremely interesting speculations which may be formulated with regard to the nature of probability and, consequently, the best method of approaching it. The nature of probability is in many respects so elusive a conception that it must necessarily be an arbitrary effort to defend a "best" approach, except within certain stated limitations (p. 179; B; 1). We shall, however, select, with reason, one approach, namely, that which leads as directly as possible to the concept of **statistical frequency.** It will be found, upon examination, to be adequate for treating the essentially practical aspects of probability with which actuaries and similar investigators are constantly brought into touch, while it avoids many—though, as will be seen, not all—of the logical dilemmas and practical difficulties of the other methods (p. 183; B; 1).

This concept of *statistical frequency* may be approached, in the first place, by considering the essential meaning of **James Bernoulli's Limit Theorem** (H:4)—the "*Law of Large Numbers*"—which, in broad terms, may be stated thus: "If p be the true probability of the happening of a certain event in a single trial, n a number of trials, and s the number of times the event is observed to happen in those n trials, then, as n increases, the probability approaches certainty that the statistical frequency, $\frac{s}{n}$, will approach p" (see p. 187; B; 2). This statement may evidently be interpreted as meaning that $\lim_{n \to \infty} \frac{s}{n} = p$. The question which therefore immediately presents itself concerns the

*These questions concerning "inverse" probability are deliberately, although reluctantly, excluded from this study. They are not essential for the purposes immediately in view—although the student will discover ultimately that he cannot escape from their due consideration if he is to grasp fully the meaning and implications of the whole subject.

situation which arises when the number, n, of trials is *not* infinitely great. We are thus brought at once to the necessity of investigating the *deviations* which are likely to occur when the number, n, of trials is limited (see also p. 263; C; 1).

In its classical form—circumscribed by certain specific and hampering assumptions—the theory of deviations so approached became known as the Theory of Errors, in which the Law of Large Numbers, the symmetrical Normal Curve of Error, and the "adjustment" of observations by the consequent Method of Least Squares represent the important subdivisions. In its more recent developments—released from the necessities of those earlier assumptions—it leads directly also to the Lexis Theory, the Theory of Random Sampling, Poisson's Law of Small Numbers, the general representation of "probability distributions" or "frequency distributions" by both symmetrical and unsymmetrical functions as in the Pearsonian Frequency Curves and Generalized Normal Curves (with the Method of Moments for fitting such curves), and to the χ^2 Test for Goodness of Fit.

III. THE CLASSICAL APPROACH

THE important distinction stated in the preceding chapter between the true *a priori* probability, p, and the statistical frequency, say, $\frac{s}{n}$, resulting from s actual observations of the success (*i.e.*, the happening) of an event in, let us suppose, n trials—where n is not infinitely great—may be crystallized in practical form (see also p. 187; B; 2) by the following question:

What is the probability, in a finite number of trials, n, in each of which the true probability of the event happening is a constant p (and the true probability of its not happening, or $1-p$, is q), that the event will actually happen $np+x$ times, i.e., will show a deviation of $+x$ from the np occurrences which would be expected (according to the Law of Large Numbers) if n were infinitely great? (See p. 264; C; 2.)

This probability is immediately expressible according to the usual rules of probability for independent trials with constant probability p. For, since the trials are independent, the probability of the event happening exactly $np+x$ times in any given order is p^{np+x}, and of its failing the other $n-(np+x)$, or $nq-x$, times, is q^{nq-x}; and since all the different sequences of the $np+x$ happenings (and the consequent $nq-x$ failures) can occur in ${}^nC_{np+x}$ different arrangements, the total probability, y_x say, is

$$ {}^nC_{np+x}\, p^{np+x}\, q^{nq-x}, \text{ where } p+q=1 \qquad \ldots.(1)$$

that is,

$$ y_x = \frac{n\,!}{(np+x)!\,(nq-x)!}\, p^{np+x} q^{nq-x} \qquad \ldots.(2)$$

This expression, of course, is the general term of the series obtained by setting out the probabilities of exactly 0, 1, 2, ... n successes, i.e., by putting $np+x=0, 1, 2, \ldots n$; that is to say, it is the general term of the series

$$ q^n + npq^{n-1} + \frac{n(n-1)}{1.2}\, p^2 q^{n-2} + \ldots + p^n = (q+p)^n \quad \ldots.(3)$$

FIGURE 1.—An Unsymmetrical Point Binomial

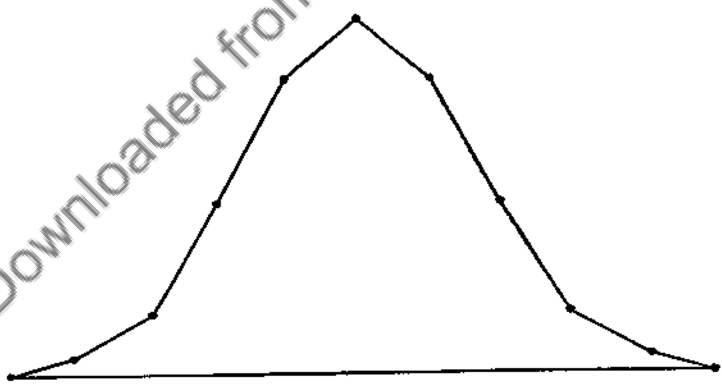

FIGURE 2.—A Symmetrical Point Binomial ($p=q$)

which is often referred to as the **Bernoullian Series** or the **Point Binomial.** It may also be useful at this stage to introduce into the terminology the idea conveyed by its alternative description as the Bernoullian or binomial **Frequency Distribution** (see p. 189; B; 3), since the series in fact obviously represents the relative frequencies (not the actual frequencies, which would be n times the relative frequencies) with which the various possible numbers of happenings are theoretically distributed over the range from no occurrences to complete success when p remains constant throughout (see p. 264; C; 2). It will be apparent that the distribution is unsymmetrical when $p \neq q$, as in the example shown in Figure 1, and that when $p = q (= \frac{1}{2})$ the shape becomes symmetrical, as in Figure 2 (see also p. 189; B; 3).

From these expressions certain functions of great importance can now be deduced easily. It will be well to set out the demonstrations *in extenso*, since exactly the same principles will be required later in more complex circumstances.

First of all, let us examine the **average**, or **mean**, expected number of happenings in the n trials. From (2) and (3) it is clear that the probability of 0 happenings is q^n, with consequently an expected number $q^n(0)$; that the probability of 1 success is npq^{n-1}, with an expected number, therefore, of $npq^{n-1}(1)$; and so on until finally we reach the last term $p^n(n)$. The total expected number of happenings in the n trials is therefore the sum of these terms, or

$$q^n(0) + npq^{n-1}(1) + \frac{n(n-1)}{1.2} p^2 q^{n-2}(2) + \ldots + p^n(n)$$
$$= np(q+p)^{n-1} = np \qquad \ldots .(4)$$

since $q + p = 1$.

What now would be the average of the expected *squares* of the number of happenings? By reasoning similar to that above it can immediately be written down as

$$q^n(0)^2 + npq^{n-1}(1)^2 + \frac{n(n-1)}{1.2} p^2 q^{n-2}(2)^2 + \ldots + p^n(n)^2$$

$$= np\,[q^{n-1} + (n-1)pq^{n-2}\,(2) + \ldots + p^{n-1}(n)]$$
$$= np\,[(q+p)^{n-1} + (n-1)\,p(q+p)^{n-2}]$$
$$= np\,[1 + (n-1)p] = np(np+q) \qquad \ldots.(5)$$

The preceding formulae can, of course, be put very simply in terms of summations—(4) being written $\sum_{t=0}^{t=n} {}^nC_t\, q^{n-t}\, p^t\,(t)$, and (5) as $\sum_{t=0}^{t=n} {}^nC_t\, q^{n-t}\, p^t\,(t)^2$. They are, moreover, in fact the mean and the **second moment** (see p. 253; B; 27) of the binomial distribution with reference to the beginning of the range.

Suppose now that, instead of (5), we investigate the expected *squares of the deviations measured from the mean np*, as defined by the formula $\sum_{t=0}^{t=n} {}^nC_t\, q^{n-t}\, p^t(t-np)^2 \qquad \ldots.(6)$

This is the **second moment about the mean** (see p. 253; B; 27), and is known as the **mean square deviation** (see p. 162; A; 6). In expanded form the expression is

$$q^n(n^2p^2) + nq^{n-1}p(1-2np+n^2p^2) + \ldots + p^n(n^2-2n^2p+n^2p^2)$$

which, by the same methods as were used to establish (4) and (5), reduces easily (for proof see p. 202; B; 4) to

$$npq \qquad \ldots.(7)$$

The **standard deviation**, σ, being defined as the square root of the mean square deviation, is consequently

$$\sigma = \sqrt{npq} \qquad \ldots.(8)$$

It will thus be seen that, up to this point, some of the most obviously important problems, as they arise naturally from the simplest examination of the theory of probability, have been dealt with by the ordinary processes of algebra. It may be well to emphasize here that, in so doing, the student has been introduced to the following basically important notions: (*a*) The distinction and relation between a true probability, *a priori*, and the observed statistical frequency, *a posteriori*; (*b*) the most elementary type of theoretical frequency distribution for the number of successes in *n* trials, i.e., the discontinuous "Bernoullian series" or "point binomial"; and (*c*) three functions derived therefrom, namely, the "mean" np, the "mean square deviation" npq, and the "standard deviation" $\sqrt{npq}$.

Now, as is shown at p. 264; C; 3, the preceding methods clearly are quite simple, and do not encounter any difficulty even if the number of trials, n, be made very large. There are, however, other functions which can be derived from the point binomial without undue effort if n is small, but which obviously would entail prohibitive labour when n becomes large. Suppose, for example, it were desired to find the probability of a deviation between certain stated limits, which would necessitate the summing of the corresponding terms of the point binomial; it would be an easy enough matter when n is small, but impossibly formidable if, as very often happens, n is large. It consequently became essential to seek a method which would bring the processes involved within the realm of practical accomplishment when the number of trials, n, and the consequent factorials, are increased greatly. We therefore now proceed to set out the classical approach to this important question.

The problem is to find a method of dealing with the fundamental probability, already given in (2), of $np+x$ successes (rather than the expected number np) in n independent trials with constant probability p, i.e., the probability

$$\frac{n!}{(np+x)!\,(nq-x)!}\,p^{np+x}\,q^{nq-x} \qquad \ldots.(2)$$

when n is large.

This can be done by Stirling's formula (p. 151; A; 1) that*

$$n! \doteqdot \sqrt{2\pi n}\,n^n e^{-n}\left(1+\frac{1}{12n}+\ldots\right) \qquad \ldots.(9)$$

Thus, replacing the factorials in (2) by this expression with the term involving $\frac{1}{12n}$ neglected (see p. 151; A; 1), taking logar-

* $\doteqdot$ denotes "is approximately equal to".

ithms, and reducing, we obtain easily (as on p. 203; B; 5) the *approximation*

$$y_x \doteq \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}} \qquad \ldots.(10)$$

which may be identified as the **Normal Law of Deviations** (P:*155*:173). Its discovery is to be attributed to De Moivre (p. 151; A; 2).

It will be remembered that $np+x$ is here restricted to integral values, so that the formula—being in fact at this stage an approximation to the point binomial—represents still a bell-shaped series of ordinates (of the type of Fig. 2). It is, however, also to be observed that, whereas (2) started out as a symmetrical or unsymmetrical series, according as $p=$ or $\neq q$, (10)—through the neglect of certain terms in the development of the approximation (as shown on p. 204; B; 5)—has emerged as a symmetrical expression in every case (whether $p=$or$\neq q$), since $y_x = y_{-x}$. For the purposes of obtaining, through this approximation, a method by which calculations can be performed when n is large, this transformation of the sometimes unsymmetrical expression (2) into the necessarily symmetrical formula (10) is not as serious a change as might at first appear. The facilities which it provides are very great, and its results are remarkably accurate, except in the comparatively unusual cases when q (or p) is so small and n sufficiently large that nq (or np) remains finite but small (less than about 10, perhaps), i.e., q (or p) is small but n is sufficiently large that the event happens only very occasionally (see p. 267; C; 4, and p. 310; C; 14).

If we now write $c = \sqrt{2npq}$ the expression remains simply an approximate formula $y_x \doteq \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$. But if its discontinuous series of ordinates, depending upon n, p, and q, be now smoothed out into a continuous bell-shaped curve (Fig. 8 on p. 69)—depending either approximately upon specific values of

n, p, and q, or upon other general and non-specific considerations—it may obviously be imagined in the form

$$f(x) = \frac{1}{c\sqrt{\pi}}\, e^{-\frac{x^2}{c^2}} \qquad \ldots .(11)$$

In this guise it has become famous as the **Normal Curve of Error,** upon which the classical theory was founded (see p. 151; A; 3), and to which modern theories—despite many efforts to break away from it—still not infrequently return.

It will be important for the student to remember that in determining formulae (4) and (5) the origin was taken at the beginning of the range, but that the deviation x in (2), (10), and (11), and formulae (6), (7), and (8), are based on the origin taken at the mean.

From the preceding method of approach it will be seen at once that there is a close analogy, but also a distinction, between the deviation in (10) and the error x as it obviously can be visualized in the more general Normal Curve of Error (11). The deviation x in (10) proceeds by *integers* (since $np+x$ is integral, although neither np nor x is necessarily integral), and depends on known *a priori* values of n, p, and q; the "error" x in (11) clearly includes also the concept of a difference, *either integral or fractional*, which may be supposed to have occurred from any cause or causes, whether expressible *a priori* in terms of n, p, and q, or to be related to the *true* value *a posteriori* by analysis of a set of observations (see p. 187; B; 2, and p. 263; C; 1). The analogy is important on account of the legitimate support which it provides for the almost simultaneous deduction of both (10) and (11), as is shown here; the distinction, however, is likewise to be emphasized, because it involves much of the philosophical debate which stimulated the attempts to "prove" the Normal Law (see p. 151; A; 3, and p. 158; A; 4).

A number of important functions can now be deduced easily from (10) and (11). Remembering that the deviation x in (10) is measured from the mean, np (as shown in (4)), and since (10) is, in fact, an approximate expression for the point binomial when n is large, we shall begin by using the calculus to establish the values of the mean, and of the mean square deviation, which have already been found for the point binomial by the use merely of simple algebra. It will be convenient and sufficient to deal here with the continuous form (11), and to direct attention to p. 206; B; 6 with regard to the restrictions necessitated in the case of (10) by the requirement of finite integration.

The probability of a deviation, or error, x being then, by (10) and (11), expressible as $\frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}$, where $c=\sqrt{2npq}$ in (10), we may take that of an error between x and $x+\delta x$ (where δx is infinitesimally small) as $\frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}\,\delta x$ (see p. 208; B; 7). In n trials, accordingly, the expected number of deviations or errors lying between x and $x+\delta x$ would be $\frac{n}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}\,\delta x$, with an expected magnitude of $\frac{n}{c\sqrt{\pi}}xe^{-\frac{x^2}{c^2}}\,\delta x$. When, therefore, a deviation or error of any magnitude and either sign is possible, as in (11), so that the range of x is from $-\infty$ to $+\infty$, it follows that in the n trials the total expected magnitude of the deviations or errors would be $\frac{n}{c\sqrt{\pi}}\int_{-\infty}^{+\infty}xe^{-\frac{x^2}{c^2}}\,dx$, and consequently that the average magnitude, with reference to a single trial, is

$$\frac{1}{c\sqrt{\pi}}\int_{-\infty}^{+\infty}xe^{-\frac{x^2}{c^2}}\,dx=\frac{1}{c\sqrt{\pi}}\left[-\frac{c^2e^{-\frac{x^2}{c^2}}}{2}\right]_{-\infty}^{+\infty}=0 \quad \ldots .(12)$$

That is to say, the *mean* here $=0$, i.e., the origin of (11) is at the mean, which obviously corresponds properly to the determination in (4) of the mean as np for the point binomial (3), for there the origin was taken at the beginning of the range (see also p. 208; B; 7).

If, instead of giving effect to the sign of x, as in the derivation of (12), we take the mean of the absolute values irrespective of sign, we find the

average or mean (expected) error, irrespective of sign, or η,

$$= \frac{2}{c\sqrt{\pi}} \int_0^\infty x e^{-\frac{x^2}{c^2}} dx = \frac{c}{\sqrt{\pi}} \qquad \ldots.(13)$$

by (*c*) and (*d*), p. 209; B; 7.

Similarly, the average of the expected *squares* of the errors, i.e., the **mean square error** or **second moment**, or **variance** (see p. 163; A; 6)

$$= \frac{1}{c\sqrt{\pi}} \int_{-\infty}^{+\infty} x^2 e^{-\frac{x^2}{c^2}} dx = \frac{2}{c\sqrt{\pi}} \int_0^\infty x^2 e^{-\frac{x^2}{c^2}} dx = \frac{c^2}{2} \qquad \ldots.(14)$$

by (*e*), p. 210; B; 7.

The **standard deviation,** σ, being defined as the square root of the mean square error as in (8), is therefore

$$\sigma = \frac{c}{\sqrt{2}} \qquad \ldots.(15)$$

The **probable error,** λ, or "quartile deviation" (see p. 192; B; 3) is the error such that there is an even chance that any error will fall short of or exceed it in absolute magnitude. That is, λ is to be determined from

$$\frac{1}{c\sqrt{\pi}} \int_{-\lambda}^{+\lambda} e^{-\frac{x^2}{c^2}} dx = \tfrac{1}{2}, \text{ or } \frac{2}{\sqrt{\pi}} \int_0^{\frac{\lambda}{c}} e^{-t^2} dt = \tfrac{1}{2} \qquad \ldots.(16)$$

from which, by tables of the *probability integral* (p. 161; A; 5),

$$\lambda = .476936c^* \qquad \ldots.(17)$$

From the above results it follows that

$$c = \sqrt{2npq} = \sqrt{\pi} \text{ (mean error)} = \sqrt{2} \text{ (standard deviation)}$$

$$= \frac{1}{.476936} \text{ (probable error)} \qquad \ldots.(18)$$

$$\text{or } c = \sqrt{2npq} = 1.772454\eta = 1.414214\sigma = 2.096665\lambda \qquad \ldots.(19)$$

*The values are stated here to 6 places for purposes of record. In practice, however, 3 places are quite sufficient.

and hence, *for the conditions of "simple sampling"* (see Chapter V) underlying (10),

$$\text{Mean Error} = .797885\sqrt{npq} = .797885\sigma = \tfrac{4}{5}\sigma \text{ approximately} \quad \ldots.(20)$$

and

$$\text{Probable Error} = .674489\sqrt{npq} = .674489\sigma = \tfrac{2}{3}\sigma \text{ approximately} \quad \ldots.(21)$$

The order of magnitude of these quantities is the probable error (λ), mean error (η), standard deviation (σ), and modulus (c—see p. 163; A; 6), as shown in Figure 3 (in which c is taken as 1, and therefore $\lambda = .48$, $\eta = .56$, $\sigma = .71$, $2\sigma = 1.41$, and $3\lambda = 1.43$, to 2 places).

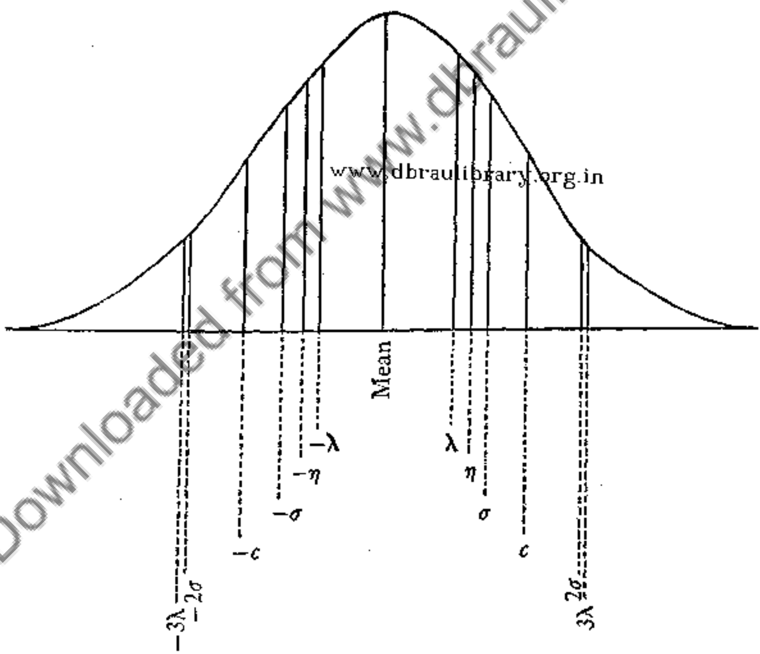

FIGURE 3.—Relative Locations of λ, η, σ, and c, and their Multiples, in the Normal Curve

It will be seen immediately from this diagrammatic presentation that λ, η, σ, and c vary in their significance. For example, remembering that, by definition—formula (16)—the probability of a deviation lying within the range $\pm\lambda$ is .5, it can easily be found similarly from tables of the "probability integral" (see p. 161; A; 5) that the probabilities of a deviation lying within the ranges $\pm\lambda$, $\pm2\lambda$, $\pm3\lambda$, $\pm4\lambda$, $\pm5\lambda$, . . . are .500, .823, .957, .993, .9993, . . . , or within the ranges $\pm\sigma$, $\pm2\sigma$, $\pm3\sigma$, $\pm4\sigma$, . . . are .6827, .9545, .9973, .99994, . . . , respectively. In other words, *in a "normal" distribution* the proportions of the total area included between the curve, the x-axis, and the ordinates through $\pm\lambda$, $\pm2\lambda$, $\pm3\lambda$, or $\pm4\lambda$, are 50%, 82%, 96%, or 99%, while the corresponding percentages for the ordinates through $\pm\sigma$, $\pm2\sigma$, or $\pm3\sigma$ are 68%, 95%, or over 99%.

It follows also, therefore, that the proportions outside $\pm\lambda$, $\pm2\lambda$, $\pm3\lambda$, or $\pm4\lambda$ are 50%, 18%, 4%, or 1%, and that those outside $\pm\sigma$, $\pm2\sigma$, or $\pm3\sigma$ are 32%, 5%, or under 1%.

These results, furthermore, may obviously—by reason of the symmetry of the Normal Curve—be stated comparably in terms of either the plus or minus halves of the discrepancies, taken separately. Thus the probabilities of a positive deviation lying beyond $+\lambda$, $+2\lambda$, $+3\lambda$, $+4\lambda$, $+5\lambda$, . . . are .250, .089, .022, .0035, .00035, . . . , with the same probabilities for negative deviations beyond $-\lambda$, -2λ, -3λ, -4λ, -5λ; and the probabilities of a positive deviation beyond $+\sigma$, $+2\sigma$, $+3\sigma$, $+4\sigma$, . . . , are .1587, .0228, .00135, .00003, and the same for a negative deviation beyond $-\sigma$, -2σ, -3σ, -4σ, . . . (see p. 269; C; 6).

It will be useful to remember from these results that, for a "normal" distribution (*i*) twice or three times $\pm\sigma$ includes an area about the same as three or four times $\pm\lambda$; (*ii*) a deviation of more than $\pm2\sigma$ or $\pm3\lambda$ is very unlikely (the percentage of occurrence being less than 5%); and (*iii*) $\pm3\sigma$ and $\pm4\lambda$ embrace over 99% of the deviations.

The preceding account of the Classical Approach will, to this stage, have familiarized the student with the following basic ideas (in which are now again included, for completeness, those already noted up to p. 13):

(*a*) The distinction and relation between a "true" probability, *a priori*, and the observed "statistical frequency", *a posteriori*;

(*b*) The discontinuous, and symmetrical or unsymmetrical, "point binomial" as the most elementary type of "frequency distribution" for the number of successes in n trials;

(*c*) The "mean" np, the "mean square deviation" npq, and the "standard deviation" $\sqrt{npq}$, for that "point binomial";

(*d*) The symmetrical "Normal Law of Deviations" (10) as a close approximation to the "point binomial" (except in special circumstances), and the symmetrical "Normal Curve of Error" (11) as a general representation of "deviations" or "errors" from the mean;

(*e*) For these expressions, the mean at the origin $=0$; $c=\sqrt{2npq}$; the mean error $=\frac{c}{\sqrt{\pi}}$; the mean square deviation $=npq$; the standard deviation $=\sqrt{npq}$; and the probable error $=.476936c$; and

(*f*) The improbability, for a "normal" distribution, of a deviation of more than $\pm 3\sigma$ or $\pm 4\lambda$.

IV. THE COMBINATION OF OBSERVATIONS

We now come naturally to a further series of basic formulae, which arise directly from the classical approach under the conditions of independent trials and constant probability. It may be asked at once, for example, what form the expressions take when the problems are viewed from the standpoint of *relative frequencies* rather than actual occurrences, or when a number of independent trials are *combined*. The formulae will therefore now be developed, with their limitations and with illustrations, for the mean square error of (*a*) a **multiple** of the number of occurrences; (*b*) a **linear compound** of n independent quantities; (*c*) any **function** of n independent quantities; and—as special cases of (*c*)—(*d*) a **product** of two independent quantities; (*e*) the **arithmetic mean**; (*f*) **the difference between two ratios**; and (*g*) a **logarithm**. These formulae are of great importance.

The expressions will be given throughout for the mean square error, by reason firstly of its descriptive character, which recalls to mind easily the nature of the problem, and secondly because it is equivalent to σ^2—a parameter much used in the modern developments of Mathematical Statistics. Since, however, by (18), the mean error (η), standard deviation (σ), and probable error (λ) are all proportional to each other, it follows that, although the formulae will be given for σ^2, all the expressions will be of exactly the same form for η^2 or λ^2.

Applications of the various formulae to the special problems encountered in actuarial work are discussed at p. 272; C; 7.

(*a*) The Mean Square Error of a Multiple

In considering *relative frequencies* instead of actual occurrences, the problem is simply that of determining the probabilities of deviations in relation to a stated number. The argument in reaching the preceding formulae (as pointed out in C;2 and illustrated in C;4, C;5, and C;6) concerned the deviations between an actual number of occurrences, s, and the mean expected number np, i.e., the deviation $s-np$, or $n\left(\frac{s}{n}-p\right)$, in which $\frac{s}{n}$ is the observed statistical frequency and p is the true *a priori* probability. If, however, the occurrences in relation to the n trials be considered, we should evidently be dealing with the deviation $\left(\frac{s}{n}-p\right)$ between the observed statistical frequency and the true *a priori* probability, p. This simply means a division throughout by the invariable factor n. We may therefore write down from (13) that the mean error in the *relative* number of successes, $\frac{s}{n}$, is $\frac{1}{n}\left(\frac{c}{\sqrt{\pi}}\right)$; from (15) that the standard deviation is $\frac{1}{n}\left(\frac{c}{\sqrt{2}}\right)$; and from (17) that the probable error is $\frac{1}{n}(.476936c)$. That is to say, since $c=\sqrt{2npq}$, we have for $\frac{s}{n}$, the observed statistical frequency, that

$$\sqrt{\frac{2pq}{n}} = \sqrt{\pi}\,\eta = \sqrt{2}\,\sigma = \frac{r}{.476936} \qquad \ldots.(22)$$

which corresponds with (18).

From the above principle it will also be apparent that the mean error, standard deviation, and probable error of *an algebraical or numerical multiple of s*, say κs, will be found from (18) by multiplying the appropriate relation therein by κ.

Correspondingly, the

Mean Square Error of κs is $\kappa^2\sigma^2$, where σ^2 is the mean square error of s $\ldots.(23)$

(*b*) The Mean Square Error of a Linear Compound

This case concerns a very important matter, of which much use is made in the theory of graduation of mortality tables. The question may be stated, in its most elementary form, as the determination of the mean square error of *the sum of two independent quantities.*

Suppose that two *independent* quantities have been observed, of which the true values are F_1 and F_2, and that their errors follow respectively the Normal Curves $\frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}$ and $\frac{1}{k\sqrt{\pi}}e^{-\frac{y^2}{k^2}}$; we seek to determine the law followed by the errors in their sum F_1+F_2.

Clearly, the simultaneous occurrence of an error between x and $x+\delta x$ in F_1, and of an error between y and $y+\delta y$ in F_2, will cause an error in F_1+F_2 which will lie between $x+y$ and $x+\delta x+y+\delta y$. If z be written for $x+y$, the error in F_1+F_2 may therefore be said to lie between z and $z+\delta z$, if δx, δy, and δz are infinitesimal increments. Consequently, if an error x, lying anywhere between $-\infty$ and $+\infty$, be committed in respect of F_1, then the remainder of the total error must be committed in respect of F_2, and must lie anywhere between $z-x$ and $z+\delta z-x$. The compound probability of these two errors occurring together (F_1 and F_2 being wholly independent) will obviously be

$$\frac{1}{c\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}\,dx\;\frac{1}{k\sqrt{\pi}}\int_{z-x}^{z+\delta z-x} e^{-\frac{y^2}{k^2}}\,dy \qquad \ldots.(24)$$

which, as shown at p. 211; B; 8, reduces to

$$\frac{1}{\sqrt{\pi}\sqrt{c^2+k^2}}e^{-\frac{z^2}{c^2+k^2}}\,\delta z.$$

Since, therefore, this expression represents the probability of an error between z and $z+\delta z$, it follows that the probability of an error z is

$$\frac{1}{\sqrt{\pi}\ \sqrt{c^2+k^2}}e^{-\frac{z^2}{c^2+k^2}}, \text{ or } \frac{1}{\gamma\sqrt{\pi}}e^{-\frac{z^2}{\gamma^2}} \text{ where } \gamma^2=c^2+k^2 \qquad \ldots.(25)$$

This remarkably elegant formula symbolizes the very important result that when the errors in F_1 and F_2 are independent and follow the Normal Curve with parameters c and k respectively, then the errors in the sum F_1+F_2 also follow the Normal Curve with parameter $\gamma=\sqrt{c^2+k^2}$.

Since here c, k, and γ are the parameters of three distinct Normal Curves of error, it will be seen that, by (14), their respective mean square errors, σ^2, are $\frac{c^2}{2}$, $\frac{k^2}{2}$, and $\frac{\gamma^2}{2}=\frac{c^2+k^2}{2}$. If, therefore, we write σ_1^2 and σ_2^2 for the mean square errors of the independent observed quantities F_1 and F_2 respectively, the mean square error of their sum F_1+F_2 is

$$\sigma_1^2+\sigma_2^2 \qquad \ldots.(26)$$

The method of proof given for (24), (25), and (26) may evidently be extended for any number of independent quantities $F_1, F_2, \ldots, F_n$, so that the errors in the sum $F_1+F_2+\ldots+F_n$ will obey the Normal Curve with parameter $\gamma=\sqrt{c^2+k^2+\ldots}$; and, correspondingly, the mean square error will be $\sigma_1^2+\sigma_2^2+\ldots+\sigma_n^2$.

Moreover, it has already been shown in (23) that the mean square error in κF is κ^2 times the mean square error in F. If, therefore, we have any linear compound, $l_1F_1+l_2F_2+\ldots+l_nF_n$, of any number of independent observed quantities $F_1, F_2, \ldots, F_n$ each obeying the Normal Curve with mean square errors $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ respectively, where $l_1, l_2, \ldots, l_n$ are multipliers (positive or negative, integral or fractional), then the

Mean Square Error of the Linear Compound is

$$l_1^2\,\sigma_1^2+l_2^2\,\sigma_2^2+\ldots+l_n^2\,\sigma_n^2 \qquad \ldots.(27)$$

(c) The Mean Square Error of a Function

It will now be convenient, in order to examine some other special cases of importance in actuarial work, to consider the general formula for the mean square error in any *function* $F=f(F_1, \ldots, F_n)$ of n independent quantities $F_1, \ldots, F_n$, where, as before, $F_1, \ldots, F_n$ are the true values, and the respective mean square errors are $\sigma_1^2, \ldots, \sigma_n^2$.

Now if errors $x_1, \ldots, x_n$ are committed in $F_1, \ldots, F_n$ respectively, the error in F will be

$$f[(F_1+x_1), \ldots, (F_n+x_n)]-f(F_1, \ldots, F_n).$$

So long as these errors $x_1, \ldots, x_n$ are so small that their squares, products, and higher powers may be neglected—an important limitation—this may be expanded by Taylor's Theorem to give approximately

$$\left(\frac{\partial F}{\partial F_1}\right)x_1 + \left(\frac{\partial F}{\partial F_2}\right)x_2 + \ldots + \left(\frac{\partial F}{\partial F_n}\right)x_n.$$

But this is a linear compound, to which formula (27) is directly applicable; it therefore follows that the

Mean Square Error of a Function $F = f(F_1, \ldots, F_n)$ is *approximately*

$$\left(\frac{\partial F}{\partial F_1}\right)^2 \sigma_1^2 + \left(\frac{\partial F}{\partial F_2}\right)^2 \sigma_2^2 + \ldots + \left(\frac{\partial F}{\partial F_n}\right)^2 \sigma_n^2 \qquad \ldots.(28)$$

$$= \sum_{t=1}^{t=n} \left(\frac{\partial F}{\partial F_t}\right)^2 \sigma_t^2.$$

It is to be noted that $\dfrac{\partial F}{\partial F_t}$ is based on the "true" values.

As special cases of the preceding general formulae which are often useful, we shall now examine the mean square error of (*d*) a *product*, (*e*) the *arithmetic mean*, (*f*) the *difference between two ratios*, and (*g*) a *logarithm*.

(*d*) The Mean Square Error of a Product

The mean square error of a *product*, F_1F_2, of two independent quantities may be written down at once from (28) as *approximately*

$$F_2^2 \sigma_1^2 + F_1^2 \sigma_2^2 \qquad \ldots.(29)$$

(*e*) The Mean Square Error of the Arithmetic Mean

If r precisely similar independent determinations are made of a single quantity F_1, and their simple *arithmetic mean* is then taken by the usual process of summing and dividing by r, the effect is the same as if one determination had been made of each of r independent quantities $F_1, \ldots, F_r$, all with the same mean square error, which were then combined by each one being multiplied by $\frac{1}{r}$. The mean square error of the arithmetic mean of r independent determinations of a single quantity is therefore obtainable directly from formula (27) for a *linear* compound, by putting $l_1 = \ldots = l_r = \frac{1}{r}$, and $\sigma_1^2 = \ldots = \sigma_r^2 = \sigma^2$, so that (writing $\sigma^2\{A.M.\}$ for the mean square error of the arithmetic mean—cf. p. 272, C; 7)

$$\sigma^2\{A.M.\} = r\left(\frac{\sigma^2}{r^2}\right) = \frac{\sigma^2}{r} \qquad \ldots (30)$$

This is often expressed verbally by the statement that the standard deviation of the arithmetic mean of r observations of a single quantity is $\frac{1}{\sqrt{r}}$ times the standard deviation of a single observation (see p. 287; C; 8).

(f) The Mean Square Error of the Difference between Two Ratios

The formula for this special case will often be met by the student in terms of n, p, and q, in which *form* it is useful for comparing rates of mortality, disability, withdrawal, retirement, etc.

Suppose that in n_1 trials, for each of which the true probabilities are p and q, a number of successes s_1 has been observed; and that in another independent set of n_2 trials, under the same conditions with the same true probabilities p and q, another number s_2 successes has occurred. Then, by (22), σ^2 for the observed statistical frequency is $\frac{pq}{n_1}$ in the first set of n_1 trials, and $\frac{pq}{n_2}$ in the second set. If we now wish to determine σ^2 for the difference between the two observed values, formula (27) may be applied directly by taking $l_1 = 1$ and $l_2 = -1$, so that we obtain

$$(1)^2\left(\frac{pq}{n_1}\right)+(-1)^2\left(\frac{pq}{n_2}\right)=pq\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \qquad \ldots .(31)$$

Illustrations of its use are shown at p. 289, in (3) of C; 9.

In this formula the true p (and q) are assumed to be known; and it is to be noted that the true p and the two observed ratios, $\frac{s_1}{n_1}$ and $\frac{s_2}{n_2}$, may all be different. When, however, p (and q) are not known it is necessary to form some estimate of their values. This may sometimes be done, as a practical matter, by the application of reasonable judgment to properly comparable data—in which case p (and q) are merely assumed on the basis of experience instead of being really known. In some cases, however, even this may be impossible; it is then essential, if the formula is to be used for practical deductions, to make some estimate merely from the data alone. Clearly, if the two samples are "random" (see Chapter V)—which is really the object of investigation—their separate ratios, $\frac{s_1}{n_1}$ and $\frac{s_2}{n_2}$, could be amalgamated to give $\frac{s_1+s_2}{n_1+n_2}$ as an estimate of p, and (31) would become (cf. P:*16*:272 and P:*32*:192)

$$\left(\frac{s_1+s_2}{n_1+n_2}\right)\left(1-\frac{s_1+s_2}{n_1+n_2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right) \qquad \ldots .(32)$$

Some further practical modifications are discussed at p. 291; C; 10.

(g) The Mean Square Error of a Logarithm

The mean square error of a logarithm, $\log_e F_1$, can be written down immediately from (28) as being *approximately*

$$\sigma_1^2\left[\frac{d}{dF_1}(\log_e F_1)\right]^2 \text{ or } \left(\frac{1}{F_1^2}\right)\sigma_1^2 \qquad \ldots .(33)$$

V. THE THEORY OF RANDOM SAMPLING

The Meaning of Simple Sampling

THE formulae of the preceding chapters have been based on the occurrences or failures of an event, in a group of n independent trials, for each of which n trials there is supposed to be the same true probability of either occurrence or failure, p or q. This assumption that the true probabilities are constant for every member of the group of n obviously means that the group is supposed to be absolutely homogeneous. Now when such a group of n persons is limited in size, so that the observed statistical frequency of deaths (let us say) can be taken only as an approximation to the true probability of death, we are in fact dealing with a **sample** of the larger **population** or **universe**—sometimes called the "parent population" or "parent universe"—of which it forms a part. Such a sample, moreover, is said to be **random** when it has been so drawn, from its parent universe, that every member of the universe has had an equal and independent chance of being chosen as a member of the sample. The case of independent trials and constant probability is generally identified in the nomenclature as one of **simple sampling.**

Modifications of the Conditions of Simple Sampling

Applications of the formulae, under these "simple sampling" conditions, to certain types of problems concerning mortality and allied statistics have been illustrated on pp. 265 to 293, in C; 4 to C; 10. Those problems have been of such a nature that it has been possible to apply directly the basic formulae already deduced. One of the questions considered has been whether the variations in the observed statistical frequencies, in a series of ν independent investigations based on $n_1, n_2, \ldots, n_\nu$ trials (i.e., cases) respectively, can have arisen from mere chance, or must be attributed to some specific cause. If the variations cannot have arisen from mere chance, so that the operation of a specific cause must be suspected, the conclusion so reached may indicate that the conditions of simple sampling have, in fact, not been fulfilled. The primary assumptions of simple sampling are that the true probabilities p and q remain unchanged throughout. If the result indicates that those assumptions may not have been fulfilled, the question arises whether the disturbance may be due to variation in the underlying true probabilities themselves.

The Theory of Lexis (see p. 163; A; 7) was the first to deal with this important question in an obviously logical manner, by modifying the primary assumption of simple sampling that p and q remain unchanged. Three different types of sampling are defined as follows in that theory:

(*a*) **Bernoulli sampling**—the simple sampling already considered—for which the true p and q remain *unchanged* in every trial of a "set" of n_1 trials, and in every trial in the next set of n_2 trials, and so on until we reach every trial in the last set of n_ν trials; i.e., *p and q remain constant in every trial in every set.*

(*b*) **Poisson sampling**—otherwise called "*stratified*" sampling—for which p and q vary in the several trials in the first set of trials, and vary in the same way in each subsequent set of trials; i.e., *p and q vary from trial to trial but are constant from set to set.*

(*c*) **Lexis sampling,** for which p and q are unchanged in every trial of the first set of trials, and are similarly constant but with a different value in each subsequent set of trials; i.e., *p and q are constant from trial to trial but vary from set to set.*

In order to develop the mathematical theory for these three types it will be assumed that the sets are all of the same size, n, i.e., that $n_1 = \ldots = n_\nu = n$. (The treatment of groups of unequal size in practice is illustrated on p. 299; C; 11).

(*a*) For the Bernoulli case, with n trials in each set, σ^2 for the number of occurrences in each set, by (7), is npq. Consequently, if this be computed for each of the ν sets, and the resulting values be summed and divided by ν to form the arithmetic mean, the result (see p. 214; B; 9) is simply $\frac{\nu(npq)}{\nu} = npq$, which is here denoted by σ_B^2.

(*b*) For the Poisson case it may be shown (see p. 215; B; 9) that the corresponding value, which may be identified as σ_P^2, is *less* than that of the Bernoulli series, being

$$\sigma_P^2 = \sigma_B^2 - n\sigma_p^2 \qquad \ldots.(34)$$

where σ_p^2 is the mean square deviation of the probabilities $p_1, \ldots, p_n$ from their mean $p = \frac{p_1 + p_2 + \ldots + p_n}{n}$.

(*c*) For the Lexis case, on the other hand (see p. 216; B; 9), the corresponding σ_L^2 is *greater* than that for Bernoulli sampling, being

$$\sigma_L^2 = \sigma_B^2 + (n^2 - n)\sigma_p^2 \qquad \ldots.(35)$$

where σ_p^2 is the mean square deviation of the probabilities $p_1, \ldots, p_\nu$ from their mean $p = \frac{p_1 + p_2 + \ldots + p_\nu}{\nu}$.

On the same principle as that which led to formula (22) for the standard deviation of a relative frequency we see that in these cases, likewise, σ^2 *for the frequency (proportion) of successes* is obtained by dividing by n^2, giving

For Bernoulli sampling, $\frac{pq}{n}$ ....(36)

For Poisson sampling, $\frac{pq}{n} - \frac{\sigma_p^2}{n}$ ....(37)

For Lexis sampling, $\frac{pq}{n} + \left(\frac{n-1}{n}\right)\sigma_p^2$. ....(38)

Two measures have been evolved in order to provide a convenient method of distinguishing between the three types. The **Lexis Ratio,** usually denoted by L, is simply

$$L = \frac{\sigma}{\sigma_B} \qquad \ldots.(39)$$

where σ is the standard deviation computed directly from the data, and σ_B is that calculated on the assumption that the data follow the Bernoulli (binomial or normal) type.

This ratio, however, is dependent on the values of the chances p_t which are involved in σ_p, and is also affected by variations in the number n. Charlier therefore suggested using a **Coefficient of Disturbancy** (or *Variability*),

$$C = \frac{100\sqrt{\sigma^2 - \sigma_B^2}}{A} \quad \ldots(40)$$

where σ and σ_B are as just defined, and A is the arithmetic mean of the data (see p. 216; B; 9).

When $L=1$, and $C=0$, the data are *normal* and of the *Bernoulli* type.

When $L<1$, and C imaginary, they are said to be *subnormal* and of the *Poisson* type.

When $L>1$, and $C>0$, they are said to be *hypernormal* (or *supernormal*) and of the *Lexis* type.

In cases where the samples are small and the observed data are used to give estimates of the true values, a correction should be introduced on the principles of Bessel's formula (42), as explained in the next section of this chapter, and as noted in C; 11.

It will be seen at once that—apart from the formal mathematical process by which the method is expressed in formulae (34) to (38)—the practical application of the principle is very easy. Simple Bernoulli sampling, in effect, is used as a standard of comparison. A computation is made of σ^2 from the actual data; if it is less than the Bernoullian σ_B^2 of simple sampling, then the conditions for Poisson sampling are indicated, i.e., that the probabilities p and q may have varied more among the individuals or groups within the sample than between samples; if it is greater than the σ_B^2 of simple sampling, then the conditions of the Lexis distribution are suggested, namely, that the probabilities may have varied more from sample to sample than within samples.

In practice, of course, the conditions are hardly definable with the precision of the above theoretical formulation, so that values which may be actually calculated for L or C must be interpreted with some care. It must be remembered that the theory deals only with average values, and that in practice deviations from such average values may occur. If the number of samples is sufficiently large, and L is found to differ considerably from 1, there is good ground for the inference that the data are not Bernoullian; but if L is near to 1 the Bernoullian hypothesis is not necessarily established definitely—it should be considered as being plausible only (cf. P:*146*:215). Furthermore, the formulae assume that the events are all independent; a value of σ^2 less or greater than σ_B^2 may therefore indicate, as an alternative interpretation, that the events are negatively or positively correlated (see P:*177*:366).

The fact that the practical application of the Lexis theory is thus based, in reality, merely on the computation of the Lexis Ratio, $L=\frac{\sigma}{\sigma_B}$, constitutes not only an important step in the classical theory of sampling—by its elucidation of the effects of removing the limitations of simple sampling—but also forms a valuable connecting link between the classical procedures and the ideas underlying the more recent development of the "χ^2 test" (for which see Chapter IX). For L^2, being $\frac{\sigma^2}{\sigma_B^2}$, where σ^2 is calculated from the data and σ_B^2 is computed on the assumption that p and q are constant, is (as may also be seen easily from the hypothetical numerical illustration of the three types given at p. 294; C; 11) simply

$$L^2=\frac{\frac{1}{\nu}\sum_{r=1}^{r=\nu}(f'_r-np)^2}{npq} \qquad \ldots.(41)$$

where $f'_1, f'_2, \ldots, f'_\nu$ are the observed occurrences; and this formula, as will become clear later from Chapter IX, is, in fact, $\frac{\chi^2}{\nu}$, under the particular conditions assumed (see p. 217; B; 9).

Modifications Necessary for Dealing with Small Samples

In Section (*f*) of Chapter IV, and at p. **291** of C; **10**, it has been remarked that when the true p (and q) are not known it may often be necessary to form some estimate of their values. When the conditions are "random", and the sample is large so that the deviation between the true p and the observed statistical frequency $\frac{s}{n}$ will be relatively small, the considerations already discussed indicate that the estimate of the true values for the universe may be obtained, for practical purposes, by adopting the observed values of the sample. This procedure clearly would become less of an approximation as n, the size of the sample, increases (cf. p. 292; C; 10). In the other direction, however, it will be equally apparent that the sample value will become less reliable as the sample becomes smaller. The investigations arising from the latter portion of this fundamental principle constitute the modern and very important **Theory of Small Samples.**

The nature of the problem may be seen from an examination of the estimates, which can be made from the sample, for the arithmetic mean and the mean square deviation of the universe. Suppose that a sample numbering n is drawn, with magnitudes $x'_1, x'_2, \ldots, x'_n$; their arithmetic mean, say $\bar{x}$, is then $\bar{x} = \frac{1}{n}\sum_{r=1}^{r=n} x'_r$; and clearly the principle on which this procedure is based is the same whether the sample is large or small. In the case, however, of the mean square deviation, the σ^2 in the universe is, by definition, the average of the squares of the deviations from the mean, so that $\sigma^2 = \frac{1}{R}\Sigma(x_r - m)^2$ where the x_r's are the values in the universe, R is the number of values of r over which Σ extends, and m is the true mean of the universe. When we come to deal with a limited sample from that universe, for example the observed values $x'_1, x'_2, \ldots, x'_n$, we could obviously make an estimate of σ^2, say σ_e^2, by calculating $\sigma_e^2 = \frac{1}{n}\sum_{r=1}^{r=n}(x'_r - m)^2$, which would approach more and more closely to the σ^2 of the universe as the size of the sample increases. This estimate, however, involves m, the mean of the universe, which usually will be unknown. If we write the preceding $\sigma_e^2 = \frac{1}{n}\sum_{r=1}^{r=n}(x'_r - \bar{x} + \bar{x} - m)^2$, where $\bar{x}$ is the mean of the sample, it will be observed that this may be put as $\sigma_e^2 = \frac{1}{n}\sum_{r=1}^{r=n}(x'_r - \bar{x})^2 + (\bar{x} - m)^2$. The first term here is the mean square deviation in the sample itself, say σ_s^2, with reference to the mean of the sample (cf. p. 299; C; 11), and it is seen to differ from the estimate σ_e^2 by the always positive quantity $(\bar{x} - m)^2$, so that clearly a biassed error would be involved if the mean square deviation as calculated from the sample alone were to be taken as an estimate of the true mean square deviation of the universe. The quantity $(\bar{x} - m)^2$, however, is the square of the deviation of the mean of the sample from the true mean of the universe, and by section (*e*) of Chapter IV its average value may be taken as $\frac{\sigma^2}{n}$ so long as the assumptions implicit therein are appropriate (see below). Furthermore $\frac{\sigma^2}{n} \doteqdot \frac{\sigma_e^2}{n}$, since σ_e^2 is the estimate of σ^2. From the original expression this course of reasoning therefore gives $\sigma_e^2 = \sigma_s^2 + (\bar{x} - m)^2 \doteqdot \sigma_s^2 + \frac{\sigma^2}{n} \doteqdot \sigma_s^2 + \frac{\sigma_e^2}{n}$

whence

$$\sigma_e^2 \doteqdot \left(\frac{n}{n-1}\right)\sigma_s^2 \qquad \ldots.(42)$$

This formula, in which the factor $\left(\frac{n}{n-1}\right)$ is known as **Bessel's Correction,** has been in use since the time of Gauss (see p. 164; A; 8).

The actual process by which (42) is obtained does not depend upon n being small, so that the formula is applicable to large as well as to small samples. The introduction of the $(n-1)$ term in the denominator, however, has an appreciable effect upon the value only when n is small, and the method is therefore usually adopted in practice especially for small samples.

It must be realized that the formula is an approximation merely. It is based essentially upon the substitution, during the argument, of $\frac{\sigma^2}{n}$ for $(\bar{x}-m)^2$, which is valid only as an average for a large number of samples, so that the resulting (42) likewise gives only an average value and may therefore provide an estimate for any particular sample which may differ markedly from the true value. The contradictions thus present have been well analyzed by Steffensen in P:*140*:12-20, where the corresponding formulae for the third, fourth, and higher moments are also discussed (see p. 219; B; 11).

Formula (42), by which an estimate, σ_e^2, for the σ^2 of the universe is obtained from the σ_s^2 of the sample, is an important example in a procedure which in general may be called **The Theory of Statistical Estimation.** Such estimated values are sometimes referred to as *Presumptive Values* (see p. 219; B; 11). The preceding derivation, moreover, being founded on a mode of reasoning which, as indicated, cannot be viewed as wholly satisfactory, serves also to introduce the student to the realization that this problem of "estimation" is surrounded by certain disagreements and controversies, accompanied often by misunderstanding. A brief reference to these difficulties is therefore essential here.

The Theory of Statistical Estimation; The Classical Approach—The Principle of Insufficient Reason

The classical method of approach to formula (42) was based upon the obvious principle that (as stated at p. 263; C; 1) the drawing of a sample, s, from a population or universe, P, by some process of selection, S, may be stated symbolically in the form $s=S(P)$. More completely, suppose that an event, E, can occur under only one of the mutually exclusive conditions $F_1, F_2, \ldots, F_n$, and that it has been observed to happen s times in a sample of n trials. Since here, in this problem of estimation, we are given merely the observed set of s occurrences, and wish to form an estimate with respect to the parent population from which they arose, let κ_r be the *a priori* probability that F_r exists, and π_r the *a priori* probability of producing the event E from F_r, where r takes the values 1, 2, . . ., n. Then the reasoning underlying the Bayes-Laplace Theorem (see p. 221; B; 12) shows that the total probability that one of the conditions $F_1, F_2, \ldots, F_n$ exists, and that E will happen s times in n, is

$$\sum_{r=1}^{r=n} \kappa_r \, {}^nC_s \, \pi_r^s \, (1-\pi_r)^{n-s} \quad \ldots\ldots(43)$$

The purpose of the Bayes-Laplace Theorem itself, as stated in formula (43a) on p. 222, is to deduce the probability, *a posteriori*, that the *particular* condition F_a (say) was the origin of the set of occurrences observed. That, however, is not exactly the problem here. We are now seeking rather an estimate as to *which* of the various possible conditions is *most likely* to have produced the set of occurrences observed (namely, the happening of E actually s times in n). Under those circumstances the "best" estimate clearly must be that which will give the greatest possible value, *a priori*, to the probability, (43), of the occurrences actually observed. [It will assist the student to note that the "model" of the Bayes-Laplace Theorem is used in this formulation of the problem of estimation; we do not, however, need to go beyond the preliminary formula (43), so that formulae (43a) and (43b) on p. 222 are here not actually required.]

In order to deduce by this means an estimate for σ^2, the classical method is to consider n observed values $x'_1, x'_2, \ldots, x'_n$, of a fixed true value X, under the conditions of the Normal Curve (11)—which, it is to be remembered, is a very close representation of the point binomial involved in (43). The *a priori* probability that all the deviations $(x'_r - X)$, where $r = 1, 2, \ldots, n$, will occur together is then $\left(\frac{1}{c\sqrt{\pi}}\right)^n e^{-\frac{\Sigma(x'_r - X)^2}{c^2}}$. This may be written

$$\left(\frac{1}{c\sqrt{\pi}}\right)^n e^{-\frac{\Sigma(x'_r-\bar{x})^2}{c^2}} e^{-\frac{n(\bar{x}-X)^2}{c^2}}$$

$=f(x'_r, X)$ say, where $\bar{x}=\frac{1}{n}\sum_{r=1}^{r=n} x'_r$.

But X, being unknown, might lie anywhere between $-\infty$ and $+\infty$; and if we accept the **Principle of Insufficient Reason** (see p. 181; (*i*) of B; 1, and p. 222; B; 12), which amounts to assuming that all the *a priori* existence probabilities, κ_r, are to be taken as equal—that is, as a constant κ, say—it follows that the total probability of the observed deviations (x'_r-X) is $\kappa\int_{-\infty}^{+\infty} f(x'_r, X)\,dX$.

To find this integral we have to evaluate $\int_{-\infty}^{+\infty} e^{-\frac{n(\bar{x}-X)^2}{c^2}}\,dX$; putting $\sqrt{n}\,(X-\bar{x})=x$, so that $dX=\frac{dx}{\sqrt{n}}$, this is

$$\frac{1}{\sqrt{n}}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}\,dx=\frac{c\sqrt{\pi}}{\sqrt{n}}$$

by (*a*) at p. 209; B; 7; and consequently

$$\kappa\int_{-\infty}^{+\infty} f(x'_r, X)\,dX=\frac{\mathrm{K}}{c^{n-1}}\,e^{-\frac{\Sigma(x'_r-\bar{x})^2}{c^2}},$$

where K is written for the constant. To determine the value of c for which this probability will be a maximum we take logarithms, differentiate with respect to c, equate to zero, and obtain

$$\frac{n-1}{c}-(2c^{-3})\,\Sigma\,(x'_r-\bar{x})^2=0, \text{ whence } \frac{c^2}{2}, \text{ which is}$$

$$\sigma^2{}_r=\frac{\Sigma\,(x'_r-\bar{x})^2}{n-1}=\left(\frac{n}{n-1}\right)\sigma_s^2 \text{ as before.}$$

The "Best Unbiassed Estimate"

The difficulties resulting from the Principle of Insufficient Reason led Gauss originally (H:*17*:49), and later the Russian

Markoff (H:*116*), to the concept of an **unbiassed estimate.** If we wish to "estimate" the value of a parameter θ, and for that purpose have at our disposal the values $x_1, x_2, \ldots, x_n$, so that any estimate of θ will be some function $F(x_1, x_2, \ldots, x_n)$ of those values, then $F(x_1, x_2, \ldots, x_n)$ is called an "unbiassed estimate" of θ if the mathematical expectation of $F(x_1, x_2, \ldots, x_n)$ is identically equal to θ. Since there are many such functions with expectations equal to θ, the **best unbiassed estimate** is taken as the one for which the "variance" (see p. 163; A; 6) is a minimum. This procedure is, of course, once more based on the use of a dogmatic "principle". Markoff, whose work has recently been examined in English by Neyman (P:*92*:130, and P:*93*:105), has dealt with those "best unbiassed estimates" which are linear functions of $x_1, x_2, \ldots, x_n$. By this method the estimate σ_e^2 is again found to be $\left(\frac{n}{n-1}\right)\sigma_s^2$.

The Method of Maximum Likelihood

The introduction of the Principle of Insufficient Reason into the demonstration on p. 38 involves, of course, all the difficulties associated with the acceptance of that principle (see also P:*92*: 128-130, and P:*29*:145 and 151). Certainly in many instances it must be very difficult to justify the supposition that the unknown X might lie anywhere between the most extreme possible limits stated, namely $-\infty$ and $+\infty$; and even if it could, it may be an even more sweeping assumption to suppose that it is equally likely to fall in any particular place in so extensive a range. There is consequently much to be said in favour of discarding that procedure—particularly as the problem of estimation can be approached logically, without the necessity of introducing the principle at all, by using the more direct **Method of Maximum Likelihood.** By this method the reasoning is based solely upon the information afforded by the actual observations, and no assumptions are made concerning *a priori* knowledge. In the case of the n observations previously considered the argument is simply that $\left(\frac{1}{c\sqrt{\pi}}\right)^n e^{-\frac{\Sigma(x'_r - X)^2}{c^2}}$, as there given, represents the *a priori* probability that the set of observed values will occur, and that the best estimates for X and c *simultaneously* (since both X and c are unknown) will be those which make this probability a maximum. Taking logarithms, then the partial differential coefficient with respect to X, and equating to zero, we find immediately that $\Sigma(x'_r - X) = 0$, whence $X = \frac{1}{n}\Sigma x'_r$; and similarly differentiating with respect to c it follows that

$$\frac{c^2}{2}, \text{ or } \sigma^2 = \frac{\Sigma(x'_r - X)^2}{n} = \sigma_s^2.$$

The estimate thus reached by the Method of Maximum Likelihood, being σ_s^2, is not the same as the $\left(\frac{n}{n-1}\right)\sigma_s^2$ obtained by the other methods. The distinction is important. It will be seen that in the classical method involving the Principle of Insufficient Reason the unknown X is dealt with by being supposed to lie anywhere, with equal probability, between $-\infty$ and $+\infty$, and then σ^2 is estimated by maximizing the resulting probability; in the Method of Maximum Likelihood no such assumption is made with respect to X, but both X and σ^2 simultaneously are estimated by maximizing the probability of the observed event without any assumption of *a priori* ignorance. The polemics which have been precipitated by these rival methods already have produced a literature far too extensive for detailed reference here. For the present purpose, however, it may be sufficient to repeat Neyman's remark (P:*92*:135) that the question as to which is preferable is, in reality, "one of taste only".

Sampling Distributions

The problems of sampling follow from the idea of drawing a random sample from a parent universe. If a group of n persons has been so derived, that group would have, with respect to any particular characteristic (such as age, height, weight, etc.), its own sample mean, its own sample standard deviation, and likewise its own sample value of any other statistical parameter which may be selected for determination. Another group of n, similarly drawn at random, would again have its own sample mean, standard deviation, etc. The process of drawing different sample groups from the parent universe could thus be continued until from a series of samples we should have found a series—a "distribution"—of sample means, a distribution of standard deviations, and distributions of other measurable characteristics —for the means of the various samples would not all be identical, the standard deviations of the samples would differ from each other, and the values of any other measurable characteristic would vary from sample to sample. The study of the forms taken by these **sampling distributions** constitutes an important section of the theories of both large and small samples.

It has already been shown in formula (30) that the standard deviation of the arithmetic mean of n independent determinations of a single quantity is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of a single observation. Viewing this result from the sampling standpoint here under consideration, the relation $\sigma\{A.M.\} = \frac{\sigma}{\sqrt{n}}$ can evidently be interpreted as giving the standard deviation of the arithmetic mean based on n samples, when the standard deviation of the parent universe is known to be σ. If, however, the universe σ is not known, the formula can be written $\sigma\ \{A.M\} \doteq \frac{\sigma_s}{\sqrt{n}}$ if σ_s can be taken as an estimate of σ, or as $\sigma\ \{A.M.\} \doteq \frac{\sigma_s}{\sqrt{n-1}}$ by substituting from (42) when Bessel's correction is required. We thus reach an important formula and method for the practical determination of the standard deviation of the arithmetic mean—for if we have a sample from which σ_s can be evaluated, and if there are n observations altogether, then

$\frac{\sigma_s}{\sqrt{n}}$, or $\frac{\sigma_s}{\sqrt{n-1}}$ for small samples, will measure the standard deviation of the arithmetic mean of such a sample, i.e., it will give the standard deviation of the distribution of sample means. An algebraical proof may be found in P:*32*:189.

The term **standard error** is often applied to the standard deviation of such a sampling distribution. A numerical application is shown on p. 300; C; 12.

The general problem of determining the standard errors of parameters (of which the formula just given for the standard deviation of the mean is a simple case) can be presented most easily by the method used by Karl Pearson in H:*93*, of which the principles are reproduced conveniently in P:*32*:187-191 and P:*177*:394-411. The objectives of this study hardly seem to require the inclusion of the extensive algebra necessary for the derivation of the various formulae there shown. It may accordingly suffice if we here state the following results, where n denotes the number of individuals (variates) in the sample, and the parent universe is assumed to be normal with standard deviation σ:

Parameter	*Standard Error*
(i) Arithmetic Mean	$\frac{\sigma}{\sqrt{n}}$
(ii) Median (see P:*155*:199 and P:*116*:134)	$\frac{1.2533\sigma}{\sqrt{n}}$
(iii) Standard Deviation	$\frac{\sigma}{\sqrt{2n}}$
(iv) Mean Square Deviation (Variance)	$\sigma^2\sqrt{\frac{2}{n}}$
(v) qth Moment about a Fixed Point	$\sqrt{\frac{\mu'_{2q}-\mu'^{2}_{q}}{n}}$
(vi) qth Moment about the Mean	$\sqrt{\frac{\mu_{2q}-\mu_q^2+q^2\mu_2\mu_{q-1}^2-2q\mu_{q-1}\mu_{q+1}}{n}}$

www.dbraulibrary.org.in

These formulae provide a method for determining the limits within which a sample value of a parameter will probably lie. When the assumption of normality is admissible, it is customary in practice to use ± 3 times the standard error as defining those limits, in conformity with the conclusions of Chapter III, and as illustrated for the case of the arithmetic mean on p. 300; C; 12.

The following points may also be noted: (*a*) The standard error of the arithmetic mean was obtained without reference to the form of the distribution; (*b*) When the parent universe is not normal, the standard error of the standard deviation should be taken as $\frac{\sigma}{\sqrt{2n}}\left(1+\frac{\beta_2-3}{2}\right)^{\frac{1}{2}}$ where $\beta_2 = \frac{\mu_4}{\mu_2^2}$, which may differ considerably from the value $\frac{\sigma}{\sqrt{2n}}$ based on the assumption of normality; (*c*) In comparing the formulæ for the standard errors of the *q*th moment about a fixed point and the mean, respectively, it should be remembered that the mean of the population is a fixed point, so that the standard error in the *q*th moment about the mean of the population can be set down from (v) at once by dropping the primes (see p. 254; B; 27) and writing $\sqrt{\frac{\mu_{2q}-\mu_q^2}{n}}$, whereas formula (vi) relating to the mean refers to the mean of the sample.

"Student's" Distribution

The distinction which has already been emphasized in respect of small samples between (*a*) the standard deviation, σ, of the parent universe, (*b*) the standard deviation, σ_s, of the sample, and (*c*) the estimate, σ_e, of the standard deviation of the universe which can be made from σ_s by formula (42), indicates immediately the importance of considering the distribution of the standard deviation of a statistical parameter as calculated from the data of the sample only. For large samples the means from a normal population are distributed normally, with standard error $\frac{\sigma}{\sqrt{n}}$ as on p. 42; and the ratio $\frac{\text{Deviation of Mean}}{\text{Standard Error}}$, by which is meant the deviation of the sample mean from the mean of the universe, divided by the standard error of the mean, or $\dfrac{(\bar{x}-m)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$, is likewise distributed normally, and $\dfrac{(\bar{x}-m)}{\left(\dfrac{\sigma_s}{\sqrt{n}}\right)}$ nearly so. For small samples, however, the ratio $\dfrac{(\bar{x}-m)}{\left(\dfrac{\sigma_s}{\sqrt{n}}\right)}$ does not follow a normal distribution. The exact form was discovered in 1908 by "Student" (p. 165; A; 10), who used $\dfrac{(\bar{x}-m)}{\sigma_s}=z$ and found (see p. 226; B; 13) for the probability of a value of z between z and $z+dz$, or $G(z)$, the expression

$$G(z) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}(1+z^2)^{-\frac{n}{2}}\,dz \qquad \ldots.(44)$$

which is known as **"Student's" Distribution.** This curve is symmetrical in z, and more sharply peaked than the normal curve; as n increases it becomes nearly normal in the centre, so that a normal curve with standard deviation of $\dfrac{1}{\sqrt{n-\frac{3}{2}}}$ provides an excellent approximation (P:*29*:140).

The important characteristic of this formula is its dependence on only one constant, m, of the parent universe; σ is not involved. Since it gives the probability, in a sample of n when m is specified, of getting a value of $\dfrac{\bar{x}-m}{\sigma_s}=z$ lying between z and $z+dz$, it will be evident that the probability in a sample of n of finding the ratio $\dfrac{\bar{x}-m}{\sigma_s}$ in absolute value as great as or greater than that arising from a stated $\bar{x}-m$ and the observed σ_s is

$$P_z = 1 - 2\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}\int_0^z (1+z^2)^{-\frac{n}{2}}\,dz \quad \ldots.(45)$$

"Student's" discovery thus provided a technique for examining, particularly for small samples, whether a value of $\frac{\bar{x}-m}{\sigma_s}$ actually observed in a given sample, when m is a specified value, is unusually large or small. A small value of P_z corresponds with a large value of z; and a value of P_z such as .05, for example, means that only 5 times in 100 trials should we obtain for the ratio $\frac{\bar{x}-m}{\sigma_s}$ a value as large as, or larger than, that actually observed. The inferences which may, or must not, be drawn from such a statement are considered later.

If we now write $\left(\frac{n}{n-1}\right)\sigma_s^2$ as the estimate, σ_e^2, of σ^2, in accordance with (42), and define t as $\frac{(\bar{x}-m)}{\left(\frac{\sigma_e}{\sqrt{n}}\right)}$, so that

$$t = \frac{(\bar{x}-m)}{\left(\frac{\sigma_s}{\sqrt{n-1}}\right)} = z\sqrt{n-1}$$

and $n-1$ is the number of **degrees of freedom** of t (since one degree is absorbed in determining $\bar{x}$ from the data, as explained in Chapter VI, p. 55), it follows from (44) that the distribution of t, say $G(t)$, is given by

$$G(t) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{n-1}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}} dt \quad \ldots .(46)$$

or by

$$G(t) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{t^2}{d}\right)^{-\frac{d+1}{2}} dt \quad \ldots .(47)$$

where, as in Chapter IX, d is the number of "degrees of freedom".

All the expressions (44), (46), and (47) will be found in the literature. For purposes of identification, (44) is often referred to as "*Student's z-distribution*", while (46) and (47) are usually called "*Student's t-distribution*" or simply "*the t-distribution*".

Tables of values of the integral of $G(t)$ have been given by "Student" (see P:*177*:536), in P:*97*, and by R. A. Fisher (P:*43*: 177), in several different forms. A very convenient nomograph for the calculation of P_z in (45), devised by Nekrassoff (H:*168*), is reproduced in P:*29*:136 (and 115).

The Assumptions and Meaning of the "Student" Method

The publication of "Student's" formula had far-reaching consequences. It seemed to release investigators from the necessity of following the earlier methods, which usually, in tests of significance, had simply taken the observed σ_s as an estimate of σ and then assumed a normal distribution (cf. H:*188*:101, and P:*156*:78). This very fact, however, that the "Student" formula avoids consideration of the parent σ altogether, by involving only σ_s directly without any apparent necessity for contemplating what σ itself may be, will show immediately that the formula must be applied with careful regard for its underlying assumptions and for the precise hypothesis which it is employed to analyze—for obviously it may lead to unjustifiable inferences if the observed σ_s is not a reasonable estimate of σ. This important point at times has given rise to much confusion; it may therefore be well here to devote some space to the meaning of the "Student" method.

Firstly, it must be emphasized (as will be seen from the proof at p. 226; B; 13) that the whole "Student" theory is based on the assumption that the distribution of the parent population is "normal". The method is therefore clearly applicable when a sample has been drawn from a universe which is known, from prior knowledge, to be approximately normal (as, for example, a distribution of the heights of men). If, however, we do not know whether a sample has been drawn from a normal or from a non-normal universe, it must be remembered that the "Student" theory, as well as the classical "normal" test, may then be open to question if the basic assumption of a normal parent population should be markedly invalid (cf. P:*118*:155-8, P:*111*: 174, and P:*7*).

Secondly, the method supposes (see again p. 223; B; 13) that the observations comprising the sample under scrutiny have been obtained by random sampling. The prime importance of this condition has frequently been overlooked. If an investigator is quite satisfied that a single sample which he has obtained has really been secured by a random process of selection, he is clearly in a position to apply the "Student" theory (so long, of course, as he also may assume that the universe is normal). If, however, he is not assured of this essential randomness, he obviously dare not apply the "Student" method to his single sample; he must then wait until further samples convince him that randomness exists; but when that point is reached he will, in fact, usually have enough data to determine the parent σ within close limits—and when he can do that he will be able to make valid tests by using the "normal" integral based on large samples rather than the "Student" small-sample theory.

Thirdly, it must not be forgotten that, in any random sampling or probability procedure, it is always possible that some most unlikely event may actually occur. If that should happen in a single random sample which an investigator has obtained, it obviously will be dangerous for him to draw any inference whatever beyond a realization that something unusual may have occurred. For this reason it will be clear, on a little consideration, that with a single sample the "Student" method may suggest a misleading inference unless all the elements inherent in the test are not unusual.

In order to see what this statement really means, we may recall that in (44) the "Student" theory sets out, on the assumptions of a normal universe and random sampling therefrom, to assign a probability to the occurrence in a single sample of the observed value $\frac{\bar{x}-m}{\sigma_s}$ when m, the population mean, is specified.

All the expressions (44), (46), and (47) will be found in the literature. For purposes of identification, (44) is often referred to as "*Student's z-distribution*", while (46) and (47) are usually called "*Student's t-distribution*" or simply "***the t-distribution***".

Tables of values of the integral of $G(t)$ have been given by "Student" (see P:*177*:536), in P:*97*, and by R. A. Fisher (P:*48*:177), in several different forms. A very convenient nomograph for the calculation of P_z in (45), devised by Nekrassoff (H:*168*), is reproduced in P:*29*:136 (and 115).

The Assumptions and Meaning of the "Student" Method

The publication of "Student's" formula had far-reaching consequences. It seemed to release investigators from the necessity of following the earlier methods, which usually, in tests of significance, had simply taken the observed σ_s as an estimate of σ and then assumed a normal distribution (cf. H:*188*:101, and P:*156*:78). This very fact, however, that the "Student" formula avoids consideration of the parent σ altogether, by involving only σ_s directly without any apparent necessity for contemplating what σ itself may be, will show immediately that the formula must be applied with careful regard for its underlying assumptions and for the precise hypothesis which it is employed to analyze—for obviously it may lead to unjustifiable inferences if the observed σ_s is not a reasonable estimate of σ. This important point at times has given rise to much confusion; it may therefore be well here to devote some space to the meaning of the "Student" method.

Firstly, it must be emphasized (as will be seen from the proof at p. 226; B; 13) that the whole "Student" theory is based on the assumption that the distribution of the parent population is "normal". The method is therefore clearly applicable when a sample has been drawn from a universe which is known, from prior knowledge, to be approximately normal (as, for example, a distribution of the heights of men). If, however, we do not know whether a sample has been drawn from a normal or from a non-normal universe, it must be remembered that the "Student" theory, as well as the classical "normal" test, may then be open to question if the basic assumption of a normal parent population should be markedly invalid (cf. P:*118*:155-8, P:*111*: 174, and P:*7*).

Secondly, the method supposes (see again p. 223; B; 13) that the observations comprising the sample under scrutiny have been obtained by random sampling. The prime importance of this condition has frequently been overlooked. If an investigator is quite satisfied that a single sample which he has obtained has really been secured by a random process of selection, he is clearly in a position to apply the "Student" theory (so long, of course, as he also may assume that the universe is normal). If, however, he is not assured of this essential randomness, he obviously dare not apply the "Student" method to his single sample; he must then wait until further samples convince him that randomness exists; but when that point is reached he will, in fact, usually have enough data to determine the parent σ within close limits—and when he can do that he will be able to make valid tests by using the "normal" integral based on large samples rather than the "Student" small-sample theory.

Thirdly, it must not be forgotten that, in any random sampling or probability procedure, it is always possible that some most unlikely event may actually occur. If that should happen in a single random sample which an investigator has obtained, it obviously will be dangerous for him to draw any inference whatever beyond a realization that something unusual may have occurred. For this reason it will be clear, on a little consideration, that with a single sample the "Student" method may suggest a misleading inference unless all the elements inherent in the test are not unusual.

In order to see what this statement really means, we may recall that in (44) the "Student" theory sets out, on the assumptions of a normal universe and random sampling therefrom, to assign a probability to the occurrence in a single sample of the observed value $\frac{\bar{x}-m}{\sigma_s}$ when m, the population mean, is specified.

Nothing is here said about what value σ_s should have; it might, indeed, be thought at first sight that since σ_s can take any positive value whatsoever (as may be confirmed from the fact that in the proof at p. 226; B; 13 the integration is performed over all values of σ_s from 0 to ∞), therefore a valid inference with respect to the parent population may be deduced regardless of what σ_s may be. The previous statement, however, that a misleading inference may be drawn unless all the elements inherent in the test are not unusual here serves to place on σ_s the requirement that it must be not unusual. For if it should be unusual, i.e., not representative of the σ of the parent population from which it came, it will be quite clear that it may lead to an unusual inference, i.e., to an inference which will not be typical of the population which is being tested on the evidence afforded by the one sample and its σ_s. The proper inference then would be merely that an unusual event (here the appearance of an unusual σ_s) had actually occurred.

As a further illustration of the importance of thus being satisfied that σ_s is not unusual, it should be remembered also that we are dealing with a ratio, of $\bar{x}-m$ to σ_s, in which both $\bar{x}$ and σ_s vary from sample to sample. Such a ratio may be small even when $\bar{x}-m$ is large, if at the same time σ_s happens to be large enough. The mere fact, therefore, that in any particular case $\dfrac{\bar{x}-m}{\sigma_s}$ is not unusual may not justify the assertion that $\bar{x}-m$ is not unusual unless also we are able to say that σ_s is not unusual.

Valuable additional remarks on these aspects of the "Student" theory may be found in P:*29*:135, 137, 139, and 141, and P:*130*:59. Graphic illustrations are there given which may be of help in grasping the importance of the statement that σ_s must be not unusual, and in realizing the kind of inference which may properly be drawn.

The Applicability of the "Student" Theory

The "Student" integral has been misinterpreted often in the literature, through failure to realize these limitations—namely, the requirements of a normal population (or, in practice, of a population which at any rate is not markedly non-normal), random sampling, and a not unusual σ_s. The manner in which at first glance it seems to avoid the earlier necessity of using the normal theory, with σ_s taken as an estimate of σ, is sometimes more apparent than real. Indeed, as observed by Deming (P:*130*:59), the numerical refinement of making probability calculations by the "Student" integral, rather than with the normal integral in which σ_s is substituted for σ, "is not so momentous as has been proclaimed by many writers; of much more importance to the statistician is the fact that, whether he uses the classical estimate . . . of σ, or 'Student's' integral, he is at the mercy of the sampling fluctuations of σ_s, even in controlled experiments". Under any circumstances, therefore, the "Student" method must be applied, and then interpreted, with care.

The most direct manner of applying "Student's" distribution, either in his z form of (44) and (45), or in the t form of (46) or (47), is to test *a hypothesized mean*, m, on the evidence afforded by the value of $\frac{\bar{x}-m}{\sigma_s}$ observed in a *single* sample of n variates—subject always to the conditions already stated concerning normality, randomness, and σ_s. Illustrations are given in examples (1), (2), and 3(*a*) at p. 301; C; 13.

Extension of the "Student" Method to Testing the Difference between the Means of Two Samples

The same procedure may also be applied immediately, with similar reservations, to test *the difference between the means of two samples*, when either (i) both samples merely exhibit the effects of two different actions upon the same set of n individuals (as in example 3(*b*) at p. 304; C; 13), or (ii) the two samples exhibit the effects of two different actions upon two different sets, each of n individuals (as in example (4) at p. 305; C; 13).

The latter use of two different sets, however, will obviously decrease the reliability of the results in comparison with method (i) which uses the same set, for there may be variations between the two sets even though both are assumed to have been drawn at random from a normal universe. For this reason, and also to provide a method of dealing with the problem when the number, n, is not the same in the two sets, R. A. Fisher (P:*39*) showed that "Student's" distribution can be extended, in order to test the difference between the means of two samples of different size, by treating the two different sets as two entirely separate series (cf. P:*43*:128-9, 130 (example 20), and 133 (second method); and P:*110*:586). Thus if $\bar{x}_1$ and ${}_1\sigma_s$ be the mean and standard deviation of the first sample of n_1 variates x'_r for $r=1, 2, \ldots, n_1$, and $\bar{x}_2$ and ${}_2\sigma_s$ those of the second sample of n_2 variates x''_r for $r=1, 2, \ldots, n_2$—both samples being random drawings from a normal universe with mean m and standard deviation σ—it follows from (27) and (30), on the assumption that the two samples are separate (i.e., independent), that the mean square error (or "variance") of the difference between the two means is $\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$. Since, however, σ^2 is unknown, an estimate of it, σ_e^2, must be made from the data; and it will be seen, on the principles of (42) as extended at p. 164; A; 8 and p. 250; B; 26 for the case of k "constraints", that we may write $\sigma_e^2 = \dfrac{\sum_{r=1}^{r=n_1}(x'_r-\bar{x}_1)^2+\sum_{r=1}^{r=n_2}(x''_r-\bar{x}_2)^2}{n_1+n_2-2}$, because k is to be taken as 2 in the denominator on account of the calculation being made from the two values, $\bar{x}_1$ and $\bar{x}_2$, which are determined from the two separate series of the data. Evidently this estimate σ_e^2 will be $\dfrac{n_1\,{}_1\sigma_s^2+n_2\,{}_2\sigma_s^2}{n_1+n_2-2}$; and substituting this value for σ^2 in $\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$, we obtain $\sigma_e^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)$ as the estimate of the variance of $\bar{x}_1-\bar{x}_2$. If now—analogously to the definition of t in (46) as the ratio of $\bar{x}-m$ to the estimated standard error of $\bar{x}$—we define t as the ratio of $\bar{x}_1-\bar{x}_2$ to the estimated standard error of $\bar{x}_1-\bar{x}_2$, we have $t=\dfrac{\bar{x}_1-\bar{x}_2}{\sigma_e\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$, where

σ_e is found as already shown, and the table of the "Student" t-distribution is then entered by taking d as the n_1+n_2-2 "degrees of freedom". An example is shown in (5) at p. 305; C; 13.

Fisher's z-Distribution

The preceding method of testing the difference between the means of two samples uses the two sample variances, ${}_1\sigma_s$ and ${}_2\sigma_s$, in obtaining σ_e^2 as an estimate of the σ^2 of the universe, but it affords no test of the question as to whether both those sample variances can justifiably be used to give estimates of one and the same σ^2. Particularly in the light of the requirement, previously stated, that σ_s in the "Student" theory must be not unusual, it is therefore important to develop a *test of significance for the difference between two sample variances*. Such a test, moreover, strictly should be applied before embarking upon the above test of the two means—for the assumptions underlying the test of the means would not be satisfied if the hypothesis that the two sample variances can be used for estimates of the same σ^2 should be refuted.

Suppose, therefore, that we have two independent estimates of σ^2, namely, ${}_1\sigma_e^2$ based as before on a first sample of n_1 variates x'_r, and ${}_2\sigma_e^2$ based on a second sample of n_2 variates x''_r. Then, by (42), ${}_1\sigma_e^2 = \dfrac{\sum_{r=1}^{r=n_1}(x'_r-\bar{x}_1)^2}{n_1-1}$ and ${}_2\sigma_e^2 = \dfrac{\sum_{r=1}^{r=n_2}(x''_r-\bar{x}_2)^2}{n_2-1}$. In order to surmount the mathematical difficulties in finding the distribution of the difference ${}_1\sigma_e^2-{}_2\sigma_e^2$, R. A. Fisher (P:*39*; cf. also P:*62*:287) based his approach on the ratio $\dfrac{{}_1\sigma_e}{{}_2\sigma_e}$. The distribution of this ratio follows easily (see (4) at p. 228; B; 13) from the distribution of σ_s found in (2) at p. 225; B; 13. Putting then $z=\log_e\left(\dfrac{{}_1\sigma_e}{{}_2\sigma_e}\right)$, and writing $n_1-1=d_1$ "degrees of freedom" and $n_2-1=d_2$ "degrees of freedom", it may be shown readily (as in (5) at p. 229; B; 13) that, if the two samples come from the same normal universe, the distribution of z, say $F(z)$, is

$$F(z) = \frac{2d_1^{\frac{d_1}{2}}\, d_2^{\frac{d_2}{2}}}{B\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \frac{e^{d_1 z}}{(d_1 e^{2z} + d_2)^{\frac{d_1+d_2}{2}}}\, dz \qquad \ldots (47z)$$

where the B-function is defined as at p. 259; B; 29.

This result is known as **Fisher's z-Distribution.** [This z, of course, must not be confused with the z used by "Student" in his distribution (44)]. It is, in fact, a very general expression, from which the normal, χ^2 (see Chapter IX), and "Student's" distributions, and others also, are obtainable easily as special cases (P:*38*:807; or see P:*111*:173, or P:*110*:605).

Since (47z) is obtained on the assumptions of random sampling and a normal universe, and as again it does not involve the σ of the universe, its interpretation (though not its derivation) is also subject to the reservation that ${}_1\sigma_s$ and ${}_2\sigma_s$ must be not unusual, just as in "Student's" distribution it was explained that σ_s must be not unusual.

In practical applications of the method the customary procedure is to test the hypothesis that ${}_1\sigma_e^2$ and ${}_2\sigma_e^2$ are estimates of the same universe σ^2. If that hypothesis is true, the ratio of ${}_1\sigma_e^2$ to ${}_2\sigma_e^2$ will fluctuate around unity from one experiment to another, and if ${}_1\sigma_e^2$ and ${}_2\sigma_e^2$ agree in giving the same estimate of σ^2 their ratio would be 1 and z would be 0. For certain values of the degrees of freedom, d_1 and d_2 (where d_1 is used with the greater estimated variance), the values of z corresponding to the probabilities .05, .01, and .001 (the "5%, 1%, and .1% points") have been tabulated by Fisher and Deming in P:*43*:250-255, and have been reproduced (with permission) in many texts.

An illustration is shown in (6) at p. 306; C; 13.

VI. GENERALIZATION OF THE BINOMIAL LAW—THE "MULTINOMIAL" DISTRIBUTION

IN THE preceding chapters the development has been founded on the "binomial law", involving the two probabilities only, p and $q\ (=1-p)$. This has been sufficient for practical application to many types of actuarial problems—for in such problems, like those concerning the occurrence of death or survival, the event under consideration either happens, or does not, i.e., the basic probabilities required are the probability of an event happening (p), and the complementary probability that it will not happen ($1-p$, or q). And whenever it has been necessary to consider a series of such occurrences (for example, a series of deaths at ν different ages, or in ν various localities), the "mathematical model" constructed has been to regard each term of the series as independent, so that, in effect, the same basic model was used for each separate term of the series, and the series was eventually covered by successive independent applications of that one-term model.

In the case, for example, of a series of observed deaths, $\theta'_1, \theta'_2, \ldots, \theta'_\nu$, which have arisen from $E'_1, E'_2, \ldots, E'_\nu$ exposed to risk, at ν different ages 1 to ν, it is assumed in this method that the terms of the series are independent, so that, if $q_1, q_2, \ldots, q_\nu$ are the "true" independent probabilities of death at the several ages, and $p_1, p_2, \ldots, p_\nu$ the corresponding "true" independent probabilities of survival, then at any age r (for $r=1, 2, \ldots, \nu$) the probability (unless $E'q<10$), by the Normal Law of Deviations (10), of getting the observed deaths θ'_r instead of the "true" θ_r is $\dfrac{1}{\sqrt{2E'_r p_r q_r}}\, e^{-\frac{(\theta'_r-\theta_r)^2}{2E'_r p_r q_r}}$, and consequently for all ages $r=1, 2, \ldots, \nu$ the probability is the product of all these independent probabilities, namely,

$$(2)^{-\frac{\nu}{2}} \sqrt{W_1 W_2 \dots W_\nu}\, e^{-\sum_{r=1}^{r=\nu}\left[\frac{(\theta'_r-\theta_r)^2}{2E'_r p_r q_r}\right]} \quad \text{where} \quad W_r = \frac{1}{E'_r p_r q_r}.$$

The exponent $\frac{1}{2}\sum_{r=1}^{r=\nu}\left[\frac{(\theta'_r-\theta_r)^2}{E'_r p_r q_r}\right]$ in this probability, which here arises from separate applications of the binomial Normal Law of Deviations (10) to each of the ν independent terms of a series of observations, should be noted carefully, for it will be encountered again in the development to be now explained (cf. p. 217).

When, however, the terms of the series cannot thus be regarded as independent of each other, it becomes necessary to extend the mathematical model so that the series of, say, ν terms can all be dealt with at once. The procedure required is clearly a generalization of the binomial into a *multinomial* formula.

Let us therefore first consider a hypothetical "parent" population only. Suppose, accordingly, that we wish to visualize the *distribution* over ν "cells" of a total hypothetical parent population of N, in the form of $f_1, f_2, \dots, f_\nu$ falling into the cells numbered 1, 2, . . . , ν respectively, where $f_1+f_2+\dots+f_\nu=N$, so that the total N of the parent population is entirely distributed (such as hypothetical deaths $D_1, D_2, \dots, D_\nu$ at successive ages 1 to ν in a parent population, where $\sum_{r=1}^{r=\nu} D_r=N$). Then—recalling, firstly, the argument used in establishing the binomial case—it is evident that, since we are here contemplating the parent population only, the "true" probability, say p_r, of one of the N items falling into the rth cell is $\frac{f_r}{N}$, and the probability of all the f_r items falling into the rth cell is $\left(\frac{f_r}{N}\right)^{f_r}=(p_r)^{f_r}$. But now—unlike the binomial model—we do not have to consider the complementary probability $\left(1-\frac{f_r}{N}\right)$ that an item will not fall into the rth cell, since that contingency is covered completely by the assignment of the other $(r-1)$ probabilities of falling into each of the other $(r-1)$ cells. It is therefore clear that, introducing the necessary permutation, the probability of getting the parent distribution $f_1, f_2, \ldots, f_\nu$ in the ν different cells is

$$\frac{N!}{f_1!\,f_2!\ldots f_\nu!}\,p_1^{f_1}\,p_2^{f_2}\ldots p_\nu^{f_\nu} \qquad \ldots.(48)$$

With this mathematical model clearly in mind, it will be apparent that if now we consider an actual case where the N items are observed to be distributed over the ν cells in the series $f_1', f_2', \ldots, f_\nu'$, instead of in the theoretical series $f_1, f_2, \ldots, f_\nu$, then the probability of getting that particular observed series (out of the many such series which are possible as variations of the "true" $f_1, f_2, \ldots, f_\nu$) is

$$\frac{N!}{f_1'!\,f_2'!\ldots f_\nu'!}\,p_1^{f_1'}\,p_2^{f_2'}\ldots p_\nu^{f_\nu'} \qquad \ldots.(49)$$

where again the "true" probabilities are $p_r = \dfrac{f_r}{N}$ for $r=1, 2, \ldots, \nu$, and in this particular case all the N items are again distributed so that $f_1'+f_2'+\ldots+f_\nu'=N$ in conformity with $f_1+f_2+\ldots+f_\nu=N$ for the parent population. This is the general term in the expansion of the multinomial $(p_1+p_2+\ldots+p_\nu)^N$.

The equality of the totals which is thus here imposed means that $\sum_{r=1}^{r=\nu} f_r' = \sum_{r=1}^{r=\nu} f_r = \sum_{r=1}^{r=\nu} (Np_r) = N$. Under these conditions, as the total N is fixed, it will be seen that the last f' is determined automatically as soon as the other $\nu-1$ of the f''s are assigned. [In the binomial case, for example, where $\nu=2$, the number of independent variables is 1, since $f_1'+f_2'=N$, so that f_2' is fixed when f_1' is known; in the multinomial case, correspondingly, there are ν variables of which, however, only $\nu-1$ are independent, since $f_1'+f_2'+\ldots+f_\nu'=N$, so that f_ν', say, is fixed when $f_1', f_2', \ldots, f_{\nu-1}'$ are known]. In the usual terminology, there is in these cases one **constraint**—here a *linear constraint* because the condition imposing the constraint, namely, that $f_1'+f_2'+\ldots+f_\nu'=N$, involves only the first powers of the f''s; and this one constraint leaves $\nu-1$ **degrees of freedom**, being the remaining $\nu-1$ of the f''s which remain free to be assigned at will. Similarly, if, as will be encountered later, the conditions of a problem impose k such constraints, then there are $\nu-k$ degrees of freedom (see p. 175; A; 19).

If now in the expression (49), we replace all the factorials by their approximations according to Stirling's formula as in the deduction of the Normal Law of Deviations on p. 203; B; 5, and write $\frac{f'_r-Np_r}{\sqrt{Np_r}}=j_r$, the expression reduces (see p. 230; B; 14) to

$$\frac{1}{(\sqrt{2\pi N})^{\nu-1}\sqrt{p_1p_2\ldots p_\nu}}e^{-\frac{1}{2}\sum_{r=1}^{r=\nu}j_r^2} \quad \ldots.(50)$$

as long as Np_r is not less than about 10.

This may be referred to as the **Multinomial Normal Law of Deviations.** Just as the introduction of the Stirling approximation into the general term of the point binomial, as given in (2), led to the symmetrical "normal" approximation (10) with the two probabilities p and q, and maximum at the mean thence decreasing to zero at both ends, so here the resulting (50) would, in ν-dimensional space, be maximum in the neighbourhood of the "point" with co-ordinates Np_1, Np_2, . . . , Np_ν, and would thence everywhere decrease symmetrically to zero.

The fact that (50) is a generalization of the binomial form (10) may be seen readily. For the binomial represents, in reality, the filling of but two cells, so that $\nu=2$, through the operation of $p_1=p$ and $p_2=q$; all the n cases are so distributed, so that $N=n$; and $f'_1=np+x$ and $f'_2=nq-x$. Putting these values in (50), the result is at once (10)—for

$$\sum_{r=1}^{r=\nu}j_r^2 \text{ is then } (j_1^2+j_2^2)=\frac{(f'_1-Np_1)^2}{Np_1}+\frac{(f'_2-Np_2)^2}{Np_2}$$

$$=\frac{(np+x-np)^2}{np}+\frac{(nq-x-nq)^2}{nq}=x^2\left(\frac{1}{np}+\frac{1}{nq}\right)=\frac{x^2}{npq},$$

since $p_1+p_2=p+q=1$.

From this analysis it will accordingly be seen that the *approximate* probability of getting a particular *observed* series $f'_1, \ldots, f'_\nu$ is (50) in which j_r is $\dfrac{f'_r - Np_r}{\sqrt{Np_r}} = \dfrac{f'_r - f_r}{\sqrt{f_r}}$. If, therefore, we here write

$$\sum_{r=1}^{r=\nu} j_r^2 = \sum_{r=1}^{r=\nu} \frac{(f'_r - Np_r)^2}{Np_r} = \sum_{r=1}^{r=\nu} \frac{(f'_r - f_r)^2}{f_r} = \chi_0^2, \text{ that is,}$$

$$\chi_0^2 = \frac{(f'_1 - Np_1)^2}{Np_1} + \frac{(f'_2 - Np_2)^2}{Np_2} + \ldots + \frac{(f'_\nu - Np_\nu)^2}{Np_\nu} \quad \ldots (51)$$

we have

$$\frac{1}{(\sqrt{2\pi N})^{\nu-1}\sqrt{p_1 p_2 \ldots p_\nu}}\, e^{-\frac{\chi_0^2}{2}} \quad \ldots (52)$$

as the probability of getting the observed data $f'_1, f'_2, \ldots, f'_\nu$, i.e., the probability of the occurrence of the deviations $(f'_r - Np_r)$ where $r = 1, 2, \ldots, \nu$.

This χ_0^2—a function of great importance, as will be seen later in Chapters VIII and IX—is thus the sum of the squares of the deviations of the observed from the expected parent values, each divided by the latter. It is not arbitrary, since it has emerged as an essential part of the derivation of the Multinomial Normal Law of Deviations, and is thus entirely consistent with the theory underlying that formula, of which the binomial is a special case.

Now for the binomial we know from (10) and (11) that the probability of a deviation x lying between $\alpha\sqrt{n}$ and $\beta\sqrt{n}$ is $\dfrac{1}{\sqrt{2\pi npq}}\displaystyle\int_{\alpha\sqrt{n}}^{\beta\sqrt{n}} e^{-\frac{x^2}{2npq}}\, dx$. Writing $p = p_1$, $q = p_2$, and $n = N$ to correspond with the multinomial case, and changing the variable by putting $\dfrac{x}{\sqrt{n}} = t_1$, this becomes $\dfrac{1}{(2\pi)^{\frac{1}{2}}\sqrt{p_1 p_2}}\displaystyle\int_\alpha^\beta e^{-\frac{1}{2}\left(\frac{t_1^2}{p_1 p_2}\right)}\, dt_1$.

Introducing now a second variable $t_2 = -t_1$, the term $\dfrac{t_1^2}{p_1 p_2}$ in the exponent can be written $\left(\frac{t_1^2}{p_1}+\frac{t_2^2}{p_2}\right)$. Hence the probability in the multinomial (50), when $\nu=2$, that the deviation $(f_1'-Np_1)$, which is $np+x-np=x$ in (10) and (11), will lie between $\alpha\sqrt{n}$ and $\beta\sqrt{n}$ (f_2' being, of course, fixed by the linear constraint $f_1'+f_2'=N$), is

$$\frac{1}{\sqrt{2\pi}\sqrt{p_1p_2}}\int_\alpha^\beta e^{-\frac{1}{2}\left(\frac{t_1^2}{p_1}+\frac{t_2^2}{p_2}\right)}dt_1, \text{ where } t_2=-t_1 \quad \ldots(53)$$

Extending this principle, therefore, for all the ν variables of (50) and (52), it will be seen that the probability that the deviations $(f_r'-Np_r)$ of the series $f_1', f_2', \ldots, f_{\nu-1}'$ will all simultaneously lie between $\alpha_1\sqrt{n}$ and $\beta_1\sqrt{n}$, between $\alpha_2\sqrt{n}$ and $\beta_2\sqrt{n}, \ldots,$ and between $\alpha_{\nu-1}\sqrt{n}$ and $\beta_{\nu-1}\sqrt{n}$ respectively (the last variable then, of course, being fixed by the linear constraint that $f_1'+f_2'+\ldots+f_\nu'=N$) is

$$\left.\begin{aligned}&\frac{1}{(\sqrt{2\pi})^{\nu-1}\sqrt{p_1p_2\ldots p_\nu}} \text{ times}\\ &\int_{\alpha_1}^{\beta_1}\int_{\alpha_2}^{\beta_2}\cdots\int_{\alpha_{\nu-1}}^{\beta_{\nu-1}} e^{-\frac{1}{2}\left(\frac{t_1^2}{p_1}+\frac{t_2^2}{p_2}+\ldots+\frac{t_\nu^2}{p_\nu}\right)} dt_1dt_2\ldots dt_{\nu-1}\end{aligned}\right\}\ldots(54)$$

where $t_\nu=-(t_1+t_2+\ldots+t_{\nu-1})$.

This expression is of great importance in the theory establishing Pearson's χ^2-test of Goodness of Fit (see Chapter IX).

VII. FREQUENCY DISTRIBUTIONS AND CURVES IN GENERAL

The Point Binomial and the Normal Curve

IT WAS observed in Chapter III that the Bernoullian, or point binomial, frequency distribution (3), i.e., $(q+p)^n$, is of the "discrete" class, and is symmetrical when $q=p=\frac{1}{2}$, but unsymmetrical when $q\neq p\neq \frac{1}{2}$—under which latter circumstances the series of ordinates, although markedly skew for small values of q and nq, rapidly approaches the symmetrical form as nq increases (see especially p. 267; C; 4). It was also shown that the ordinates of this distribution, when measured from the mean np, may be represented very closely by the always symmetrical Normal Law of Deviations

$$y_x \doteq \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}} \qquad \ldots .(10)$$

except when the asymmetry of (3) would be very marked by reason of q (or p) being so small and n sufficiently large that nq (or np) remains finite but small, i.e., when q (or p) is very small but a sufficient number of trials, n, is made that the event does happen occasionally. Finally, (10) was expressed in the continuous form of the **Normal Curve of Error**

$$y_x = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} \qquad \ldots .(11)$$

which again is an always symmetrical bell-shaped curve with relationship to (10), and parameters, expressible as

$$c=\sqrt{2npq}=\sqrt{\pi}\text{ (mean error, }\eta) = \sqrt{2}\text{ (standard deviation, }\sigma)$$
$$=\frac{1}{.476936}\text{ (probable error, }\lambda) \qquad \ldots .(18)$$

The "Skew-Normal" Curve

Notwithstanding the early belief that this symmetrical Normal Curve might, indeed, represent a universal law of nature (see p. 151; A; 3), it was realized in the classical dissertations that the symmetrical forms (10) and (11) had been reached by a particular method of approximation, in which certain terms had been neglected. In the demonstration at p. 204; B; 5, for example, it is shown that (10) and (11) are, in fact, obtained by neglecting the last term in the **Skew-Normal Curve**

$$y_x = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} e^{\frac{x(p-q)}{c^2}} \qquad \ldots (ii)$$

which, therefore, is in theory more strictly applicable when $p \neq q$. The effect of the last slightly skew term in this expression, however, is very small except in extreme cases (see p. 265; C; 4), so that the Skew-Normal curve is of little practical importance.

Poisson's Exponential—The "Law of Small Numbers"

Another formula which has commanded much greater interest, and is certainly of major importance, was given so long ago as 1837 (H: 30. 205) by the French mathematician Poisson, and has become known as the **Law of Small Numbers** (see p. 166; A; 11). The conditions for its applicability are precisely those under which the approximations inherent in the Normal Law are not satisfied, i.e., when q (or p) is so small but the number of trials, n, is sufficiently large that nq (or np), while small, is finite, so that the event only happens occasionally. Let us examine, therefore, the effect upon the fundamental

$$\frac{n!}{(np+x)!\,(nq-x)!} p^{np+x} q^{nq-x} \qquad \ldots (2)$$

which represents the probability of $(np+x)$ successes and $(nq-x)$ failures in n trials, of supposing that q is very small but n still large enough that $nq = m$, where m is small but nevertheless finite. Then (writing r for $nq-x$), it follows, as shown at p. 230; B; 15, that the probability of r occurrences of such a rare event in n trials is given by

$$y_r \doteq \frac{m^r e^{-m}}{r!} \qquad \ldots (55)$$

where $m = nq$.

This **Poisson exponential,** of which several tables are available (see p. 234; B; 15), is a discrete function, existing only for integral values $r = 0, 1, 2, \ldots, n$. The single parameter, m, is very easily determined from

$$\text{Mean} \doteq m \qquad \ldots.(56)$$

It may also be shown (p. 234; B;15) that

$$\sigma^2 \doteq m \qquad \ldots.(57)$$

The marked skewness of its frequency polygon for small values of m, and the rapidity with which it approaches the symmetrical "normal" form, may be seen from Figure 4.

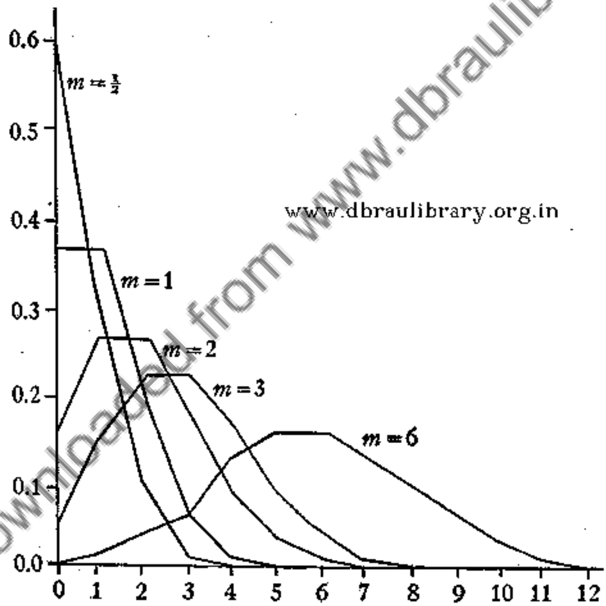

FIGURE 4.—Frequency Polygons of the Poisson Exponential

While it is inadvisable to attempt an entirely specific answer to the question as to the exact statistical conditions under which the Poisson exponential is to be preferred to the "normal" form, it may, nevertheless, be indicated that the application of the

Normal Curve should be made with caution for values of q (or p) below about .03 when nq (or np)—to which m in the diagram on p. 61 is an approximation—is 10 or less. The admissibility of the procedure based on the Normal Curve may, of course, also depend upon the type of statistical enquiry under consideration; the application of the normal, or the Poisson, theory to a distribution shaped as in the preceding diagram might be uncertain in respect of the calculation of some particular ordinate or an area on one side only of the mean, but yet might well give a close approximation if an area of the curve were required in order to examine certain limits of both positive and negative deviations taken together (see p. 310; C; 14).

Edgeworth's Generalized Law of Error

Another method of taking into account the terms neglected in the derivation of the normal forms (10) and (11) was developed in England by Edgeworth, and may be noted conveniently at this point on account of the manner of its approach, although chronologically it succeeded the basic investigations of the Scandinavian school (see p. 167; A; 12). The **Generalized Law of Error** which Edgeworth reached (see p. 234; B; 16) may be written

$$y_x = \left[e^{-\left(\frac{k_1}{3!}\right) D^3 + \left(\frac{k_2}{4!}\right) D^4 - \ldots + (-1)^l \frac{k_l}{(l+2)!} D^{l+2} + \ldots} \right] \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} \qquad \ldots (58)$$

where $k_1, k_2, \ldots$ are constants and D represents differentiation with regard to x.

When all the k's are zero, this expression obviously reduces at once to the Normal Curve.

If, however, we retain k_1, but neglect k_2, etc., the result in the form used by Edgeworth is (see also p. 205; B; 5)

$$y_x \doteqdot \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} \left[1 - 2j\left(\frac{x}{c} - \frac{2}{3}\frac{x^3}{c^3} \right) \right] \text{where } j = \frac{\mu_3}{c^3} \quad \ldots (59)$$

Similarly, retaining terms with k_1, k_2, and k_1^2, but neglecting the rest, the approximation becomes (see H:*162*:45, and p. 206; B; 5) in Edgeworth's notation

$$y_x = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}} \left[1 - 2j\left(\frac{x}{c} - \frac{2}{3}\frac{x^3}{c^3}\right) + j^2\left(-\frac{5}{3} + 10\frac{x^2}{c^2} - \frac{20}{3}\frac{x^4}{c^4} + \frac{8}{9}\frac{x^6}{c^6}\right) + i\left(\frac{1}{2} - \frac{2x^2}{c^2} + \frac{2}{3}\frac{x^4}{c^4}\right)\right] \quad \ldots .(60)$$

where $j = \frac{\mu_3}{c^3}$ and $i = \frac{1}{4}\left(\frac{\mu_4}{\mu_2^2} - 3\right)\cdot$

The practical application of these curves is discussed briefly on p. 311; C; 15. For later comparison with the Gram-Charlier Type A series it may be useful to note (see P:*32*:133) that by expanding the exponential Edgeworth's form (60) can be put as

$$\frac{N}{\sigma}\{\tau_1(h) + .81649658\sqrt{\beta_1}\,\tau_4(h) + .98601330\,\beta_1\,\tau_7(h) + \ldots + .45643546\,(\beta_2 - 3)\,\tau_5(h) + \ldots\} \quad \ldots .(61)$$

where N is the total frequency, $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$, $\beta_2 = \frac{\mu_4}{\mu_2^2}$, and $\tau_{n+1}(h) = \frac{(-1)^n}{\sqrt{n!}} \cdot \frac{d^n}{dh^n}\left(\frac{1}{2\sqrt{\pi}} e^{-\frac{h^2}{2}}\right)$ as used in "Tables for Statisticians" (P:*97*:Part II).

The Gram-Charlier (Type A) and Poisson-Charlier (Type B) Series

Edgeworth's general expression (58), as noted on p. 234; B; 16, can be established as the distribution of a magnitude depending on a number of independently varying elements, and as such can evidently be written (see P:*155*:168-172)

$$y_x = \left[1 + A_3\left(\frac{d}{dx}\right)^3 + A_4\left(\frac{d}{dx}\right)^4 + A_5\left(\frac{d}{dx}\right)^5 + \ldots\right]\varphi(x) \quad \ldots .(62)$$

where A_3, A_4, A_5, ... are constants, and $\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$.

This method of using the "normal" function, $\varphi(x)$, as a *generating function* of a series was first applied to the representation of skew frequency distributions by Gram, of Denmark (H:*59*), and has since been developed by several Scandinavian mathematicians—notably Charlier (H:*100*), Wicksell (H:*146*), and Jörgensen (H:*125*). The values of the constants have been determined by several methods (see p. 235; B; 17), and in terms of moments enable (62) to be expressed as

$$y_x=\varphi(x)-\frac{1}{3!}\left(\frac{\mu_3}{\sigma^3}\right)\varphi_3(x)+\frac{1}{4!}\left(\frac{\mu_4}{\sigma^4}-3\right)\varphi_4(x)-\frac{1}{5!}\left(\frac{\mu_5}{\sigma^5}-\frac{10\mu_3}{\sigma^3}\right)\varphi_5(x)+\dots \quad \dots(63)$$

where $\varphi_n(x)$ is written for $\frac{d^n\varphi(x)}{dx^n}$. These derivatives of $\varphi(x)$ are easily obtained, and their values are available in many standard tables and texts (for example, in P:*97*, P:*47*, and P:*114*:209; see also P:*36*:214 and 280-1). This expansion is known as the **Gram-Charlier,** or **Type A,** series.

For comparison with Edgeworth's series, and in computation, it is convenient to put (63) in the form (see P:*32*:130)

$$\frac{N}{\sigma}\left\{\tau_1(h)+.81649658\sqrt{\beta_1}\,\tau_4(h)+.45643546\,(\beta_2-3)\,\tau_5(h)+\dots\right\} \quad \dots(64)$$

in which the notation follows that already stated for (61).

The practical utility of the Gram-Charlier Type A series, which is thus based on the use of the Normal Curve as a generating function, is evidently dependent upon rapid convergence in order that a few terms only shall be required.

For cases of marked skewness, Charlier has also employed as a generating function the Poisson exponential (55), which can assume a very skew shape, instead of the symmetrical Normal Curve, in the form

$$y_x=B_0\psi(x)+B_1\psi^{I}_{(x)}+B_2\psi^{II}_{(x)}+\dots\dots,$$

where

$$\psi(x)=e^{-m}\ \frac{\sin\pi x}{\pi}\left[\frac{1}{x}-\frac{m}{1!(x-1)}+\frac{m^2}{2!(x-2)}-\dots\right]$$

which in the limit when m is an integer becomes $\frac{m^x e^{-m}}{x!}$, and where also $\psi^{N}_{(x)}=\Delta\psi^{N-1}_{(x-1)}$. $\quad\dots(65)$

This expression proceeds by differences instead of derivatives because the Poisson exponential is a discrete function (see also P:*36*:268-9, and P:*140*:38). It is usually called the **Poisson-Charlier,** or **Type B,** series.

Four different methods of fitting are suggested by Charlier, for which the formulae are easily accessible in P:*32*:131-2 (see also P:*36*:271 et seq.). References to practical illustrations are given on p. 311; C; 16 here.

Pearson's System of Frequency Curves

Many frequency distributions which are encountered in practice are either symmetrically or unsymmetrically bell-shaped, and thus rise from zero to a maximum and thence decrease. This characteristic has led to the extensive use of a highly valuable system of frequency curves which was originated and developed by Karl Pearson. For if a curve is asymptotic to the x-axis (i.e., has "high contact") at one end we must have $\frac{dy}{dx}=0$ when $y=0$; if the maximum occurs at, say, $x=-a$, we must again there have $\frac{dy}{dx}=0$; and accordingly a very general expression for such a frequency distribution (which may evidently include other types of curves as well) can be written down immediately as

$$\frac{dy}{dx}=\frac{y(x+a)}{F(x)}, \text{ or } \frac{d \log y}{dx}=\frac{x+a}{F(x)} \qquad \ldots.(66)$$

As an alternative method of approach it may be recalled that the Normal, Skew-Normal, and Poisson expressions were obtained as developments of the "point binomial", in which the probability of success, p, remains constant. If, however, this probability is not constant, but depends on the previous occurrences in a set of trials, we must use a different probability series, namely, the *hypergeometrical*, based on the supposition of drawing, say, r balls one at a time, without replacements, from a bag containing np white and nq black balls. The probability that s balls will be white out of the r balls so drawn is

$$\frac{{}^{r}C_{s}\,{}^{np}P_{s}\,{}^{nq}P_{r-s}}{{}^{n}P_{r}}, \text{ or } \frac{{}^{np}C_{s}\,{}^{nq}C_{r-s}}{{}^{n}C_{r}}.$$

Putting then $s=0, 1, 2, \ldots, r$ we obtain the ordinates of the distribution at unit intervals—the series taking (see P:*32*:39-41, and P:*116*:50-3) the general form

$$\frac{1}{y}\frac{dy}{dx} = \frac{x+a}{b_0+b_1x+b_2x^2} \quad \ldots\ldots(67)$$

This expression is that already written down at (66) when $F(x)$ is expanded in ascending powers of x.

The various curves which arise from the integration of (67) evidently depend upon the forms taken by the denominator, i.e., they depend upon the nature of the roots of $b_0+b_1x+b_2x^2=0$, for which the criterion is, of course, $\frac{b_1^2}{4b_0b_2}$. Too much space would be required to give the integrations here. They are, however, quite straightforward, and are available so clearly in Elderton's book (P:*32*:38 et seq.) that actuarial students may be referred to it without hesitation.

In Pearson's numbering there are 13 curves, called Types I to XII, and the Normal curve. Types I, IV, and VI are the *Main Types*; the others are *Transition Types* which arise as limiting cases when the main types change into each other, and embrace not only the Normal Curve but also a straight line, a geometrical progression, and J-shaped, twisted J-shaped, and U-shaped curves.

The distinctions between the three main types and the ten transition types are indicated in the following table, where, for ease of reference, Pearson's own numbering is used and the sequence adopted by Elderton in P:*32* is retained.

Typical shapes of the various curves are next shown for certain positive values of the parameters.

Number of Type	Equation	*Shape*; and whether *Limited* (in both directions), *Limited One Way* (i.e., in one direction only), or *Unlimited* (in both directions)
	MAIN TYPES	
I	$y=y_0\left(1+\frac{x}{a_1}\right)^{\nu a_1}\left(1-\frac{x}{a_2}\right)^{\nu a_2}$	Usually skew bell-shaped. J-shaped when νa_1 negative. Twisted J-shaped when both νa_1 and νa_2 are arithmetically <1 and one of them negative. U-shaped when both νa_1 and νa_2 negative. Limited.
IV	$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m} e^{-\nu \tan^{-1}\frac{x}{a}}$	Skew bell-shaped. Unlimited.
VI	$y=y_0(x-a)^{q_2}x^{-q_1}$	Usually skew bell-shaped. J-shaped when q_2 negative. Limited one way.
	TRANSITION TYPES	
Normal Curve	$y=y_0e^{-\frac{x^2}{2\sigma^2}}$	Symmetrical bell-shaped. Unlimited.
II	$y=y_0\left(1-\frac{x^2}{a^2}\right)^{m}$	Usually symmetrical bell-shaped. U-shaped when m negative. Limited.
VII	$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}$	Symmetrical bell-shaped. Unlimited.
III	$y=y_0\left(1+\frac{x}{a}\right)^{\gamma a} e^{-\gamma x}$	Usually skew bell-shaped. Geometrical progression (exponential) when $\gamma a=0$. J-shaped when $\gamma a<0$. Limited one way.
V	$y=y_0x^{-p}e^{-\frac{\gamma}{x}}$	Skew bell-shaped. Limited one way.
VIII	$y=y_0\left(1+\frac{x}{a}\right)^{-m}$	From infinite ordinate at $x=-a$ to finite ordinate at $x=0$; m lies between 0 and 1. Equilateral hyperbola when $m=1$.
IX	$y=y_0\left(1+\frac{x}{a}\right)^{m}$	From zero ordinate at $x=-a$ to finite ordinate at $x=0$ when $m>0$; straight line when $m=1$.
X	$y=y_0e^{-\frac{x}{\sigma}}$	Exponential (special case of Type III); Laplace's "First Law of Error" (see p. 159; A; 4).
XI	$y=y_0x^{-m}$	J-shaped (special case of Type VI).
XII	$y=y_0\left(\frac{a_1+x}{a_2-x}\right)^{p}$	Twisted J-shaped (special case of Type I).

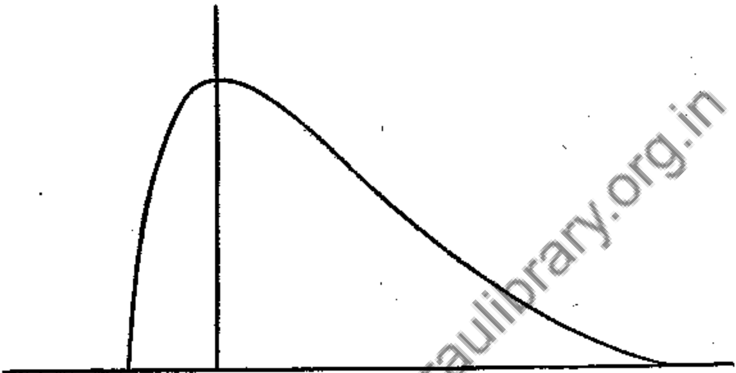

FIGURE 5.—Main Type I

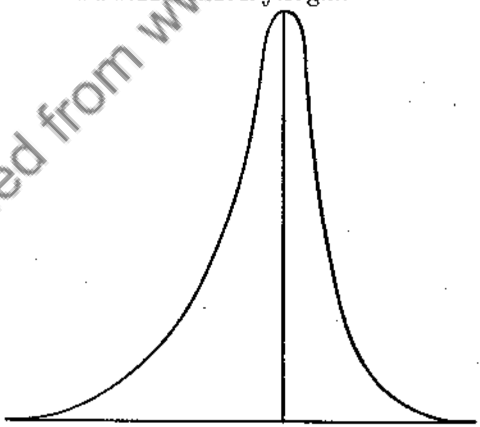

FIGURE 6.—Main Type IV

FIGURE 7.—Main Type VI

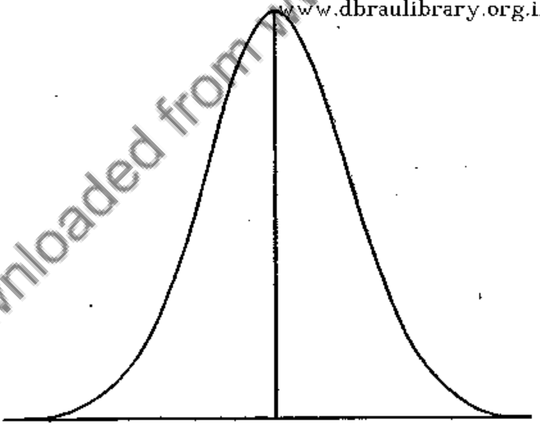

FIGURE 8.—Transition Type—"Normal Curve"

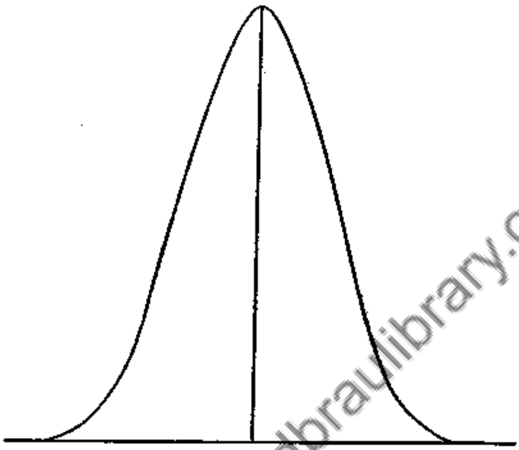

FIGURE 9.—Transition Type II

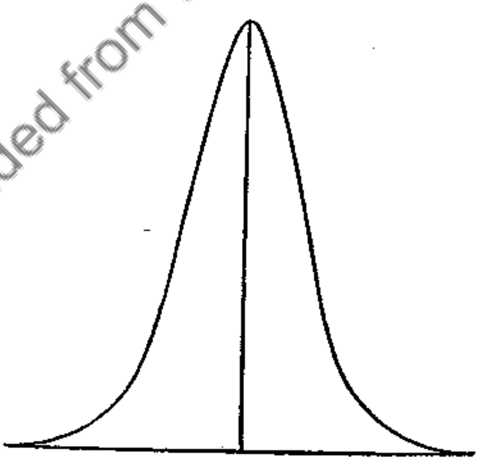

FIGURE 10.—Transition Type VII

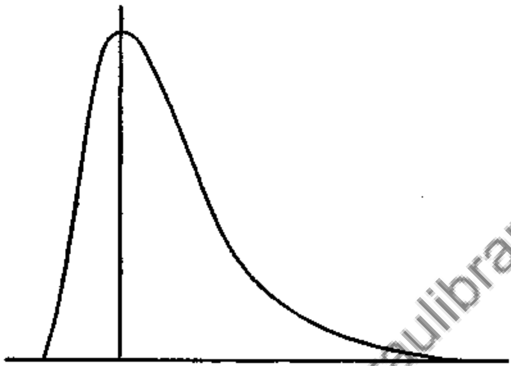

FIGURE 11.—Transition Type III

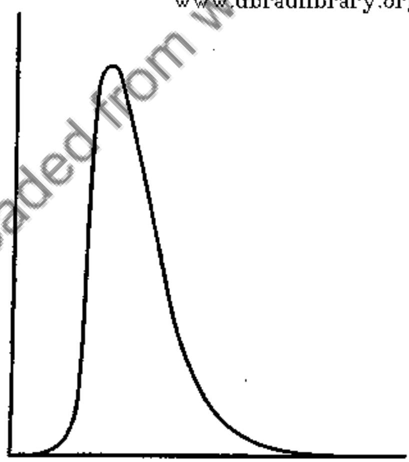

FIGURE 12.—Transition Type V

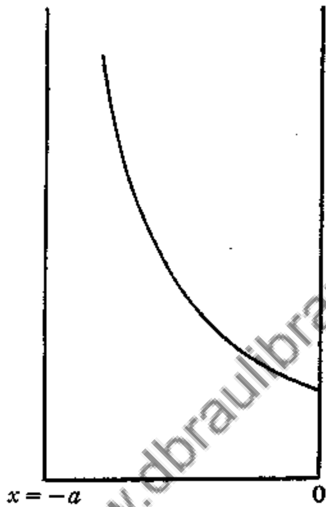

FIGURE 13.—Transition Type VIII

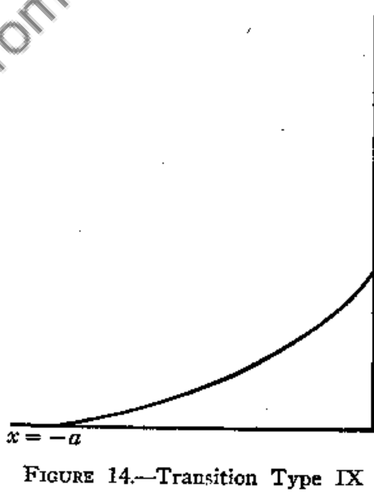

FIGURE 14.—Transition Type IX

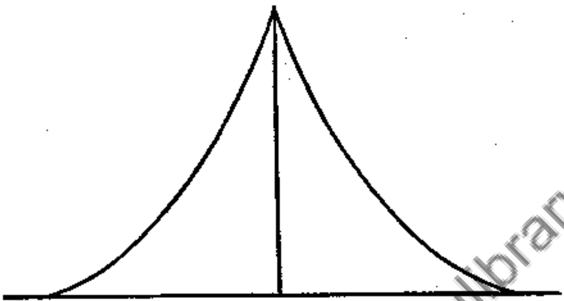

FIGURE 15.—Transition Type X

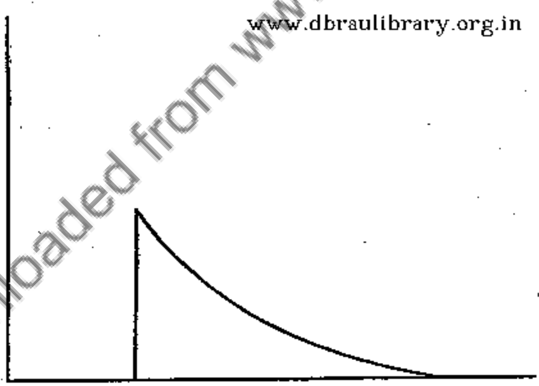

FIGURE 16.—Transition Type XI

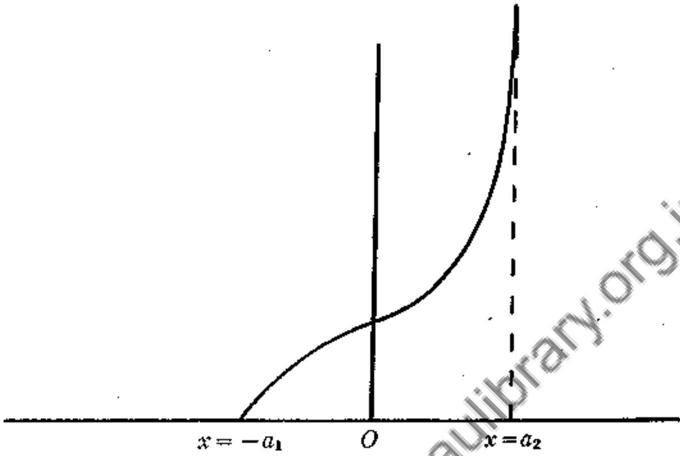

FIGURE 17.—Transition Type XII

Since the form of the curve depends upon $\frac{b_1^2}{4b_0b_2}$, the selection of the appropriate type in any particular case can be made readily by computing the numerical value of this *criterion*. In terms of moments it may be written easily (see P:*32*:41-45) as

$$\frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)} \text{ where } \beta_1=\frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2=\frac{\mu_4}{\mu_2^2} \quad \ldots.(68)$$

from which the curve can be selected by the table in P:*32*:51. Alternatively, the type can be chosen by means of a diagram, to be found in P:*97*, showing the regions of each type for values of β_1 and β_2. Another rapid method is to calculate from the data the values of $\Delta^2 \log y$, whence the type may be indicated as suggested in P:*51*:50.

The fitting of these curves to statistical data is accomplished by the "method of moments", which was developed by Karl Pearson for that special purpose—see p. 97 here.

Their practical applicability in actuarial work has been explored extensively, and is discussed at p. 312; C; 17.

Further Modifications of the Normal Curve, and of Pearson's System

Although they are not of immediate practical importance, it may be well to note here—as a matter of theoretical interest only—that several other methods of representing frequency distributions have also been investigated.

(*i*) It may sometimes happen in dealing with certain statistics that the frequencies for a variable x do not accord with those of the Normal Curve (11), but that the corresponding frequencies for some function of x, say $f(x)=z$, may be so distributed. Under those circumstances it follows, as shown at p. 236; B; 18, that the *transformed* frequency function is

$$\varphi(x)dx = \frac{1}{c\sqrt{\pi}} f'(x)\, e^{-\frac{[f(x)]^2}{c^2}} dx \qquad \ldots .(69)$$

This device for dealing with frequency distributions was used extensively by Edgeworth, who called it the *Method of Translation* (see H:*162*:65), and has been explored also by Kapteyn (H:*91*), Wicksell, and Rietz (H:*139*, *161*, and *175*). One important case which has been widely discussed (for bibliography see P:*175*:73) results from the transformation $z = \frac{1}{\sqrt{2}} \log \left(\frac{x-a}{b}\right)$, and leads (see p. 236; B; 18) to the *logarithmic frequency function*

$$\phi(x) = \frac{1}{c\sqrt{2\pi}(x-a)}\, e^{-\frac{1}{2c^2}\left[\log\left(\frac{x-a}{b}\right)\right]^2} \qquad \ldots .(70)$$

The properties and practical application of this distribution are examined in P:*175*.

(*ii*) The logarithmic frequency function (70), or an equivalent form, has also been employed as the generating function of a series by the Scandinavians Charlier, Jörgensen, and Wicksell.

(*iii*) Romanovsky (H:*150*) has explored the possibility of using still other generating functions, such as Pearson's Types I, II, and III (cf. P:*116*:75 and P:*114*:116).

(*iv*) Carver has suggested (H:*129*) the use of a difference equation, instead of Pearson's differential equation (67), in the form

$$\frac{\Delta y}{\Delta x}=\frac{y(a-x)}{b_0+b_1x+b_2x^2} \qquad \ldots .(71)$$

The resulting formulae, with examples of the graduation of frequency distributions and stumps of such distributions, are given in H:*129* and P:*114*:111.

(*v*) Since the "mode" (the position of the maximum ordinate) is often of primary importance in economic data, Mouzon (in H:*167*) has given the curves resulting from assuming that the value of the constant a in (66) is first determined from the observed data and equated to the value of the mode in the theoretical distribution, and that the polynomial in the denominator is of the third degree or lower (instead of being taken of the second degree as in the fundamental equation (67) of Pearson's system).

(*vi*) Hansmann (H:*182*) also has examined the six main and fourteen transition type curves which emerge from expanding $F(x)$ in (66) to the fourth degree, with zero coefficients for x and x^3, so that $F(x)$ is taken as $b_0+b_2x^2+b_4x^4$. His conclusion is that these fourth order symmetrical curves fitted by moments give improved results, and so justify the additional work entailed. It must be remembered, however, that moments up to μ_8 are involved, with large sampling errors, and that in practical curve-fitting work it is usually preferable to avoid the use of such high moments (cf. P:*141*:754). Heron previously had taken $F(x)$ as far as $b_0+b_1x+b_2x^2+b_3x^3$, but had not published his investigations because the additional b_3x^3 term did not seem to effect any practical improvement over Pearson's curves.

(*vii*) Pearson's curves are fitted by the use of moments up to μ_4 (see Chapter VIII). R. A. Fisher (P:*37*:355) has observed that the system of curves for which the method of moments thus applied is the "best" method of fitting would be given by an exponential with a fourth degree argument, namely,

$$y = e^{-a^2(a_0+a_1x+a_2x^2+a_3x^3+x^4)} \qquad \ldots .(72)$$

—"the convergence of the probability integral requiring that the coefficient of x^4 should be negative, and the five quantities a, a_0, a_1, a_2, a_3 being connected by a single relation, representing the fact that the total probability is unity" (see p. 237; B; 19). Putting $a^2=1$ (since a depends only on the unit of measure of x), and replacing x by $x-\frac{a_3}{4}$, this exponential takes the form

$$y = e^{-(c_0+c_1x+c_2x^2+x^4)} = Ce^{-(c_1x+c_2x^2+x^4)} \text{ where } C = e^{-c_0} \qquad \ldots .(73)$$

Since Pearson's curves cannot produce a double hump, and it is doubtful whether Edgeworth's or the Gram-Charlier Type A can do so conveniently (cf. P:*32*:139 and 140), it is interesting to note that the classes of curves which arise from (73) are typically bimodal. They have been investigated by O'Toole, who shows an example (H:*180*:28) of a double-humped distribution which can be so represented (see also P:*65*:115).

Other Curves of Interest to the Actuary

From the preceding it will have been realized that not even the very varied capacities of the Normal, Skew-Normal, and Poisson curves, or of the Edgeworth, Gram-Charlier, Poisson-Charlier, and Pearson and other generalizations, will always handle some of the frequency distributions or series of statistical ratios with which the actuary is specially concerned. A great deal of investigation and ingenuity, in fact, has been devoted for many years to the invention of other curves which might be suitable for actuarial statistics. This chapter would therefore be seriously incomplete unless some attempt were made to catalogue the various proposals which have been advanced—although space will not permit in many cases more than a statement of each formula, with references to sources where additional details may be found.

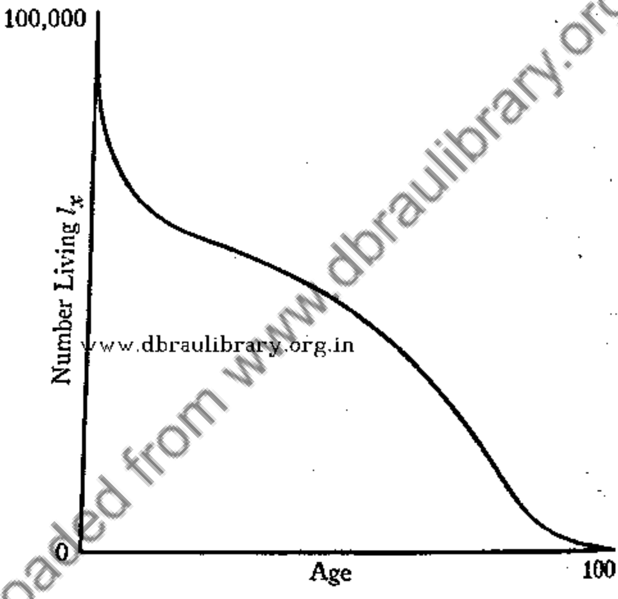

FIGURE 18.—Typical Curve of l_x

(1) Frequency Distributions

For the representation of frequency distributions proper, such as the "exposed to risk" or deaths, Hardy has pointed out (P:*51*:50) that, when m and n are numerically unequal, a skew curve vanishing when $x=-a$ or $-b$ is given by

$$y=ke^{-\left(\frac{m}{a+x}+\frac{n}{b+x}\right)} \qquad \ldots.(74)$$

while he has also suggested for the same purpose

$$y=kx^{\alpha}(1-x^{\beta}) \qquad \ldots.(75)$$

where x represents a proportionate part of the range of the curve so that x varies between 0 and 1 (see P:*51*:135).

An interesting and much earlier attempt to find a "law" (cf. p. 168; A; 13) followed by the "entrants" into mortality experiences led Chandler (H:*45*) to the empirical expression

$$y=ab^{-x^k}c^{x^k \sin x\varphi} \qquad \ldots.(76)$$

for the representation of a skew frequency distribution.

(2) The Curves of l_x and log l_x

Much attention has been given, notably some years ago (see p. 168; A; 13), to the possibility of discovering a formula which would describe adequately the twisted descending curve (as indicated in Figure 18) of l_x—the "number living" in the hypothetical "life-table", which results from multiplying an arbitrarily selected number of births, l_0, by the successive values of the probability of survival $p_x(=1-q_x)$. Since these explorations are of value now only to the extent that they have led in certain instances to workable expressions for the force of mortality $\mu_x\left(=-\frac{d\log_e l_x}{dx}\right)$, it must suffice here to refer to Elston's paper, H:*141*, for the record of most of the various attempts. It should be noted, nevertheless, that some of the curves which have attained wide recognition—especially Makeham's first and second formulae, (83) and (84)—have often been fitted to the data in their l_x form, and that Hardy (P:*51*:88) gives an illustration of the application of those formulae to enumerated census populations.

Another curve used with success by Hardy (loc. cit., 89 and 68) is the representation of log l_x, or the logarithms of the numbers living above age x, by

$$y=k+ma^x+nb^x \qquad \ldots.(77)$$

and the form $$y=A+Hx+Bc^x+Cx^2 \qquad \ldots.(78)$$

was employed for the populations and deaths at and above age x in the life tables constructed for the British National Insurance Act, 1911 (see H:*121*:554).

For the years of infancy and childhood up to age 12, Hardy has also used

$$l_x=A+Hx+Bc^x+\frac{m}{nx+1} \qquad \ldots.(79)$$

where the last term gives effect to the heavy mortality of the early ages and thereafter becomes practically negligible (see H:*122*:329 and 391).

This problem of evolving a formula which would be able to take into account the rapid change in the values during the infantile ages has more recently been examined again by Steffensen (H:*169*), who suggests that the "uniform seniority" property (see p. 319; C; 18) of the Makeham function (83) can be preserved by adding at the infantile ages the expression

$$\log l_x=10^{a\sqrt{x}+b} \qquad \ldots.(80)$$

It has also been pointed out (H:*185*) that even more satisfactory results may be obtainable if (80) is modified into

$$\log l_x=10^{a\sqrt{x}+b+kx}+d \qquad \ldots.(81)$$

(3) The Curves of q_x, m_x, μ_x, and colog p_x

Of all the formulae which have been proposed, however, those of the greatest practical importance to actuaries are of course the expressions which endeavour to represent the rate of mortality by age x, namely, $q_x \left(= \frac{l_x - l_{x+1}}{l_x} = \frac{d_x}{l_x} \right)$, or—more easily —the force of mortality $\mu_x \left(= -\frac{d \log_e l_x}{dx} \right)$, the central death rate $m_x \left(= \frac{d_x}{\int_0^1 l_{x+t}\, dt} \doteq \mu_{x+\frac{1}{2}} \right)$, $\text{colog}_e\, p_x \left(= -\log_e p_x = -\log_e \frac{l_{x+1}}{l_x} \doteq m_x \doteq \mu_{x+\frac{1}{2}} \right)$, or $\text{colog}_{10} p_x$ or $\log_{10} \mu_x$. The numerical values of these ratios, naturally, are not like bell-shaped frequency distributions; from infancy to old age they follow a contorted U-shaped curve, with the minimum in the neighbourhood of age 11, so that from about age 11 to the end of life the curve increases steadily (often with two minor undulations in the thirties and seventies) with its convexity towards the x-axis (cf. P:*102*:41, 45, and 55, and H:*141*:88). If, then, the values up to about age 10 are dealt with separately, the remainder of the curve from the region of age 11 upwards will usually be found to change slowly at first, and at the older ages to resemble more nearly a geometrical progression—a circumstance which means that often the logarithms of the values there approximate to an arithmetical progression. The points of inflexion, however, introduce great difficulties into the problem of finding any expression which will represent mortality rates over the whole period of life. A typical curve for q_x is shown in Figure 19 on the next page.

For a detailed catalogue of the many early attempts to find a "law" (cf. p. 168; A; 13) of mortality in mathematical form the reader may again be referred to H:*141*, as it will suffice to include here only those which have proved to be practically useful. The first which attained wide recognition was *Gompertz's* simple geometrical progression (H:*18*)

$$\mu_x = Bc^x, \text{ from which } l_x = kg^{c^x} \qquad \ldots .(82)$$

Next followed *Makeham's First Modification* (H:*31*:303) where the first differences of μ_x follow a geometrical progression,

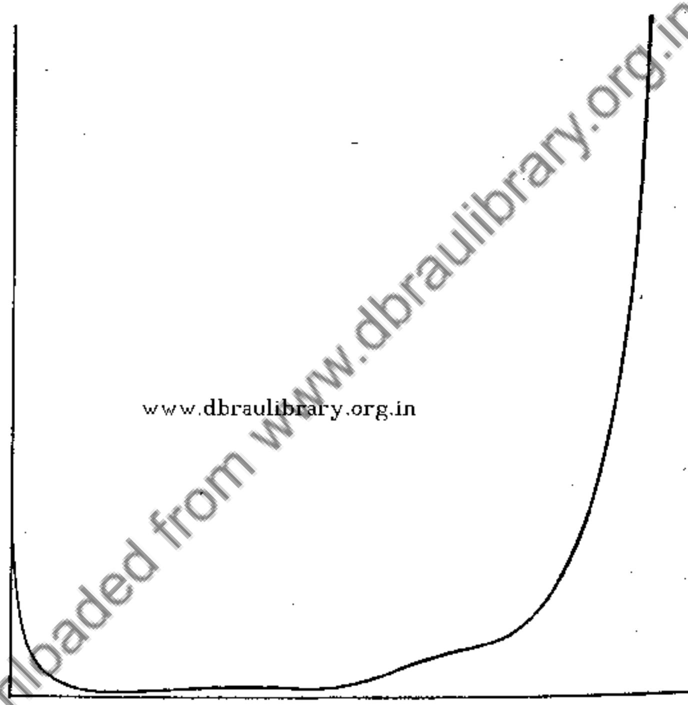

FIGURE 19.—Typical Curve of q_x

$\mu_x = A + Bc^x$, whence $l_x = ks^x g^{c^x}$, and m_x or colog $p_x = \alpha + \beta c^x$ ....(83)

which possesses the valuable property of "uniform seniority" for the actuarial computation of joint-life annuities (see p. 319; C; 18).

Later (H:*69*:191) *Makeham's Second Modification* was suggested, with the second differences of μ_x following a geometrical progression as

$\mu_x = A + Hx + Bc^x$, whence $l_x = ks^x w^{x^2} g^{c^x}$, and m_x or colog $p_x = \alpha + \gamma x + \beta c^x$ ....(84)

which leads to a less convenient method of "uniform seniority".

These formulae are, in fact, particular cases of an interesting generalized expression given by *Quiquet* (H:*72*, and H:*96*), namely,

$$\log l_x = A + Bx + \sum e^{r_i x} f_i(x) \qquad \text{....(85)}$$

where $f_i(x)$ is a polynomial, and the constants r_i are the roots of $A_0 + A_1 r + \ldots + A_n r^n = 0$. Several other formulae which have been suggested by various authors may also be derived from (85) when $n = 0$, 1, 2, or 3, as is indicated in H:*141*:85 (cf. also P:*102*:54).

From Gompertz's formula (82) we have

$$\log_e \mu_x = \log_e B + x \log_e c = a + bx \text{ where } a = \log_e B \text{ and } b = \log_e c,$$

so that the formula can be put into the exponential form $\mu_x = e^{a+bx}$. With a view to introducing greater elasticity, *H. L. Trachtenberg* (P:*144*) extended this into three improved expressions

$$\mu_x = e^{a_0 + a_1 x + a_2 x^2} \qquad \text{....(86)}$$

$$\mu_x = e^{a_0 + a_1 x + a_2 x^2}(1 + hx) \qquad \text{....(87)}$$

$$\text{or} \quad \mu_x = e^{a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4} \qquad \text{....(88)}$$

and pointed out that the last form provides for the two points of inflexion which are often found in the curve of $\log \mu_x$, since (88) arises from $\frac{d^2}{dx^2} \log \mu_x = a(x-b)(x-c)$ where the points of inflexion are at b and c (see also P:*102*:45 and 56).

Another method of generalizing Makeham's formula has been examined and illustrated by *J. Buchanan* (H:*153*) in the form

$$\text{colog } p_x = K + Lr^x \cos(x\theta' - \theta_0) \qquad \ldots.(89)$$

and by *G. J. Lidstone* (P:*85*:419) as

$$\mu_x = (A + Bc^x) + Mxc^x \qquad \ldots.(90)$$

and

$$\mu_x = A + (B \cos x\varphi + M \sin x\varphi)c^x \qquad \ldots.(91)$$

Noting that many attempts to employ Makeham's first expression (83) gave values too high at the older ages, and emphasizing the importance of being able to take account of the points of inflexion, *W. Perks* (P:*102*) has investigated the expressions

$$\frac{q_x}{1-q_x} = A + Bc^x \qquad \ldots.(92)$$

$$q_x \text{ or } \mu_x = \frac{A + Bc^x}{1 + Dc^x} \qquad \ldots.(93)$$

and

$$q_x \text{ or } m_x \text{ or } \mu_x = \frac{A + Bc^x}{Kc^{-x} + 1 + Dc^x} \qquad \ldots.(94)$$

It was pointed out by *G. F. Hardy* (P:*51*:68) that formula (77), when used for $\log l_x$ so that it means that

$$\mu_x = Bc^x + Mn^x \qquad \ldots.(95)$$

preserves a modified "uniform seniority" principle. With the addition of a constant, so that

$$\mu_x = A + Bc^x + Mn^x \qquad \ldots.(96)$$

(which is sometimes referred to as the "double geometric law"), the application of the modified seniority method is applicable as shown in P:*85*:413.

This "double geometric" curve is the Makeham expression

(83) with another geometric term added. The incorporation of still another such term leads to a "triple geometric" expression

$$\mu_x = (A + Bc^x) + (Mn^x + Rr^x) \qquad \ldots.(97)$$

which has been discussed and illustrated in H:*187*:539 and P:*94*.

Another formula which has attained some prominence in practical work as a means of representing mortality over the whole range of life is *Wittstein's* expression (H:*65*:164)

$$q_x = a^{-(M-x)^n} + \frac{1}{m} a^{-(mx)^n} \qquad \ldots.(98)$$

for which the reader may be referred also to H:*141*:82 and P: *59*:100.

(4) The Curves of e_x, $\mathring{e}_x$, and $\bar{a}_x$

The mathematical representation of the curtate or complete "expectation of life" (e_x or $\mathring{e}_x$) has received some attention—firstly, because the "expectation of life" has captured unjustifiably (see P:*164*:401, and P:*171*:281) the minds of many laymen, and secondly, because it may be claimed that in the calculation of that function a certain amount of graduation has been implicitly done (H:*110*:93). Being a gradually decreasing curve, it can often be reproduced closely (as suggested by *G. F. Hardy*, P:*51*:79) by

$$\log_{10} e_x = a + bx + cx^2 + dx^3 + fx^4 \qquad \ldots.(99)$$

Several other forms which have been used are noted in P:*165*:153.

Steffensen (H:*105* and H:*127*) has experimented also with its reciprocal (to produce an increasing series—see also P:*102*:40) in the Makeham form

$$\frac{1}{\mathring{e}_x} = A + Bc^x \qquad \ldots.(100)$$

(5) The Curves of $\frac{l_x}{l'_x}$, $\frac{T_x}{T'_x}$, $\log \frac{T_x}{T'_x}$, $\log \frac{p_x}{p'_x}$, etc., with Reference to a Standard Table

Instead of attempting the direct representation of functions such as l_x, or T_x (the population at age x and beyond), a useful device sometimes is to deal with the simplified curves which depict the ratios of such functions to their corresponding values, l'_x, or T'_x, in a standard table. Thus it was found by *H. G. W. Meikle* (P:*88* and P:*168*) that the census data for certain sections of India in 1921 could be represented by fitting 3rd degree parabolas to values of $\log \frac{T_x}{T'_x}$; the use of $\frac{l_x}{l'_x}$ and $\frac{T_x}{T'_x}$ is noted in P:*167*:101; and the advantages of $\log \frac{p_x}{p'_x}$ as a function of small values which progress slowly is indicated in P:*81*:213.

(6) The Curve of Sickness Rates

An early examination of the possibility of representing rates of sickness, s_x, by an analytical formula led Makeham (H:*47*) to suggest again the use of the function $A+Bc^x$, which Hardy (P:*51*:90) accordingly has noted might be applied in the form

$$\log (N-s_x)=A+Bc^x \qquad \ldots.(101)$$

where N is 52 or a value determined by trial. Discussions of the possibility of graduating sickness rates by Makeham's formula are also to be found in H:*83* and H:*177*.

(7) Curves for the Retirement and Depreciation of Physical Property

Though of course outside the realm of this study, it is interesting to note that a number of the methods outlined in the preceding sections of this chapter for representing mathematically the mortality of human beings have been applied (see H:*176* and H:*184*) also to the estimation of the rates of depreciation and retirement of physical property (such as telephone or telegraph poles, cables, coils, etc., railroad culverts, cross-ties, cars and locomotives, automobiles, electric power equipment, etc.).

(8) Curves of Population Growth

Another problem of some interest to the actuary, but often of more concern to the vital statistician and economist, is that of computing "intercensal" or "mean" populations between successive census enumerations, and of estimating—i.e., "predicting"—future populations for groups, localities, countries, etc. (The questions which arise in preserving consistency between such calculations for the constituent parts of a total and for the total itself are incidental to the main problem, and need not be discussed here; they are dealt with in P:*161*:219, P:*167*:63-67, P:*171*:283-289, and P:*105*:732.) As an alternative to the assumptions of arithmetical or geometrical progression, or of a parabolic trend (with an assigned order of differences constant)—which are useful particularly for the estimation of values lying between the known values of successive censuses (for which see the references just given)—much discussion has surrounded a curve-fitting method based on the *Verhulst-Pearl-Reed* (otherwise known as the *Logistic*) curve of population growth (see p. 169; A; 14). According to this theory the "law" of growth over time, t, of a self-contained population, P_t, undisturbed by migration should (see p. 238; B; 20) be capable of representation as

$$P_t = \frac{A + Be^{k(t-\tau)}}{1 + e^{k(t-\tau)}} \qquad \ldots . (102)$$

where τ denotes the abscissa of the point of inflexion of the curve (which follows the symmetrical shape of Figure A6 at p. 197; B; 3), A and B are the ordinates of the asymptotes, and k is a constant.

Pearl and Reed have also suggested (see P:*96*:575 and P:*105*) that the essential symmetry of formula (102) can be modified, in order particularly to make allowance for cyclical growth, by using the generalized form

$$P_t = d + \frac{k}{1 + me^{a_1t + a_2t^2 + a_3t^3}} \qquad \ldots . (103)$$

The "logistic" curve within recent years has been fitted, by several different methods, to a wide variety of populations (see p. 320; C; 19). It is undoubtedly capable of representing many series of known past populations with reasonable accuracy—particularly in the earlier stages of a population's growth, when the curve and the data often follow closely a geometrical progression (cf. P:*176*:7 and 46). The influences which determine the progression of a population, however, in reality are numerous, varied, and unstable; they are actually much more complex and unpredictable than are the simple principles on which the symmetrical logistic curve is based. The formula accordingly cannot have much claim to be accepted as a "law" of growth, and it is now generally conceded that its application to the problem of long-range prediction is surrounded inevitably by all the uncertainties of extrapolation (cf. P:*65*:110).

(9) Curves for Forecasting Mortality

A natural development of the preceding methods of extrapolating population data is their application to the forecasting of mortality rates. Since changes over time in the mortality rates of measurable populations are usually gradual, it has generally been found sufficient to exhibit the relationships between q_x^z and q_x^a (where z and a denote calendar years) on some simple basis like a parabola, a geometrical progression, a Makeham expression, or a "logistic" curve. Such applications therefore need not be detailed here, as they do not involve any new curve-fitting processes. The history of their evolution, however, is indicated at p. 169; A; 15, with references to the literature.

VIII. THE FITTING OF CURVES, AND GRADUATION

IN ORDER to appreciate the importance of the methods of mathematical statistics in connection with the problem of "fitting" curves and "graduating" data, it will be of assistance to remember that the latter questions are related essentially to that of "random sampling".

For suppose that we have a set of data $f'_1, f'_2, \ldots, f'_\nu$, such as the deaths $\theta'_1, \theta'_2, \ldots, \theta'_\nu$, at ages 1 to ν, which have been obtained by observation of a supposedly homogeneous group of men "exposed to risk". Even under the assumed conditions of homogeneity this series, having been secured by observation of a group necessarily limited in number, will of course contain irregularities attributable to that limitation. The observed series $f'_1, f'_2, \ldots, f'_\nu$, if it is to be viewed as unbiassed, must consequently have been brought into existence by some method of random selection operating upon a hypothetical "parent population". Any attempt either to "fit" a mathematical curve to such a series, or by some other process to remove the irregularities due to the paucity of the data and the method of selection, is therefore in effect an attempt to determine the hypothetical "parent population" of which the observed $f'_1, f'_2, \ldots, f'_\nu$ can be considered to be a random sample.

Now the ideal accomplishment would obviously be the determination of the "true" parent series, which may be called $f_1, f_2, \ldots, f_\nu$; and this determination must indeed be the real objective, in the sense here discussed (cf. p. 239; B; 21). If, however, a curve is "fitted" to the given series $f'_1, \ldots, f'_\nu$ in any such attempt to find the true "parent" series $f_1, \ldots, f_\nu$, it is obvious that the best we can do is to determine, on some criterion, a series of "fitted" values specified by the calculated parameters $\alpha, \beta, \gamma, \ldots$ of a fitted curve $y''_x = f''(x; \alpha, \beta, \gamma, \ldots)$ *of appropriate form*, by which the observed series will be represented, and from which the characteristics of the parent series may be estimated.

With such a curve of appropriate form the problem is consequently one of (*a*) determining the unknown parameters $\alpha, \beta, \gamma, \ldots$; (*b*) thence computing the "graduated" values; and (*c*) applying tests of "goodness of fit" to the whole curve, in order to decide, in conformity with the hypotheses, whether the results should be accepted.

It is thus important to realize that there are three concepts in the problem—the hypothetical "parent" series $f_1, \ldots, f_\nu$ which is unknown, the actual series $f'_1, \ldots, f'_\nu$ which has been "observed", and the "fitted" or "graduated" series $f''_1, \ldots, f''_\nu$, say, which is to be determined. In the ideal state—unlimited material, complete homogeneity, absolute randomness, and perfect fit—the three would be the same. Under the limitations of practice, however, it is necessary to formulate some process of fitting the graduated series $f''_1, \ldots, f''_\nu$, through determination of its parameters $\alpha, \beta, \gamma, \ldots$ in $y''_x = f''(x; \alpha, \beta, \gamma, \ldots)$, so that certain conditions arising from the theoretical requirements will be satisfied. The methods of establishing such processes of fitting will now be discussed.

The Principle of Maximum Likelihood

As explained on p. 39 (Chapter V), and at p. 239; B; 21, the specification of the "parent" series $f_1, \ldots, f_\nu$ from the observed data $f'_1, \ldots, f'_\nu$ can evidently be based logically upon the principle of **maximum likelihood,** by which the greatest possible value is assigned to the probability of the observed data having been drawn by chance (cf. p. 239; B; 21). When we come to the problem of determining a series of "fitted" values, $f''_1, \ldots, f''_\nu$, which may be considered to be the best possible representation of a series of observed data $f'_1, \ldots, f'_\nu$, it must therefore be remembered still that the characteristics of the parent series $f_1, \ldots, f_\nu$ are usually unknown; it is consequently necessary to assume that we are fitting a curve which is, in fact, of appropriate form, so that the fitted values (f''_r) will be estimates of the true parent values (f_r). Under these conditions, accordingly, we can view the deviations, $f''_r - f'_r$, between the fitted and observed series as being due to chance fluctuations arising from the limited size of the sample of f'_r's which has been drawn from the parent f_r's.

The Method of Least Squares

Since the deviations $f''_r - f'_r$ can thus be considered to have arisen from chance, they may be assumed to follow the Normal Curve (11), in which the parameter c will depend on the parent population f_r's from which the sample is drawn, and is accordingly independent of the observed f'_r's and the fitted f''_r's. It will therefore be clear that c will be a constant, either known or to be estimated, for any given process of selection by which the observed f'_r's are derived from the parent f_r's.

If now we assume, firstly, that c is the same for each of the observed f'_r's, i.e., that every one of the f'_r's has been determined by an equally good process of selection, then the probability of any particular deviation $f''_r - f'_r$ occurring, by (11), will be $\frac{1}{c\sqrt{\pi}} e^{-\frac{(f''_r - f'_r)^2}{c^2}}$, and the probability of the whole series of deviations $(f''_1 - f'_1), \ldots, (f''_\nu - f'_\nu)$ occurring together will be

$$\left(\frac{1}{c\sqrt{\pi}}\right)^\nu e^{-\frac{\sum\limits_{r=1} (f''_r - f'_r)^2}{c^2}} \qquad \ldots.(104)$$

In the more general case, when c is not the same for each of the observed f'_r's, i.e., if the f'_r's have not been equally well determined, suppose that the methods of selection have been such that c_r is the value appropriate to f'_r; then the probability of any particular deviation $f''_r - f'_r$, by (11), becomes $\frac{1}{c_r\sqrt{\pi}} e^{-\frac{(f''_r - f'_r)^2}{c_r^2}}$, and the probability of the whole series of deviations is

$$\frac{1}{(c_1 c_2 \ldots c_\nu)(\sqrt{\pi})^\nu} e^{-\sum\limits_{r=1}^{r=\nu}\left[\frac{(f''_r - f'_r)^2}{c_r^2}\right]} \qquad \ldots.(105)$$

Employing now, in connection with this general case (105), the same principle of "maximum likelihood" as before we see that, in order for this probability to be a maximum,

$$\sum_{r=1}^{r=\nu}\left[\frac{(f''_r-f'_r)^2}{c_r^2}\right] \text{must be a minimum} \qquad \ldots(106)$$

This is the simple fundamental condition of the **Method of Least Squares** (see p. 170; A; 16). The name arises, obviously, from the fact that the condition (106) requires that the least possible value be assigned to the sum of the squares of the deviations between the fitted and observed values, each divided by the square of the appropriate parent parameter c_r.

If, now, $\dfrac{1}{c_r^2}$ be called the **weight** appropriate to the observed f'_r, and if it be symbolized by W_r, we have

$$W_r=\left(\frac{1}{c_r}\right)^2 \qquad \ldots(107)$$

It also follows from (18) and (19), since c, η, σ, and λ are all proportionate, and because any constant factor can be omitted from the minimization of (106), that in practice we may take for W_r $\left(\text{in addition to } \dfrac{1}{c_r^2} \text{ or } \dfrac{1}{\eta_r^2}\right)$ the very usual forms (as they are often stated, and where λ denotes the "probable error")

$$W_r=\frac{1}{\sigma_r^2} \text{ or } \frac{1}{\lambda_r^2} \qquad \ldots(108)$$

The principle of Least Squares consequently may be formulated in the simple statement that

$$\sum_{r=1}^{r=\nu}\left[W_r(f''_r-f'_r)^2\right] \text{must be a minimum} \qquad \ldots(109)$$

The "Normal" Equations

When, therefore, the fitted values f''_r are to be those given by a curve $y''_x=f''(x;\alpha,\beta,\gamma,\ldots)$, in which the parameters $\alpha,\beta,\gamma,\ldots$ have to be determined, the condition (109) means that

$\Sigma[W_x\{f''(x;\ \alpha, \beta, \gamma, \ldots)-f'_x\}^2]$ must be minimized (where now the more usual variable x is written instead of r, and it will henceforth be understood that the Σ indicates summation over all the values of x which are included in the data). This evidently will be effected (cf. P:*51*:118, footnote) if we equate to zero the partial differential coefficients with respect to the unknowns $\alpha, \beta, \gamma, \ldots$, and solve the resulting equations, which then will clearly be the same in number as the unknowns.

The application of this very simple and logical principle may be illustrated first for the fitting of a general equation of the nth degree, $y''_x=\alpha+\beta x+\gamma x^2+\ldots$. We then have to minimize $\Sigma[W_x\{(\alpha+\beta x+\gamma x^2+\ldots)-f'_x\}^2]$. The partial differential coefficients with regard to $\alpha, \beta, \gamma, \ldots$ respectively, when equated to zero, give immediately (the common factor 2 being omitted)

$$\begin{aligned}\Sigma[W_x\ \{(\alpha+\beta x+\gamma x^2+\ldots)-f'_x\}]&=0\\ \Sigma[xW_x\{(\alpha+\beta x+\gamma x^2+\ldots)-f'_x\}]&=0\\ \Sigma[x^2W_x\{(\alpha+\beta x+\gamma x^2+\ldots)-f'_x\}]&=0\\ \text{etc.}\end{aligned} \qquad \ldots\ldots(110)$$

These Normal Equations, as they are called, are in this case simultaneous equations linear in $\alpha, \beta, \gamma, \ldots$, which therefore can be solved easily. It will be seen that they can be written down at once by a rule which is often stated in the following terms: "Set down the 'observation equation' $(\alpha+\beta x+\gamma x^2+\ldots)-f'_x=0$ for each value of x, noting its weight. Form the normal equation for the unknown α by multiplying each observation equation by the coefficient of α in that equation, and also by its weight, and adding the results; similarly form the normal equation for β by multiplying each observation equation by the coefficient of β in that equation, and also by its weight, and adding the results; and likewise form a normal equation for each of the other unknowns, γ, etc. The solution of these normal equations in the usual manner will give the 'best' values for the unknowns $\alpha, \beta, \gamma, \ldots$" (see also p. 322; C; 20).

This rule for the formation of the normal equations in respect of a parabolic function $y''_x=\alpha+\beta x+\gamma x^2+\ldots$ is very simple, and leads to normal equations which can be solved immediately since they are linear with respect to the unknowns $\alpha, \beta, \gamma, \ldots$. It has assumed much importance in the practical application of the method of least squares, because it has also been employed widely in a method of successive approximation for cases when the curve to be fitted is not parabolic and the resulting normal equations are not linear with respect to the unknowns. Under these circumstances the classical method is to find, firstly, approximate values, say, $\alpha', \beta', \ldots$, of the unknowns $\alpha, \beta, \ldots$, and then to suppose that $\alpha=\alpha'+\delta\alpha$, $\beta=\beta'+\delta\beta, \ldots$, where the corrections $\delta\alpha, \delta\beta, \ldots$ which are now to be determined may be assumed to be so small that their squares and higher powers may be neglected. Expansion by Taylor's Theorem thereupon immediately reduces the procedure to that already given, with the small corrections $\delta\alpha, \delta\beta, \ldots$ appearing as the unknowns in linear normal equations (see p. 241; B; 22).

This method of approximation could, of course, be repeated until a satisfactory fit is obtained if it were found that the approximate values $\alpha', \beta', \ldots$, with their first corrections $\delta\alpha, \delta\beta, \ldots$, did not provide sufficiently good results. The numerical work involved, however, is considerable even when the corrections are determined adequately at the first attempt. In many cases it is therefore preferable to use other devices, such as those stated for Makeham's formula and the "logistic" curve at p. 325; C; 21, where the practical application of the method of least squares is discussed.

The Weights

In the preceding statement of the principle of least squares it must be noted that the assignment of proper values to the weights, W_x, is of essential importance. Being defined as $\frac{1}{c_x^2}$, or $\frac{1}{\sigma_x^2}$, or $\frac{1}{\lambda_x^2}$, by (107) and (108), they will be taken in practice in accordance with the requirements of the mathematical model which is in fact being used.

Thus in a graduation of a series of observed rates of mortality, q_x', by fitting a mathematical expression to the values of q_x' by the method of least squares, we see, as explained in (iii) on p. 274; C; 7, that $\sigma^2\{q'_x\} = \dfrac{p_x q_x}{E'_x}$; writing therefore $W\{q'_x\}$, as in C; 7, to denote the weight of the observed q'_x, it follows from (108) that

$$W\{q'_x\} = \frac{E'_x}{p_x q_x} \qquad \ldots .(111)$$

Similarly, as in (iv) of C; 7, and since $m'_x \doteqdot \mu'_{x+\frac{1}{2}}$,

$$W\{\mu'_{x+\frac{1}{2}}\} \doteqdot W\{m'_x\} = \frac{E'_{x+\frac{1}{2}}}{m_x(1-m_x)} \doteqdot \frac{E'_x q_x}{(m_x)^2} \qquad \ldots .(112)$$

Also, by (vi) of C; 7,

$$W\{\text{colog}\, p'_x\} \doteqdot \frac{E'_x p_x}{q_x} \qquad \ldots .(113)$$

Instances of the practical use of these formulae are given at p. 326; C; 21.

If, on the other hand, the method of least squares were being applied to the fitting of a curve to a frequency distribution of actual deaths, θ'_x (instead of to a graduation of ratios such as $q'_x = \dfrac{\theta'_x}{E'_x}$, etc.), it will likewise be seen, from (ii) of C; 7, that

$$W\{\theta'_x\} = \frac{1}{E'_x p_x q_x} \qquad \ldots .(114)$$

Since $p_x \doteqdot 1$ at most age groups, this formula (as noted also at p. 293; C; 10) may sometimes be taken as

$$W\{\theta'_x\} \doteqdot \frac{1}{E'_x q_x} \qquad \ldots .(115)$$

In certain cases, such as the fitting of an exponential $y''_x = e^{f''(x)}$ to an observed series f'_x, it is obvious that the problem may be reducible to an easier form by taking logarithms and so fitting $\log_e y''_x = f''(x)$ to the series $\log_e f'_x$. There is one very important matter, however, which is frequently overlooked in the use of this device, but which must not be forgotten. For, from (33), with the notation of p. 272; C; 7, $\sigma^2\{\log_e f'_x\} \doteq \left(\frac{1}{f_x}\right)^2 \sigma^2\{f'_x\}$, whence by (108) and similar notation for the weights, $W\{\log_e f'_x\} \doteq (f_x)^2 W\{f'_x\}$; that is to say, the weight of $\log_e f'_x$ must be taken as $(f_x)^2$ times the weight of f'_x. Consequently, even if the weights in the fitting of $y''_x = e^{f''(x)}$ to f'_x can be taken as the same throughout and therefore all as unity and negligible, the weights to be used in the fitting of $\log_e y''_x = f''(x)$ to $\log_e f'_x$ will be $(f_x)^2$, and therefore, not being uniform, must not be neglected (cf. P:*28*:140, 144-5, and p. 326; C; 21 here).

In all these formulae for weights the undashed symbols indicate that the functions are, strictly, the "true" values of the parent population (as pointed out in C; 7—which conforms with the rationale of the fitting process explained at the commencement of this chapter, and with the consequent statement in connection with (107) that c_r is there the parameter of the appropriate parent population). For practical purposes these "true" values may be taken as the approximately adjusted values given by some simple method of preliminary graduation (cf. P:*51*:37 and P:*135*).

The proper introduction of these weights is of great theoretical and practical importance, for the whole foundation of the method of least squares rests—as has been shown—upon the condition (109) of which the weights form an essential part. They should not be assumed to be of the same value for the whole range of x (and therefore constant and negligible in the normal equations) until proper investigation has shown that assumption of uniformity to be justifiable.

In those fields where the curves to be fitted lead to normal equations which are not linear with respect to the unknowns, a great amount of criticism has been heaped upon the method of least squares because of the necessity of then using either the method of approximation outlined, or some other device which will bring the problem to a linear form. That the classical method of approximation may require heavy arithmetical work is undoubtedly true, though with present-day modes of computation this objection is not as serious as might appear. There would seem, however, to be very little justification for the extent to which the method of least squares has at times almost been shunned (see, for example, P:*33*:255, and, *per contra*, cf. P:*28*:42). As will be pointed out later, it is often a very satisfying method in comparison with the "method of moments", so long as an appropriate device is applied to deal with any constants which may be involved non-linearly in the normal equations. This study will therefore emphasize both the theoretical justification and the practicability of the method of least squares. Furthermore, it should be remembered, quite apart from its evolution from the theory of errors embodied in the Normal Curve, that the principle of minimizing the sum of the squares of the deviations (duly prepared by a system of weighting) must evidently be expected to produce very good results, since the squaring of the deviations gives the same influence to positive and negative departures of equal amount, and a large error has a greater effect than a small one (cf. P:*51*:119).

The Method of Moments

It has already been seen that the fundamental condition (109) of the method of least squares, namely that $\sum_{r=1}^{r=\nu}\left[W_r(f''_r - f'_r)^2\right]$ must be a minimum, will be satisfied when the partial differential coefficients with respect to the unknowns in f''_r are equated to zero. This means that

$$\sum_{r=1}^{r=\nu}\left[W_r(f''_r - f'_r)\frac{\partial f''_r}{\partial a_n}\right] = 0 \qquad \ldots.(116)$$

where $\dfrac{\partial f''_r}{\partial a_n}$ represents the partial differential coefficient with respect to the unknown a_n $(n=0, 1, \ldots)$ of the curve which is to be fitted, so that $a_0, a_1, \ldots$ are the several unknowns, and there is one "normal" equation thus derived for each of them.

The application of this method to the general parabolic function, $y''_x = \alpha + \beta x + \gamma x^2 + \ldots$, has also been shown to lead immediately to the set of normal equations (110), which, writing $a=a_0$, $\beta=a_1$, $\gamma=a_2$, . . . , may be put shortly as

$$\Sigma\left[x^n W_x\left\{(a_0+a_1x+a_2x^2+\ldots)-f'_x\right\}\right]=0 \quad \ldots.(117)$$

for $n=0, 1, 2, \ldots$. These equations are evidently based on the principle of **moments,** by which the observation equations, *duly weighted*, are simply multiplied by the successive powers of the variable, i.e., by 1, x, x^2, . . . (cf. p. 253; B; 27).

It is thus clear that, in the case of a parobolic function of the form $y''_x=a+\beta x+\gamma x^2+\ldots$, the method of least squares leads to exactly the same equations as the method of moments, so long as *both* sets of equations are properly weighted in accordance with the requirements of the method of least squares.

Reference has already been made to the fact that the normal equations of least squares are usually difficult of solution in cases other than that of the parabolic function just discussed. Moreover, if the weights, W_x, can be taken throughout as uniform, and therefore negligible, in the identical normal and moment equations (117) of that case, we reach a simplified set from which W_x has disappeared, namely,

$$\Sigma\left[x^n\left\{(a_0+a_1x+a_2x^2+\ldots)-f'_x\right\}\right]=0 \quad \ldots.(118)$$

for $n=0, 1, 2, \ldots$. This *unweighted* procedure would mean, in general, that

$$\Sigma\left[x^n(f''_x-f'_x)\right]=0 \quad \ldots.(119)$$

In this unweighted form it has become known as the **Method of Moments,** and is so used very widely as an easily applicable process for the fitting of frequency curves and other formulæ to observed data.

The relationships just shown between (*i*) the strictly weighted equations of least squares, (*ii*) the weighted equations of moments (identical with those of least squares in the case of a parabolic function), and (*iii*) the unweighted equations of moments, are highly important. They have been rather seriously overlooked, however, on some occasions, with resulting disparagement of the method of least squares, and corresponding implications that unweighted moments may be applied with assurance of success under almost any circumstances. It should therefore be noted that, in general, the strictly weighted equations (116) of least squares will be reproduced by the unweighted equations (119) of moments when, for $n=0, 1, 2, \ldots,$

$$W_x\left(\frac{\partial f''_x}{\partial a_n}\right)=x^n \qquad \ldots.(120)$$

This relation has been examined by Steffensen in a valuable paper (P:*137*:357), and in effect also by Hardy in P:*51*:129. It is there pointed out that if, as is generally possible, an exponential function $f''_x=e^{a_0+a_1x+a_2x^2+\ldots+a_{n-1}x^{n-1}}$ be used to represent approximately a frequency distribution of observed values f'_x, for which we can take $W_x \doteq \frac{1}{f''_x}$ (see p. 242; B; 23), then the weighted equations of least squares will lead to practically the same results as the unweighted equations of moments (loc. cit.). The meaning of this important conclusion is well expressed by Steffensen (P:*137*:358) in the following words: "Most frequency-curves can more or less approximately be represented by an expression of the form $f''_x=e^{a_0+a_1x+\ldots+a_{n-1}x^{n-1}}$. This is probably the true reason why the method of moments has proved such a powerful instrument for determining the constants of frequency-curves. But the same consideration is a warning against using that method indiscriminately, with or without weights, outside its natural scope. The cases where it is said that the *weighted* method of moments (yet with incorrect weights) has been successfully applied to curves (such as the force of mortality) which are *not* frequency-curves, will, on inspection, dissolve into cases where some artifice has been brought in (such as working directly on the exposures and deaths) by which the problem has been transformed into that of applying the *unweighted* method of moments to a *frequency-curve*. Hence the success."

These observations emphasize in an important manner the distinction which should be kept in mind between the weighted and unweighted applications of the method of moments which have appeared in practice. From the preceding discussion it is to be anticipated, evidently, that (1) the strictly weighted least square equations will give a "best" fit according to the particular (and very defensible) definition of "best" involved therein; (2) approximately weighted least square equations will lead to very good results (since the weights are in reality only relative values); (3) strictly weighted, or even approximately weighted, equations based on moments will produce values closely comparable to those of least squares for parabolic functions; (4) the weighted equations of least squares and the unweighted equations of moments may be expected to give fairly similar results only in certain cases such as those which are in reality the representation of a frequency distribution; but that (5) unweighted least squares should be applied only under circumstances which appear to be justifiable after due examination in accordance with the principles on which the use of weights is based; and (6) the unweighted equations of moments should not be accepted as being necessarily or universally applicable, and therefore should be employed with some caution.

A brief indication of the practical significance of these principles is given at p. 328; C; 22, in respect of certain applications of the method of moments in actuarial work. References to numerical examples of the fitting of curves by the method of moments may also be found at pp. 256-257; B; 27, and pp. 312-319; C; 17.

The Minimum-χ^2 Principle

The rationale of the method of least squares is based, in fact, on a mathematical model which contemplates (as explained in the first paragraph of Chapter VI) that each term of the series to be dealt with, namely, $f'_1, \ldots, f'_\nu$, is independent of every other term, so that each deviation, $f''_r - f'_r$, between the fitted and observed values (and each corresponding parent value, f_r) is to be treated separately as conforming with the concepts of the binomial case represented by the Normal Law of Deviations (10).

As shown in Chapter VI, however, this binomial model may be extended easily in order to deal with all the terms from f'_1 to f'_ν at once by means of the Multinomial Normal Law of Deviations (50). Under these circumstances it is evident from (52) that the probability for the observed series, $f'_1, \ldots, f'_\nu$, when the parent population is specified by the "true" values $p_1, \ldots, p_\nu$, in respect of a total of N, is a maximum when

$$\chi_0^2 = \sum_{r=1}^{r=\nu}\left[\frac{(f'_r - Np_r)^2}{Np_r}\right] = \sum_{r=1}^{r=\nu}\left[\frac{(f'_r - f_r)^2}{f_r}\right]$$

is a minimum. Similarly, if it be assumed (as in the previous discussions of the methods of maximum likelihood, and of least squares) that a curve is being fitted which is of appropriate form, so that the fitted values (f''_r) will be estimates of the true parent values (f_r), then the observed series $f'_1, \ldots, f'_\nu$ may be viewed as having been drawn as a sample from a fitted series $f''_1, \ldots, f''_\nu$, and the preceding condition for the maximum probability becomes the condition that

$$\sum_{r=1}^{r=\nu}\left[\frac{(f'_r - f''_r)^2}{f''_r}\right] \text{ must be a minimum} \qquad \ldots.(121)$$

The application of this method to graduations which arise in actuarial work has been discussed by Cramér and Wold in P:*23*:172.

The relation between this principle of minimizing χ^2, and the least squares condition (109), namely that $\sum_{r=1}^{r=\nu}[W_r(f''_r - f'_r)^2]$ must be a minimum, will be seen readily. For this least squares condition requires the use of the weights W_r; in the case of the fitting of a curve to a frequency distribution it is pointed out at p. 242; B; 23 that W_r may be taken roughly as $\frac{1}{f''_r}$; for a frequency distribution, therefore—being the case dealt with by the mathematical model of the Multinomial Law of Deviations—the method of least squares may be applied approximately by minimizing $\sum_{r=1}^{r=\nu}\left[\frac{(f''_r - f'_r)^2}{f''_r}\right]$; and this is precisely the minimum-χ^2 principle just stated. It may therefore be anticipated that the strict equations of least squares, and the minimum-χ^2 method, would produce closely comparable results in the case of a frequency distribution (see also p. 244; B; 24).

The process of minimizing (121), however, is usually difficult, for the unknown parameters in the fitted values f''_r are involved in the denominator as well as in the numerator, so that the differentiations become very complex (cf. p. 243; B; 23). Cramér and Wold, therefore, have suggested (P:*23*:173) that a close approximation will evidently be obtained if the unknown f''_r in the denominator be replaced by the observed f'_r—for the f'_r series is supposed to have been derived as a random sample from f''_r, and the method is based on the assumption that the f'' function to be fitted is of appropriate form. The principle then would become that

$$\sum_{r=1}^{r=\nu}\left[\frac{1}{f'_r}(f''_r - f'_r)^2\right] \text{ must be a minimum} \quad \ldots .(122)$$

As a practical method of approximation this may be expected to give results close to those of the method of least squares in the case of the fitting of a curve to a frequency distribution, since then $W_x \doteq \frac{1}{f''_x} \doteq \frac{1}{f'_x}$ as noted at p. 242; B; 23, and with these substitutions the strict least squares condition (109) becomes the approximate minimum-χ^2 condition (122). An application of the principle in this form is noted at p. 331; C; 23.

Other Methods of Fitting Curves

The methods of curve fitting which have been discussed in the preceding paragraphs—namely, the methods of least squares, moments, and minimum-χ^2—all make some use, although through slightly different formulations, of the *whole* of the available information provided by the series of observed data, since every value from f'_1 to f'_ν enters into the relations from which the unknown parameters are eventually determined. Several other processes which similarly employ the whole of the observed material have been suggested; but as they are all based upon principles less satisfying theoretically than those of least squares, moments, or minimum-χ^2, it will be sufficient only to note them, as of historical interest, at p. 171; A; 17.

Other much simpler methods, furthermore, have sometimes been employed, which base the equations for solution merely upon *isolated values* of the observed data, in order to use only as many equations as the number of unknown parameters. Evidently, however, they discard entirely all the information which might be derived from the data which are ignored, and thus cannot be expected to give anything beyond roughly approximate values of the parameters (cf. p. 173; A; 17).

Graphic methods which employ "semi-logarithmic" or "double-logarithmic" paper may also permit simple procedures in some cases—*semi-logarithmic paper* being ruled on the y-axis according to the logarithms of the numbers (the x-axis being spaced arithmetically), and *double-logarithmic paper* giving log x as well as log y (see P:*27*:224 and 237). Thus an exponential curve $y=be^{ax}$, or Laplace's First Law of Error $y=\frac{k}{2}e^{-k|x|}$ as at p. 159; A; 4, or a geometrical progression $y=ar^x$, or Gompertz's formula (82) when written $\mu_x=Bc^x$, are all of the form log $y=K+Ax$ and consequently appear as straight lines on semi-logarithmic paper ruled for log y and x. The parabola $y=ax^b$ when written $\log y=\log a+b\log x$, and being thus of the form $\log y=K+A\log x$, will similarly be represented by a straight line on double-logarithmic paper ruled for log y and log x. On these principles Gerhard (P:*46*) has given a very simple method for Makeham's formula (83) in the form colog $p_x=\alpha+\beta c^x$; for the second term is a geometrical progression which will be a straight line on semi-logarithmic paper, so that, when observed values of colog p_x are plotted as points on semi-logarithmic paper, the fitting will be given when a straight line (representing βc^x) can be located at a constant distance α below those points. (See also H:*183*).

Other graphical devices are noted at p. 173; A; 17.

IX. THE TESTS OF GOODNESS OF FIT

When statistical data, such as observed rates of mortality (q'_x) at various ages x, have been subjected to "graduation" with the objects and by the methods explained in Chapter VIII, it at once becomes essential to determine whether such graduated values give a proper representation of the material at hand.

As a practical device it may often be sufficient simply to examine the differences between the data and the graduated results, at individual ages, and in groups of ages, and for all ages combined, with due regard to (*i*) the frequency of changes of sign, and (*ii*) the standard deviations or "probable errors".

(*i*) The Frequency of Changes of Sign

Periodic Series. The changes of sign can be examined according to the principle that, if the distribution of + and − signs in N terms of a periodic series (in which the first and last terms are consecutive) has occurred merely by chance, then (*a*) the average number of isolated sequences of r signs alike will be $\frac{N}{2^{r+1}}$; (*b*) the number of signs which fall within groups of one or two like signs will be approximately equal to the number which fall within groups of more than two; and (*c*) the average number of sequences of all orders ($r=1, 2, 3, \ldots$) will tend to the limit $\frac{N}{2}$ (see p. 246; B; 25).

Non-Periodic Series. When a series, as in the case of rates of mortality, is not periodic, the preceding rules may still be applied so long as the first and last signs are treated as consecutive (so that they will be considered as belonging to the same group if they are alike).

If, however, the series is dealt with directly as being non-periodic (without the first and last signs being placed in the same group if they are alike), then the average number of isolated sequences of r signs alike becomes $\frac{N-r-1}{2^{r+1}}$ when the first and last signs are omitted (see p. 249; B; 25, and p. 332; C; 24), or $\frac{N-r+3}{2^{r+1}}$ if the first and last signs are included (see p. 250; B; 25).

(*ii*) Standard Deviations or Probable Errors

The comparison in the light of the standard deviations or probable errors is effected easily, on the assumption that the graduated values, q''_x, are estimates of the parent values, q_x, by setting out the actual deviations between the ungraduated and graduated values, and comparing them (so long as q or p is not so small that nq or np is less than about 10) with those expected according to the criteria summarized at the end of Chapter III. The actual deviations can then be regarded as being due only to accidental fluctuations if they are within 2 (or 3) times the standard deviations, or within 3 (or 4) times the probable errors; when the actual deviations are within these limits, that is to say, the graduated rates may be accepted as a permissible representation of the ungraduated values. Since, however, the $\pm 3\sigma$ or $\pm 4\lambda$ limits embrace over 99% of the chance deviations, and $\pm 2\sigma$ or $\pm 3\lambda$ include practically 95%, it will be clear that such comparisons are useful merely as a test of the hypothesis that the graduated rates constitute an admissible representation—they do not establish an admissible graduation as a good one, and certainly not as a "best" one. It is therefore necessary to narrow the limits so that, instead of examining the widest possible deviations, we may test, for example, the average deviations which would have arisen by chance alone. By formula (20), therefore, a graduation may be considered satisfactory (not merely permissible) if the actual deviations (irrespective of sign) are not greater than approximately $\frac{4}{5}\sigma$. Illustrations of these criteria are noted at p. 333; C; 24.

(*iii*) Comparisons of Financial Functions

Another practical and sometimes comprehensive test—of value to the actuary in view of his responsibility for the financial appropriateness of the results—is a comparison between the ungraduated and graduated values of a financial function of the basic rates, such as the annuity values $a_x\left(=\sum_{t=1}^{t=\omega} v^t\,{}_tp_x\right.$, where ${}_tp_x=p_xp_{x+1}\ldots p_{x+t-1}$, $p_x=1-q_x$, $v=\dfrac{1}{1+i}$, in which i is the rate of interest and ω is the final age$\Big)$.

The three preceding methods, however, practically useful and often sufficient though they are, clearly proceed little further than a mere comparison between the numerical observations as they are given, and the graduated values obtained therefrom. They afford criteria of the "admissibility" of the graduation, and of its "satisfactory" nature for practical purposes; but they do not reach the concept of a criterion for a "best" graduation, and they do not attempt any formal recognition of the fundamental objective of a graduation, namely, the determination of a hypothetical population from which the data may be supposed to have been drawn by chance. We shall therefore now consider the manner in which a test may be based on (*iv*) the criterion of fit on which the "best" graduation by least squares is founded, and (*v*) the χ^2 function as it emerges from the theory of determining the hypothetical parent population.

(*iv*) The "Least Squares" Criterion

When an observed series, f'_r, for values of r from 1 to ν, has been drawn from a parent series, f_r, the *true error* of each observed term is $f_r-f'_r$. If a graduation were now made, and if the graduated values, f''_r, were a perfect representation of the true values, f_r, then each true error, $f_r-f'_r$, would be precisely equal to the corresponding *residual*, $f''_r-f'_r=v_r$. The expected square of each true error (i.e., the mean square error) would thus be, *a priori*, σ_r^2, whereas the equivalent squared residual, *a posteriori*, would be v_r^2. The mean value of the ratio $\dfrac{v_r^2}{\sigma_r^2}$ would therefore be unity; and for the mean values over the whole series of ν terms we should consequently have $\sum_{r=1}^{r=\nu}\left(\frac{v_r^2}{\sigma_r^2}\right) = \sum_{r=1}^{r=\nu}(1)=\nu$. Since $\frac{1}{\sigma_r^2}=W_r$ and $v_r^2=(f''_r-f'_r)^2$, this means that in a "perfect" graduation we should have

$$\sum_{r=1}^{r=\nu}\left[W_r(f''_r-f'_r)^2\right]=\nu \qquad \ldots.(123)$$

In order to test the "goodness" of fit of a graduation which is not "perfect" we may accordingly say that the condition (123) should be satisfied approximately. It will be noted that the relation $\frac{v_r^2}{\sigma_r^2}=1$, or $W_r(f''_r-f'_r)^2=1$, expresses the fact that the mathematical expectation of the weighted squared residual is unity; and (123) states that the mathematical expectation of $\sum_{r=1}^{r=\nu}\left[W_r(f''_r-f'_r)^2\right]$, which is the function (109) to be minimized by the method of least squares, is equal to the number of terms, ν (see p. 252; B; 26 for this terminology, and cf. P:*23*:178).

In the case of a mortality table it has been shown at p. 244; B; 24 that for the least squares graduation of an observed rate of mortality, q'_x, the function to be minimized is $\sum\left[\frac{E'_x}{p_xq_x}(q''_x-q'_x)^2\right]$, which is $\sum\left[\frac{(\theta''_x-\theta'_x)^2}{E'_xp_xq_x}\right]$, and that the same expression is to be minimized in a least squares graduation of the observed deaths, θ'_x. If, therefore, a graduation of either q'_x or θ'_x has been made, a measure of the closeness of fit would be given by testing the agreement between $\sum\left[\frac{(\theta''_x-\theta'_x)^2}{E'_xp_xq_x}\right]$ and the number of ages, ν. Since the parent values p_x and q_x are usually not known, the computation might be made in practice by using the observed p'_x and q'_x as estimates of p_x and q_x (as in H:*48*:336), or values obtained from an approximate preliminary graduation (see p. 96 here), or the finally graduated p''_x and q''_x themselves (as in H:*48*:323-325 and P:*23*:178). This test of a mortality table graduation was discussed and used extensively by *De Forest* as long ago as 1873 (see p. 175; A; 18).

The above process is founded on a comparison of the mean square errors over the whole of the observed and graduated series, and in dealing with the graduated series every term enters into the calculations. When, however, the graduation is performed by the least squares fitting of a curve in which there are, say, k unknowns, it may be argued (as in the analogous case of Bessel's correction (42) for one unknown and uniform weights) that in effect k of the terms are fixed, so that only $\nu-k$ of them are free. Under these circumstances, as is shown by (*ii*) at p. 252; B; 26, the mean square error of an observation of unit weight is $\dfrac{\sum\limits_{r=1}^{r=\nu}\left[W_r(f''_r-f'_r)^2\right]}{\nu-k}$, so that then (123) becomes modified to

$$\sum_{r=1}^{r=\nu}\left[W_r(f''_r-f'_r)^2\right]=\nu-k \qquad \ldots.(124)$$

This principle has been used since the time of Gauss (see p. 175; A; 18). While it recognizes, in effect, the concept of "degrees of freedom", it should be noted that, like (42), its derivation is based on the Principle of Insufficient Reason, or alternatively, upon other assumptions which are open to some argument (see p. 252; B; 26).

In the case of a least squares graduation of q'_x or θ'_x this relation for goodness of fit means that

$$\sum\left[\frac{(\theta''_x-\theta'_x)^2}{E'_x\,p_x\,q_x}\right]=\nu-k \qquad \ldots.(125)$$

where again in practice p_x and q_x would be taken as p'_x and q'_x, or as approximately graduated values, or as p''_x and q''_x. This test was first employed by *Thiele* in 1871 (see p. 175; A; 18), with p''_x and q''_x being used for p_x and q_x, so that Thiele's form was

$$\sum\left[\frac{(\theta''_x-\theta'_x)^2}{E'_x\,p''_x\,q''_x}\right]=\nu-k \qquad \ldots.(126)$$

(v) The χ^2 Test

It has already been explained in Chapter VI that an observed series $f'_1, f'_2, \ldots, f'_\nu$, totalling N, may be regarded as having fallen into the ν "cells" through the operation of the true "parent" probabilities $p_1, p_2, \ldots, p_\nu$, in accordance with the Multinomial Normal Law (52), that is, $\frac{1}{(\sqrt{2\pi N})^{\nu-1}\sqrt{p_1 p_2 \ldots p_\nu}} e^{-\frac{\chi_0^2}{2}}$ where $\chi_0^2 = \sum_{r=1}^{r=\nu} \frac{(f'_r - Np_r)^2}{Np_r}$; and it was also pointed out that there are under those circumstances $\nu-1$ "degrees of freedom", since one linear "constraint" upon the possible values is imposed by the condition that $f'_1 + f'_2 + \ldots + f'_\nu = N$. It was further shown that the probability of getting a series with $\nu-1$ of its values exhibiting deviations simultaneously lying between $\alpha_r\sqrt{n}$ and $\beta_r\sqrt{n}$ (where $r = 1, 2, \ldots, \nu-1$), with the νth fixed by the "constraint", could be expressed as the multiple integral (54), and that both formulae are merely extensions of the corresponding (10) and (11) for the binomial case when $N = n$, $\nu = 2$, $p_1 = p$, and $p_2 = q$. Since the process by which the multinomial expressions (52) and (54) were reached is thus a straightforward generalization of the binomial formulation, it may be of assistance again here to return to the binomial for the start of the development in order to establish clearly the basic principles of the χ^2 **Test for Goodness of Fit.**

It may therefore be recalled that in the binomial case it was shown (see p. 56) that when the deviation is any quantity x, then $\chi^2 = \frac{x^2}{npq}$. Suppose, therefore, that a particular observation has shown a deviation, ϵ, for which χ^2 has the particular value $\chi_0^2 = \frac{\epsilon^2}{npq}$. This may be considered as an improbable (i.e., "poor") observation if in a large proportion of other similar observations the deviations arising from chance alone would be *smaller than* ϵ, so that χ^2 would be *less than* χ_0^2. But by (10), as shown on p. 162; A; 5, the probability of a deviation less than ϵ in absolute magnitude, i.e., lying between $-\epsilon$ and $+\epsilon$, is $\frac{1}{\sqrt{2\pi npq}} \int_{-\epsilon}^{+\epsilon} e^{-\frac{x^2}{2npq}}\,dx.$

The region of integration here is for values of $x \leqq \epsilon$, which means values of $\frac{x^2}{npq} \leqq \frac{\epsilon^2}{npq}$, that is, values of $\frac{x^2}{npq} \leqq \chi_0^2$, being values of $\chi^2 \leqq \chi_0^2$.

Now it has been seen in (53) that this integral may be written $\frac{1}{\sqrt{2\pi}\sqrt{p_1p_2}}\int e^{-\frac{1}{2}\left(\frac{t_1^2}{p_1}+\frac{t_2^2}{p_2}\right)}dt_1$, where the limits may still be defined as the region for which $\chi^2 \leqq \chi_0^2$. Alternatively, by putting $\frac{x^2}{npq} = v^2$, say, that same integral becomes $\frac{1}{\sqrt{2\pi}}\int e^{-\frac{v^2}{2}}dv$, where the integration is again over the same region, which is here that for which $v^2 \leqq \chi_0^2$.

It therefore follows (P:*146*:326) that in the multinomial case the multiple integral (54), which corresponds to the binomial (53), can be written

$$\frac{1}{(\sqrt{2\pi})^{\nu-1}}\int\int\ldots\int e^{-\frac{1}{2}\left(v_1^2+v_2^2+\ldots\ldots+v_{\nu-1}^2\right)}dv_1dv_2\ldots dv_{\nu-1}$$

where the domain of integration extends over all the values for which $v_1^2+v_2^2+\ldots+v_{\nu-1}^2 \leqq \chi_0^2$; and this will give the probability of getting, by chance alone, a set of deviations which will produce a value of χ^2 equal to or less than the χ_0^2 actually observed.

The reduction of this multiple integral is heavy but not difficult. The proof is too long for inclusion here; it is given excellently, however, in P:*146*:331 and P:*21*:II, 302, and leads to the simple integral

$$\frac{1}{(\sqrt{2\pi})^{\nu-1}}\left[\frac{\pi^{\frac{\nu-1}{2}}}{\Gamma\left(\frac{\nu-1}{2}\right)}\int_0^{\chi_0^2}e^{-\frac{u}{2}}u^{\frac{\nu-3}{2}}du\right]$$

in which the "Gamma Function" is defined as stated on p. 259; B; 29. Putting $u=z^2$, this becomes

$$\frac{1}{2^{\frac{\nu-3}{2}}\Gamma\left(\frac{\nu-1}{2}\right)}\int_0^{\chi_0}e^{-\frac{z^2}{2}}z^{\nu-2}dz.$$

As already noted (and explained in Chapter VI) the ν variables here are subject to one "linear constraint" (imposed by the condition that $\Sigma f'_r = N$), so that there are $\nu - 1$ "degrees of freedom". Writing $\nu - 1 = d$ to represent the degrees of freedom, the expression therefore is

$$\frac{1}{2^{\frac{d-2}{2}}\Gamma\left(\frac{d}{2}\right)}\int_0^{\chi_0} e^{-\frac{z^2}{2}} z^{d-1}\,dz.$$

This formula, however, measures the probability of getting by chance alone a value of χ^2 as small as, or smaller than, the χ_0^2 actually given by the observations, and is thus based on an appraisal of the poorness of the result. The probability of the opposite inequality, $\chi^2 \geqq \chi_0^2$, will accordingly measure the "goodness of fit". Identifying that measure by P we therefore have

$$P = \frac{1}{2^{\frac{d-2}{2}}\Gamma\left(\frac{d}{2}\right)}\int_{\chi_0}^{\infty} e^{-\frac{z^2}{2}} z^{d-1}\,dz \qquad \ldots.(127)$$

where d is the number of degrees of freedom.

In the preceding deduction there are $\nu - 1$ degrees of freedom, being the number of the variables, ν, less one constraint (imposed by the requirement that the total frequencies falling into the ν cells shall be N). Similarly, if there are, in any particular problem, further conditions which restrict the values which can be assigned at will (as, for example, an additional requirement that the mean, as well as the total, of the frequencies shall be equal in the observed and the parent series), the degrees of freedom will be reduced by one for each such condition. The degrees of freedom, d, that is to say, will be taken as $\nu - k$, where ν is the number of cells and k the number of constraints.

In order to apply this formula it is therefore necessary to compute χ_0^2 from the data, to settle the number of degrees of freedom d, to determine the numerical value of P—which may be done easily from tables which have been prepared—and then to interpret the result. The history of this χ^2 Test (or "Chi-

Squared Test"), as it is usually called, and the available tables of P, are recorded at p. 175; A; 19.

The following precautions are imposed, by the preceding theory, upon the calculation and interpretation of P: (*a*) As already noted in the derivation of the Multinomial Normal Law of Deviations (50), the total N must be fairly large (probably at least 50), and no value of Np_r should be less than about 10—the latter restriction being surmountable in practice by amalgamating the values of one or more adjacent cells, where necessary, in order to produce a frequency not less than 10; (*b*) Since P gives the probability, under random sampling, of getting a value of χ^2 equal to or greater than the observed χ_0^2, it is to be concluded, when P is small, that the observed χ_0^2 may have arisen from significant causes other than the merely chance variations of random sampling; (*c*) When P is large, however, the inference cannot be drawn that the observed χ_0^2 has arisen from random sampling alone—the proper inference is the negative one only, namely, that the existence of significant causes has not been proved; (*d*) Moreover, if P is found to be very close to unity, the result must be viewed with suspicion, because so large a value of P will ordinarily have arisen from a value of χ_0^2 so small as to raise doubts concerning the sampling technique employed.

Now it is to be remembered that a value $P = .05$, for example, means that only in 5% of trials should we obtain a value of χ^2 as large as or larger than the observed χ_0^2. Any smaller value of P, accordingly, indicates an even smaller percentage of trials. It is, of course, a matter of individual preference to select a value of P at or below which all values would be considered small, in the sense that the investigator would view them as undoubtedly significant. The values $P = .05$, $P = .01$, and $P = .001$ thus chosen are often said to define the 5%, 1%, and .1% **levels of significance**, and smaller values than those designated are spoken of as lying below the particular level used. While judgment must naturally enter into the decision to be made in any particular case, experience with the practical use of the χ^2 test has led to the following suggestions which have been widely adopted: (*i*) If P is found to lie between .1 and .9, there is no reason to suppose that significant causes have produced the value χ_0^2 observed; (*ii*) If, however, P is less than .05, there is good reason to conclude that a real discrepancy exists between the observed and theoretical values, and if P is below .02 such an indication is strong; (*iii*) In testing the "fit" of a curve (i.e., in comparing observed and theoretical frequencies) it will generally be found that, when the data are very large, P will be small even though the fit appears to be very good—a feature which may be due to the inability of the χ^2 test to distinguish between heterogeneity and merely accidental variations, and to the fact that the basic assumptions of the theory are revealed, when the data become very large, as being not wholly appropriate (see P:*32*:204 and P:*10*:526, and p. 342; C; 25); (*iv*) Values of P above .95 should be viewed with suspicion; and (*v*) When d exceeds 30, P can be found with sufficient accuracy (as suggested by R. A. Fisher, P:*43*; see also p. 176; A; 19) by computing $\sqrt{2\chi_0^2}$, and thence (from tables of the "probability integral" with unit standard deviation—see p. 176; A; 5) the area of the Normal Curve beyond the ordinate which is $\sqrt{2\chi_0^2}-\sqrt{2d-1}$ units from the mean—the value of P so determined then being interpreted as above; or, more simply, by computing $\sqrt{2\chi_0^2}-\sqrt{2d-1}$, from which it may be inferred, when this quantity is considerably greater than 2, that the observed χ_0^2 differs significantly from that expected.

From the basis of the χ^2 test as a method of examining goodness of fit, and the foregoing suggestions concerning its practical application, it will be realized that the process often becomes, in effect, an attempt to compress the results of a wide variety of controlling influences and resultant deviations within the compass of a single figure, P. The precise inferences to be drawn therefrom are consequently often obscure; the discovery of a value of P which does not lie between about .1 and .9 may indicate the need of caution—but while it will then prompt a search for causes of disturbance, it generally can indicate little of the nature of those causes. To that extent the method, at least in many instances with which actuaries are concerned, suffers from the disabilities of any "index number". For this reason the calculation of χ_0^2, and its resulting P, as a test of goodness of fit, will often—especially in work with large blocks of mortality data where, as already pointed out, P may be small even though the fit is good—remain somewhat of a theoretical criterion, and in the hands of a practical actuary will not replace an analysis of the deviations group by group. In fact, the analysis by groups will have to be performed whenever P is small or large, and even when P lies between .1 and .9 an actuary would generally not be content to forego such a enquiry.

Notwithstanding these limitations of the χ^2 method as a test of goodness of fit of mortality table graduations, it should be emphasized again, however, that the χ^2 function occupies a fundamentally important place in the philosophical and mathematical formulation of the theory of sampling, and should therefore be clearly understood by every student.

X. RECENT RESEARCHES, AND MISCELLANEOUS PROBLEMS

In this chapter we shall indicate very briefly—mainly, indeed, by a series of references only—a number of matters which, although of interest to actuaries and vital statisticians, are not of sufficient practical utility in connection with this study to require more extended treatment.

(1) "Confidence" or "Fiducial" Limits

In the problem of statistical estimation an important question concerns the limits, or "interval", or "belt", within which an estimate of an unknown parameter may be expected to lie.

For the simplest case of the point binomial, for example, an estimate of the parent probability p may be taken as $\frac{s}{n}$ (cf. p. 291; C; 10) where, however, the confidence to be reposed in $\frac{s}{n} = p'$ as an estimate of p will, on the principles of Bernoulli's theorem, evidently increase as n increases. With p thus taken as p', and when n and p' are not too small, so that the assumptions of the Normal Curve are permissible, the relations deduced at the end of Chapter III indicate that p will lie between $p' - 2\sigma\{p'\}$ and $p' + 2\sigma\{p'\}$ in approximately 95% of cases, between $p' - 3\lambda\{p'\}$ and $p' + 3\lambda\{p'\}$ in about 96%, and between $p' - 3\sigma\{p'\}$ and $p' + 3\sigma\{p'\}$ in over 99%. The values 95%, etc., are spoken of in modern terminology as **confidence coefficients.** These conclusions are employed, for instance, in the method noted at p. 174; A; 17 for setting up vertical bars to mark the confidence interval in a graphic graduation. The limits for values of n up to 1000, and for confidence coefficients of .95 and .99, are shown in a very convenient diagram form in the point binomial case by Clopper and E. S. Pearson in P:*18*:410-411.

Indications of the use of these principles are given at p. 343; C; 26.

When p (or q) is less than about .03 and np is less than about 10, the use of the Poisson distribution is indicated (see p. 60), and the confidence limits may consequently then be taken with advantage from tables or diagrams based on Poisson's exponential (55). This modification has been discussed in P:*45*, and again clearly by Ricker in P:*108*. An example from the latter paper is given at p. 344; C; 26.

In the case of small samples (say $n<30$) the problem of confidence limits has also received increasing attention in recent years through the development of "Student's distribution" (44)—see, for examples, P:*112*:89-91.

(2) The Theory of Estimation, and the Testing of Hypotheses

The nomenclature, theoretical formulations, and practical inferences involved in modern statements of the theory of estimation and the testing of hypotheses have been examined and re-examined in a great number of recent papers. For the purposes of this study it is not necessary to attempt even any classification of these contributions. The philosophical intricacies of the discussions have led to abstractions which have proved difficult to grasp, and misunderstandings which have produced criticisms and sharp controversies. Some idea of these disagreements may be gathered from the long but extremely interesting interchanges in P:*42*, P:*91*, and P:*92*, where Bowley, Isserlis, Jeffreys, R. A. Fisher, Neyman, E. S. Pearson, Greenwood, and others have explored the foundations of various logical approaches. A connected and very comprehensive mathematical presentation is that of Wilks in P:*156*.

(3) Orthogonal Polynomials

In the fitting by least squares of the parabolic expression $y''_x=\alpha+\beta x+\gamma x^2+\ldots$ for which the "normal equations" have been discussed in Chapter VIII, it is sometimes desirable to ascertain whether the use of an additional term will effect an improvement in the fit. In this connection it is to be remembered that the values of α and β determined by least squares for the straight line will not be the same as those for a similar fitting of a second-degree curve which requires γ as well as α and β; the problem of passing to a curve of higher degree therefore is not that of simply appending a term to others which have been found before (see, for example, P:*177*:321-4). By using the **orthogonal polynomials** of the famous Russian mathematician Tchebycheff, however, it is possible to follow a systematic procedure which utilizes all previous work when the parabolic expression to be fitted is extended term by term.

Recent investigations by A. C. Aitken have contributed materially to the development of this method. Actuarial students may be referred conveniently to P:*84* and P:*3*:115-120 for the theory of these polynomials, and to the same sources and also P:*65*:92 for their practical applications.

(4) Regression and Correlation

The elementary text-books providing clear descriptions and worked examples of linear and non-linear, multiple, and partial "regression", and of linear and non-linear, multiple, and partial "correlation", are now so numerous and accessible that we shall here simply refer the student to several of the most recent of such publications. For elementary discussions, P. R. Rider's volume (P:*112*:27 et seq.), W. D. Baten's (P:*8*:145-200 and 223-232), and C. H. Goulden's (P:*48*:52-87 and 219-246), are excellent; more mathematical approaches are given by B.H. Camp (P:*16*:129-179 and 286-347), Elderton (P:*32*:141-180 and 210-230), and H. L. Rietz (P:*116*:77-113); and a detailed treatment is provided by Yule and Kendall (P:*177*:196-308).

The numerical illustrations of the fitting of linear and curved (polynomial) regression lines by the method of least squares (unweighted) which are shown in these texts will be of interest also in connection with Chapter VIII (and cf. p. 324; C; 21). R. A. Fisher's convenient summation method of fitting so that at any stage a further term may be added to a fitting already made without disturbing the previous calculations is described and illustrated in P:*43*:148-176; P:*48*:234; P:*143*:156; and P:*131*:324.

It may be useful also to note here that the least squares fitting of polynomials, which usually becomes laborious when the data comprise a large number of terms and the polynomial is of degree above the 3rd, has been facilitated by the publication by H. T. Davis of tables of the required coefficients as far as 51 equidistant terms for degrees up to the 7th (P:*25*). Another valuable contribution describing a method which is even simpler by reason of the small figures required in the solution (cf. P:*28*:117) is given by Birge and Shea in P:*11*.

Certain applications of correlation theory to actuarial problems are noted at p. 344; C; **27**.

(5) The Analysis of Variance

The name "analysis of variance" (cf. p. 163; A; 6) is a term now used to indicate a process—arising from the Lexis theory (pp. 30-33 here, and cf. P:*3*:54), based in fact upon the "correlation ratio" introduced by Karl Pearson in 1905, and since developed widely by R. A. Fisher—for dividing the total sum of the squared deviations of a variate from its sample mean into those distinct sums of squares, corresponding to supposed or real causes of variation, which give estimates of the variance from each such cause. The method is employed principally in connection with designs in biological and agricultural experiments, and is noted here because the student will encounter the name frequently in the perusal of modern texts on the general application of mathematical statistics. The subject has been covered in the first instance by Fisher in P:*43*:216-306; the mathematical basis is described clearly in P:*177*:444-448 and P:*3*:54 and 136-142; and additional explanations and numerical examples are given conveniently in P:*112*:117 et seq., P:*131*:179 et seq., and P:*48*:114 et seq.

XI. AN OUTLINE OF A COURSE IN GRADUATION

THE main objective of the preceding chapters has been to assemble those portions of the theory and applications of Mathematical Statistics which are required by actuaries in their studies and their daily practice. Amongst the various applications which have been indicated, the theory and practice of "graduation"—the fundamental concept of which is discussed in Chapter VIII—represents one of the most important subjects, as will be evident from the discussions of curve forms and fitting methods in Chapters VII, VIII, and IX. Other modes of graduation, however, are available, and are widely used; and all of them involve, in varying degrees, the principles of mathematical statistics considered in this volume.

In the Preface and the Introduction it was pointed out that one of the chief difficulties encountered in the teaching of actuarial mathematics has always arisen from the hiatus which has existed between the elementary studies of "probability" and the student's subsequent encounters with the advanced methods necessary for a proper understanding of graduation processes. The aim of the preceding chapters has been to bridge that gap, so that the student may be enabled to understand the mathematical concepts of the various graduation methods with greater ease. Co-ordination of the reading from many sources will still be essential, however—for although the underlying theories of mathematical statistics are to be found here, the expositions of several important graduation devices must yet be sought elsewhere.

This final chapter will consequently indicate a course of reading, bringing the appropriate sections of this volume into relation with the available discussions of particular graduation methods. An outline only will be given, with the essential references, but without detailed explanations. By this means it is hoped that the student will be able to plan his reading systematically, and with so satisfactory an understanding of the necessary fundamentals that he will then experience no difficulty in grasping the theories and the practices of graduation as he proceeds.

(I) The Nature and Objects of Graduation

1. The basic concept of the graduation of a statistical series is explained at the beginning of Chapter VIII as, in effect, the determination of the "true" hypothetical "parent population", $f_1, f_2, \ldots, f_\nu$, of which the "observed" $f'_1, f'_2, \ldots, f'_\nu$ can be considered to be a random sample. In practice, however, it is usually possible to find only the "fitted" or "graduated" values, $f''_1, f''_2, \ldots, f''_\nu$ as "estimates" of the true values $f_1, f_2, \ldots, f_\nu$, by some process which will satisfy some predetermined criterion of a "satisfactory" or "best" graduation.

2. When the graduation is accomplished by fitting a mathematical formula (by any of the methods discussed in Chapter VIII) it will be clear that the main criterion must be a test of goodness of *fit*, since the values derived from the graduation process will lie on a mathematical curve, and therefore will be inherently "smooth". If, however, some other method of graduation is employed which does not necessarily place the graduated values upon an inherently smooth curve, it will evidently be necessary to test the results for *smoothness* as well as for goodness of fit. In this connection it should be noted also, at this stage, that in such a graduation of irregular data it will obviously not be practicable to secure a best possible fit and greatest possible smoothness at the same time—for the ultimate interpretation of a "best possible fit" would require the precise reproduction of the original data, without any smoothness having been attained. It is therefore necessary in such cases to settle the criteria for fit and smoothness so that the practical results may be *satisfactory*, rather than *best*, in both respects.

(II) The Tests of Fit

1. After a graduation has been performed, the *admissibility*, i.e., the "satisfactory" character, of the results secured by the graduation process may be examined by the following tests:

(*i*) The frequency of changes of sign—see Chapter IX, section (*i*).

(*ii*) The standard deviations (or probable errors)—see Chapter IX, section (*ii*).

(*iii*) Comparisons of financial functions—see Chapter IX, section (*iii*).

2. The *goodness of fit* may be tested by an examination of the extent to which the results satisfy the "least squares" criterion —see Chapter IX, section (*iv*).

3. The *hypothesis* that the observed values may be supposed to have been drawn, by chance alone, as a random sample from the graduated series (so that the graduated values may be acceptable as a representation of the "true" values of the parent population) may be investigated by the χ^2 test—see Chapter IX, section (*v*).

(III) The Tests of Smoothness

1. As already stated in (I), an enquiry into the smoothness of the graduated values is necessary only when the graduation has been performed by some method other than the fitting of an inherently smooth mathematical formula. In the graduation of most actuarial data it is sufficient (and customary) to assume that differences beyond the 3rd in the true values may be neglected (although in some cases this assumption should be examined carefully—P:*166*:81, 107, and 110). On this basis the smallness of the 3rd differences of the graduated values would constitute a practical indication that the differences of higher orders would be almost negligible, and the sum of the squares of the 3rd differences would afford a comparative measure of the smoothness of the results (cf. P:*59*:8).

(IV) Graduation by the Fitting of a Mathematical Formula

Because a mathematical formula is inherently smooth, it is natural to place the single problem of fitting such a formula as the first method of graduation to be listed here—particularly since the principles have been largely discussed in the earlier chapters of this volume.

1. In actuarial work the selection of the form of curve which is likely to be appropriate may be made from a consideration of the underlying hypotheses and types set out in Chapter VII.

2. The methods of fitting the selected form are described in Chapter VIII.

3. Special devices for the graduation by Makeham's formula (83) of the rates of mortality during the select period are noted in section (IX) hereafter.

4. Particularly in the compilation of annuitants' experiences where the annuity values are a most important function, or sometimes if it seems advisable in any experience to merge the select into the ultimate values at a duration earlier than that strictly indicated by the data, it may be desirable to adopt a special method of graduation which will reproduce the annuity values (rather than the basic rates such as q_x) as closely as possible. G. F. Hardy devised such a process, on the assumption of Makeham's formula (83), for the British Offices' Annuitants' Experience. [The employment of the method in that experience, however, was occasioned not so much by the use of an arbitrarily short period of selection as by the basing of the ultimate table upon "aggregate" data, from which duplicates had been eliminated in a manner different from that adopted for the "select" data. It has been pointed out in H:*106*:289 and H:*98*:361 that if the ultimate table had been founded on "select" data (as was done with the $O^{[M]}$ table where the method here under discussion was not used), this special method would not have been required even though the select period would still (on the evidence of the material) have been arbitrarily short.] His own brief description in H:*90*:127-131, and that given in P:*59*:97-99, may be elucidated by the detailed explanations in P:*77*, while the proofs of the formulae required are amplified in P:*162*:548-551. [The references in this last discussion to pars. 88-91 of Henderson's

Actuarial Study No. 4 (P:*59*) are to the first edition—the corresponding paragraphs of the second edition being numbered 7.91. The symbol $(IA)_x$ in the top line of p. 551 should be $(I\bar{A})_x$].

(V) Graduation by the Graphic Method

Instead of selecting and then fitting a mathematical formula as in section (IV), it is evident that a simpler alternative in many cases would be the drawing of a smooth freehand curve amongst the points which represent the data.

1. The elementary principles to be employed in a direct graphic graduation are discussed in P:*59*:11-16 and P:*167*:98. The limits within which the graduated curve should lie may be indicated as at p. 174; A; 17. The use of splines is described in P:*152*.

The graphic method obviously can be applied in a manner to secure whatever compromise between fit and smoothness the graduator may desire. While this flexibility may be a disadvantage (in that there is no *a priori* criterion of the degree of compromise to be secured), it is nevertheless a useful characteristic under some conditions—such as those explained in par. 10 of section (VI) here.

2. As it is sometimes difficult to apply the graphic method directly in the graduation of mortality and similar material, a useful device is to adopt an approximate graduation by a mathematical formula, or the values from some standard table, as a basis, and hence to graduate graphically only the ratios or differences between the data and the selected base. An illustration of this device is given in P:*59*:17-19. The student should refer also to P:*81*, and to Chapter VII, section (5), p. 85 here.

3. A graphic method of determining the constants in Makeham's formula (83) in the form colog $p_x = \alpha + \beta c^x$, by the use of "semi-logarithmic" paper, is noted at p. 103 of Chapter VIII.

4. Methods of handling the graphic graduation of select rates of mortality are set out in section (IX) hereafter.

(VI) Graduation by Finite Difference Interpolations

The assumption that differences in the true values, f_r, may be neglected beyond a certain order, j, means that $f_x = A + Bx + Cx^2 + Dx^3 + \ldots + Jx^j$. In most actuarial data it is legitimate to take $j=3$ (although, as already noted in section (III), this supposition should be examined carefully in some cases).

1. The first step in approaching the graduation of observed values by finite difference methods obviously would be to choose the value of j, to select *single points* in the data equal in number to the constants $A, B, \ldots, J$ in the above polynomial, to solve the simultaneous equations for $A, B, \ldots, J$, and thence to compute the interpolated values (see P:*166*: par. 6). The selected points, however, remain unadjusted in this process.

2. Instead of using single values, therefore, *groups of values* might be selected, so that consecutive sums rather than single points would be retained—for example, by using $\Sigma \log p_x$ in decennial sums, i.e., $\log ({}_{10}p_x)$, the original decennial probabilities of living would be undisturbed, but the yearly values p_x would be redistributed (P:*166*: pars. 7 and 12). The formulae of this method may be reached easily as in P:*166*: pars. 8-11. This process may be viewed as one of "fitting" in which the criterion of "fit" is the identity of the graduated and ungraduated consecutive sums.

3. When ordinary interpolations are made from selected points, or separate abutting sums, however, breaks of continuity occur at the points of division. An early empirical method widely used for the smoothing of these breaks was a double interpolation based on two overlapping series and then blended by the factors of the *Curve of Sines* (P:*167*:101). Other overlapping methods of ordinary interpolation also may give good results (P:*166*:102, footnote).

4. The problem of securing smoothness in the interpolations at the points occupied by the original data is dealt with more satisfactorily, however (P:*167*: par. 99), by employing the method of *osculatory interpolation*. The principles embodied in the var-

ious formulae embraced by that designation may be understood easily by tracing their development in the following order, which is mainly chronological.

(*i*) Sprague's original 5th difference osculatory formula is given, with references for its assumptions and demonstrations, in P:*167*: par. 100.

(*ii*) The corresponding Karup-King 3rd difference osculatory formula is stated similarly in P:*167*: par. 101.

(*iii*) The methods of derivation on which (*i*) and (*ii*) are based have been generalized by Reilly for differences of odd order $2h+1$ and contact of order k with the partial curves—Sprague's form of proof being covered in H:*149*, and Lidstone's in H:*154*. Thus with $h=2$ and $k=2$ Sprague's formula (*i*) is given; the expressions for 5th differences and 3rd order contact ($h=2$, $k=3$), 7th differences and 2nd order contact ($h=3$, $k=2$), and 7th differences and 3rd order contact ($h=3$, $k=3$) are also shown. For practical purposes, however, these extensions are not usually required.

(*iv*) Another 5th difference formula founded on Sprague's basic assumptions, but with the variation that the differential coefficients are determined from the mean of their values in the partial curves, was evolved by Buchanan (H:*114*:372-4; see also P:*71*:198 and P:*69*:90). Sprague's formula (*i*) is preferable, however, because of its more convenient numerical coefficients (P:*14*:124).

(*v*) In Sprague's assumptions the partial curves for determining the differential coefficients are taken as of one degree less than that of the osculatory curve. Henderson pointed out (H:*108*:215-7) that this restriction is not necessary, and accordingly produced an improved formula given, with references, in P:*167*: par. 103.

(*vi*) The use of partial curves in the preceding applications of Sprague's assumptions is, in fact, somewhat arbitrary. Henderson accordingly suggested (H:*108*:219, and P:*54*:190) that it is preferable to discard the partial curves, and instead simply to impose the condition that the differential coefficients at the points of junction should be continuous. The 5th difference formula which he thus reached is given, with references, in P:*167*: par. 104.

(*vii*) The preceding formula (*vi*), however, as observed by Henderson (H:*103*:221 and P:*54*:186-7 and 191), is not truly osculatory, for there is a discontinuity in the first differential coefficient of $\frac{1}{36}\delta^6$ resulting from the fact that the formula is based on only an approximate solution of the difference equation involved. He therefore gave (H:*149*:24) an exact method as stated in P:*167*: par. 105 (see also P:*59*:119 and 121-4 for a further mathematical discussion).

(*viii*) The development of a whole set of exact osculatory formulae has further been shown clearly by Jenkins in P:*69*, on the same assumptions as Henderson's in (*vi*), namely, (*a*) that the two curves must take the given value at the common point, and (*b*) that the corresponding derivatives of each curve at the common point must be equal to each other (though not necessarily equal to any predetermined value as in Sprague's assumptions). The Karup-King formula (*ii*) forms one of the set; the other formulae reached were new. For the 5th difference case, assumption (*a*) is met by taking $\varphi(x)=x(x-1)\psi(x)$, and $\psi(x)=a_0+a_1x+a_2x^2$, which gives Jenkins' formula; $\psi(x)=a_0+a_1x+a_2x^2+a_3x^3$ (which is redundant to the extent of one degree in x since the term a_3x^3 is not necessary) produces Sprague's formula (*i*) when $a_2=-\frac{7}{24}$, and Buchanan's (*iv*) when $a_2=-\frac{1}{3}$; and $\psi(x)=a_0+a_1x$ gives Henderson's formula (*vi*).

In a subsequent paper (P:*70*) Jenkins also developed another set of truly osculatory formulae of minimum degrees in all cases.

In P:*72*:24 and 30 he further produced the formulae based on even instead of odd differences.

(*ix*) All the preceding formulae deal with the usual problem of osculatory interpolation from data at equal intervals. The case of unequal intervals on Sprague's assumptions is considered by Ackland (H:*124*) and Reilly (H:*159*).

5. The principle of osculatory interpolation was first generally applied in the preparation of mortality tables from the death registers and census returns of population statistics, for which in earlier years the data were given in age-groups only—tabulations by single years of age not being available (P:*167*:100 and 102).

It was realized, moreover, even when the data can be secured by single years, that the errors in the statements of age which are characteristic of such population material (P:*167*:26-57, and P: *172*) may often be dealt with satisfactorily by grouping in suitable age classes, on the principle that the totals in quinquennial or decennial age-groups may be assumed to be correct although the values within each such group may require redistribution. The problem of interpolation therefore was to effect a smooth redistribution of the grouped data into the values at each age, so that the irregularities at individual ages would be removed without unduly disturbing the totals in each group.

This reproduction of the grouped data being fundamental, Shovelton (H:*123*) therefore investigated the effect of introducing that requirement directly into the determination of the osculatory curve, and found a formula given in P:*167*: par. 102. As pointed out in P:*14*:121, this method imposes one condition less than in a truly osculatory formula.

6. All the formulae derived by Sprague's assumptions which use partial curves, or from Henderson's which discard the partial curves and simply require that the differential coefficients shall be continuous, have in common the other basic condition that (being formulae for interpolation) the two curves must take the given value at the common point. When such osculatory formulae have been used, as in many of the population tables, to fill in the values between certain predetermined points, it has been found—unless the values at those points themselves lie upon a smooth curve—that the whole curve which finally results will show many undulations and points of inflexion, even though it will be free from discontinuities (P:*14*:124). In order to meet this weakness Jenkins (P:*71*) therefore released the two curves from the requirement that they must take the given value at the common point, and instead permitted, in effect, that the interpolated value shall differ from the given value by a fraction of its 2nd difference in the 3rd difference formula, or of its 4th difference in the 5th difference formula (see P:*71*:201 and 202). An excellent demonstration of the precise assumptions involved is given by Lindsay in P:*86*:211, where it is shown that, in the 5th differ-

ence formula, "the ordinates and second derivatives must pass through points differing from the predetermined points by the values of certain functions, while the first derivatives must exactly assume the predetermined values" (cf. also P:*106*:189 and 220, and P:*14*:150-1). Since the ordinates at the predetermined points are thus not to be reproduced, the resulting *modified* formulae will evidently effect some adjustment of those values in addition to performing an osculatory interpolation for the intermediate values. Jenkins gave a general expression, and the 3rd and the 5th difference formulae, in P:*71*:199-206; and the general form for even orders of differences, and the 4th difference formula, in P:*72*:10-12 and 24-30 [see also P:*14*:138 for complete proof of the 5th difference formula by Jenkins' method (noting the comment at P:*106*:189), and P:*86*:211 for Lindsay's elegant alternative demonstration]. In practice it is important to choose the formula which extends into an order of differences alternating in sign, for otherwise the graduated series will lie everywhere above or below the observed points; in mortality experiences, therefore, where the 2nd differences are usually not negligible while the 4th differences change in sign frequently, it is found that the 4th or 5th difference "modified" formulae are suitable but the 3rd difference one is unsatisfactory (P:*71*:202 and P:*72*).

Since all the preceding osculatory formulae interpolate in the middle major interval, special treatment is necessary for the initial and final intervals. It may therefore be noted here that Jenkins gives the formulae for the two intervals at each end with his 5th difference "modified" formula of par. 6 preceding (P:*71*:209; see also P:*14*:135 and 140). It has also been pointed out by Lidstone (P:*82*) that precisely the same results can be obtained by inserting the value 0 in the places of the missing values of δ^4 preceding the first and succeeding the last known δ^4, completing the difference table by addition to get the artificial values of the missing differences, and then applying the usual osculatory formula throughout.

7. Observing that Jenkins' "modified" 5th difference formula is not completely determinate, but only one of many satisfying the necessary conditions, Reid and Dow (P:*106*) have remarked

that one arbitrary constant is avoided by the implicit condition that no differences beyond the 5th are to appear in the final formula. By relaxing this condition, therefore, they obtain a general 5th degree "modified" formula, which completely satisfies the osculatory conditions, with 3 instead of only 2 arbitrary constants—Jenkins' expression being the special case of this general formula when the new constant, b, is zero. In practical application, Reid and Dow accordingly suggest that the values should first be calculated by Jenkins' formula, and that the flexibility afforded by the b term should be used to improve the results as may seem desirable. The procedures which they actually adopted are noted in par. 10 hereafter, since they are of a type belonging to the methods there stated.

8. All the foregoing interpolation processes are designed to secure smoothness of junction in the final results at the predetermined points of division of the observed data. In some of the practical applications the data at these points have been left unchanged—only the intermediate values being interpolated by one of the osculatory formulae (P:*167*: par. 99). In order to graduate the values at the points of division, therefore, George King proposed a simple method in which adjusted quinquennial *pivotal* values are first calculated centrally by ordinary 3rd difference interpolation from three quinquennial sums of the data, whence the intermediate values are supplied by 3rd difference osculatory interpolation (P:*167*: par. 107 and 110). King applied his method separately to the deaths and populations; in more recent examples (P:*170*:334 and P:*14*:129 and 149) a single application to q_x has given equally good results. In H:*155* it was found convenient to compute the pivotal values at the first instead of the central ages of each quinquennial group, by the formula resulting from putting $n=1$ and $x=-2$ in (7) of P:*166*:87.

If it is thought desirable to determine the pivotal values from more than three groups, the corresponding formulae based on four groups and $j=3$, or on five groups and $j=4$ or 5, may be used as given in P:*167*: par. 108. Fifth difference osculatory interpolations have also been used widely for the subsequent intermediate values.

9. The pivotal values in the method of the preceding paragraph are all found by ordinary interpolation from groups. Since the real objective in the calculation of those values, however, is to obtain reliable points which represent the original data adequately but yet remove any undue fluctuations, and also because the subsequent osculatory interpolations passing through such predetermined points have a tendency to show undulations and points of inflexion (cf. par. 6 here), it has been suggested that the pivotal values might be determined from the quinquennial sums by using the formulae of osculatory instead of ordinary interpolation—the intermediate values thereafter being supplied, of course, by a corresponding osculatory process. This proposal was made first by Buchanan; a subsequent investigation by Jenkins, however, re-examined the idea with certain earlier stages of its development which will therefore be given here in their natural order:

(*i*) The pivotal value formula based on Sprague's original 5th difference osculatory expression (see (*i*) of par. 4 here) is given in P:*72*:14.

(*ii*) The pivotal formula based on Jenkins' modified 4th difference method (par. 6) is demonstrated in P:*72*:13.

(*iii*) The pivotal formula based on Jenkins' modified 5th difference process is derived in P:*14*:128-9.

The practical effects of these variations in procedure have been tested in P:*14*: and P:*72*. Since (as Buchanan observed) the pivotal values (which he called "guiding "values) would tend to lie on a smoother curve when they are computed by the "modified" formulae, it was found (as would be expected) that, with regard to smoothness, Sprague's osculatory interpolation as a basis for both the pivotal and subsequent calculations was improved slightly by the modified 4th difference method, which in turn was improved by the modified 5th difference process, while of course the order was reversed in respect of fit.

For the determination of the first two and last two pivotal values with these methods, Lidstone has suggested (P:*82*:278) the method of inserting zero values for the missing differences as in par. 6 of this section. Buchanan, however, objected that this has the effect of assigning definite magnitudes to the pivotal values sought, and that unless they are reasonable the curve may be distorted; he therefore prefers the insertion of reasonable values for the missing terms, calculation of the resulting differences, and then interpolation so that the curve would not pass through the pivotal values (P:*15*:209).

10. At the beginning of the preceding paragraph it was indicated that the objectives to be borne in mind in determining the pivotal values should be the adequate representation of the original data and, in the final analysis, the derivation from them of satisfactorily smooth results. The pivotal values, consequently, should fit the data in accordance with some adequate criterion, which, however, must still permit the eventual construction of a smooth curve. The necessary compromise between fit and smoothness thus inherent in the process may therefore be facilitated by determining the pivotal values by a method which will specifically recognize the requirement of fit as well as of smoothness—the subsequent interpolations then being made by one of the osculatory formulae already considered.

(*i*) This evidently might be done by using a graphic method (P:*71*:206) to find the pivotal values, since that method can easily produce any compromise between fit and smoothness that may be desired (see par. 1 of section (V) here).

(*ii*) The criterion of fit adopted could be that of least squares, by which the "best" fitting pivotal values would be computed from, say, the quinquennial groups in order to effect the greatest possible reduction of the mean square error (P:*145*:368). The formula for five symmetrical groups is deduced as (51) in P:*167*: 112-3 (and in P:*166*:105, where $w_n = \sum_{x=n-2}^{x=n+2} u_x$). The principles underlying this method are summarized at p. 282; C; 7, section (*xii*) (cf. also par. 2 (*i*), section (VII) here).

(*iii*) The flexible b term of the general modified formula of Reid and Dow (par. 7) may be brought into the calculations. In P:*106* they illustrate the following alternatives: (*a*) b taken numerically to make the formula resemble Everett's formula closely; (*b*) the pivotal values found as in (*a*), but in the interpolations b taken to secure maximum smoothness by making

$\Sigma(\Delta^3 q_x)^2$ a minimum; (*c*) a double application of method (*b*); and (*d*) *b* determined for the pivotal values to make the expected equal to the actual deaths, and in the interpolations to yield maximum smoothness therefrom. As would be expected, the results of (*d*) were the most satisfactory.

(VII) Graduation by Linear Compounding (including "Summation" Formulae)

1. The basic principle of this method has already been stated at p. 282; C; 7, section (*xii*). The manner in which it was first developed, from the interpolation methods of pars. 1, 2, 4(*i*), and 4(*ii*) of the preceding section (VI) here, is set out fully in P:*166*: pars. 1-20, and the detailed references there stated. The linear compounding formulae so evolved were based on the preliminary selection of certain interpolations, and were therefore somewhat fortuitously dependent on the particular selection made. The very important pioneer work of De Forest (P:*166*) should be noted here particularly, since his investigations, beginning in 1871, on the theory of reduction in the mean square error, by which the capacities of the various formulae could be compared, *a priori*, in respect of fit and smoothness, constitute the earliest and most complete treatment of the subject—a fact which still seems to be not fully recognized.

2. Having thus established the method of comparing the *a priori* fitting and smoothing abilities of linear compounding formulae, De Forest then—between 1871 and 1880—gave a remarkably clear and exhaustive examination of the whole subject, which unfortunately remained unrecognized until the appearance of P:*166* in 1924. Equipped with this measure of the reduction in the mean square error of the graduated term itself, or its differences, De Forest published the following important series of formulae:

(*i*) The symmetrical formulae, up to 25 terms for $j=2$ or 3, which give the greatest possible reduction in the mean square error of the graduated term itself. He also indicated the corresponding unsymmetrical formulae. [De Forest noted that Schiaparelli had made an independent investigation in 1867; and

Woolhouse had considered the problem with a slightly different objective in 1865. The formulae, with the additional cases of $j=4$ or 5, 6 or 7, and 8 or 9, were restated many years later by Sheppard and Sherriff, without knowledge of De Forest's work]. See P:*166*: pars. 1 and 22-26. These formulae give the best possible fit, according to the criterion of least squares.

(*ii*) The most important formulae which he published were the symmetrical series, up to 25 terms when $j=3$, which give the greatest possible reduction in the mean square error of the 4th differences (P:*166*: pars. 27 and 30). De Forest also gave the 7-term formula for the maximum reduction in the mean square error of the 6th differences when $j=5$, for use when a series varies so rapidly that $j \neq 3$ (loc. cit., 107, footnote). These formulae, which first appeared in 1873, antedated in their fundamental conception a large number of very similar investigations (dealing with the 2nd or 3rd differences instead of the 4th) by much later writers (see par. 3 below). They are designed to give the greatest possible smoothness, according to the criterion of the reduction to be anticipated in the mean square error of the 4th differences.

(*iii*) Recognizing that it is not possible to secure the best fit and maximum smoothness at the same time (cf. par. 2, section (I) here), De Forest gave the formulae up to 15 terms which emerged from an examination of the curve to be anticipated for the flow of the linear compounding coefficients. These formulae have good fitting as well as smoothing power. Again here his conclusions preceded by many years the work of subsequent investigators (P:*166*: par. 32).

(*iv*) Another series with good values in respect of both fit and smoothness which De Forest gave is shown in P:*166*: par 33.

(*v*) He also made an extensive investigation of the effects of applying some of his formulae repeatedly, and in this instance also reached some important conclusions on a matter which was suggested, but not closely examined, by others in later years (P:*166*: par. 34, and footnote). De Forest noted clearly the manner in which, when any linear compounding formula with $j=2$ or 3 is repeated a large number of times, the curve of the coefficients ultimately tends to a central bell-shaped portion with an infinite number of small undulations at each end.

3. As stated in (*ii*) of the preceding paragraph, De Forest determined his formulae for maximum smoothness when $j=3$ by minimizing the mean square error in the 4th differences. Many years later the precisely analogous problems with respect to the 3rd or 2nd differences (when $j=3$) were investigated in several independent enquiries—all again without knowledge of De Forest's contributions. The formulae for the 2nd differences were considered by Hardy and Sheppard; those for the 3rd differences by Henderson and Larus. The details are given in P:*166*: pars. 28 and 29.

4. All the formulae which are derived by the methods of pars. 1-3 emerge in the linear compound form given at (*a*) and (*b*), p. 283; C; 7, section (*xii*). Numerical calculation by such expressions offers no difficulty with modern calculating machines—a device explained in P:*79*:20-21 being useful in the work.

Before the widespread use of calculating machines, however, the computations were greatly facilitated by putting the formulae into a special "summation" form, i.e., in sums $[p][q][r]\ldots$ where $[p]u_s$ denotes the sum of p u's of which u_s is the middle term; and that method is of course still very convenient if the formula can be so expressed. The early literature consequently gives many examples of this form. In some instances the linear compounds were changed into the summation type by trial (H:*109*:543); in others the governing summation could often be selected from inspection of the central coefficient in the linear compound, whence the remainder of the formula followed easily (H:*111*: vol. XLII, 133).

Since, however, it was sometimes difficult to find these summations, a great deal of attention was paid to the direct production of summation formulae correct to 3rd or 5th differences, without first deriving the corresponding linear compound, and without any attempt to secure any maximum or other specified degree of reduction of mean square error in either the term or any order of its differences. References to the long list of researches are given in P:*166*: par. 21; the résumé in P:*87*:260-267, and the table at P:*59*:53, will now generally be sufficient for the student's purposes. It should be realized that the foundation of these methods is the simple formula for $[n]u_s$ as first employed by Hardy to 3rd differences in P:*50*:371 (noting the misprint of $\frac{24}{n^2-1}$ for $\frac{n^2-1}{24}$ in formula (1) on p. 371 thereof), and deduced also to 5th differences in H:*109*:534-7 and H:*111*: vol. XLII, 137-8.

Amongst these methods the name "wave-cutting" (H:*111*: vol. XLII, 111) has been applied to certain formulae characterized by the use of very unequal summations, resulting in a curve for the linear compounding coefficients which has a broad flat top. A formula of this type first suggested by Hardy in P:*50*:375 has been discussed in H:*111*: vol. XLII, 131 and 109-111. Another, composed entirely of even summations, has been deduced by Vaughan, and illustrated, in P:*147*:434-440. The whole problem is also discussed further in P:*148*:477-487.

5. (*i*) While the investigations covered by the references in the preceding paragraph dealt almost entirely with the direct production of summation formulae correct to 3rd or 5th differences, it was remarked by Hardy that an increase in smoothing power would sometimes result from changing the summations of a 3rd difference formula even though the change would introduce a 2nd difference error (see H:*110*:69, and cf. H:*68*:277 and P:*50*: 374).

(*ii*) For a rough preliminary adjustment only (when it is desired merely to obtain results in general conformity with the data), even simpler formulae of this character, using summations without any operand, may occasionally be useful (see P:*51*: 35, footnote, and P:*147*:431).

(*iii*) These formulae with 2nd difference errors have also been employed as a basis for evolving special formulae for the graduation of colog p_x. Thus Spencer (H:*112*:402-7) derived an expression by providing that the errors due to the 2nd and 4th differential coefficients were practically counterbalanced. This principle was further developed, and additional formulae given, in P:*150*, P:*151*, and P:*148*:465-470.

6. It was observed in par. 4 that the various summation formulae correct to 3rd or 5th differences have usually been

deduced without any attempt to secure, *a priori*, any specified degree of reduction of mean square error in either the graduated term or any order of its differences. In P:*51*:29, however, Hardy suggested the possibility of selecting a convenient set of summations, and then determining the rest of the formula (the "operand") to minimize the mean square error in one of the orders of differences. This idea was investigated by Vaughan in P:*147*:443-5. Neither Hardy nor Vaughan, however, seems to have had any knowledge of De Forest's more exhaustive work.

7. Since almost all the linear compound, and all the summation, formulae are symmetrical, it is necessary to use corresponding unsymmetrical formulae, or a formula of short range, or special devices, to reach certain terms (depending on the range of the symmetrical formula) at the beginning and end of the series to be graduated. In practice special devices are often employed. Their purposes and methods are described sufficiently in P:*59*:60-61 (noting that Ackland's process is explained in H:*63*:357, and that examples of other methods may be found in H:*97*:340 and H:*112*:378; H:*110*:90; H:*112*:373; and H:*171*:113-117).

8. The graduation of rates of mortality during the select period may be dealt with by the device noted in section (IX) hereafter.

It is important to realize, both for an understanding of this section (VII) and the next, that all the linear compounding graduation formulae (whether they can be thrown into a convenient "summation" form or not) effect, in reality, some particular *a priori* degree of reduction in the mean square error of the graduated term itself, and also some other *a priori* degree of reduction in the mean square error of each order of differences of that graduated term. The degrees of reduction thus attained vary, of course, according to whether the formula is designed to secure (*a*) the greatest possible reduction in the mean square error of the graduated term, say f''_r, itself (accompanied by some other degree of reduction in respect of the differences)—these being the "fitting" formulae of par. 2(*i*); or (*b*) the greatest possible reduction in the mean square error of one of the orders of differences, say $\Delta^z f''_r$, of the graduated term (accompanied by some other degree of reduction in respect of the term itself)—these being the "smoothing" formulae of par. 2(*ii*) when $j=3$ and $z=4$, or those of par. 3 when $j=3$ and $z=3$ or 2; or (*c*) such degrees of reduction (not in any case being a greatest possible reduction) in the mean square errors of the graduated term and its differences as may emerge fortuitously from the particular linear compounding factors chosen—these being the formulae of pars. 2(*iii*), (*iv*), and (*v*), and pars. 4, 5, and 6.

The first main defence of the formulae of category (*c*) lies in the fact (already emphasized) that it is not possible, either in constructing a formula *a priori*, or in testing a graduation *a posteriori*, to secure at the same time the best possible fit and the best possible smoothness. In many cases—though not all—it may therefore be advisable to sacrifice something by way of fit in order to improve the smoothness, or *vice versa*, and so to choose deliberately a formula which combines both fit and smoothness in reasonable proportions. The second main defence of all the linear compounding formulae of section (VII) is that each graduated value is calculated from only a limited range of the ungraduated data, so that in effect a redistribution is secured on the basis of the information supplied by the adjacent terms (depending on the range selected), without recourse to very distant terms which can hardly have any proper bearing on the value being graduated.

(VIII) Graduation by the Difference-Equation Method

1. If the whole range of the observed values, f'_r, extends from $r=1$ to $r=\nu$, the linear compounding formulae of section (VII) perform their graduations by progressive applications of the same formula over the successive partial ranges covered by the number of terms in the formula selected (P:*166*:94). The problem, however, of deriving all the best "fitting" values, f''_r, according to the least squares criterion, from a series of observed data, f'_r, may be viewed (cf. Chapter VIII) as the problem of minimizing

$\sum_{r=1}^{r=\nu} (f''_r - f'_r)^2$ over the whole range, when the weights are taken as uniform; and this process is equivalent to effecting the greatest possible reduction in the mean square error of f''_r (cf. P:*166*: pars. 1, 22, and 23). The best "smoothing" linear compounding formulae (which were considered in pars. 2(*ii*) and 3 of section (VII) when $z=4$, 3, or 2 for $j=3$) likewise effect, for the selected range, the greatest possible reduction in the mean square error of one of the orders of differences, $\Delta^z f''_r$, and in section (III) here it was noted that $\Sigma(\Delta^z f''_r)^2$ would afford a measure of the comparative smoothness of the results. Since, however, it is impossible to secure at the same time the maximum of fit and smoothness, it is evident that a compromise may be obtained by combining those desiderata in some stated proportion, so that the blended function to be made a minimum could be taken on analogous principles, over the whole range, as $k\sum_{r=1}^{r=\nu} (f''_r - f'_r)^2 + \sum_{r=1}^{r=\nu-z} (\Delta^z f''_r)^2$, where k is the arbitrary proportion, and the upper limit for the smoothing term is $r=\nu-z$ because Δ^z is not available for higher values in a series ending at $r=\nu$. [A mathematical derivation of this function, by employing the principles of probability and Bayes' formula (43b) of p. 222; B; 12, has been given by Whittaker in P:*155*:304-6.]

2. Since this function to be minimized is thus only a combination and extension of the principles already dealt with in linear compounding, it will be clear that the result must be expressible as a linear compounding formula, and that it will cover the whole range from $r=1$ to $r=\nu$ instead of a limited range only. This relation between the methods of section (VII) and those to be now considered here is important, as will appear. It may indeed be useful, at this stage, to remark that the shape of the curve taken by the linear compounding coefficients for this treatment over the whole range resembles closely that stated by De Forest for the formulae of limited range (par. 2(*iii*) of section (V) here, and P:*166*: par. 32), with the numerous small undulations at each end which he also observed when any linear com-

pounding formula is repeated frequently (par. 2(*v*) of section (V) here, and P:*166*: par. 34, footnote).

3. The relationship pointed out in the preceding paragraph may be seen readily by minimizing the expression of par. 1, and then following the method of solution first shown by Aitken (H:*152*) and demonstrated very clearly by Spoerl (P:*134*:423-5). The minimizing is effected easily by differentiating with regard to each of the unknown values, f''_r, in turn, and equating to zero—remembering, since all the f''_r's are independent, that $\frac{d}{df''_x}(f''_x)=1$, and that $\frac{d}{df''_x}(f''_{x+t})=0$ when $t\neq 0$. Expanding the expression as

$$k\sum_{r=1}^{r=\nu}\left[(f''_r)^2-2f''_rf'_r+(f'_r)^2\right]+\sum_{r=1}^{r=\nu-z}\left[f''_{r+z}-zf''_{r+z-1}+\ldots+(-1)^zf''_r\right],$$

and for differentiation with regard to f''_x therefore discarding all terms not involving f''_x, the ν equations for the solution of the ν unknown f''_r's emerge at once (as shown in slightly different notation, for $z=3$, at P:*134*:404 and P:*26*:300-1—noting that in lines 3 and 4 at the commencement of the latter discussion on p. 300 the Σ should cover both the terms, and that in the second line of formula (2) on p. 301 the subscripts of the last two terms should be $a+1$ and a). Except for the first z equations for $f''_1, f''_2, \ldots$, and f''_z, and likewise the last z at the other end, all the equations are of the same form $k(f''_r-f'_r)+(-1)^z\Delta^{2z}f''_{r-z}=0$; and the first z and the last z differ only in the successive omission of certain differences. The problem now is to solve these ν equations, and so to find the unknown graduated f''_r's.

4. In seeking this solution the first point to be noticed is that if, instead of covering only values from 1 to ν, the series were of indefinite length from $-\infty$ to $+\infty$, so that the equation $k(f''_r-f'_r)+(-1)^z\Delta^{2z}f''_{r-z}=0$ held unchanged throughout, the expression to be minimized would be $k\sum_{r=-\infty}^{r=+\infty}(f''_r-f'_r)^2+\sum_{r=-\infty}^{r=+\infty}(\Delta^zf''_r)^2$; and this would reduce to the expression for the range 1 to ν if all the values before $r=1$ and after $r=\nu$ were zero—that is, if $f''_r=f'_r$ and $\Delta^zf''_r=0$ for $r<1$ or $>\nu$. But when $f''_r=f'_r$, it follows that

$\Delta^z f''_r = \Delta^z f'_r$; it is therefore only necessary at each end to supply new ungraduated terms, f'_r, such that $\Delta^z f'_r = 0$. This means that at each end additional ungraduated terms should be supplied with their values of $\Delta^{z-1} f'_r$ equal. Some method must therefore be settled for annexing those terms, and then completing the solution for all the ν values of f''_r required.

5. The problem has been examined in a number of papers, of which the following brief chronological account will give the necessary references.

(*i*) The first statement of the whole principle here under discussion was given by E. T. Whittaker (H:*144*; see also P:*155*: 303). An approximate solution was suggested at that time; but it may now be discarded as greatly improved methods have been developed (cf. P:*24*:24).

(*ii*) A complete theoretical solution was published next, again by Whittaker (H:*145*); but it also may be discarded on account of the extremely unwieldly figures which it involved (cf. P:*24*:18).

(*iii*) A method when $z \ngtr 3$ was then given by Henderson (P:*56*), which depended on resolving the difference equation (of par. 3 here) into two factors. The final result is thus reached by constructing an "intermediate series"—the calculations being performed in two steps. The process can be applied for any value of k, although the factors are more easily handled when k takes certain special forms. When $z=3$, for example, the difference equation for all but the first 3 and last 3 terms is

$k(f''_r - f'_r) - \Delta^6 f''_{r-3} = 0$, which may be written $\left(1 - \dfrac{\Delta^6 E^{-3}}{k}\right) f''_r = f'_r$

where $E = 1 + \Delta$; and when k takes the form $\dfrac{16(2n+3)^2}{n(n+1)^3(n+2)^3(n+3)}$

the operator $\left[1 - \dfrac{n(n+1)^3(n+2)^3(n+3)}{16(2n+3)^2}\Delta^6 E^{-3}\right]$ can be resolved

into the two cubic factors

$$\left[1 - n\Delta + \frac{n(n+1)}{2}\Delta^2 - \frac{n(n+1)^2(n+2)}{4(2n+3)}\Delta^3\right]$$
$$\left[1 + n\Delta E^{-1} + \frac{n(n+1)}{2}\Delta^2 E^{-2} + \frac{n(n+1)^2(n+2)}{4(2n+3)}\Delta^3 E^{-3}\right].$$

In order to start the process the terms at the beginning are first estimated by using constant values of Δ^{z-1}; those at the end are derived from the results obtained from the first step. The problem of estimating approximate terms at the beginning, in the manner originally suggested by Henderson, is discussed in P:*56*:36-38, P:*12*:7-10 (see also P:*61*:516-517), and with a complete numerical example (for $z=3$ and $k=.009$) in H:*171*:151-156. A numerical illustration where $z=2$ and $k=\frac{1}{3}$ is also set out in P:*12*:9-10.

More recently the correction of these initial terms has been examined in P:*134*:409-413, P:*58*:51, and P:*9*:52, and a mathematical summary of a method proposed by Henderson is stated in P:*59*:28, par. 4.2.

A useful résumé of the 2nd difference case ($z=2$) when $n=1$ ($k=\frac{1}{3}$)or $n=3$ ($k=\frac{1}{80}$), and of the 3rd difference case ($z=3$) when $n=1$ ($k=\frac{25}{54}$) or $n=3$ ($k=.009$), is given in P:*17*:512-514, with a convenient form of card for performing the calculations.

[Note that when $z=2$ the difference equation is

$$k(f''_r - f'_r) + \Delta^4 f''_{r-2} = 0, \text{ or } f'_r = f''_r + \frac{1}{k}\Delta^4 f''_{r-2},$$

and when $z=3$ it is

$$k(f''_r - f'_r) - \Delta^6 f''_{r-3} = 0, \text{ or } f'_r = f''_r - \frac{1}{k}\Delta^6 f''_{r-3}.]$$

This method is sometimes referred to as the "Whittaker-Henderson Formula A", from Professor Whittaker's first enunciation of the principles and Henderson's subsequent solution.

(*iv*) In applying this difference-equation method to the graduation of mortality rates, Henderson (P:*57*) has suggested that, instead of using the rates of mortality in the fitting portion of the function by taking $f'_r = q'_r$ and $f''_r = q''_r$, it may be advisable to make allowance for the extent of the data at each age by weighting those rates approximately with E'_r, so that the fitting portion, $\Sigma(f''_r - f'_r)^2$, might be taken as

$$\Sigma\,[E'_r(q''_r - q'_r)^2] = \Sigma\left[E'_r\left(q''_r - \frac{\theta'_r}{E'_r}\right)^2\right].$$

Moreover, when $z=3$ the effect of the smoothing term is to make the graduated values close to a series with constant 2nd differences, whereas at the younger ages the 2nd differences should increase; and this discrepancy might be corrected by adding $.1(\Delta^2 q_1'')^2$ at the youngest age. Under these particular conditions the whole blended function to be minimized would become

$$k\sum_{r=1}^{r=\nu}\left[E_r'\left(q_r''-\frac{\theta_r'}{E_r'}\right)^2\right]+\sum_{r=1}^{r=\nu-3}(\Delta^3 q_r'')^2+.1\,(\Delta^2 q_1'')^2.$$

By differentiating with regard to the unknown q_r'' and equating to zero, as before, a set of simple equations equal in number to the unknowns emerges as stated in P:*59*:40 (see also P:*133*:61).

This method is sometimes referred to as the "Whittaker-Henderson Formula B". The form of calculation is indicated in P:*59*:41. While it avoids the difficulty with the initial terms, the arithmetical work is heavy. The selection of k may be made from the fact that in the Formula A method (par. (*iii*) preceding) $k \doteq .01$ usually gives a satisfactory graduation, and since this factor is replaced by kE_r' in the B method it follows that kE_r' should be about .01 over the range of ages where the data are heaviest (P:*119*:289); or k may be taken as inversely proportional to the square root of the maximum or average E_r' (P:*59*:43).

(*v*) At almost the same time as the appearance of Henderson's solution, Aitken (as already noted in par. 3) examined the difference equation in linear compound form (H:*152*). When $z=3$, for example, the equation (see par. 3) is $kf_r''-kf_r'=\Delta^6 f_{r-3}''$, or $kf_r'=(k-\Delta^6E^{-3})f_r''$, where $E=1+\Delta$, whence $f_r''=\left(1-\dfrac{\Delta^6E^{-3}}{k}\right)^{-1}f_r'$

and the linear compound follows immediately by expansion (see also P:*134*:418, 443, and 517). Although these linear compounding coefficients, which are symmetrical, evidently cover the whole range from $r=1$ to $r=\nu$ and also continue indefinitely but diminish rapidly beyond each end (instead of dealing only with a limited range as in the formulae of section (VII) here) it is to be noted that they follow closely the curve first indicated by De Forest for the case of repetitions of a limited-range formula (see par. 2).

Aitken calculated the linear compounding coefficients when $z=3$ for certain values of k (being ϵ in his and Whittaker's notation), which for $k=.01$, .02, .05, .1, .25, and 1 are given also in P:*1*:34. When $z=3$, values for other convenient special cases $k=.46296$, .12738, .04537, .01906, .009, and .004639 are to be found (in the columns headed k_z) in P:*134*:457-462. When $z=2$ the coefficients for $k=\frac{1}{3}$ are worked out similarly in P:*12*:516 and 517.

Aitken further established the important and simple fact that an accurate solution may be reached by extending the ungraduated data, f'_r, for $(z+1)$ terms at each end (in order to determine the required terminal zero values of $\Delta^z f'_r$) by the use of a set of unsymmetrical coefficients (depending on the value of k), so that at each end the constant values of Δ^{z-1} given by these extended terms may then be used to build up easily as many more extended terms as may be required to permit the application of the symmetrical linear compounding form over the whole range of the data. The method, and the unsymmetrical coefficients for the extensions at each end when $z=3$ and $k=.01$, .02, .05, .1, .25, or 1, are given clearly and accessibly by Aitken in P:*1*:31-36; the mathematical analysis, and the coefficients for the extensions when $z=3$ and k has the particular values .46296, .12738, etc., noted above, are available conveniently in P:*134*:423-5 and 457-462. A complete numerical illustration may be found in P:*1*: 34-36.

(*vi*) From the preceding development of the linear compound form it is clear that a closely approximate solution will be reached by using only the significant portion of the expansion. Davidson and Reid (P:*24*) accordingly investigated the formulae of "summation" type which could be found by taking $\left(1-\frac{\Delta^6 E^{-3}}{k}\right)^{-1}$ only as far as $\left(1+\frac{\Delta^6 E^{-3}}{k}\right)$, and throwing the linear compound so retained into the "summation" form of par. 4, section (VII) here. They gave (loc. cit., 6) both the linear compound coefficients and the equivalent summation formulae for 17, 21, and 25 terms with $k=.01$, .05, .06, .1, and .2, and also (loc. cit., 12) summation formulae of 15, 17, 21, 23, and 27 terms when $k \doteq .02$, .05, .1, .3, and .8.

(*vii*) Finally, an excellently compact method of solution, which also has the great merit of accuracy and simplicity, has been evolved by Spoerl (P:*134*:414-6). He gives a systematic and adequate process of finding z additional terms at each end by applying Aitken's unsymmetrical linear compounding multipliers, and then computing any graduated term directly and very easily from the original data and the $2z$ added terms only, by applying the linear compound formula to the original series of data, and also z special multipliers to the z added terms at each end. A clear summary of the method when $z=2$ is given by Spoerl at P:*12*:515.

The mathematical analysis of Spoerl's formulae is treated fully in P:*134*:425-7 and 443-9. The multipliers required in the work are tabulated in P:*134*:456-462 for $z=3$ and $k=$[illegible], .12738, .04537, .01906, .009, and .004639, and at P:*12*:516 for $z=2$ and $k=\frac{1}{3}$.

Spoerl's process is considerably shorter than Aitken's procedure of par. (*v*), since it requires at each end the calculation and use of only z added terms, whereas Aitken's method employs a long extension of the original data depending on the value of k. Spoerl's method, moreover, has the very important practical advantage that when the $2z$ additional terms have been computed, any desired graduated values at selected points (such as every 5th or 10th value) can be found with the greatest ease in order to see whether the graduation is likely to be satisfactory with the particular z and k adopted.

(IX) The Graduation of "Select" Mortality Tables

When it is desired to construct a "select" mortality table, in which the rates of mortality will be tabulated for a certain number of years of duration, t, since entry at select age $[x]$, the problem of graduation is to adjust the table consistently for both the variables, and to run the values of the select period smoothly into the "ultimate" rates. Using the standard notation already stated on p. 296; C; 11, the following special methods have been employed on various occasions.

(*a*) ***With Makeham's Formula***

Since the property of "uniform seniority" (see p. 319; C; 18) in Makeham's formula (83) depends only on the constant c, G. F. Hardy preserved its applicability to select tables (H:*73*: 359; H:*76*:493; H:*90*:126, 134-5, and 158; H:*95*:508) by writing $\mu_{[x]+t}=A+f(t)+[B+\varphi(t)]c^{x+t}=$ (say) $A_t+B_t c^{x+t}$, or correspondingly $\text{colog}_{10}\, p_{[x]+t}=\alpha_t+\beta_t c^{x+t}$. Treating separately the data for each year of duration, t, the values of A_t and B_t (or α_t and β_t) are first determined by one of the fitting methods of Chapter VIII. These values, however, will generally require some further adjustment in order to secure consistently smooth junctions between the select and ultimate portions of the tables. The shape of the curves is shown at P:*51*:74-5; the particular expressions used by Hardy for their representation are stated in H:*90*:135 and 158; H:*95*:508; and P:*51*:76-7. His methods of determining the constants therein are referred to in H:*90*:127 and 157-161, and H:*95*:507, and are explained more clearly in H:*106*:292-5 and 322-3.

Hardy based much of his analysis on the fact that when $\mu_{x+t}=A+Bc^{x+t}$, and $\mu_{[x]+t}=A+f(t)+[B+\varphi(t)]c^{x+t}$, it follows immediately by integration that

$$\log_{10} l_{[x]+t}=\log_{10} l_{x+t}-\frac{1}{\log_e 10}\,[F(t)+c^x F'(t)].$$

Similarly in terms of the constants of colog p the formula may be written $\log_{10} l_{[x]+t}=\log_{10} l_{x+t}-f_t-\beta c^x \psi_t$. It should be noted that the relation $\beta_t=[1-2n(10-t)c^{-t}]\beta$ given in P:*51*:77 ought to be $\beta_t=[1-n(19-2t)c^{-t}]\beta$, as pointed out in H:*118*:474.

(*b*) ***With the Graphic Method***

(*i*) The process adopted originally by T. B. Sprague for the H[M] data was firstly to graduate $q_{[x]}$. For the first year of duration he then graduated the unadjusted $\dfrac{q_{[x]+1}}{q_{[x]}}$ for each value of x, and multiplied the graduated value of this ratio by the graduated $q_{[x]}$ to obtain the required value of $q_{[x]+1}$. A similar method was used for each subsequent year of duration (H:*61*:244-6).

(*vii*) Finally, an excellently compact method of solution, which also has the great merit of accuracy and simplicity, has been evolved by Spoerl (P:*134*:414-6). He gives a systematic and adequate process of finding z additional terms at each end by applying Aitken's unsymmetrical linear compounding multipliers, and then computing any graduated term directly and very easily from the original data and the $2z$ added terms only, by applying the linear compound formula to the original series of data, and also z special multipliers to the z added terms at each end. A clear summary of the method when $z=2$ is given by Spoerl at P:*12*:515.

The mathematical analysis of Spoerl's formulae is treated fully in P:*134*:425-7 and 443-9. The multipliers required in the work are tabulated in P:*134*:456-462 for $z=3$ and $k=.46296$, .12738, .04537, .01906, .009, and .004639, and at P:*12*:516 for $z=2$ and $k=\frac{1}{3}$.

Spoerl's process is considerably shorter than Aitken's procedure of par. (*v*), since it requires at each end the calculation and use of only z added terms, whereas Aitken's method employs a long extension of the original data depending on the value of k. Spoerl's method, moreover, has the very important practical advantage that when the $2z$ additional terms have been computed, any desired graduated values at selected points (such as every 5th or 10th value) can be found with the greatest ease in order to see whether the graduation is likely to be satisfactory with the particular z and k adopted.

(IX) The Graduation of "Select" Mortality Tables

When it is desired to construct a "select" mortality table, in which the rates of mortality will be tabulated for a certain number of years of duration, t, since entry at select age $[x]$, the problem of graduation is to adjust the table consistently for both the variables, and to run the values of the select period smoothly into the "ultimate" rates. Using the standard notation already stated on p. 296; C; 11, the following special methods have been employed on various occasions.

(*a*) *With Makeham's Formula*

Since the property of "uniform seniority" (see p. 319; C; 18) in Makeham's formula (83) depends only on the constant c, G. F. Hardy preserved its applicability to select tables (H:*73*: 359; H:*76*:493; H:*90*:126, 134-5, and 158; H:*95*:508) by writing $\mu_{[x]+t}=A+f(t)+[B+\varphi(t)]c^{x+t}=$ (say) $A_t+B_t c^{x+t}$, or correspondingly $\text{colog}_{10}\, p_{[x]+t}=\alpha_t+\beta_t c^{x+t}$. Treating separately the data for each year of duration, t, the values of A_t and B_t (or α_t and β_t) are first determined by one of the fitting methods of Chapter VIII. These values, however, will generally require some further adjustment in order to secure consistently smooth junctions between the select and ultimate portions of the tables. The shape of the curves is shown at P:*51*:74-5; the particular expressions used by Hardy for their representation are stated in H:*90*:135 and 158; H:*95*:508; and P:*51*:76-7. His methods of determining the constants therein are referred to in H:*90*:127 and 157-161, and H:*95*:507, and are explained more clearly in H:*106*:292-5 and 322-3.

Hardy based much of his analysis on the fact that when $\mu_{x+t}=A+Bc^{x+t}$, and $\mu_{[x]+t}=A+f(t)+[B+\varphi(t)]c^{x+t}$, it follows immediately by integration that

$$\log_{10} l_{[x]+t}=\log_{10} l_{x+t}-\frac{1}{\log_e 10}\,[F(t)+c^x F'(t)].$$

Similarly in terms of the constants of colog p the formula may be written $\log_{10} l_{[x]+t}=\log_{10} l_{x+t}-f_t-\beta c^x \psi_t$. It should be noted that the relation $\beta_t=[1-2n(10-t)c^{-t}]\beta$ given in P:*51*:77 ought to be $\beta_t=[1-n(19-2t)c^{-t}]\beta$, as pointed out in H:*118*:474.

(*b*) *With the Graphic Method*

(*i*) The process adopted originally by T. B. Sprague for the H$^{[M]}$ data was firstly to graduate $q_{[x]}$. For the first year of duration he then graduated the unadjusted $\dfrac{q_{[x]+1}}{q_{[x]}}$ for each value of x, and multiplied the graduated value of this ratio by the graduated $q_{[x]}$ to obtain the required value of $q_{[x]+1}$. A similar method was used for each subsequent year of duration (H:*61*:244-6).

(*ii*) An alternative method used by J. Chatham in an experimental adjustment of the British Offices' Annuitants' Experience was to graduate $q_{[x]}$ first, and then for each value of x to graduate the series $q_{[x]}$ graduated, $q_{[x]+1}$, $q_{[x]+2}$, . . . ungraduated (H:*88*).

(*iii*) As noted at p. 296; C; 11 here, Jastremsky concluded in 1912 from an investigation of Austro-Hungarian mortality that the ratio $\frac{q_{[x-t]+t}}{q_x^{(5)}}$ might be considered to be independent of the attained age x, which means that $q_{[x]+t}=k_t\,q_{x+t}^{(n)}$ where k_t depends only on the duration t, and n denotes the number of years in the select period. In the Japanese Three Offices' Tables (see P:*74*: 103) it was similarly thought justifiable to assume that $q_{[x]}=.62q_x^{(5)}$, $q_{[x]+1}=.87q_{x+1}^{(5)}$, $q_{[x]+2}=.95q_{x+2}^{(5)}$, and $q_{[x]+3}=.97q_{x+3}^{(5)}$ and that $q_{[x]+t}=q_{x+t}^{(5)}$ when $t=4$ or more. The rates during the 10-year select period of the O$^{[M]}$ experience, and in other data also, likewise approximate closely to the same type of relationship except for the first year of duration (P:*78*:332 and 336). George King therefore suggested (P:*78*: 289-292, et seq.) that a short and often satisfactory method of determining the graduated values during the select period would be simply to apply such factors, k_t, to the graduated ultimate rates—the values of k_t being adjusted, graphically or otherwise, as might seem necessary.

(*iv*) Another ratio process (P:*169*:136) employed in the 1926 Life Tables of the German Life Assurance Companies was first to graduate the ultimate rates of mortality, q_x, and also $q_{[x]}$; then the data for each separate year of duration were divided into five sections, each embracing 9 ages attained, from age 20 to age 64, and interpolation factors $\lambda^a{}_{[x-t]+t}$, where a denotes the number of ages in the group, were computed as

$$\lambda^a{}_{[x-t]+t}=\frac{\sum^a E_{[x-t]+t}\,q_x-\sum^a E_{[x-t]+t}\,q_{[x-t]+t}}{\sum^a E_{[x-t]+t}\,q_x-\sum^a E_{[x-t]+t}\,q_{[x]}};$$

these factors were next graduated horizontally and vertically; and finally they were calculated for each age, after the average age to which they applied had been determined by weighting.

(c) *With Finite Difference Interpolation Methods*

The use of elementary finite difference formulae to obtain smooth junctions between the various sections of the data was discussed by George King (P:*78a*:108 et seq.—see also P:*127*:247) as a development of his method already noted in (*b*) (*iii*).

(d) *With Linear Compounding (or Summation) Methods*

A useful device for the graduation of select tables by these methods is described conveniently in P:*59*:62.

(e) *With Difference-Equation Methods*

A process based on the simplest form of the Whittaker-Henderson A method is given by Wells in P:*153*.

(X) Conclusion

From the number and variety of the methods and formulae noted in this outline the student will readily appreciate that there is no "royal road" to knowledge and proficiency in the field of graduation. The selection of the process to be tested in any particular case will depend partly on the character and extent of the material, which obviously may eliminate certain procedures as inapplicable or clearly inadvisable; the method should be determined also with due regard for the practical situations in which the graduated results will be employed; and it is not scientifically inappropriate to add that the personal preference of the graduator may be allowed to have some influence upon his choice. All these considerations, however, must in the end be circumscribed by the requirement that the numerical results must satisfy those tests of fit and smoothness which shall be thought suitable.

The student accordingly should not accept any argument—however persuasive—or any opinion—from whatever source—which claims preeminence or universality for any special process. No single method has yet been proposed which can rightly claim an ability to reach a best, or a satisfactory, result under all circumstances, or even under most; and no such method is likely to be proposed. The student of graduation should therefore be very careful to preserve an impartial and wide outlook, so that under any set of practical conditions he may select, apply, and test those methods which, on close consideration, appear to him most suitable for dealing with the case in hand.

SECTION
A

HISTORY

A;1. Stirling's Formula

Stirling's formula (9) was published in 1730 (H:*6*:137). De Moivre in 1718, in the first edition of his "Doctrine of Chances" (H:*5*), nearly reached the formula; in his second edition (1738) he credited the completed expression to Stirling, who had obtained it by using Wallis' formula (P:*80*:67, and P:*36*:90) which gives a relation, in the limit, between π and factorials (H:*2*). The usual proof, which is lengthy, need not be reprinted here; it may be found in P:*155*:138, P:*80*:65, or P:*146*:349, while a shorter demonstration by Cesàro, based on inequalities, is given in P:*36*:93 (see also H:*132*).

Since the relative error involved by dropping the term $\frac{1}{12n}$ is only about $\frac{8}{n}$ per cent, the formula is very generally used simply as $\sqrt{2\pi n}\, n^n e^{-n}$. This elegant expression shows remarkably accurate results, even for small values of n—for example, its value when $n=5$ is 118.1, whereas $5!=120$; when $n=10$ it gives 3,598,699 in comparison with $10!=3,628,800$.

A;2. The Discovery of the Normal Curve

Until recently the names of Bernoulli, Laplace, and Gauss, either separately or in combination, have been associated with the functions (10) and (11). In 1924, however, Karl Pearson (P:*99*) announced his discovery of De Moivre's remarkable "Approximato ad Summam Terminorum Binomii $\overline{a+b}\vert^n$ in Seriem Expansi" (H:*7*), which was published in 1733. In that work—of which two copies only are extant—the first statement of the Normal Curve is clearly given (see P:*149*:13-18, P:*116*:47, and H:*164*:566 for the complete English version).

A;3. History and "Proofs" of the Normal Curve

This Normal Curve, and the Theory of Errors and Method of Least Squares (to be discussed later) which resulted from it, presented an effective process (highly systematized by Gauss especially, and by astronomers and physicists in several countries) for dealing with the "errors" of unbiassed observations made by precision instruments and skilled observers. The technique so developed still has that same utility.

In view of this wide use in an era of expanding mathematical research, it is not surprising that the genesis and symmetry of the curve led for many years to a search for other proofs which might release the derivation from any dependence upon the *a priori* supposition of repeated trials, and would instead permit the formulation to be based upon the concept of "errors of observation", measured from an average or mean, as such "errors" might occur in nature. Some of the "proofs" were quite unsatisfactory, others were more rigorous; all, however, are interesting still in the history of thought, and should not be treated lightly by any student who wishes to grasp thoroughly the full significance of the development of more recent methods.

The search for a convincing "proof" was accompanied often by the belief that all mass phenomena would show variations occurring symmetrically about a mean, and that the Curve of Error (11) was the expression of that fundamental law of nature. It was not realized for many years that the quest for uniformities in nature is not necessarily predicated best on an essential symmetry; nor was it appreciated fully that the Normal Curve would fail as a mode of representing many types of chance distributions. When, however, attempts were made, by Quetelet and others, to apply the method to problems of biology and sociology, it was found that the statistics frequently exhibited a defiance of the "Normal Law". At first this was thought to be merely the result of paucity of data; but later it was recognized to indicate the existence of skew variation even in ample and homogeneous material. When these facts did finally appear they led quickly to the development of unsymmetrical (skew) curves in addition to the elegant symmetry of the "Normal Law", and thus created a much wider and more general theory (see also P:*149*:17 and 28-49, and P:*36*:149 and 179-181).

The very great importance which was thus attached to the discovery of the "Normal Law", and the numerous attempts to place it upon a logical foundation, are illustrated by the following summary of the "proofs", "explanations", and discussions which were published by many of the greatest mathematicians between 1733 and 1872.

(i) In 1733 the original discovery was made by De Moivre (see p. 151; A; 2) although it does not seem to have attracted any attention, and later investigators proceeded without any knowledge of it.

(ii) In 1778, Laplace appears to have recognized the existence of the formula in an evaluation of $\int_0^\infty e^{-t^2}dt$ (see P:*149*:18 and 20), and in his later work showed clearly that he had anticipated Gauss (P:*80*:8). His method (see H:*15*, and the English translation in H:*164*:588), which was based largely on the view that an "error" may be supposed to be produced by the combination of a vast number of very small independent "elementary errors", and on the principle of minimizing the mean value of the "error" thus committed, was improved and extended by many subsequent writers—notably Poisson (H:*19*) and Ellis (H:*25*)—and is well re-stated by Airy (H:*32*:7), Glaisher (H:*46*), Todhunter (H:*33*:464), Whittaker and Robinson (P:*155*:168-173), and partly by Arne Fisher (P:*36*:197). Notwithstanding their profoundly important character, to which all later investigators, to this day, pay tribute, the great difficulty of mastering Laplace's intricate analyses caused his methods for many years to be largely superseded by the less imaginative procedure of the German Gauss.

(iii) Robert Adrain in 1808 was led independently to its discovery. The two deductions which he gave (H:*13*), however, cannot be considered satisfactory (see H:*46*, and P:*149*:20 and 94). The second was essentially that given much later by Herschel (see (vii) here), and is open to the same question with regard to the validity of assuming the probabilities of the x and y deviations as independent (H:*53*).

(iv) In 1809, Gauss—acknowledging his indebtedness to Laplace (see P:*149*:22)—published his first "Theoria Motus" (H:*14*:§177) derivation (for which see P:*155*:218, P:*80*:118-120, and P:*13*:22), involving the "postulate of the arithmetic mean", i.e., that when any number of equally good direct observations of an unknown quantity x have been made, the "most probable" value is their arithmetic mean.

The precise extent to which Gauss really depended on this postulate, and the postulate itself, have been subjected to critical discussion for many years. In 1834, Encke attempted unsuccessfully to prove it (H:*21*). In 1844, Ellis (H:*25*), supposing that a is the true value, and $x_1, x_2, \ldots$ the n observed values with errors $e_1, e_2, \ldots$, so that $x_1 - a = e_1$, $x_2 - a = e_2$, etc., pointed out that the rule of the arithmetic mean is easily deducible from very simple *a priori* considerations—for if, in the long run, there is no permanent cause tending to make the sum of the positive differ from that of the negative errors, then $\Sigma e = 0$, that is $\Sigma(x-a) = 0$, or $a = \frac{1}{n} \Sigma x$, namely, the arithmetic mean. But he also drew attention to the important fact that these suppositions are in reality too simple—for, instead of only $\Sigma e = 0$, we could have $\Sigma f(e) = 0$ where $f(e)$ is any function such that $f(e) = -f(-e)$, that is to say we could have $\Sigma f(x-a) = 0$—"and no satisfactory reason can be assigned why . . . the rule of the arithmetic mean should be singled out from the other rules which are included in the general equation $\Sigma f(x-a) = 0$", for "we are perfectly sure that in different classes of observations the law of probability of error must vary". He laid stress also on the ambiguity of the words "most probable" in Gauss' treatment—remarking that "there is no reason for supposing that because the arithmetic mean would give the true result if the number of observations were increased without limit, it must give the 'most probable' result, the number of observations being finite", for "by losing sight of this distinction we are led to the inadmissible conclusion that a principle recognized as true *a priori* necessarily implies a result, viz., the universal existence of a special law of error, [which is] not only not true *a priori*, but not true at all" (cf. also his further remarks in H:*27*).

Glaisher remarked in 1872 that "what was in effect Gauss' view, viz., that the arithmetic mean is *practically* the best mode of combining simple observations . . . , was quite reasonable and consistent—but he was very far from asserting that the arithmetic mean is the most probable value of the quantity observed" (H:*46*).

More recently the Italian mathematicians have re-examined the axiomatic foundations from which the postulate may be deduced (see P:*155*:215-7), while for certain types of observations it has been shown to be invalid (see P:*155*:217-8). It must be added that Gauss himself did not present his method as "other than tentative and hypothetical" (H:*42*), and that subsequently he expressed a preference for his second proof in (v) below (see P:*155*:224 and 228).

(v) Although it concerned more directly the establishment of the "Method of Least Squares", Gauss' "Theoria Combinationis" (H:*17*) investigation must be included here. Following largely the methods of Laplace, and thus abandoning the necessity of introducing the "postulate of the arithmetic mean", he based his treatment on the principle of minimizing the probable value of the square of the error (instead of Laplace's postulate that the importance of the error is proportional simply to its magnitude)—see H:*25*, and P:*155*:227-8. In illustration of the various opinions which these several treatments have evoked, it may be noted that Ellis (H:*25*) observed that "nothing can be simpler or more satisfactory" than Gauss' "Theoria Combinationis" demonstration; Glaisher, however, differed, saying (H:*46*) "with this remark I cannot at all agree"; Merriman (H:*53*) and others have expressed the opinion that, in its relation to the Method of Least Squares, "it is but little more than a begging of the question to assume that the mean of the squares of the errors is a measure of precision"; and Crofton (H:*42*:183) stated that "it is of infinitesimal importance whether, with Laplace, we estimate the importance of an error by its mean value (irrespective of sign), or, with Gauss, by its mean square"—for both approaches lead to the same result (see P:*155*:227-8).

(vi) In 1837, Hagen (H:*22*), again basing his treatment on the assumption that an "error" may be viewed as the sum of a large number of infinitesimally small errors, deduced the Normal Curve by a method which has been adopted for text-book presentation by Merriman (P:*90*:17) and Brunt (P:*13*:11)—the latter giving also a generalized proof due to Eddington (P:*13*:15).

(vii) In 1850, Sir John F. W. Herschel put forward independently a different form of demonstration (H:*28*), apparently without the knowledge that Adrain had employed essentially the same method in 1808 (see (iii) here). Supposing "a stone to be dropped with the intention that it shall fall upon a given mark," any "deviation from this mark is error" which may be expressed by the function $f(r^2)$ or $f(x^2+y^2)$. In his first presentation he proceeded by means of an assumption to which Ellis objected strongly in a paper (H:*27*) written primarily as a criticism of Herschel's suggestion, namely, that an observed deviation, being equivalent to two deviations parallel respectively to the co-ordinate axes, "is a compound event [so that] its probability will be the product of their separate probabilities". Ellis' criticism was that "it is not true that the probability of a compound event is the product of those of its constituents unless the simple events . . . are independent of one another, and there is no shadow of reason for supposing that the occurrence of a deviation in one direction is independent of that of a deviation in another". Boole, however (in the Trans. Royal Soc. Edinburgh, XXI, 628), attempted to relieve the demonstration of this difficulty by asking at the outset that "it be assumed that the actual deviation is a compound event", and that the two component deviations "are independent events"; and Herschel himself later (in 1857, in a footnote to a reprint of the greater portion of his article in J.I.A., XV, 179), took the position that "the increase or diminution in one [component] may take place without increasing or diminishing the other"—adding that "on this, the whole force of the proof turns". Again in 1870, Crofton (H:*42*) characterized these assumptions as "bold", and held that the method "can hardly be seriously viewed as a demonstration", while Glaisher (H:*46*) referred to "the unwarrantable character of the assumption of equally probable x and y deflections". It has, however, been used in a number of text-books, with further explanations as to the import of the assumptions made—e.g., in Thomson and Tait's "Natural Philosophy", and by Brunt (P:*13*:14), Levy and Roth (P:*80*:121), and Scarborough (P:*122*:288).

(viii) Sir George F. Hardy, in P:*51*:5, has more recently suggested that "we may, perhaps, see a logical basis" in the supposition that a "deviation" (from the mean) results from an infinity of minute superimposed causes, of which the nature is unknown although "any one [of them] may produce a minute positive or negative deviation from the average . . . [and] we may without loss of generality assume them of equal magnitude". Then, "if the number of possible causes of deviation is $2n$, and if the extent of each indefinitely small deviation is k (n being indefinitely large, but $k\sqrt{n}$ finite), the probability or frequency of a total deviation lying between x and $x+k$ will depend on our having $\left(n+\frac{x}{2k}\right)$ positive values of k and $\left(n-\frac{x}{2k}\right)$ negative values. The probability of this occurring will be represented by the appropriate term in the expansion of the binomial

$(\frac{1}{2}+\frac{1}{2})^{2n}$ or $\dfrac{2n!}{\left(n+\frac{x}{2k}\right)!\left(n-\frac{x}{2k}\right)!}(\frac{1}{2})^{2n}$ . . . ." This expression,

n being indefinitely great, takes [by Stirling's Theorem (p. 151; A; 1)] the form $\dfrac{1}{(n\pi)^{\frac{1}{2}}}e^{-\frac{\left(\frac{x}{2k}\right)^2}{n}}$, which is the Normal Curve (11) when c is written for $\sqrt{n}$ and x for $\dfrac{x}{2k}$.

The intellectual struggle which revolved about the validity and consequences of the "postulate of the arithmetic mean", and the attempts to establish the universality of the Normal Curve, comprise a most illuminating chapter in the history of philosophic thought and the gradual evolution of critical analysis. It will repay any student well to picture De Moivre—living in London as a fugitive from France—enunciating, with then no recognition, his remarkable discovery; to realize the extraordinary elegance of Laplace's analytic power, and the stimulus of his great mind which shows so clearly in the work of Poisson and the later Frenchmen; to follow the careful and systematic Gauss—born humbly, like Laplace—and the contributions also of his German students and compatriots—Bessel, Encke, and Hagen; to evaluate the interpretations of the English school—Ellis and Glaisher, the critical De Morgan, and Todhunter; and to appreciate the great practical importance of the sequelae and wider theories which in later years have been developed by the Russian, Scandinavian, and English writers—indebted still, as they must always be, to Laplace and Gauss above all others. Much will be lost by any student who essays now to grasp the technique of this subject without a close examination of its history. The philosophic contemplations—even the religious questionings—by which its evolution has for so long been accompanied have represented largely the soul of man's search for understanding. The material assembled here has accordingly been given in the hope that even so mathematical a subject may continue to challenge the reader with something of its romance.

A;4. Early Attempts to Establish a "Law of Error"

It may be of interest to note that the varied deductions of the Normal Curve by De Moivre, Laplace, Gauss, and others, as set out in A; 3, were not accomplished without other suggestions being made which led to entirely different formulae.

It must be realized, in the first place, that the problem is, fundamentally, that of finding an equation representing the actual error resulting in a function, in terms of the component errors to which its several independent variables may be subject. Mathematically, if a quantity Q (such as a rate of mortality) be a function $f(v, w, z \dots)$ of several independent variables $v, w, z, \dots$ (such as age, height, domicile, . . . , here assumed to be independent), which are subject respectively to component errors $\Delta v, \Delta w, \Delta z, \dots$, then the resulting error, say ΔQ, in Q will be given by $f(v+\Delta v, w+\Delta w, z+\Delta z, \dots)-f(v, w, z, \dots)$, that is, by $\frac{\partial Q}{\partial v}\Delta v+\frac{\partial Q}{\partial w}\Delta w+\frac{\partial Q}{\partial z}\Delta z+\dots$ (from Taylor's Theorem) if the component errors are sufficiently small that the terms involving their squares and the differential coefficients of second and higher orders may be neglected. The "law of error" thus emerging, say $y_x=\varphi(x)$, expressing the relative frequency of the occurrence of the resultant error x, or ΔQ, in Q, will obviously take a form which is dependent upon the laws of propagation of the various component errors.

With this principle clearly in mind, therefore, it is not surprising that very early—in 1755 and 1757—Thomas Simpson (H:*9*), assuming simply that errors of different magnitudes are equally probable (as in the case of the errors of tabular logarithms, etc.), so that $\varphi(x)=c$, found in that case a rectangle (see P:*173*:32). He likewise proposed an isosceles triangle when the probability of a positive error is $\varphi(x)=-mx+c$, and positive and negative errors are equally likely—a case discussed also by Lagrange (H:*33*:309, and P:*155*:167).

In 1778 Daniel Bernoulli, objecting to the common use of the arithmetic mean as assuming that all the observations are of equal weight, supposed that small errors are more probable than large ones, and proposed $\varphi(x)=+\sqrt{(r^2-x^2)}$ where r is a constant—thus reaching a semi-circle (H:*33*:237, and P:*173*:33). Laplace himself brought out and then discarded several others, such as $\frac{1}{2a}\log_e\frac{x}{a}$, and $\frac{k}{2}e^{-k|x|}$ where $|x|$ indicates that x is always taken positively. This last expression, suggested in 1774, is usually referred to as *Laplace's First Law of Error*, and results from assuming that the median (the middle value in a sequence arranged in order of magnitude), rather than the mean, is the most probable value of the unknown quantity (see P:*155*:188 and P:*13*:27). Examples have been given very recently (P:*157*) of certain data thus distributed, with the suggestion that the applicability of the curve may be tested by plotting on "semi-logarithmic" paper (sometimes called "arith-log" or "ratio" paper, on which a plotting gives x and $\log y$ instead of x and y), for then the points by this formula lie on a straight line. It has also been noted by P. R. Rider (P:*109*) that this "First Law" of Laplace is the case when $m=1$, and the Normal Curve when $m=2$, of a "generalized law" $\dfrac{mk^{\frac{1}{m}}}{2\Gamma\left(\frac{1}{m}\right)}e^{-k|x|^m}$, although in practice "it would doubtless be sufficiently accurate in most instances to use $m=1$ or $m=2$".

Under yet other circumstances, such as where the quantity Q, or $f(v, w, z, \ldots)$, is known to be determined by some definite relation, for example, $Q = kv + mw^n$, it is clear that the law of the propagation of errors in Q is fixed by the laws governing the generation of errors in v and w. It must be realized, therefore, as Woodward states succinctly (P:*178*:31), that "every investigator in work of precision should have a clear notion of the error-equation $\left[\text{of the type } \Delta Q = \frac{\partial Q}{\partial v}\Delta v + \frac{\partial Q}{\partial w}\Delta w + \ldots\right]$ appertaining to his work—for it is thus only that he can distinguish between the important and unimportant sources of error". It will also be evident from these considerations why it is, when the quantity Q is related to a *very large number* of independent variables, instead of only two in such a case as $Q = kv + mw^n$ above, that the resultant errors in Q may be visualized as fortuitous—for then they can hardly be determined by fixed conditions, and indeed in many cases their propagation can only be imagined as the combined operation of innumerable small influences of closely equal effect, for each of which some reasonable assumption, in the absence of knowledge, would appear to be legitimate. This, of course, is the hypothesis by which the Normal Curve is approached in many of the deductions given in A; 3. The train of thought here indicated may also emphasize the importance of the statement often made, that the failure of some observed functions—particularly in economic and sociological data—to follow the Normal Curve is probably due largely to the predominance of relatively few influences out of the many which determine the values of the function.

A; 5. Tabulations of the "Probability Integral" or "Error Function"

Because of its great importance in the theory of errors of observation, numerical tables of the area of the *Probability Integral* or *Error Function*, $\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$, and of the ordinates, have been prepared by many computers. For an account of the various tables which have been published since Kramp's first tabulations of 1799, see P:*149*:58.

The methods of computation depend on the expansion of e^{-t^2} and its integration for small values of x, and on an asymptotic series when x is large (see P:*155*:179, P:*13*:19, and P:*122*:297).

The tables are to be found in several different forms (see P:*114*:14), which must be watched carefully in practical applications. One method arises by substituting $c=\sqrt{2npq}=\sigma\sqrt{2}$, so that (11) becomes $\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$. Another form frequently adopted results from using σ as a unit of measure, thus obtaining $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$. Convenient arrangements of extensive values have been given in comparatively recent years by W. F. Sheppard (P:*97* and P:*129*), and by J. W. Glover (P:*47*:392). Actuaries will find short tables of the areas easily accessible in P:*174*:50 and P:*51*:138, and in most text-books on statistics, while the ordinates may be found in P:*27*:397 and P:*16*:384. The following specimen values of $\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt=\text{Erf}(x)$ will give an idea of the areas:

x	Erf(x)	x	Erf(x)	x	Erf(x)
0.10	0.11246	1.10	0.88021	2.10	0.99702
0.20	0.22270	1.20	0.91031	2.20	0.99814
0.30	0.32863	1.30	0.93401	2.30	0.99886
0.40	0.42839	1.40	0.95229	2.40	0.99931
0.50	0.52050	1.50	0.96611	2.50	0.99959
0.60	0.60386	1.60	0.97635	2.60	0.99976
0.70	0.67780	1.70	0.98379	2.70	0.99987
0.80	0.74210	1.80	0.98909	2.80	0.99992
0.90	0.79691	1.90	0.99279	2.90	0.99996
1.00	0.84270	2.00	0.99532	3.00	0.99998

Since it is frequently desirable to have a ready means of referring to the "error function" for particular values of the limit x, the custom has arisen of using Erf x or Erf(x) to denote

$\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt.$ Thus $\text{Erf}(\infty)=1$; $\text{Erf}(0)=0$; formula (16) is $\text{Erf}\left(\frac{\lambda}{c}\right)=.5$; the probability that a deviation will lie in the range $\pm\sigma$ is, by (11), $\frac{1}{c\sqrt{\pi}}\int_{-\sigma}^{+\sigma} e^{-\frac{x^2}{c^2}}\,dx$, which, by putting $\frac{x}{c}=t$, becomes $\frac{2}{\sqrt{\pi}}\int_0^{1/\sqrt{2}} e^{-t^2}\,dt = \text{Erf}\,(1/\sqrt{2}) = .6827$; and so forth. Similarly, the probability that a deviation will not exceed $\pm d$ (i.e., d in absolute magnitude) is

$$\frac{1}{c\sqrt{\pi}}\int_{-d}^{+d} e^{-\frac{x^2}{c^2}}\,dx = \frac{2}{\sqrt{\pi}}\int_0^{\frac{d}{\sigma\sqrt{2}}} e^{-t^2}dt = \text{Erf}\left(\frac{d}{\sigma\sqrt{2}}\right);$$

consequently, the complementary probability that a deviation will not be less than $\pm d$ (i.e., will be as much as, or more than, d in absolute magnitude) is $1-\text{Erf}\left(\frac{d}{\sigma\sqrt{2}}\right)$.

It is likewise sometimes convenient to use $\text{Erf}^{-1}(x)$ as an operator (like $\sin^{-1}x$), to be read as "the number of which the error function is x."

A; 6. The Nomenclature Associated with the Normal Curve, and its History

Considerable variation, unfortunately, is to be found in the English terminology associated with the Normal Curve, and the student therefore must be careful to identify the names used in any of the classical—and even in some of the modern—works which he may peruse. Detailed examinations of the discrepancies, and the origins of the terms, are given in P:*149*:49 and 175.

Attention may here be drawn to the following matters in particular:

(i) Although the "standard deviation" is the usual name for σ —the square root of the mean of the squares of the deviations (or "errors") measured from the mean—many Continental writers (following Gauss) refer to it as the "mean error", which is certainly not descriptive, and creates confusion with the "average or mean error (irrespective of sign)" of formula (13).

(ii) The "standard deviation" was called by Lexis the "dispersion".

(iii) The term "variance" for the mean square deviation—being the square of the standard deviation—is now used extensively by R. A. Fisher and his followers.

(iv) Brunt (P:*13*:29) adopts a confusing procedure in using "mean square error, or M.S.E." for the standard deviation.

(v) The "probable error"—unquestionably a misleading phrase—is gradually being supplanted, although it is still widely used.

(vi) The parameter c in (11) has been called by some writers (notably Airy) the "modulus".

(vii) By putting $\frac{1}{c} = h$, (11) becomes $f(x) = \frac{h}{\sqrt{\pi}} e^{-h^2x^2}$, in which h is often called the "precision" (particularly by astronomers and physicists) since the cluster of the values about the mean becomes closer as h increases.

(viii) The "weight" is usually defined, in the classical texts, as the reciprocal of the square of the probable error—or, as in this study (see (107) and (108) in Chapter VIII) as the reciprocal of the square of the modulus c, the probable error λ, or the standard deviation σ. Woolhouse (P:*174*:46) and G. F. Hardy (P:*51*:118), however, define the weight as the reciprocal merely of the probable error (not of its square)—a procedure which must be remembered carefully in connection with their development of the "normal equations" in the Method of Least Squares (see p. 323; C; 20).

In this study the terms adopted are those now most generally employed in the literature with which actuaries are mainly concerned.

A; 7. The Lexis Theory

The publication (H:*52*) by Lexis in 1877 of his method of analyzing "dispersion" (as he called the standard deviation) forms an outstanding landmark in the development of mathematical statistics. It is of interest for actuaries to note that his theory was devised with particular reference to variations in mortality and in the sex ratio at birth.

Charlier, of Sweden, illustrated the fundamental Lexis theory and the "Lexis ratio" by a wide variety of card-drawing experiments (see P:*36*:137), and developed the "Charlier Coefficient of Disturbancy".

Since their original statement the ideas of Lexis have undergone much elaboration and many refinements, which are now embodied in the modern techniques associated with the χ^2 test (see Chapter IX) and the "analysis of variance" (see Chapter X).

A; 8. Bessel's Formulae

The use of the factor $\frac{n}{n-1}$ in (42) is generally referred to as "Bessel's Correction", after the astronomer Bessel (1784-1846), who in 1815 originated the term "probable error" and contributed largely to the theory of errors and least squares (see P:*149*:24 and 186). The precise history of its first derivation has sometimes appeared doubtful, and it has been ascribed variously to Bessel, Gauss, and Encke (P:*29*:116 and 145). It now appears to be clear, however, that the method should be credited to Gauss, since he set forth the correction in H:*17*: art. 38 (see P:*28*:18).

In the classical presentations the formula is often stated in terms of the probable error rather than for σ^2 (see, for example, P:*155*:206). The student will also meet there a corresponding expression in terms of the deviations without regard to sign, which is known as *Peters' formula* (see P:*155*:206, and P:*13*:38).

The denominator $(n-1)$ in reality allows for the loss of one "degree of freedom" in thus estimating σ_c^2 from the data. The comparable formulae when there are k unknowns (see p. 175; A; 18 and p. 250; B; 26) which are treated in the classical discussions of the Method of Least Squares will be found similarly to employ a denominator $(n-k)$ when there are k "constraints" (P:*90*:59, 82, and 86, and P:*28*:18). The principle involved in the concept of degrees of freedom is accordingly (as stated at p. 176; A; 19) to be credited in the first instance to Gauss.

A; 9. The Bayes-Laplace Theorem

The "probability of causes" was first examined by an English dissenting minister, Thomas Bayes, F.R.S., whose famous paper (H:*10*) on the subject was submitted posthumously to the Royal Society in 1763 by Dr. Richard Price (well known to actuaries as the author of the Northampton Mortality Table).

The great importance of Bayes' contribution, however, has often been confused through misstatements of its original form, and inadmissible applications. Many text-book treatments, moreover, have failed to recognize the facts that Bayes dealt actually with the special case when the *a priori* existence probabilities, κ_r, are equal, and that it was Laplace who gave (H:*15*) the generalized form when the κ_r's are not all equal (see H:*166*:14, and H:*173*:29 and 32). As in the text here, the name "Bayes' Theorem" should therefore be confined to the former case, with the term "Bayes-Laplace Theorem" as a preferable designation for the latter.

A complete and very convenient reproduction of Bayes' original argument, which followed a geometric process, is available in H:*190*, and partly also in H:*33*:294 and H:*165*:48. Excellent discussions of the misstatements concerning both Bayes' and Laplace's methods, which have become so prevalent, may be found in H:*166*:10-16 and P:*36*:54.

A; 10. Helmert's Distribution of σ_s, and "Student's" Distribution

"Student's" original derivation (H:*117*) of formula (44) was effected in 1908 through an empirical process of first finding the distribution of σ_s by means of algebraic expressions for the first four moments and fitting Pearson's Type III curve. In so doing he was unaware that the distribution of σ_s (given in (2) at p. 225; B; 13) had been obtained by Helmert in 1876 (H:*50*)—this fact not being generally known until Karl Pearson brought Helmert's prior work to light in 1931 (H:*174*).

The first rigorous proof of the distribution of "Student's" z

was given by R. A. Fisher in 1926 (H:*158*), and is followed on p. 226; B; 13 here.

The importance of "Student's" work was not immediately realized. It has now, however, assumed a position of great prominence in the theory of small samples, largely through Fisher's wide extensions.

A memoir of "Student", who was an English statistician, W. S. Gosset, may be found in H:*189*.

A; 11. Poisson's Exponential "Law of Small Numbers"

Although this exponential was discovered by Poisson in 1837, it remained for the Russian Bortkiewicz (H:*75*), in 1898, to draw attention to its importance through an illustration—of interest to actuaries—based on the mortality in the Prussian army from the kicks of horses (see p. 308; C; 14).

For this reason, apparently, it is sometimes referred to as the "Poisson-Bortkiewicz", or even as the "Bortkiewicz", function. It seems much more proper, however, to credit so important a formula only to its discoverer.

Some difference of opinion, likewise, is manifest with regard to the term "Law of Small Numbers"—a phrase suggested by Bortkiewicz; some modern writers have questioned its appropriateness, and have suggested instead the "Law of Small Probabilities". It is to be observed, however, that Bortkiewicz was undoubtedly right in using the word "numbers" rather than "probabilities"—for the function is applicable in particular to circumstances in which both q (or p) and nq (or np) are small, while it is not necessarily preferable to the Normal Curve when only the "probabilities" q (or p) are small without nq (or np) being small as well (see p. 267; C; 4).

A; 12. The Generalizations of the "Normal Law"

The dependence of the normal forms (10) and (11) on a particular method of approximation (see p. 203; B; 5), and their consequent inability to deal with skew distributions, was realized early in the classical literature. The nature of the approximation, for example, was pointed out by Laplace (see H:*33*:548-552) in a form equivalent to the *j* term of Edgeworth's Generalized Law of Error (see H:*107*:329, footnote). Poisson, moreover, and many years later De Forest independently (H:*64*), found the *j* term of Edgeworth's series (see H:*82*:290).

Gram, of Denmark (P:*36*:182-3, and H:*81*), however—under the inspiration of a Danish actuary, Opperman (see P:*32*:130, footnote)—was the first mathematician, in 1879, to represent a general system of skew frequency curves by a series of which the Normal Curve is a special case (H:*59*). Subsequently Thiele—likewise a Danish mathematician, who is known to actuaries also for other contributions (H:*43*, and see H:*81*)—reached Gram's series by the use of his half-invariants (P:*36*:183, and H:*94*). The German astronomer Bruns next gave a comprehensive treatment in H:*102*. More recently Charlier, the Swedish astronomer, has discussed the derivation and application of these series representations in much detail, and has so far systematized their classification and methods of fitting that two general types are now usually named the Gram-Charlier, or Type A, and the Poisson-Charlier, or Type B.

Edgeworth's important work in England appeared between 1883 and 1920 in the midst of these Continental researches, and was characterized by an unusually wide knowledge of the significant steps which had been taken by De Forest in the United States, and by the Scandinavian and German mathematicians. His method of approach, which was always more philosophical than definitely practical, laid great emphasis upon the metaphysical concept of probability, and the necessity of deducing from *a priori* considerations a universal law which would represent the frequency distribution of a magnitude arising from a number of independently varying elements. For that reason, and because his extensive contributions—numbering 74 papers altogether—were scattered in many different scientific journals, his work has been overshadowed by the less philosophical researches of the Scandinavians and the more practical investigations of Pearson and his followers. Edgeworth's writings, nevertheless, constitute an outstanding presentation of a viewpoint of great interest. The main principles of his work should therefore certainly be known—a task which has been assisted notably by Bowley's explanation listed at H:*162*.

A; 13. The Search for a "Law" of Mortality

Perhaps to some extent as the result of an over-estimation of the importance of the hypothetical "life-table" (cf. P:*164*:401, and P:*171*:282), and consequently too little emphasis upon the fact that the function of primary significance in the measurement of mortality must be the rate of mortality, q_x, only (or an analogous function such as the force of mortality, μ_x, or the central death rate, m_x, or colog p_x), the classical literature is replete with many interesting attempts to reach a formula which might represent the values of l_x from infancy to old age, or during a part only of the entire span of life. Unquestionably, also, the belief persisted for many years that a "law" of mortality, exhibiting itself as the ordered progress of the life-table's theoretical cohort of l_x persons marching through time, would ultimately be revealed to some lucky or diligent enquirer—a "law" which again would confirm the faith of those deeply religious philosophers who sought uniformities in Nature as a manifestation of the Divine Will. Nor was this as fanciful in those days as it might seem now; for scientists were discovering the fundamental "laws" of physics and astronomy, and everywhere the destinies of men as well as of matter were contemplated as probably resulting from the unwavering influences of an Unseen Power. The failure to discover so inflexible a "law" for the mortality of the human race merely brought gradually to light the unfathomable complexities which may at any time determine the mortality experience of an isolated group; it could not of course disturb the faith with which those earlier investigators conceived of the whole Universe proceeding to its destiny upon a "law" of progress.

This search for a "law" of mortality was, of course, supported by many philosophical and analytical discussions of the features which such a law might be expected to exhibit. Stimulating though they are, it is hardly necessary to give here any summary of those speculations. The main references for the assistance of any interested reader, however, would be H:*43*; H:*141*; H:*18*; H:*31*; H:*69*; H:*47*; H:*74*; H:*169*; H:*65*; and P:*102*.

A; 14. The Verhulst-Pearl-Reed (the "Logistic") Curve of Population Growth

The influences which may be expected to direct the growth of populations have naturally been the subject of speculation and enquiry ever since the rudiments of a scientific approach to current problems first made their appearance. The original edict of Malthus (H:*11*)—to the effect that a population unaffected by migration would soon face starvation if it grew in geometrical progression while its food supply increased only in arithmetical progression—is of course now obvious enough; in his day it was sufficiently alarming, and ever since that time it has re-appeared constantly in popular discussions of the world's politico-economic problems. The Malthusian principle, however, did not lead to any constructive mathematical formulation. That was, some years later, the accomplishment of the Belgian Verhulst, for in 1838 (H:*23*) he reached a curve of type (102) with $A=0$, and in 1845 (H:*26*) named it the "logistic". This important contribution lay dormant, however, until the curve was re-discovered independently by Pearl and Reed in 1920 (H:*133*)—Verhulst's prior work not coming to their knowledge until 1922 (see P:*176*:5 and P:*96*:569).

A; 15. The Development of Curves for Forecasting Mortality

For many years it was the custom to base the calculation of mortality rates for any experience upon the age x as the sole variable, so that the function investigated was q_x—the rate of mortality in the year of age x to $x+1$. Later the statistics derivable from the records of insurance companies showed that where "selection" can be exercised with respect to the group of lives under observation (whether by the medical examination of applicants for insurance, or the self-selection of buyers of annuities) the rates of mortality in the year of age x to $x+1$ must be denoted as $q_{[x-t]+t}$ and viewed as a function of two variables—the age at entry, $[x-t]$, and the years of duration since entry, t. During about the last fifteen years the realization has been growing that a progressive change has been occurring in these rates, and that consequently it may often be necessary to take into account a third variable—either the calendar year, z, in which the rates of mortality occurred, or the year of birth of the "generation" from which they are derived.

When rates of mortality are thus analyzed by age and generation, and if a change in the rates with time is thus revealed, it evidently may become important to "forecast" the rates which may be anticipated in the calendar years to come. A number of tentative forecasts based on graphic methods or the use of straight lines or parabolas have been published by Scandinavian and German investigators since Gyldén's (H:*57*) first effort in 1875 (see P:*23*). Knibbs' advocacy of "fluent" or "projected" life tables appeared in 1917 in his remarkably exhaustive work published as an Appendix to the reports on the Australian census of 1911 (H:*126*:380). The problem has received serious attention, however, only since the compilation of the British Government and British Offices' annuitants' experiences of 1900-1920 (H:*151*, and H:*142*—see also H:*157*). A very useful summary of the various formulae which have been employed is to be found in P:*23*:165. Reference may also be made to Greenwood's paper P:*49*.

A; 16. The History of the Method of Least Squares

The problem of solving ν observation equations for unknowns numbering less than ν seems to have engaged the attention of astronomers since about 1750. The principle of "least squares" as a means of accomplishing this solution appears to have been used originally by Gauss in 1795 (H:*164*:576). It was named and first published in 1805, however, by the French mathematician Legendre (H:*12*)—an English translation of his exposition being available in H:*164*:576.

The appearance of this important method led quickly to an enormous literature (see H:*53*, and P:*90*:213), particularly in Germany and France, where Gauss, Encke, Hagen, Bessel, Helmert, Laplace, and others explored its theoretical foundations and its applicability to astronomical problems, and examined fully its relation to the postulate of the arithmetic mean and the "curve of errors" (cf. p. 151; A; 3). The debt which the mathematicians of today thus owe to Gauss in particular, however, has not always been fully recognized (cf. p. 164; A; 8, and p. 176; A; 19); indeed it may be well here to quote Deming's sound opinion that Gauss "accomplished most of what we think we know now about least squares" (P:*28*:1). Later Kummel (H:*60*), whose work has been overlooked until recently, in 1879 "filled in a good share of what was missing", as Deming (loc. cit. pp. 2, 42, and 47) has also pointed out, and papers by Stewart (H:*134*) and Uhler (H:*143*) have clarified greatly the questions which arise when the independent variable x as well as the dependent variable y is subject to error (see P:*122*:380, and P:*28*).

A; 17. Early Methods of Fitting Curves, and Methods other than those of Least Squares, Moments, and Minimum-χ^2

A brief summary of the three following methods, with references, is given in P:*155*:259.

(*i*) The method apparently first used by Tobias Mayer as early as 1748, by which the equations of condition are merely summed in sets, was later used extensively.

(*ii*) A method of "minimum approximation", the object of which is to determine the smallest value which can be assigned to the absolute value of the greatest discrepancy between the data and the fitted function, was solved by a laborious process by Laplace in 1799, named by Goedseels in 1907, and solved more easily by C. J. de la Vallée-Poussin in 1911 (see also P:*121*: 14).

(*iii*) Edgeworth's method, which was suggested in 1887, would minimize the sum of the absolute values of the residuals, and thus is based on "Laplace's First Law of Error" (cf. p. 159; A; 4), i.e., on the hypothesis that the error function is of the form $y = \frac{k}{2} e^{-k|x|}$ where $|x|$ is taken positively in both directions, so that the median is the most probable value. The solution for the case of two unknowns is explained and illustrated in H:*162*: 103-109; when there are more than two unknowns, however, it is difficult to minimize the sum of the absolute values of the residuals, because, when all the terms are to be taken as positive, it is not known until the work is completed which are naturally positive or negative (see H:*87*:49-50).

Another proposal has been advanced more recently by H. S. Will (H:*170*), and is called by him descriptively a "method of mean difference functions", or more shortly a "method of differences". Arguing that the criterion of least squares is not entirely satisfying in respect of "data in which the errors of observation are small in comparison with the analytic deviations from trend", and hence discarding the least squares process of minimizing the sum of the squares of the residuals, he remarks that his method gives a sum of the absolute values of the residuals less than those resulting from least squares or moments, although it does not rigorously satisfy Edgeworth's desideratum that the sum of the absolute values should be a minimum. The procedure—for which simplicity of computation and satisfactory practical results are claimed—is to develop a series of expressions for each parameter by differencing, and thence to take the mean value. For example, in the fitting of a straight line, where α and β are to be determined from a set of ν observation equations $f'_x = \alpha + \beta x$ for $x = 1, 2, \ldots, \nu$, the increment $k\Delta x$ in x (where k is settled from a suggested rule) would give an increment in f'_x, say $\Delta_k f'_x$, of $\beta k \Delta x$; hence $\beta = \frac{\Delta_k f'_x}{k\Delta x}$; proceeding thus for each of the $\nu - k$ such increments, the $\nu - k$ values are obtained for β, and the value sought is to be taken from the mean of all those values as $\frac{\sum_{x=1}^{x=\nu-k} \Delta_k f'_x}{(\nu - k) k\Delta x}$; and α thereafter would be found similarly. In the paper referred to (H:*170*) the expressions by this method are also given for fitting polynomials, hyperbolic, logarithmic, and exponential series, and the logistic curve. The process, however, is of course arbitrary, even though in some cases it may give reasonable practical results.

Will's method is closely analogous to a procedure which has in fact been used for many years. Woolhouse, for example (H:*40*:403), in determining the constants of Makeham's formula (83) in the logarithmic form $\log l_x = \log k + x \log s + c^x \log g$, differenced twice over interval t, whence (cf. P:*59*:92) immediately $c^t = \frac{\Delta^2 \log l_{x+t}}{\Delta^2 \log l_x}$; proceeding similarly for several values of x he then took the mean of the resulting values of $\log c$, and based the final values for the other unknowns on mean values also. Woolhouse's method was extended later by King and Hardy (H:*62*:200—see also P:*59*:94, and H:*85*:81 for a complete numerical illustration) into one based on the use of four sums (instead of single values) of $\log l_x$ from a to $a+t-1$, $a+t$ to $a+2t-1$, $a+2t$ to $a+3t-1$, and $a+3t$ to $a+4t-1$, by which the subsequent construction of the mean values was avoided.

This employment of sums over consecutive portions of the data has also been used in Cantelli's "method of areas" (H:*99*), except that he bases his procedure on integrals instead of finite sums (see H:*101*, and H:*113*:444).

Two *graphical devices* of some interest which have been put forward may also be included in this record of curve-fitting processes, since they both employ principles beyond the mere drawing of a smooth curve through the unadjusted data.

The first is Calderon's invention of a mechanical contrivance for obtaining $\log c$ in Makeham's formula graphically (see H:*79*:173), although the proposal is now only of historical significance.

The other, however, is more immediately practicable, since it concerns the prescribing of limits, in accordance with the theory of errors, within which the graphically graduated values of observed data should lie. Calderon again (in H:*79*:170) seems to have been the first in actuarial literature to entertain the idea, which was well described by G. F. Hardy in the following words (loc. cit., 193): "The probable error curve, by means of which Mr. Calderon represented the unadjusted values as a sort of river instead of a single line, was a useful suggestion for graphically graduating a series of observations. The great danger in graphic graduation was that they had not constantly before them in different parts of the curve any guide as to the extent to which they were justified in departing from the original facts in drawing their graduated curve. This method provided them with such a guide, in the form of limiting curves, between which, on the whole, the graduated curve should lie." In the graduation of the observed q'_x, for example, we have, as at p. 274; C; 7, that $\sigma\{q'_x\} = \sqrt{\frac{p_x q_x}{E'_x}}$, and in Chapter III it is shown that, under the conditions of a normal distribution (which here means that $E'_x q_x$ should not be less than about 10) 95% of the area will be included between $\pm 2\sigma$ (or 96% between $\pm 3\lambda$). Furthermore, if E'_x is large enough that the observed p'_x and q'_x can be substituted as estimates for p_x and q_x (cf. p. 291; C; 10), it follows that the upper and lower limits $q'_x \pm 2\sqrt{\frac{p'_x q'_x}{E'_x}}$ will embrace about 95% of the values of q'_x which will be obtained by observation. On this principle Orloff (P:*95*) has suggested plotting on the graph a vertical bar extending between these limits for each observed q'_x, and then performing the graphic graduation by drawing the curve smoothly, but so that it will cut all the bars, or as many of them as practicable. [Calderon's proposal, although reached by a less easy analysis, was made for a graduation of m'_x, using in effect $\pm\sigma$ (embracing 68%) instead of $\pm 2\sigma$ as limits, and thus taking $m'_x \pm \sqrt{\frac{m'_x(1-m'_x)}{E'_{x+\frac{1}{2}}}}$, which were set up on the graph (see H:*79*:172) as being approximately $m'_x \pm \frac{\sqrt{\theta'_x}}{E'_{x+\frac{1}{2}}}$ since $m'_x = \frac{\theta'_x}{E'_{x+\frac{1}{2}}}$, and $\sqrt{1-m'_x} \doteq 1$ at most ages.] The process follows the ideas inherent in the "confidence" or "fiducial" limits mentioned in Chapter X.

A; 18. The Mean Square Error of an Observation, when there are ν Observation Equations and k Unknowns

The expression $\dfrac{\sum_{r=1}^{r=\nu}\left[W_r(f''_r - f'_r)^2\right]}{\nu - k}$ derived at p. 250; B; 26, which in the classical notation (see p. 324; C; 21) is usually written $\dfrac{[pvv]}{\nu - k}$ where p denotes the weight W_r and v the "residual" $(f''_r - f'_r)$, was deduced originally by Gauss in H:*17*: arts. 37-39 (cf. p. 164; A; 8, and the references there given). Being thus firmly established as part of the formal procedure in determining the mean square error of an observation when the number of unknowns is k, it is to be found throughout the text-books on the method of least squares, and is there applied frequently in examining the agreement which has been obtained between a fitted curve and the observed data. It is so used systematically by Merriman, for example, in P:*90*:134, 137, and 197, and by many later writers (see P:*124*:145, 153, and 168).

In actuarial literature the formula was employed first, in the form (126), by Thiele in 1871 (H:*43*:321), as a test of the goodness of fit of a mortality table graduation. The only writer—at least in English publications—who subsequently noticed Thiele's procedure seems to be De Forest. His exhaustive series of papers began to appear in 1871 (see P:*166*); in 1873 he referred to Thiele's methods (H:*48*:334), and in H:*49*:14 and H:*54*:6 he discussed the conditions under which (126) should be applied. Recently attention has again been drawn to Thiele's work by Seal in P:*125*:6.

A; 19. The χ^2 Test for Goodness of Fit, and "Degrees of Freedom"

The mathematical expression for the distribution of χ^2, in a form equivalent to (127), was given originally by Pizzetti in 1892 (H:*71*:267). The χ^2 test, however, was first developed for practical use by Karl Pearson in 1900 (H:*80*), and constitutes undoubtedly one of the most important of his many contributions to Mathematical Statistics. In reality, as pointed out in Chapte V, it completed the theory underlying the Lexis method.

For some years the formula was applied incorrectly in certai cases, since it was not then realized that the number of variables ν, must be diminished by 1 for each linear "constraint". Certair discrepancies in the results of the test had been noted by Brownlee (H:*147*), and Greenwood and Yule (H:*140*); the necessity of allowing for the "degrees of freedom", d, was finally established clearly by R. A. Fisher (H:*138*) in 1922. Many of the statistical applications of the χ^2 test prior to that date are therefore erroneous. It should be noted, moreover, that the correct principle has always been recognized as the accepted procedure which emerges in the analogous Method of Least Squares, and is to be credited in the first instance to Gauss (see P:*123*; P:*28*:18; and p. 164; A;8).

Tables of the integral P corresponding to values of χ_0^2 and $d+1$ $(=\nu)$ were first computed by Elderton, to 6 places, and, after publication in "Biometrika", were reprinted in P:*97*. More recently, R. A. Fisher (in P:*43*) has given a table of χ_0^2 to 3 places corresponding to selected values of the integral to 2 places only, which in some respects is more convenient since 2 places are ordinarily sufficient. The values are there shown for values of d from 1 to 30 (with the useful suggestion that for large values of d a close enough approximation is afforded by assuming that $\sqrt{2\chi_0^2}$ is "normally" distributed about a mean $\sqrt{2d-1}$ with unit standard deviation). For most practical purposes, indeed, sufficient accuracy is attained simply by reading P from a diagram (see P:*177*:418, 422, and 540).

The range is indicated by the following specimen values:

Degrees of Freedom, d	$P=.99$	$P=.95$	$P=.5$	$P=.05$	$P=.01$
	Value of χ_0^2				
1	.000	.004	.455	3.841	6.635
10	2.558	3.940	9.342	18.307	23.209
20	8.260	10.851	19.337	31.410	37.566
30	14.953	18.493	29.336	43.773	50.892

SECTION

B

MATHEMATICS

AND

INTERPRETATIONS

B; 1. The Nature of "Probability"

It has been pointed out by many writers (see, for example, P:*36*:3-10, 54-55, and 146-149, and P:*80*:15 et seq.) that the theory of "probability" can be viewed as (*i*) a mathematical theory of arrangements; (*ii*) a study of the actual statistical frequencies of observed occurrences; or (*iii*) a branch of logic. It will be convenient to speak of these three approaches as the *mathematical*, *statistical*, and *logical* respectively—using these terms, of course, merely for purposes of identification, since the three methods are not in reality so sharply separable as this abbreviated nomenclature might at first seem to imply.

In the purely "mathematical" sense, a "true" value for the probability is either known, or assumed to be known, *a priori*, and the problems then concern the "arrangements" which may arise (or any set of consistent hypotheses may be assumed, and the consequent propositions are to be deduced therefrom). The "statistical" viewpoint considers primarily the actually observed occurrences, when the true probability is not known *a priori* but is, instead, to be "estimated" *a posteriori* from the statistical frequencies observed. The consideration of probability as a branch of "logic" involves questions which are partly psychological rather than wholly logical.

It will thus be at once apparent that we are dealing with a subject which comprises the pure mathematics of the theorist, the statistical observations of the experimentalist, and also—inseparably interlocked with both—the contemplations of the philosopher. The variations of approach are thus so wide that it has even been suggested, with more than a little justification, that any attempt to set forth "the theory of probability" is largely an attempt at "a description of a state of mind".

Even though the student need not ordinarily concern himself too deeply with these often fine and sometimes trivial distinctions, nevertheless he should be thoroughly aware that the possible variations in that "state of mind", as they may influence the fundamental ideas essential to an understanding of the nature of probability, have of course been discussed at great length in many languages, and are to be found conveniently in English in such presentations as Balfour's elegant "Defence of Philosophic Doubt—An Essay on the Foundations of Belief" (H:*58*), and in Keynes' "Treatise on Probability" (P:*75*). A reading of those speculations is perhaps more likely to confirm than to resolve the uncertainty as to what may constitute a "best" approach; any attempt to explore thoroughly the meanings of "rational belief", "inference", the principle of "insufficient reason" (Boole's "equal distribution of ignorance" and Keynes' "principle of indifference"), and the opposite principle of "cogent reason" (being an "unequal distribution of ignorance"), must even impose caution upon the very use, for example in a context such as this, of the words "perhaps", "likely", "uncertainty", and "best". Positive assertions are indeed discouraged by such contemplations, and we are brought inevitably to seek an approach to the applications of probability theory which will recognize the practical demands of any system of "statistical or scientific inference", while yet conforming with the obvious theoretical requirements.

For the theory of probability is intended really to provide a basis of measurement for scientific inference. That different people draw different inferences regarding a proposition from the same set of data is one of the recognized characteristics of a practical world. It illustrates, in every-day life, the relationship between the inference which the logical situation would demand, the measure of the probability concerning the inference derivable *a priori* from the data or *a posteriori* from the observations, and the effect of the person's state of mind. Jeffreys (although in a somewhat different connection) has put the matter well: "One person, reading the proof of Euclid's fifth proposition, is completely convinced; another is entirely unable to grasp it; while there is at any rate one case on record when a student said that the author had rendered the result highly probable" (P:*68*:10).

It may therefore be advisable at this stage to classify very briefly the schools of thought which have arisen from these three main interrelated approaches, namely, (*i*) the purely *mathematical*; (*ii*) that based on the observed *statistical frequencies*; and (*iii*) that in which the principles of academic *logic* form the starting-point.

(*i*) *Mathematical*

This was the classical approach, and was developed principally as a mathematical theory of arrangements. The "true" *a priori* probabilities either are clearly known, as in direct problems concerning dice, coins, cards, urns, etc., or are to be assumed in some identifiable form.

The notion of probability seems to have been mentioned first in China by Sun-Tze about 200 B.C. in connection with the probability that a birth would be that of a boy or girl. No really serious attention, however, was given to problems in probability until the fifteenth century in Europe, and then the discussions concerned mainly games of chance and gambling. After certain preliminary references by Pacioli, Cardan, Galileo, and Kepler, interest became greatly intensified by a series of questions propounded by a gambler, the Chevalier de Méré, to Pascal, amongst which was the famous "Problem of Points" concerning the equitable division of the stakes of two gamblers who discontinue the game prior to its completion. The fundamental letters which passed between the Frenchmen Pascal and Fermat in 1654 (see P:*149*:6), and the first printed work on the subject by the Dutchman Huygens in 1657 (H:*1*), led to the great "Ars Conjectandi" of James Bernoulli in Switzerland in 1713, and thereafter to the contributions of the French Montmort (H:*3*), De Moivre, Legendre, and Laplace, and the German Gauss.

The fact that some of the modern analyses of these classical dissertations have placed certain of their premises and conclusions in better perspective does not alter the opinion that the mathematical foundations thus laid must, to a very large degree, form the basis of any reasoned discussion of the theory of probability and mathematical statistics. It is true, of course, that much sharp criticism has been directed against Laplace and others in connection with their use of the "Principle of Insufficient Reason". That principle—which is perhaps more clearly described by Boole's phrase "the equal distribution of ignorance"—asserts that the unknown *a priori* probabilities must be equal when our state of ignorance precludes the assignment of unequal values; i.e., to quote Jeffreys (P:*68*:20), "if we have no means of choosing between alternatives, the probabilities attached to those alternatives are equal"; or, to quote Keynes (P:*75*:372), "when the probability of an event is unknown, we may suppose all possible values of the probability between 0 and 1 to be equally likely *a priori*". This "principle", however, will be found upon examination to be in reality an arbitrary assumption; and, unless it is accompanied by a careful analysis and statement of the circumstances in which it is applied, it may lead—as indeed it led Laplace and others in the case of the so-called "Rule of Succession"—to some strange and paradoxical results. Most of the criticism, nevertheless, has been undiscriminating, and much of it unmerited and far too bitter.* These difficulties, however, clearly arise from that one special arbitrary assumption, and should not be taken as justification for discarding many of the fundamental mathematical concepts or for belittling (as is sometimes done today) the monumental contributions of the Laplacian school.

Within recent years some attention has been attracted by the efforts of von Mises (H:*163*), which have been examined in English particularly by Copeland (H:*186*), to establish the ideas of probability upon an "axiomatic" basis, from which the entire theory would be deducible by strictly mathematical deductions. According to this method, the "mathematical" and "statistical" approaches are, in effect, related through a definition of probability based on the notion of "sequences". If there is a set, or "Kollektiv", of n objects, then the "probability" of an object with a specified characteristic which occurs m times in the first n will be $\frac{m}{n}$ when n is increased indefinitely; but von Mises requires that two conditions must be satisfied: (*a*) the limit just mentioned must exist, and (*b*) a "principle of irregularity (or disorder)" must also exist such that the limit remains unaltered in any sub-sequence—for example, if the sequence H, T, T, H, H, H, T, . . . is the heads and tails of throws of a coin, and in the limit H appears in one-half of the throws, then also in any subsequence such as T, H, H, H, T, . . . , selected from any position in the sequence, and forming all or only some of the terms of the sequence, H must likewise appear in one-half of the throws. Great mathematical and logical difficulties, however, have been

*The controversy mainly involves the theory of "inverse probability," and consequently, for the reasons already stated in the footnote on p. 8, need not be pursued in this present study.

encountered in the attempts to establish by this means the fundamental theorems of mathematical probability, while as a practical method a major obstacle, arising from the definition of a "Kollektiv" restricted by condition (*b*), must be that the set of objects cannot apparently be identified as a "Kollektiv" until n has been increased indefinitely (see P:*80*:30; P:*30*; P:*22*:4; and P:*156*:2).

Another strictly mathematical treatment, based on certain axiomatic foundations, has been given by Kolmogoroff (H:*178*), and recently has been developed in an English presentation by Cramér (P:*22*). In the latter will be found also many of the modern purely mathematical contributions of Borel, de la Vallée Poussin, Lévy, Fréchet, Cantelli, Khintchine, Liapounoff, Markoff, Romanovsky, Radon, and others. In view of the essentially practical requirements of this volume, however, it will be sufficient here to give the preceding references for any reader who may wish to pursue the complexities of these formal mathematical investigations.

(ii) Statistical

Keynes has pointed out (P:*75*:92) that Ellis (H:*24*) was apparently the first to state the importance of considering probabilities on the basis of observed frequencies. It was Venn, however (H:*35*), who elaborated that viewpoint and really established the course of reasoning which has led to its wide acceptance by the modern English school. Venn's original and involved presentation, of course, has undergone much critical examination, and in its details would hardly now prove satisfactory to many of his followers.

The discussions again have involved attacks upon the "principle of insufficient reason"—for, as Ellis expressed the matter, "mere ignorance is no ground for any inference whatever; *ex nihilo nihil*". But since this position, which is equivalent to saying (as previously noted) that the "principle" is merely an arbitrary assumption, must obviously lead in many cases simply to a complete inability to reach a solution, attempts have been made to formulate conditions under which the "principle of insufficient reason" can be invoked. Von Kries, for example, in his discussion of the opposite "principle of cogent reason" (H:*67*), elaborates the view that "the arrangement of the equally likely cases must have a cogent reason and not be subject to arbitrary conditions" (see P:*36*:7, and P:*75*:42 and 87). In this connection Arne Fisher (P:*36*:9) suggests that, since "a rigorous application of the principle of cogent reason seems impossible", a compromise between that principle and the principle of "insufficient reason" may be effected "by the following definition of equally possible cases: 'Equally possible cases are such cases in which we, after an exhaustive analysis of the physical laws underlying the structure of the complex of causes influencing the special event, are led to assume that no particular case will occur in preference to any other'." Keynes also (P:*75*:41), in a long discussion of the subject, during which he seeks to clear up the difficulties and formulate the "principle of insufficient reason" in a more precise and workable form, observes that it is not a "sufficient" condition, and reaches the conclusion that the difficulties have arisen "when the alternatives, which the principle . . . treated as equivalent, actually contain or might contain a different or an indefinite number of more elementary units"—that is to say, the principle can be utilized only so long as it is recognized that it "is not applicable to a pair of alternatives if we know that either of them is capable of being further split up into a pair of possible but incompatible alternatives of the same form as the original pair".

Faced with this questionable applicability of the "principle of insufficient reason", and recognizing at the same time the difficulties of formulating and using any correction of it in the form of a "principle of cogent reason", the modern supporters of the statistical frequency approach have therefore taken the position that "it should be possible to draw valid conclusions from the data alone, and without *a priori* assumptions" (P:*41*)—they "disclaim knowledge *a priori*, or prefer to avoid introducing such knowledge as we possess into the basis of an exact mathematical argument" (ibid.), so that it may be possible to construct a theory of statistical inference without the use of *a priori* probability (P:*67*).

The possibility of thus assembling a consistent body of doctrine based solely upon observed statistical frequencies, without recourse to the concept of a true *a priori* probability, seems first to have been examined rigorously by T. N. Thiele of Denmark (H:*94*) in 1884. But in so doing, of course (just as in postulating the existence, in the mathematical method, of the *a priori* probabilities which are now to be avoided), care must be taken in formulating the approach. It is obviously quite insufficient merely to suggest that the observed statistical frequency in "a large number of trials" is to be taken as the "probability" (as has been done so often in algebraical text-books and elsewhere)—for the question, "What is a large number?" again at once asserts itself. It is therefore clearly essential to adopt for the concept of "probability", as Thiele suggests, "the limiting value of the relative [statistical] frequency of an event, when the number of observations amongst which the event happens approaches infinity as a limit". It will be noted here that, in effect, the emphasis is laid upon the statistical frequency—the concept of "probability" not emerging until the ultimate hypothetical state of an infinite number of observations is reached—whereas in the mathematical approach the true "probability", as an *a priori* concept, is supposed to be given, and we find ourselves at once faced with the question as to how closely, and under what conditions, it can be measured from the statistical frequency observed in a stated number of trials (see p. 263; C; 1). In neither the mathematical nor in the statistical approach, therefore, can we avoid, in reality, the essential problem of establishing the connection between the true "*a priori* probability" and the "statistical frequency" observed.

The modern developments in the theory of "sampling" (see Chapter V) which have been founded on this viewpoint have, of course, a special interest for actuaries and vital statisticians on account of their evident plausibility and their practical nature. The mathematical and logical bases of the method, however, are still undergoing critical examination and discussion, particularly from those whose mental processes lead them to prefer a rigorously mathematical or logical, instead of a plainly empirical, approach. (See, for example, Keynes' comments, P:*75*:92-110, and the interesting controversy and misunderstandings between

Jeffreys and R. A. Fisher, P:*67*, and P:*41*—partially summarized in P:*64*:138-142, and eventually resolved as stated in P:*68a*:323.)

(*iii*) *Logical*

The first critic to assail the philosophical position of the mathematicians of Bernoulli's time was Hume (H:*8*). To quote Keynes (P:*75*:272): "Argument by induction—inference from past particulars to future generalisations—was the real object of his attack. Hume showed, not that inductive methods were false, but that their validity had never been established and that all possible lines of proof seemed equally unpromising". D'Alembert, also—partly known for his discussions of the famous "St. Petersburg Paradox" (see H:*33*:258; P:*36*:51; and P:*75*:316) originally considered by Daniel Bernoulli—is remembered (notwithstanding many demonstrated errors in his reasoning) for the scepticism which he likewise expressed.

Within recent years Keynes (P:*75*)—acknowledging his indebtedness to Leibniz, "who in the dissertation, written in his twenty-third year, on the mode of electing the kings of Poland, [first] conceived of Probability as a branch of Logic"—has dealt at great length with the description of probability as "comprising that part of logic which deals with arguments which are rational but not conclusive". Following the English tradition of Locke, Berkeley, Hume, Mill, and later Bertrand Russell—"who, in spite of their divergencies of doctrine, are united in a preference for what is matter of fact, and have conceived their subject as a branch rather of science than of the creative imagination"—Keynes sets out, from the fundamentals of Logic, to develop, first, "the characteristics and the justification of probable Knowledge"; second, the deduction, by the methods and symbolism of Formal Logic, of the usual theorems for the addition, multiplication, etc., of probabilities; third, the methods of Induction and Analogy; and fourth, the foundations of "Statistical Inference". Except so far as mathematics are essential in some parts to illustrate his criticisms of other methods, Keynes accordingly bases his entire approach upon the principles of Logic. The fact that probability when so treated remains essentially non-measurable must, under these circumstances, be dealt with by a variety of terminological refinements. The stimulating discussions to which this method leads, however, are of course important, and may be found in the works of Ramsey (P:*104*) and Jeffreys (P:*68* and *68a*), as well as in Keynes' treatise (see also P:*55*).

B; 2. Bernoulli's Limit Theorem

While the statement on p. 8 conveys the essential meaning of Bernoulli's famous Limit Theorem—the Law of Large Numbers—in a convenient verbal form, it is not exactly what Bernoulli said. It is to be noted especially, therefore, that Bernoulli's Theorem in more precise terms was this:

If in a set or series of n trials of an event, in each of which trials the "true" *a priori* probability of success is a constant p, the actual number of successes is observed to be s (so that the "statistical frequency" of success observed is $\frac{s}{n}$) then the probability, say P, that the discrepancy $\left|\frac{s}{n}-p\right|$ between the observed statistical frequency, $\frac{s}{n}$, and the true *a priori* probability, p, is less than a previously assigned quantity (say $\epsilon > 0$), approaches 1 or certainty (that is to say, will be greater than $1-\eta$, where η is a previously assigned positive quantity) as the number of trials, n, is increased indefinitely.

Or, symbolically, given two positive numbers ϵ and η, the probability P of the inequality $\left|\frac{s}{n}-p\right| < \epsilon$ will be greater than $1-\eta$ if n exceeds a certain limit.

The interpretation ordinarily placed on these precise statements is that the observed statistical frequency, $\frac{s}{n}$, tends to coincide with the theoretical or true probability, p, as the number of trials, n, is increased indefinitely. Symbolically, $\lim_{n\to\infty} \frac{s}{n} = p$. In other words, by postulating the existence of the true probability, p—even if we are unable to determine its value *a priori* (as can be done in the conventional urn and coin-tossing experiments, but under other circumstances frequently cannot be done)—it thus becomes possible to examine the degree of approximation afforded by an observed statistical frequency, $\frac{s}{n}$. The great importance of Bernoulli's discovery therefore lay in the facilities which it afforded for the examination of mass phenomena, through the observation of statistical frequencies when the numerical value of the *a priori* probability was in fact unknown.

Bernoulli's original proof of his theorem—upon which he states that he spent upwards of twenty years—is to be found in modernized notation in P:*146*:96, together with a translation of his explanations, which are important as clear evidence of his intention to investigate the degree of approximation afforded by observed statistical frequencies based upon a limited number of trials. The theorem was also established later by the investigations of De Moivre and Laplace, as indicated herein and set out in detail in P:*146*:119, and may be derived easily from the Bienaymé-Tchebycheff criterion (see p. 218; B; 10).

It may be of interest to note, even at this point, that a good deal of attention has been given recently to certain aspects of the proof and conditions of Bernoulli's Theorem and the Law of Large Numbers. It will be seen that the theorem states that there exists a number, n, such that, for any *single* instance of a number of trials greater than n, the probability of a discrepancy (between $\frac{s}{n}$ and p) less than a given amount (ϵ) will exceed a given quantity $(1-\eta)$. In 1916, therefore, *Cantelli's Theorem* raised the question whether there exists a number, n', such that for *all* numbers of trials greater than n', the probabilities of *all* the infinity of *simultaneous* discrepancies (between $\frac{s}{n}$ and p, where n takes all values greater than n') less than the given amount (ϵ) will still exceed the given quantity $(1-\eta)$. The mathematical investigation of this important extension of the original Bernoullian Law of Large Numbers gave an affirmative answer, which was later perfected by Kolmogoroff and Glivenko. It has become known as the *Uniform* or *Strong Law of Large Numbers*, and, as will be seen, shows the probability that the statistical frequencies, $\frac{s}{n}$, will differ from p by less than ϵ in the n'th and all the following trials is greater than $1-\eta$ (cf. P:*35*:717-8, and P:*146*:101-3).

Much work has also been done on the problem of establishing the limits and inequalities involved—particularly by Khintchine, Lévy, Glivenko, Kolmogoroff, and De Finetti. It will be sufficient here to refer to a valuable review of these investigations by Fieller (P:*35*:721), and to Uspensky's observations in P:*146*:204, in both of which the inequalities are expressed in a form which has been called *The Law of the Repeated Logarithm*. The numerous original papers dealing with these researches are listed conveniently in P:*35*:766-8.

B; 3. Diagrammatic Representations

An understanding of the nature of the problems of practical statistics will be assisted materially by visualizing the types of diagrams, distributions, curves, and surfaces represented by the various statistical tables or mathematical formulae. A summarized description is therefore given here, and is referred to throughout the text.

The collation, into a statistical table or diagram, of a series of observations taken under essentially uniform conditions, but in which one or more characteristics are subject to variation, will give a picture of the *observed distribution* with respect to the one or more *independent variables* or *variates* (or "characteristics") $x_1, x_2, \ldots$.

(*i*) For example, the actual numbers of deaths, y, occurring in a particular community (and under essentially uniform circumstances) on each exact birthday would give an observed distribution of deaths with respect to one variable—exact birthday x_1; the numbers of deaths similarly observed on each exact birthday for persons of each exact height would give an observed distribution of deaths with respect to two variables—exact birthday x_1 and exact height x_2; and so on for any number of such variables. If for one variable, x_1, such values be plotted on squared paper with x and y axes, they will be represented simply by a series of *points*—because the variable x_1 clearly can assume only integral ("discrete") values (since only the deaths at exact integral ages are included in the observations); and if the points be joined together by straight lines (although in this instance of a series of integral values such straight lines have no interpretative significance except as a first approximation to hypothetical intermediate, i.e., non-integral, values) the resulting outline will be a series of two-dimensional jagged peaks if the data are very irregular (Figure A1), or of jagged humps (Figure A2).

For two variables, x_1 and x_2, we may, with x_1, x_2, and y axes, picture the same process by erecting verticals at each point on the base (the $x_1\,x_2$ plane); and if again the tops of the verticals be joined by straight lines, the picture obtained is composed of three-dimensional jagged (saw-toothed) peaks (Figure B1).

(*ii*) In many observational procedures, however, it will clearly be either impracticable or inadvisable (because of the extent to which the observed data are in fact reduced thereby) to observe, as in (*i*), the values of y only for integral values of $x_1, x_2, \ldots$. In the examples cited, for instance, the natural process would be to include all the data, whether at integral or fractional values of $x_1, x_2, \ldots$, in order to make use of the observed deaths not only at the exact ages, heights, etc., but also at all intermediate ages, heights, etc. Since, however, it would be unduly laborious to attempt the tabulations for every such intermediate value of the variables $x_1, x_2, \ldots$, we are at once brought to the idea of *grouping*. The obvious procedure in the case of the distribution by age alone would be to classify together all those aged x last birthday, which means grouping the observed deaths in groups by single years of age (a to $a+1$ exact, $a+1$ to $a+2$ exact, etc.)—although a broader grouping (as for each span of 5 ages, a to $a+5$, $a+5$ to $a+10$, etc.) could often be used. The groupings a to $a+1$, etc., or a to $a+5$, etc., are spoken of as the *class intervals*.

In now plotting such data we can either (a) assume (sometimes) the middle points $x+\frac{1}{2}$ (or $x+2\frac{1}{2}$ as the case may be) and erect the ordinates thereon, or (*b*) employ the age-group (1 or 5) as the bases for a series of rectangles. Joining by straight lines the points of (*a*) we obtain again a series of two-dimensional jagged peaks or humps, or a *frequency polygon* (the lines in Figure A3); (*b*), on the other hand, gives a diagram of vertical columns representing areas, or a *histogram* (the rectangles in Figure A3).

For two variables, similarly, the same processes lead to (α) again an Alp-like series of three-dimensional jagged peaks (Figure B1), or (β) a series of three-dimensional square pillars—like a city built exclusively of square pillars, with bases of uniform area but of various heights (Figure B2).

(*iii*) If now, instead of the preceding *discontinuous* classifications by discrete values or groups, we suppose that the variations are *continuous*, then the two-dimensional jagged peaks would obviously become smoothed out into a curve—undulating (Figure A4) or not (Figure A5) according to the nature of the data—while the three-dimensional Alps or the pillar-like city would likewise be transformed into a solid mountain with curved slopes (Figure B3, in which the curves are symmetrical). The former is an observed *frequency curve*, the latter is an observed *frequency surface*.

If the observed curve, or the observed surface, can be postulated *a priori*, or described *a posteriori*, by a mathematical formula, the observed distributions will thereby be replaceable by a theoretical, or by a "fitted", frequency curve $y=f(x_1)$ or surface $y=f(x_1, x_2)$.

With due allowances in the parameters of the expressions adopted, these functions can of course be made to represent the distribution either of the numbers of cases or of their relative frequencies. The curve relating to the relative frequencies is generally referred to as a *probability curve*. For such a probability curve, say $y=\varphi(x)$, the term *frequency function* or *probability density* is given to $\varphi(x)$; the numbers of cases would be expressed by $y=N\varphi(x)$ where N is the total number of cases; the number of cases in the interval between $x=a$ and $x=b$ would be $N\int_a^b \varphi(x)\,dx$; the number between x and $x+dx$ will be given by $N\varphi(x)dx$; and if the whole distribution is included within $x=h$ and $x=k$, then $\int_h^k \varphi(x)dx=1$, and $\varphi(x)\geqq 0$ for all such values of x, while if the distribution extends indefinitely $\int_{-\infty}^{+\infty}\varphi(x)\,dx=1$ and $\varphi(x)\geqq 0$ for all values of x—since the total number of cases must equal N, and a probability, of course, cannot be negative.

If the ordinates of a frequency curve are added successively from its commencement to form a series of "integrated frequencies"—thus giving, in the continuous case when the curve starts, for example, at $-\infty$, the function $\int_{-\infty}^{t}\varphi(x)\,dx$ for successive values of t—the result will be a function (never decreasing) varying between 0 and 1, and representing the probability for a value of x not exceeding t (like the smooth shoulder of a mountain in the form of Figure A6). It is sometimes called the *Distribution Function of Probability*, and the curve is generally identified as a *Cumulative Probability Curve.* If the number of cases N is introduced, the curve would, of course, be referred to as a *Cumulative Frequency Curve.* In the discontinuous case of a series of observed values, we obtain clearly a series of progressively rising points, as in Figure A7, while when groups are used they become a stairway with steps of varying height as in Figure A8.

[Although the method is not of special importance for the purposes of this study, it may be well to note that the "cumulative frequency" principle may be exhibited graphically by the application of Galton's *Method of Percentiles.* If, after a cumulative frequency diagram has been constructed in the form of a smooth curve as in Figure A6, the terminal ordinate is divided in half, and a horizontal line then drawn to cut the curve at M (Figure A9), the abscissa of M is the *median*—being, in fact, the value of the variable in the corresponding non-cumulative curve of Figure A5 such that its ordinate divides the area into two equal portions. The three *quartiles* likewise mark the quarters, as shown; the nine deciles show the results of proceeding similarly from the points which divide the terminal ordinate into ten portions; and the *percentiles* mark the percentage divisions—so that, for example, the 5th decile, 2nd quartile, and 50th percentile all coincide with the median. If we now take an axis marked off in ten equal parts, and erect ordinates corresponding to the deciles, the curve becomes of the type shown in Figure A10—generally known as Galton's *Ogive*.]

The concept of distribution has been illustrated in the preceding paragraphs for the cases of one or two independent variables, which can be represented diagrammatically by the direct use of our every-day knowledge of space of not more than three dimensions. The *explicit* functional relationship $y=f(x_1)$ between the *dependent variable* y and the single *independent variable* x_1 is pictured by a graph in two dimensions with two co-ordinate axes x_1 and y, and a point is fixed by its co-ordinates (x_1, y). It is convenient to think of this as an *implicit* function relating the two variables x_1 and y in the form $F(x_1, y)=0$, so that the case is one of two variables being represented by a graph in two dimensions. For the explicit function $y=f(x_1, x_2)$ it was necessary to use three-dimensional space, with three co-ordinate axes x_1, x_2, and y; here we have an implicit function of the form $F(x_1, x_2, y)=0$ relating three variables (one being a function of the other two) and requiring three-dimensional space for its depiction.

In dealing with more than three variables, as is often necessary, there is little difficulty in extending the mathematical principles; actual diagrammatic representation of more than three dimensions, however, is not possible. It is convenient, nevertheless, to continue the use of certain geometrical terms—thus a particular set of the ν variables $x_1, x_2, \ldots, x_\nu$ is still called a "point" in ν-dimensional space, with ν co-ordinates referred to ν mutually perpendicular axes; relations between the variables are *hyper-surfaces*; and in such ν-dimensional space the $\nu-1$ independent variables will represent a $\nu-1$ *dimensional surface*. As the student will realize from the preceding explanations, these concepts of hyperspace can usually be dealt with most clearly by founding the course of reasoning upon analogy from the corresponding concepts in the 1, 2, or 3 dimensions with which we are visually familiar (cf. the method of proof used in B; 13).

In this connection one device of considerable value is the use of *contours* (or "contour lines"). If we are dealing with an explicit function $y=f(x_1, x_2)$ of three variables x_1, x_2, and y, which represents a surface, and we take all the points in the x_1x_2-plane for which $y=f(x_1, x_2)$ has a constant value, k, then these points will lie on the contour line for the constant value, k, of the function. Alternatively, a given plane cutting a surface will have points lying on a curve designated a "plane section" of that surface; a "horizontal section" by a plane $y=k$ parallel to the x_1x_2-plane, when projected perpendicularly on to the x_1x_2-plane, will therefore give the "contour". The contour thus indicates simply the variation of x_1 and x_2 for the given value, k, of y, and shows the shape of the surface. For example, if the parabola $y=x^2$ is rotated about the y axis, a "paraboloid" results, which is represented by $y=x^2+z^2$ in Figure C1. Giving k the values 1, 2, 3, . . ., the contours of this paraboloid are obtained at heights $y=1, 2, 3,$ . . . ; and when projected onto the xz-plane as in Figure C2 they are shown to be concentric circles with the centre at the origin, from which the steepness or flatness of the surface is indicated by the closeness or otherwise of the contours. A change in the independent variables has the effect of moving a point (x, z) in the xz-plane. Consequently, if an actual path on the surface is projected onto the xz-plane, it will appear as a line or a curve across the system of contours, and will thus show how y changes with changes in x and z; in fact, y increases as the point moves from lower to higher contours, and vice versa.

Vertical sections can be used similarly. The contour system, moreover, can be extended to functions of three independent variables; the contour lines then become "level surfaces" $f(x_1, x_2, x_3)=k$; for the function $y=x^2+z^2+v^2$, for example, they are concentric spheres about the origin.

The student who may wish to pursue these matters will find excellent discussions in P:*76*:317, P:*4*:270, and P:*21*:I, 460.

FIGURE A1

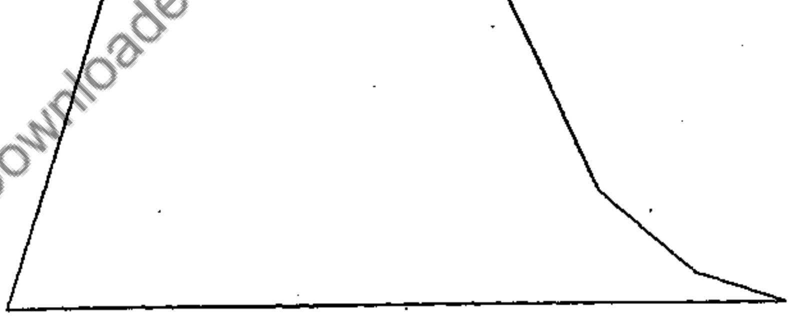

FIGURE A2

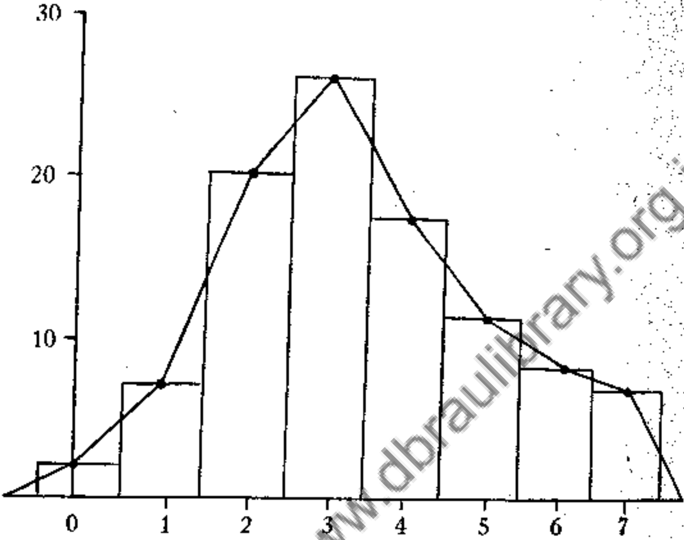

FIGURE A3

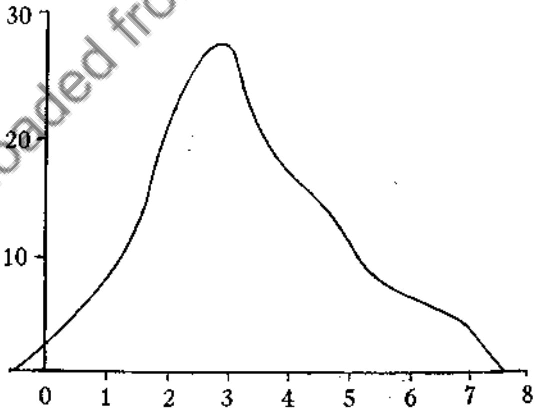

FIGURE A4

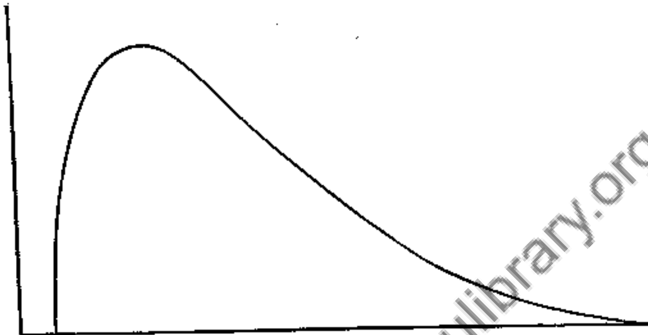

FIGURE A5

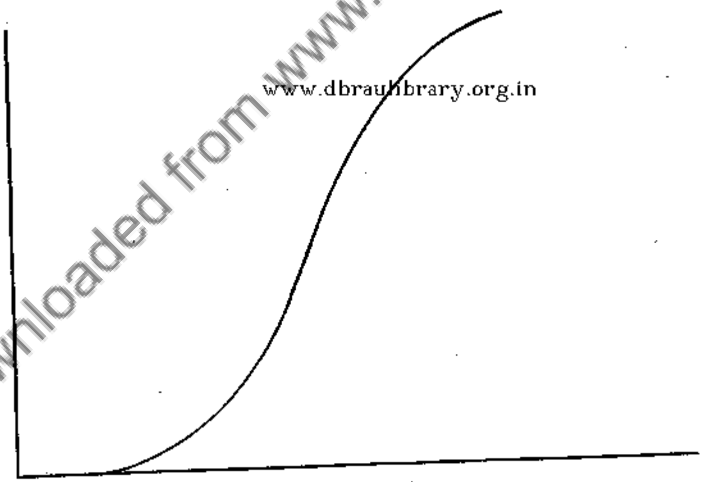

FIGURE A6

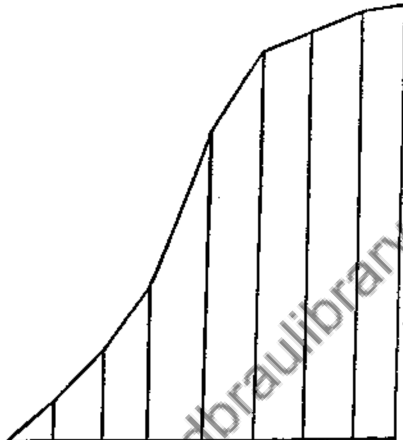

FIGURE A7

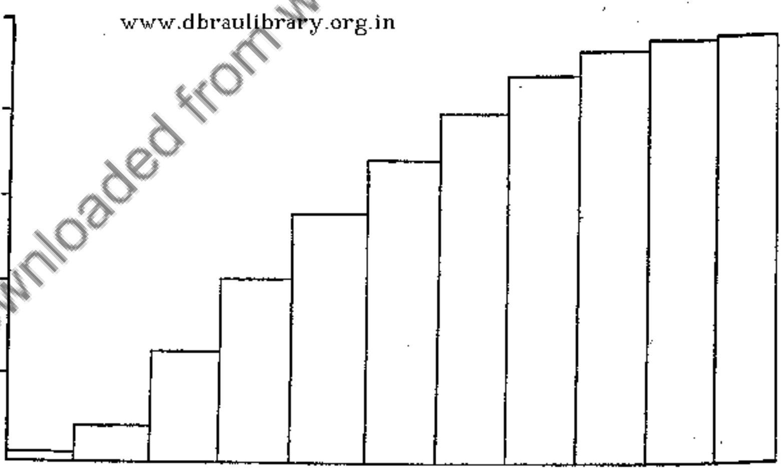

FIGURE A8

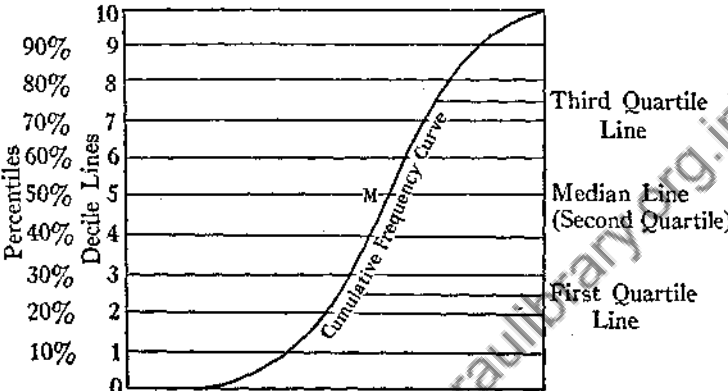

FIGURE A9

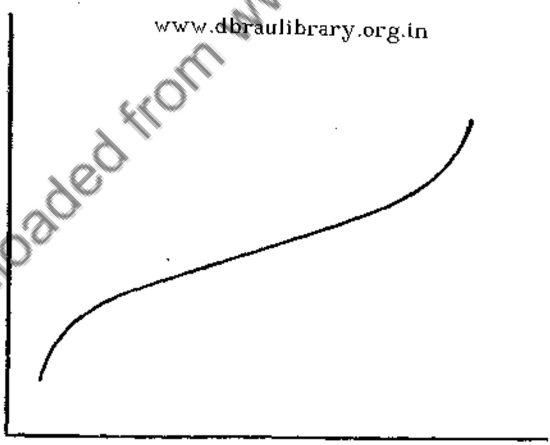

FIGURE A10

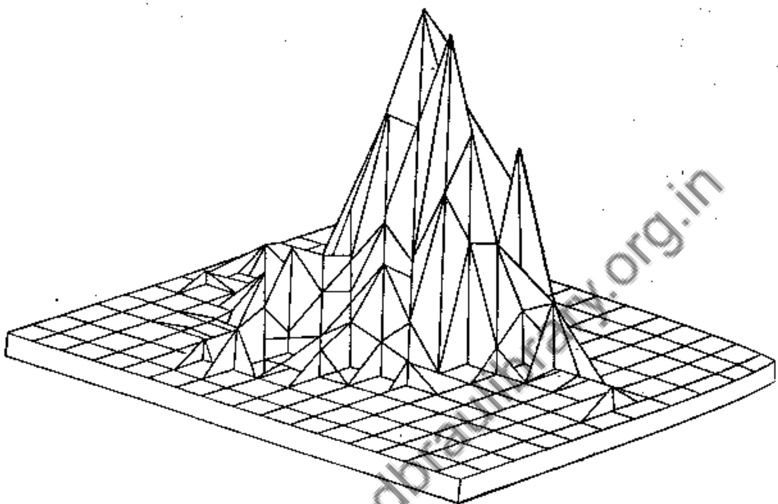

Figure B1

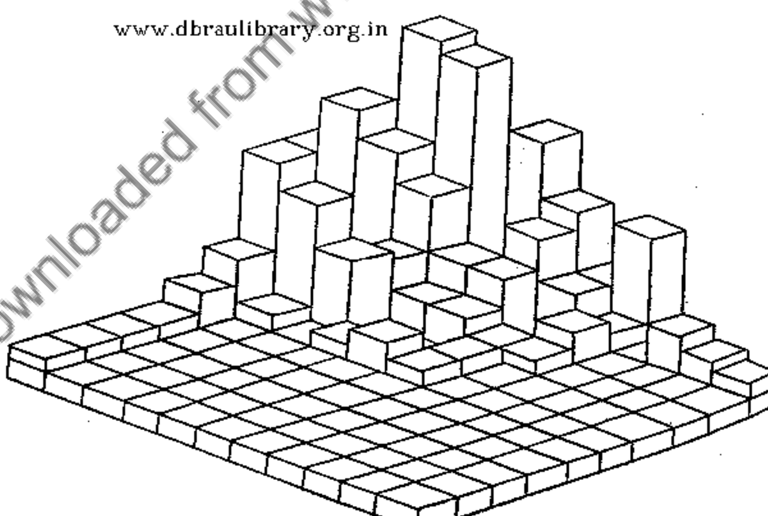

Figure B2

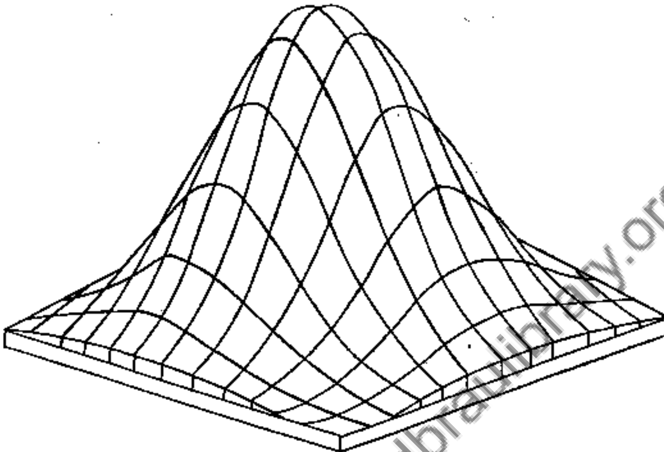

FIGURE B3

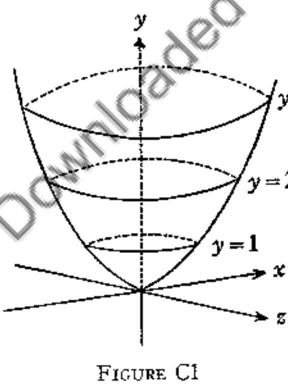

FIGURE C1

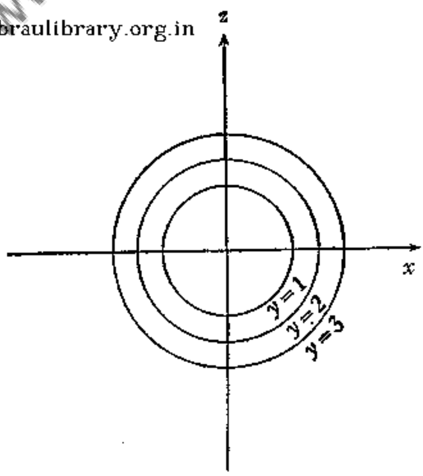

FIGURE C2

B; 4. The Mean Square Deviation of the Point Binomial

The expression (6) is clearly

$$n^2p^2(q^n+nq^{n-1}p+\ldots+p^n)-2np(nq^{n-1}p+\ldots+np^n)$$
$$+(1^2nq^{n-1}p+\ldots+n^2p^n).$$

The first of these brackets is $n^2p^2(q+p)^n=n^2p^2$. The second, from (4), is $-2np(np)$. The third may be written

$$np\left[q^{n-1}+(n-1)2pq^{n-2}+\frac{(n-1)(n-2)}{2!}3p^2q^{n-3}+\ldots+np^{n-1}\right]$$

$$=np\left[q^{n-1}+(n-1)(1+1)pq^{n-2}+\frac{(n-1)(n-2)}{2!}(1+2)p^2q^{n-3}+\ldots\right.$$
$$\left.+\{1+(n-1)\}p^{n-1}\right]$$

$$=np\left[q^{n-1}+(n-1)pq^{n-2}+\frac{(n-1)(n-2)}{2!}p^2q^{n-3}+\ldots+p^{n-1}\right]$$
$$+np(n-1)p[q^{n-2}+(n-2)pq^{n-3}+\ldots+p^{n-2}]$$

$$=np(q+p)^{n-1}+n(n-1)p^2(q+p)^{n-2}$$
$$=np[1+(n-1)p]=np(np+q).$$

The three brackets together therefore give

$$n^2p^2-2np(np)+np(np+q)=npq.$$

A simple alternative proof (given in P:*80*:141, and based on Laplace's original method of differentiation set out in P:*36*:104) is the following:

The expression (6) can at once be written—omitting the limits $t=0$ and $t=n$ for convenience—

$$\Sigma\,{}^nC_t q^{n-t}p^t t^2-2np\,\Sigma\,{}^nC_t q^{n-t}p^t t+n^2p^2\,\Sigma\,{}^nC_t q^{n-t}p^t \quad \ldots(a)$$

Now $(q+p)^n=\Sigma\,{}^nC_t p^t q^{n-t}$. Differentiating this identity with respect to p, we have $n(q+p)^{n-1}=\Sigma\,{}^nC_t\, t\, p^{t-1} q^{n-t}$, whence

$$np(q+p)^{n-1}=\Sigma\,{}^nC_t\, t\, p^t q^{n-t} \quad \ldots(b)$$

And now differentiating (b) with respect to p,

$$n(q+p)^{n-1}+n(n-1)p(q+p)^{n-2}=\Sigma\,{}^nC_t\, t^2 p^{t-1} q^{n-t} \quad \ldots(c)$$

Substituting from (*b*) and (*c*) in (*a*), and using $q+p=1$, we get (*a*) as $$np+n(n-1)p^2-2n^2p^2+n^2p^2=np-np^2=npq.$$

The same type of proof can be used in the other cases also.

It may be useful at this stage to observe, in the notation of moments as stated at p. 254; B; 27, that the relationships established in the text, and above, are

$$\mu_1'=np \quad \ldots.(4)$$

and $$\mu_2'=np(np+q) \quad \ldots.(5)$$

or $$\mu_2=npq \quad \ldots.(6)\text{–}(7)$$

By the same methods of proof we may also obtain easily

$$\mu_3'=n^3p^3+3n^2p^2q+npq(q-p)$$

or $$\mu_3=npq(q-p)$$

and $$\mu_4'=n^4p^4+6n^3p^3q-n^2p^2q(4p-7q)+npq(1-6pq)$$

or $$\mu_4=npq(1-6pq)+3n^2p^2q^2$$
$$=npq[3(n-2)pq+1].$$

Other methods of derivation are given in P:*59*:109-110 and P:*51*:107-110. In the proofs shown in the last reference, and in some others to be found in the text-books, these moments of the binomial are given for the expansion $(p+q)^n$, instead of $(q+p)^n$ as used here, with the result that p and q are interchanged in the results. The student should accordingly note, for example, that in P:*51*:110, Hardy thus states $\mu_3=npq(p-q)$ for the binomial $(p+q)^n$, in contrast with $\mu_3=npq(q-p)$ for $(q+p)^n$ shown in P:*59*:110, and in this study.

B; 5. Derivation of the Normal Law of Deviations, and Skew-Normal Forms

Substituting (9) as $\sqrt{2\pi}\,n^{n+\frac{1}{2}}e^{-n}$ for $n!$ in the factorials of (2), the latter becomes

$$y_x \doteqdot \frac{1}{(2\pi)^{\frac{1}{2}}} n^{n+\frac{1}{2}}(np+x)^{-np-x-\frac{1}{2}}(nq-x)^{-nq+x-\frac{1}{2}} p^{np+x} q^{nq-x}$$

$$\doteqdot \frac{1}{(2\pi npq)^{\frac{1}{2}}}\left(1+\frac{x}{np}\right)^{-np-x-\frac{1}{2}}\left(1-\frac{x}{nq}\right)^{-nq+x-\frac{1}{2}}$$

Taking the natural logarithm, and expanding, we then obtain

$$\log [y_x(2\pi npq)^{\frac{1}{2}}] \doteqdot -(np+x+\tfrac{1}{2})\left[\frac{x}{np}-\frac{x^2}{2n^2p^2}+\frac{x^3}{n^3}\varphi_1(x)\right]$$
$$-(nq-x+\tfrac{1}{2})\left[-\frac{x}{nq}-\frac{x^2}{2n^2q^2}-\frac{x^3}{n^3}\varphi_2(x)\right]$$

where $\phi_1(x)$ and $\phi_2(x)$, the remainder terms, are convergent series.

Now if it be assumed that n is so large that the terms involving in their denominators powers of n above the first may be neglected, this becomes

$$\frac{x}{2n}\left(\frac{1}{q}-\frac{1}{p}\right)-\frac{x^2}{2n}\left(\frac{1}{p}+\frac{1}{q}\right)=\frac{x}{2npq}(p-q)-\frac{x^2}{2npq}.$$

Hence
$$y_x \doteqdot \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}} e^{\frac{x(p-q)}{2npq}} \quad \ldots.(i)$$

In this expression the exponent $\frac{x(p-q)}{2npq}$ differs but little from zero when n is so large that $\frac{x}{np}$ and $\frac{x}{nq}$ may be neglected, and becomes zero when $p=q=\frac{1}{2}$. On those conditions of approximation, however, $\frac{x^2}{2npq}$ will not be negligible. We therefore reach finally

$$y_x \doteqdot \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}} \quad \ldots.(10)$$

as a representation of the probability of a deviation of $+x$ under the particular conditions assumed.

Numerical illustrations are given at p. 266; C; 4.

It will be noted that (i), involving x as well as x^2, represents a slightly unsymmetrical (skew) series of ordinates with $p \neq q$, and that when $p=q=\frac{1}{2}$ it reduces to the symmetrical form (10) involving x^2 only and not x. When written, like (11), as a continuous curve

$$f(x)=\frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}e^{\frac{x(p-q)}{c^2}} \qquad \ldots.(ii)$$

it is sometimes called appropriately (P:*27*:181) the *Skew-Normal Curve.* As will be seen from the illustrations on p. 266; C; 4, however, the effect of the skew term $e^{\frac{x(p-q)}{c^2}}$ is very slight, and may generally be neglected except when q (or p) is so small and n sufficiently large that nq (or np) remains finite but small. Even in those cases it will be preferable to use the Poisson exponential (55), which is developed later in Chapter VII, rather than this Skew-Normal form. The rationale of the expression, nevertheless, is of interest in connection with the development of generalized frequency curves, as set out in Chapter VII.

It may also be noted here (for later use in Chapter VII) that the retention of certain terms neglected in the derivation of (i) leads to $y_x \doteq \frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2npq}+\frac{x(p-q)}{2npq}+\frac{x^3(q-p)}{6(npq)^2}}$. The proof is given easily in H:*87*:33. This expression evidently may also be written approximately as

$$y_x \doteq \frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2npq}}\left[1-\frac{x(q-p)}{2npq}+\frac{x^3(q-p)}{6(npq)^2}\right],$$

which is $y_x \doteq \frac{1}{c\sqrt{\pi}}e^{-\frac{x^2}{c^2}}\left[1-2j\left(\frac{x}{c}-\frac{2x^3}{3c^3}\right)\right]$ where $c=\sqrt{2npq}$, and

$j=\frac{q-p}{2\sqrt{2npq}}=\frac{\mu_3}{c^3}$, since $\mu_3=npq(q-p)$ as stated at p. 203; B; 4.

[This last form, which is given later in the text as Edgeworth's formula (59) of Chapter VII, may be encountered by the student in Bowley's discussions of Edgeworth's work. It should therefore be remarked that Bowley's analyses in H:*162*:33, 47, etc., and in H:*107*:329 et seq., use $j=\frac{\mu_3}{c^3}=\frac{q-p}{2\sqrt{2npq}}$ as above, basing the formulae on the point binomial $(q+p)^n$ as in (3) et seq.; in H:*87*: 33-36, however, his analysis is based on $(p+q)^n$, so that $j = \frac{\mu_3}{c^3} = \frac{p-q}{2\sqrt{2npq}}$ in that case, as pointed out at p. 203; B; 4 here.]

Retaining still further terms, the result is Edgeworth's (60) in Chapter VII, where $i = \frac{1}{4}\left(\frac{\mu_4}{\mu_2^2} - 3\right)$, which is $i = \frac{1-6pq}{4npq}$ for the point binomial $(q+p)^n$, as stated in H:*162*:47 (cf. p. 203; B; 4).

An extended algebraic analysis of the form taken by the expressions in the point binomial case has been given recently by Shannon in P:*126*:380-1. The Skew-Normal approximation is there deduced for the point binomial $(p+q)^n$, so that (i), which is based on $(q+p)^n$, becomes $\frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}} e^{\frac{x(q-p)}{2npq}}$. The statement of this function, moreover, is given separately for negative and positive values of the deviation x, so that the numerator of the exponent of the skew term appears as $t_2(q-p)$ on the positive side for a positive deviation t_2, but as $t_1(p-q)$ on the negative side for a negative deviation t_1.

B; 6. Finite Integrations of the Normal Law of Deviations

The dependence of the Normal Law of Deviations, (10), upon integral values of $np+x$ imposes the restriction that any functions involving summation derivable therefrom shall be determined by finite integration.

The probability, for example, that in the n trials the number of actual occurrences of the event will lie between $np-k$ and $np+k$ will be the summation $\sum_{x=-k}^{x=+k} y_x$. For the Normal Curve of Error, however, x (and c) may take any values; the corresponding probability deducible from (11) is therefore the integral $\int_{-k}^{+k} y_x dx$.

The algebraic processes of finite integration, accordingly, should be employed for the former, and the infinitesimal calculus for the latter; and the demonstrations (to be given later) of certain functions derivable from these two expressions should either recognize the distinction by the use of parallel proofs based respectively on finite and infinitesimal summation, or should employ a sufficiently accurate approximation between $\sum_{x=-k}^{x=+k} y_x$ and $\int_{-k}^{+k} y_x dx$.

The establishment of parallel proofs, of course, is quite simple when the summations or integrations are performed over the entire range. Under those circumstances the resulting formulae are identical, as will be seen from the derivation of the mean number of successes, np, for the point binomial in (4), and for the Normal Law of Deviations (10)—or for the Normal Curve of Error (11) when $c=\sqrt{2npq}$—and similarly for the mean square deviation, npq. The demonstration of the average deviation (irrespective of sign), which is given in this study for the Normal Curve as $\frac{c}{\sqrt{\pi}}$, may likewise easily be shown to be $\sqrt{\frac{2npq}{\pi}}$ for the integral case of the point binomial—see, for example, P:*51*:110 (from J.I.A. XXVII, 214).

When the entire range is not covered, an approximation between $\sum_{x=-k}^{x=+k} y_x$ and $\int_{-k}^{+k} y_x dx$ is required. It can of course be effected to any desired degree of accuracy by using the well-known formulae for approximate summation, as illustrated in P:*51*:122 and P:*16*:375. (The problem has also been examined elaborately in H:*39*:249, and rules for computation are given in P:*16*:377.) Remembering, however, that, by neglecting the differential coefficients which arise in the formulae of approximate summation, the ordinate y_n may be regarded practically as representing $\int_{n-\frac{1}{2}}^{n+\frac{1}{2}} y_x dx$, it will generally be sufficient to extend the limits of integration by $\frac{1}{2}$ at each end and so take the integral as $\int_{-k-\frac{1}{2}}^{k+\frac{1}{2}} y_x dx = 2\int_{0}^{k+\frac{1}{2}} y_x dx$ (since the expressions (10) and (11) are symmetrical), or to leave the limits of integration unchanged but to add $\frac{1}{2}(y_k+y_{-k})$, which is y_k, and thus employ $\int_{-k}^{+k} y_x dx + y_k = 2\int_0^k y_x dx + y_k$. (See p. 268; C; 5 here for the method and examples of numerical computation, and also P:*174*:48, H:*39*:249, and P:*116*:35.) When the whole range is covered by taking $k=\infty$, y_{-k} and y_k, of course, vanish, so that $\int_{-\infty}^{+\infty} y_x dx$ may be substituted for $\sum_{x=-\infty}^{x=+\infty} y_x$, and then, as already noted, the expressions appropriate to the integral case are those resulting from the integrations of (11) when $c=\sqrt{2npq}$.

In calculating such values (see p. 160; A; 5) the integral $\int_0^k y_x dx$, where $y_x = \frac{1}{\sqrt{2\pi npq}} e^{-\frac{x^2}{2npq}}$, is taken, by changing the variable to $t=\frac{x}{c}$, as $\frac{1}{\sqrt{\pi}}\int_0^{\frac{k}{c}} e^{-t^2} dt$ where $c=\sqrt{2npq}$. Similarly,

$$\int_0^{k+\frac{1}{2}} y_x dx = \frac{1}{\sqrt{\pi}} \int_0^{\frac{1}{c}(k+\frac{1}{2})} e^{-t^2}\, dt.$$

B; 7. The Integrations of the Normal Curve

In the following demonstrations based on the use of the integral calculus, the probability of a "deviation" or "error" between x and $x+\delta x$, where δx is infinitesimally small, will, of course, be taken as $y_x \delta x$ if y_x is the probability of a deviation or error x. While this will be entirely clear to those students who are thoroughly familiar with the concepts of the calculus, it may nevertheless be advisable to observe that the principle follows at once from the fact that y_x represents a probability curve, so that the probability of an error between given limits is equal to the area under the probability curve between those limits. When the limits are infinitesimally small, this area is represented simply by a rectangle of which the height is y_x and the base the infinitesimal δx.

Alternatively, the derivation of (12) may be viewed simply as the calculation of the average value of the deviation or error x, from the continuous curve (11), and hence—omitting $\frac{1}{c\sqrt{\pi}}$ from both numerator and denominator—being

$$\frac{\int_{-\infty}^{+\infty} xe^{-\frac{x^2}{c^2}}\,dx}{\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}\,dx} = \frac{0}{c\sqrt{\pi}} \text{ by } (c) \text{ and } (a) \text{ here, } = 0.$$

The integrals in this expression, and in (12), (13), (14), and (16) are found as follows:

(*a*) In $\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}\,dx$, put $\frac{x}{c} = t$ and hence $dx = c\,dt$. The integral therefore $= c\int_{-\infty}^{+\infty} e^{-t^2}\,dt = 2c\int_0^{\infty} e^{-t^2}\,dt = c\sqrt{\pi}$ by (*b*) here.

(*b*) $\int_0^{\infty} e^{-t^2}dt = \frac{\sqrt{\pi}}{2}$, which may be shown thus: Call $\int_0^{\infty} e^{-t^2}\,dt = u$. Then, substituting at for t, $\int_0^{\infty} e^{-a^2t^2}\,a\,dt = u$, whence (multiplying each side by e^{-a^2}), $\int_0^{\infty} e^{-a^2(1+t^2)}\,a\,dt = ue^{-a^2}$. Integrating both sides now with regard to a, $\int_0^{\infty}\int_0^{\infty} e^{-a^2(1+t^2)}a\,da\,dt = \int_0^{\infty} ue^{-a^2}\,da$ $= u^2$. But $\int_0^{\infty} e^{-a^2(1+t^2)}\,a\,da = \left[-\frac{e^{-a^2(1+t^2)}}{2(1+t^2)}\right]_0^{\infty} = \frac{1}{2(1+t^2)}$. Hence $u^2 = \frac{1}{2}\int_0^{\infty}\frac{dt}{1+t^2} = \frac{1}{2}\left[\tan^{-1}t\right]_0^{\infty} = \frac{1}{2}\left(\frac{\pi}{2} - 0\right) = \frac{\pi}{4}$, so that $u = \frac{\sqrt{\pi}}{2}$.

(*c*) $\int xe^{-\frac{x^2}{c^2}}\,dx$ is obtainable by direct integration, and is $-\frac{c^2e^{-\frac{x^2}{c^2}}}{2}$.

(*d*) In (13) we deal with twice the integral from 0 to ∞ when sign is disregarded, instead of the integral from $-\infty$ to $+\infty$ when sign is taken into account as in (12).

(*e*) To evaluate $\int_0^\infty x^2 e^{-\frac{x^2}{c^2}}\,dx$, integrate by parts with $xe^{-\frac{x^2}{c^2}}$ as one part, and thus obtain $\left[\frac{-xc^2e^{-\frac{x^2}{c^2}}}{2}\right]_0^\infty + \frac{c^2}{2}\int_0^\infty e^{-\frac{x^2}{c^2}}\,dx$. Since the bracket vanishes, the remaining integral becomes $\frac{c^3}{2}\int_0^\infty e^{-t^2}\,dt$ by putting $\frac{x}{c}=t$, and this, by (*b*) here, is $\frac{c^3\sqrt{\pi}}{4}$.

B; 8. The Integration in the Case of Two Independent Observed Quantities

The evaluation of

$$\frac{1}{c\sqrt{\pi}}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}\,dx\,\frac{1}{k\sqrt{\pi}}\int_{z-x}^{z+\delta z-x} e^{-\frac{y^2}{k^2}}\,dy \quad \ldots.(24)$$

is effected as follows (cf., for instance, H:*32*:28-33; P:*80*:128; or P:*122*:292-5).

When δa is an infinitesimally small increment we have, by the geometrical principles of the integral calculus (the "mean ordinate rule" for integration), that $\int_a^{a+\delta a}\varphi(x)\,dx = \frac{1}{2}[\varphi(a+\delta a)+\varphi(a)]\delta a$, since the area depicted by the integral is a rectangle with base δa and mean height $\frac{1}{2}[\varphi(a+\delta a)+\varphi(a)]$. In (24) above, $\int_{z-x}^{z+\delta z-x} e^{-\frac{y^2}{k^2}}\,dy$ may accordingly be written

$$\frac{1}{2}\left[e^{-\frac{(z+\delta z-x)^2}{k^2}}+e^{-\frac{(z-x)^2}{k^2}}\right]\delta z = e^{-\frac{(z-x)^2}{k^2}}\,\delta z,$$

since δz is infinitesimally small.

In now replacing the second integral of (24) by this value it is to be remembered that x may take any value from $-\infty$ to $+\infty$; the expression $e^{-\frac{(z-x)^2}{k^2}}\,\delta z$ must therefore be brought under the first integral, so that (24) becomes

$$\frac{1}{ck\pi}\int_{-\infty}^{+\infty} e^{-\frac{x^2}{c^2}}e^{-\frac{(z-x)^2}{k^2}}\,\delta z\,dx = \frac{\delta z}{ck\pi}\int_{-\infty}^{+\infty} e^{-\left[\frac{x^2}{c^2}+\frac{(z-x)^2}{k^2}\right]}\,dx \ldots(24a)$$

Now

$$\frac{x^2}{c^2}+\frac{(z-x)^2}{k^2}=\frac{1}{c^2k^2}[x^2k^2+c^2(z-x)^2]$$
$$=\frac{1}{c^2k^2}[x^2(k^2+c^2)-2xzc^2]+\frac{z^2}{k^2}=\frac{k^2+c^2}{c^2k^2}\left[x^2-\frac{2xzc^2}{k^2+c^2}\right]+\frac{z^2}{k^2}.$$

Adding within the bracket $\left(\frac{zc^2}{k^2+c^2}\right)^2$ to make a perfect square, and subtracting correspondingly outside, this becomes

$$\frac{k^2+c^2}{c^2k^2}\left(x-\frac{zc^2}{k^2+c^2}\right)^2+\frac{z^2}{k^2}-\frac{z^2c^2}{k^2(k^2+c^2)}$$
$$=\frac{k^2+c^2}{c^2k^2}\left(x-\frac{zc^2}{k^2+c^2}\right)^2+\frac{z^2}{k^2+c^2}.$$

Putting $\frac{\sqrt{k^2+c^2}}{ck}\left(x-\frac{zc^2}{k^2+c^2}\right)=t$, so that $dx=\left(\frac{ck}{\sqrt{k^2+c^2}}\right)dt$,

and remembering that z is to be regarded as constant, (24a) may therefore be written

$$\frac{\delta z}{ck\pi}e^{-\frac{z^2}{k^2+c^2}}\int_{-\infty}^{+\infty}e^{-t^2}\left(\frac{ck}{\sqrt{k^2+c^2}}\right)dt$$
$$=\frac{\delta z}{\pi\sqrt{k^2+c^2}}e^{-\frac{z^2}{k^2+c^2}}\quad(\sqrt{\pi})\quad\text{by } (b) \text{ on p. 209; B; 7}$$
$$=\frac{1}{\sqrt{\pi}\sqrt{k^2+c^2}}e^{-\frac{z^2}{k^2+c^2}}\delta z.$$

A useful geometrical interpretation of the preceding proof has been given by Crofton (H:*42*), and is reproduced (with slight changes) in P:*13*:25.

The derivation of (24) may be explained in an even more elementary manner by setting out the various combinations of errors which might occur when $x+y=z$, i.e., $x-nh, \ldots, x-h, x, x+h, \ldots, x+nh$ in F_1, and $y+nh, \ldots, y+h, y, y-h, \ldots, y-nh$ in F_2, then assigning to each its probability, and combining—as is shown in H:*32*:29.

In many texts the student will meet an explanation of (26) which may be extended as follows: Any error x in F_1, and y in F_2, producing an error $x+y$, say X, in F_1+F_2, will give a square of error $X^2=x^2+y^2+2xy$. This being true for all values of x and y, the mean values of the two sides of the last equation will be equal; and since it is known that the means of the squares of the errors in F_1 and F_2 are σ_1^2 and σ_2^2, it follows that the mean square error, say E^2, in F_1+F_2 will be given by $E^2=\sigma_1^2+\sigma_2^2+2$ (mean value of the product xy). But the errors x in F_1 follow a Normal Curve, in which particular positive and negative values of x are equally likely, while similarly the errors y in F_2 obey another Normal Curve with particular positive and negative values of y again equally likely. Whatever parameters those two normal laws have, therefore (i.e., whether the normal curves show wide or narrow distributions of their respective error systems), there is an equal probability that a particular value of x will be associated with a negative or with a positive value of y, and vice versa. Since positive and negative values thus cancel each other throughout, the mean value of the product of errors, xy, must be zero. Consequently, $E^2=\sigma_1^2+\sigma_2^2$.

B; 9. The "Dispersion" (Standard Deviation) under Bernoulli, Poisson, and Lexis Sampling

The derivation of the formulae will be seen most easily by following three urn schemata (cf. P:*36*:117-122, P:*27*:323-8, and P:*116*:146-155).

(*a*) For the Bernoulli sampling, suppose that we have ν urns, containing white and black balls, knowing that in each urn the true probability of drawing a white ball is p. We draw a set of n balls (one at a time, the balls being replaced each time) from the first urn, U_1. Suppose we get α_1 white balls. Then from the second urn, U_2, we similarly draw n balls—remembering that the probability of a white ball is again p; and suppose we get α_2 white balls. If we continue in exactly the same way with each of ν urns—p having the same value for every urn (so that $p_1=p_2=\ldots=p_\nu=p$)—the sequence $\alpha_1, \alpha_2, \ldots, \alpha_\nu$ forms a sampling of the *Bernoulli* type.

This is clearly the same as if we had ν different districts, each populated by *a single group* of n males of a certain age, for whom the probability of survival (represented by the proportion of white balls) is under examination, and in each of which the "true" probability of survival is a *constant* p.

(*b*) For the Poisson case, suppose that there are n urns with white and black balls—the proportions of white balls being different in each urn, or $p_1, p_2, \ldots, p_n$ respectively. By drawing one ball from each urn we get a set of n balls altogether; suppose that a_1 of them are white. Having replaced them, the process is repeated—giving a second set, of which a_2 are white. Making ν sets of drawings thus, the sequence $a_1, a_2, \ldots, a_\nu$ forms a sampling of the *Poisson* type.

This scheme can be visualized as ν districts, each populated by *n separate* males (of n different ages) with *different* true probabilities of survival $p_1, p_2, \ldots, p_n$. The average chances are the same for each universe from which a sample is drawn, but they vary from group to group within the universe.

(*c*) For the Lexis case, suppose that there are ν urns, as in the Bernoulli scheme, with white and black balls—the proportions of white balls again being different in each urn, or $p_1, p_2, \ldots, p_\nu$ respectively, as for the Poisson case. But balls are now drawn as for the Bernoulli scheme. That is, we draw a set of n balls (one at a time, the balls being replaced) from the first urn, U_1—getting, say, a_1 white balls; then another set of n balls from U_2, giving a_2 white balls; and so on up to the final set of n from U_ν, getting a_ν white balls. Then the resulting $a_1, a_2, \ldots, a_\nu$ form a sampling of the *Lexis* type.

Here we can imagine ν different districts, each populated by *a single group* of n males of a certain age, but each group having a *different* true probability of survival—$p_1, p_2, \ldots, p_\nu$ respectively. The probabilities vary from universe to universe from which the samples are drawn.

Now if we were presented with a sequence $a_1, a_2, \ldots, a_\nu$ of white balls, drawn by any one of these methods, there would be no *a priori* reason for choosing any particular a rather than any other; consequently the arithmetic mean might be taken, namely, $\frac{a_1+a_2+\ldots+a_\nu}{\nu}$. What, then, would be the expected value of this arithmetic mean, and what would be the average of the mean square deviations over all the sets, i.e., σ^2; under these three urn schemes?

In order to determine these values, and at the same time show the three cases in parallel forms, we can use the same principles as those adopted for the Bernoulli case in reaching formulae (4) and (7).

(*a*) To illustrate *Bernoulli* sampling again here in the simple urn-drawing scheme, we find

In Set No. 1, the expected value of a_1 is np
" " " 2, " " " " a_2 is np
.
.
.
In Set No. ν, the expected value of a_ν is np.

Consequently the expected value of

$$\frac{a_1+a_2+\ldots+a_\nu}{\nu}=\frac{\nu(np)}{\nu}=np.$$

Similarly for the mean square deviation—remembering that, as shown in (7), the expected value of the squares of the deviations measured from the mean in the Bernoulli case is npq—we have

In Set No. 1, the expected value of $(a_1-np)^2$ is npq
" " " 2, " " " " $(a_2-np)^2$ is npq
.
.
.
In Set No. ν, the expected value of $(a_\nu-np)^2$ is npq. Summing, therefore, and averaging by dividing by ν, we find

$$\sigma_B^2=\frac{\nu(npq)}{\nu}=npq.$$

(*b*) For the *Poisson* case the same method gives

In Set No. 1, the expected value of a_1 is $(p_1+p_2+\ldots+p_n)$
" " " 2, " " " " a_2 is $(p_1+p_2+\ldots+p_n)$
.
.
.

In Set No. ν, the expected value of a_ν is $(p_1+p_2+\ldots+p_n)$. The expected value of

$$\frac{a_1+a_2+\ldots+a_\nu}{\nu} \text{ is therefore } \frac{\nu(p_1+p_2+\ldots+p_n)}{\nu}=np$$

where
$$p=\frac{p_1+p_2+\ldots+p_n}{n}.$$

In examining the mean square deviation here it must be noted that one ball is first drawn from U_1 with the knowledge that the probability of getting a white ball is p_1; this is the same as n trials with probabilities p_1 and q_1 when $n=1$; the mean square deviation for this single trial is therefore np_1q_1 where $n=1$, or p_1q_1; and similarly for each subsequent ball drawn to make the first set of n. Consequently, since the trials are all independent, σ^2 for their sum is to be obtained, by (27), from the sum of the various values of σ^2, so that in Set No. 1, the mean square deviation is $p_1q_1+p_2q_2+\ldots+p_nq_n=\sum_{t=1}^{t=n} p_tq_t$, and in every other set up to Set No. ν we get the same result. Summing, therefore, and dividing by ν, as in the Bernoulli case, we find

$$\begin{aligned}\sigma_P^2=\sum_{t=1}^{t=n} p_tq_t &= \sum_{t=1}^{t=n}\left[\{p+(p_t-p)\}\{q-(p_t-p)\}\right]\\ &=\sum_{t=1}^{t=n}\left[pq-(p_t-p)(p-q)-(p_t-p)^2\right]\\ &=npq-\sum_{t=1}^{t=n}(p_t-p)^2, \text{ since } \sum_{t=1}^{t=n}(p_t-p)=0,\\ &=\sigma_B^2-n\sigma_p^2 \text{ where } \sigma_p^2 \text{ is } \frac{1}{n}\sum_{t=1}^{t=n}(p_t-p)^2,\end{aligned}$$

being the mean square deviation of the probabilities $p_1, p_2, \ldots, p_n$ from their mean p.

(*c*) In the *Lexis* case it is evident that

In Set No. 1, the expected value of a_1 is np_1
" " " 2, " " " " a_2 is np_2
.
.
.

In Set No. ν, the expected value of a_ν is np_ν, whence the expected value of $\frac{a_1+a_2+\ldots+a_\nu}{\nu}$ is $\frac{n(p_1+p_2+\ldots+p_\nu)}{\nu} = np$ where $p = \frac{p_1+p_2+\ldots+p_\nu}{\nu}$.

Consider now the mean square deviation. In drawing n balls from U_1, when the probability of a white ball is p_1, the mean square deviation of white balls from np_1 is np_1q_1; similarly for U_2, when the probability of a white ball is p_2, the mean square deviation of white balls from np_2 is np_2q_2; and likewise for each drawing, until finally the mean square deviation of white balls from np_ν in drawing n balls from U_ν is $np_\nu q_\nu$. These several mean square deviations are measured from the means, np_1, np_2, ..., np_ν, respectively. But they are here, in conformity with the Bernoulli and Poisson cases, to be measured from np (where $p = \frac{p_1+p_2+\ldots+p_\nu}{\nu}$ as above); we must therefore (see p. 255; B; 27) add to each of the preceding the respective values $(np_1-np)^2$, $(np_2-np)^2$, ..., $(np_\nu-np)^2$. We thus have $np_tq_t+(np_t-np)^2$ for the mean square deviation of white balls from np in drawing n balls from U_t $(t=1, 2, 3, \ldots, \nu)$. Summing and dividing by ν as in the derivation of the corresponding expressions for the Bernoulli and Poisson series we find that for the Lexis series $\sigma_L^2 = \frac{n}{\nu}\sum_{t=1}^{t=\nu} p_tq_t + \frac{n^2}{\nu}\sum_{t=1}^{t=\nu}(p_t-p)^2$. It was shown incidentally, however, in the Poisson case that $\sum_{t=1}^{t=\nu} p_tq_t$ can be written as $\nu pq - \sum_{t=1}^{t=\nu}(p_t-p)^2$; consequently $\sigma_L^2 = npq + \frac{n^2-n}{\nu}\sum_{t=1}^{t=\nu}(p_t-p)^2$ $=\sigma_B^2+(n^2-n)\sigma_p^2$ where σ_p^2 is $\frac{1}{\nu}\sum_{t=1}^{t=\nu}(p_t-p)^2$, being the mean square deviation of the probabilities p_1, p_2, ... p_ν from their mean p.

Another type of proof employing "generating functions" is given conveniently in P:*3*:16-25 and 49-53. A further discussion of the assumptions underlying these Poisson and Lexis cases may also be found in P:*146*:208-234.

The *Charlier Coefficient of Disturbancy* is founded on the last formula. Dividing by p^2 we have

$$\frac{\sigma_p^2}{p^2} = \frac{\sigma_L^2 - \sigma_B^2}{(n^2-n)p^2} \div \frac{\sigma_L^2 - \sigma_B^2}{n^2p^2} = \frac{\sigma_L^2 - \sigma_B^2}{A^2}$$

where A is written for the mean np. Charlier's Coefficient C is taken as $\frac{100\sqrt{\sigma^2 - \sigma_B^2}}{A}$.

The Relation between the Lexis and χ^2 Methods

The equivalence of L^2 and $\frac{\chi^2}{\nu}$, in testing the Bernoullian hypothesis of constant probability, as remarked by Irwin (P:*63*: 507), "has perhaps been insufficiently appreciated". Even though the student will not encounter the χ^2 method until later in this text, it may therefore be well to establish the relationship here, in order to emphasize R. A. Fisher's statement (P:*43*:83) that "in many references in English . . . it has not . . . been noted that the discovery of the distribution of χ^2 in reality completed the method of Lexis".

The demonstration—following Irwin's discussion in P:*63*: 507—may be based on the mathematical model used in (ii) on p. 245; B; 24. If we are given, for example, an observed series f'_r for values of r from 1 to ν, which may be regarded as samples from a population of n in which the true probability of occurrence is p (and of failure, q), so that the true mean is np, the Lexis function L^2 is found by (41) as $\frac{\frac{1}{\nu}\sum_{r=1}^{r=\nu}(f'_r - np)^2}{npq}$. In the χ^2 method, on the other hand (as explained further in Chapter VI, and in (ii) on p. 245; B; 24 and (4) on p. 337; C; 25), the hypothesis of constant probability is tested by computing

$$\chi^2 = \sum_{r=1}^{r=\nu}\left[\frac{(f'_r - np)^2}{np} + \frac{\{(n-f'_r)-(n-np)\}^2}{n-np}\right]$$

with ν "degrees of freedom." This expression, since $p+q=1$, reduces to $\frac{\sum_{r=1}^{r=\nu}(f'_r - np)^2}{npq}$; and this is νL^2, so that L^2 and $\frac{\chi^2}{\nu}$ are equivalent.

B; 10. Tchebycheff's Inequality and its Extensions

In 1853 Bienaymé (H:*29*) suggested certain fundamental ideas in connection with the Law of Large Numbers, which in 1867 were established independently by Tchebycheff (H:*37*), and since that time have undergone development at the hands of Tschuprow, Markoff, Bernstein, Khintchine, Guldberg, Meidell, Pearson, Camp, and others.

The *Theorem, or Inequality, of Tchebycheff*, or the *Bienaymé-Tchebycheff Criterion*—as the basic inequality is variously named—states that the probability is $\leq \frac{1}{a^2}$ that a value of x, taken at random from n values $x_1, x_2, \ldots, x_n$, will differ from their mean by as much as or more than $a\sigma$ (where a is >1, and σ is the standard deviation). For example, the probability is not greater than $\frac{1}{16}$ that a variate taken at random from that "population" will deviate from the arithmetic mean by at least 4 times the standard deviation. Similarly, it follows that the probability is greater than $\frac{15}{16}$ that the deviation will be less than 4 times the standard deviation. For, writing m for the mean, we have $m = \frac{1}{n}\sum_{r=1}^{r=n} x_r$

and

$$\sigma^2 = \frac{1}{n}\sum_{r=1}^{r=n}(x_r - m)^2 \qquad \ldots (a)$$

and by Tchebycheff's inequality not more than $\frac{n}{a^2}$ of the x's can deviate from m by more than $a\sigma$. In order to establish this, suppose that $\frac{n}{a^2}$ of them do so deviate by more than $a\sigma$ each; the sum of the squares of their deviations would then be greater than $\frac{n}{a^2}(a\sigma)^2$, that is, greater than $n\sigma^2$, which is inconsistent with (a), and is therefore impossible.

The derivation of Bernoulli's Theorem from this inequality proceeds by investigating the probability (see p. 187; B; 2) that the deviation $\left|\frac{s}{n} - p\right|$ will be numerically greater than $a\left(\sqrt{\frac{pq}{n}}\right)$, since, by (22), $\sigma\left\{\frac{s}{n}\right\} = \sqrt{\frac{pq}{n}}$. By Tchebycheff's criterion, this probability does not exceed $\frac{1}{\alpha^2}$. Writing $\alpha\left(\sqrt{\frac{pq}{n}}\right)=\epsilon$, it follows that the probability in question does not exceed $\frac{pq}{\epsilon^2 n}$. Here pq can never exceed $\frac{1}{4}$, and for any assigned ϵ the expression approaches zero as the number of trials, n, is increased indefinitely—that is to say, the probability of obtaining a deviation greater than ϵ tends ultimately to 0, and consequently the probability of a deviation less than any assigned positive quantity ϵ approaches 1 as the limit.

Tchebycheff's inequality places an upper limit on the probability that a variate will deviate from the mean by at least a given multiple of the standard deviation, and it will be observed that no restriction is placed upon the nature of the distribution of the population values $x_1, x_2, \ldots, x_n$. This important feature led Karl Pearson (H:*128*) to the extension that the probability is $\leqq \frac{\mu_{2r}}{\alpha^{2r}\sigma^{2r}}$ that a value of x represented by a continuous function will deviate from the mean by at least $\alpha\sigma$ (which reduces to Tchebycheff's inequality when $r=1$), and Camp has evolved (H:*136*) a still further generalization which includes both of the preceding cases (see P:*116*:143, and also P:*75*:353 and P:*146*:182 for the contributions of the Russian mathematicians).

B; 11. "Presumptive" Values

Formula (42) provides an estimate or "presumptive" value, σ_e^2, of the σ^2 in the universe, from the σ_s^2 of the sample, by the relation $\sigma_e^2=\left(\frac{n}{n-1}\right)\sigma_s^2$. In addition to the proof given on p. 35, and the alternative demonstration (p. 38) based on the Principle of Insufficient Reason, the student may be referred to P:*52*:35 (or 434-5 in the J.I.A. reprint), to P:*140*:11-20, and to P:*83*:345-352 for further analyses.

Since the standard deviation is defined, as shown in formulae (6) to (8), as the square root of the second moment about the mean, the expression for σ_e^2 may be put in terms of moments as

$\overline{m}_2 = \left(\frac{n}{n-1}\right) m_2$, where m_2 is the second moment about the mean as derived from the sample and $\overline{m}_2$ is the estimated value for the universe. (Steffensen's notation, $\overline{m}_2$ and m_2, is used here for simplicity in view of the references to his discussion given above). The corresponding relations for the estimates of the mean ($\overline{m}_1$) and for the third and fourth moments about that mean ($\overline{m}_3$ and $\overline{m}_4$), give the series of formulae

$$\left.\begin{aligned} \overline{m}_1 &= m_1 \\ \overline{m}_2 &= \left(\frac{n}{n-1}\right) m_2 \\ \overline{m}_3 &= \frac{n^2}{(n-1)(n-2)} m_3 \\ \text{and} \quad \overline{m}_4 &= \frac{n^3}{(n-1)(n^2-3n+3)}\left[m_4 - \frac{3(2n-3)}{n(n-1)} m_2^2\right] \end{aligned}\right\} \ldots (42a)$$

Algebraic proofs may be found in P:*52*: loc. cit. and P:*83*:350-2. [It should be noted, as pointed out by Tschuprow (Biometrika, XII, 187) and Lidstone (J.I.A., LXI, 346), that in P:*52* the formula for the fourth moment is incorrect, being too small by $\frac{3(n-1)}{n^2}\mu_2^2$ in the notation there employed, due to the omission of 2 before $\Sigma x_1^2 x_2^2$ in line 15 of T.A.S.A., VIII, 35, and line 9 of J.I.A., XLI, 435.]

These expressions appear (see P:*140*:14) to have been given first, in terms of half-invariants, by Thiele (H:*94*:48), who derived the series of formulae as far as the 8th half-invariant.

The criticisms which, as pointed out on p. 36, may be levelled at these "presumptive" values led Tschuprow to another set of formulae which agree with those of Thiele stated above for $\overline{m}_1$, $\overline{m}_2$, and $\overline{m}_3$, but for $\overline{m}_4$ give

$$\overline{m}_4 = \frac{n}{(n-1)(n-2)(n-3)}\left[(n^2-2n+3)m_4 - 3(2n-3)m_2^2\right].$$

Reference may be made to P:*140*:18 for further details.

The values of the m's in formulae (42a) are in reality based (as noted on pp. 35-36) on mean values derived from a large number of samples. The practical problem of which a solution is required, however, is the estimation of the values in the universe from the data furnished by a single sample. This important distinction has been emphasized by R. A. Fisher, who has derived (P:*40*, and P:*43*:75) a series of *k-statistics* for this purpose, where the k's are estimates of the population (i.e., universe) half-invariants, m' denotes the moment about the mean derived from a single sample, and M'_1 is that mean:

$$k_1 = M'_1$$

$$k_2 = \left(\frac{n}{n-1}\right) m'_2$$

$$k_3 = \frac{n^2}{(n-1)(n-2)} m'_3$$

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)} \left[(n+1)\, m'_4 - 3\,(n-1)\,(m'_2)^2\right].$$

The actuarial student may be referred conveniently to P:*160* for additional comments.

B; 12. The "Probability of Causes"; Bayes' Theorem, and Laplace's Generalization

In the mathematical model stated on p. 37, κ_1 (for example) is the *a priori* probability of the existence of F_1; and π_1 is the *a priori* probability that, when F_1 exists, the event E will happen. The product $\kappa_1\pi_1$ is therefore the probability that the event E originated from F_1. Now we are here concerned with n trials, and s successes; if the event E originates from F_1, the probability of its happening exactly s times out of n is ${}^nC_s\pi_1^s(1-\pi_1)^{n-s}$; and the probability that F_1 exists, and that then E happens s times in n trials, is consequently $\kappa_1\,{}^nC_s\,\pi_1^s(1-\pi_1)^{n-s}$. There must clearly be a similar expression for each one of the conditions $F_1, F_2, \ldots, F_n$; and since they are mutually exclusive, the total probability that one of the conditions $F_1, F_2, \ldots, F_n$ exists, and that then E will happen s times in n, is

$$\sum_{r=1}^{r=n} \kappa_r\,{}^nC_s\,\pi_r^s(1-\pi_r)^{n-s} \qquad \ldots.(43)$$

It accordingly follows, when we know, by observation, that E has actually happened s times out of n, that the probability, *a posteriori*, that the particular condition F_a was the origin is

$$\frac{\kappa_a \pi_a^s (1-\pi_a)^{n-s}}{\sum_{r=1}^{r=n} \kappa_r \pi_r^s (1-\pi_r)^{n-s}} \quad \ldots .(43a)$$

since the constant factor ${}^nC_s$ cancels from the numerator and denominator.

This is the general form of the method which is usually referred to loosely as "Bayes' Rule". It is extremely important to realize that the *a priori existence probabilities*, κ_r, as well as the *a priori productive probabilities*, π_r, must be known if the formula is to be applicable in practice. Failure to appreciate this fact has led on many occasions to misapplication of the rule, and consequently to paradoxical results.

The chief source of these paradoxes has been the tacit assumption—the "Principle of Insufficient Reason"—that all the values of κ_r can be supposed to be equal when their real values are unknown. If that assumption were sound, (43a) clearly would reduce to

$$\frac{\pi_a^s (1-\pi_a)^{n-s}}{\sum_{r=1}^{r=n} \pi_r^s (1-\pi_r)^{n-s}} \quad \ldots .(43b)$$

This simplified expression, however, obviously must be used with very great care. For Ellis' remark must always be remembered, that "mere ignorance is no ground for any inference whatever; *ex nihilo nihil*" (H:*24*).

The general formula (43a) represents a perfectly sound and logical argument, and leads to unexceptionable results when it can be applied rigorously. The famous theorem of Bayes-Laplace, in other words, cannot be challenged successfully when it is properly employed; the doubts which have been thrown upon it have arisen from the improper assumption that (43b) should be adopted whenever specific values cannot be attached to the *a priori* existence probabilities, κ_r.

Bayes' original treatment of the problem was restricted to the case where the values of κ_r are all equal. The generalization when the κ_r's are not all equal was given first by Laplace (see p. 165; A; 9).

B; 13. The Proofs of the Distributions of (1) $\bar{x}-m$; (2) σ_s; (3) "Student's z" $=\dfrac{\bar{x}-m}{\sigma_s}$; (4) $\dfrac{{}_1\sigma_e}{{}_2\sigma_e}$; and (5) "Fisher's z" $=\log_e\left(\dfrac{{}_1\sigma_e}{{}_2\sigma_e}\right)$.

Suppose that we have a parent population, N in number, distributed normally with mean m and standard deviation σ according to the Normal Curve $y=\dfrac{N}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$, which is another form of (11). If a random sample numbering n, with magnitudes $x_1', x_2', \ldots, x_n'$, is then drawn, the probability that the members of the sample will lie between x_1' and $x_1'+dx_1'$, x_2' and $x_2'+dx_2'$, ..., x_n' and $x_n'+dx_n'$, is clearly

$$\frac{N^n}{\sigma^n(2\pi)^{\frac{n}{2}}}e^{-\frac{\Sigma(x_r'-m^2)}{2\sigma^2}}dx_1'dx_2'\ldots dx_n',$$

which may be written (as already noted on p. 38)

$$Ae^{-\frac{[\Sigma(x_r'-\bar{x})^2+n(\bar{x}-m)^2]}{2\sigma^2}}dx_1'dx_2'\ldots dx_n' \qquad \ldots.(a)$$

where A is a constant and $\bar{x}$ the sample mean.

Now consider the members of the sample as a point, P', in a space of n dimensions with rectangular coordinates $x_1', x_2', \ldots, x_n'$. Then, just as with two coordinates, x and y, we find that the line $y=x$ is inclined equally to the two axes, so in space of n dimensions (cf. p. 193; B; 3) the line $x_1'=x_2'=\ldots=x_n'$ is inclined equally to the n axes. The perpendicular from P$'$ upon this line may similarly be seen to be given by $(P'M)^2=(x_1'-\bar{x})^2+(x_2'-\bar{x})^2+\ldots+(x_n'-\bar{x})^2$ where M is the point $(\bar{x}, \bar{x}, \ldots, \bar{x})$. But this expression is $n\sigma_s^2$; hence $P'M=\sigma_s\sqrt{n}$. The point P', also, must lie upon the plane $x_1'+x_2'+\ldots+x_n'=n\bar{x}$ in order to satisfy the definition of its mean; and as this plane is normal to the line $x_1' = x_2' = \ldots = x_n'$ at M, it follows that P' must lie on a sphere of $n-1$ dimensions with radius $\sigma_s\sqrt{n}$ and centre at M.

In this space an element of volume, dv, may be expressed in terms of the variation of $\bar{x}$, that is, $d\bar{x}$, and the variation $d\sigma_s^{n-1}$ in surface area, since $d\bar{x}$ represents a perpendicular distance increment above the plane, and $k\,d\sigma_s^{n-1}$ represents an increment to area enclosed by the plane within the sphere. [This may be seen from the 3-dimensional case where the plane is $x_1'+x_2'+x_3'=3\bar{x}$ and the sphere is $(x_1'-\bar{x})^2+(x_2'-\bar{x})^2+(x_3'-\bar{x})^2=3\sigma_s^2$; then $d\bar{x}$ is the perpendicular distance increment above the plane, and $d\sigma_s^2$ is proportional to the increment of area of the circle enclosed on the plane by the sphere, since the area $=3\pi\sigma_s^2$ and $d\,(\text{area}) = 3\pi d\sigma_s^2$.] Hence dv is proportional to $d\sigma_s^{n-1}d\bar{x}$, that is, to $\sigma_s^{n-2}d\sigma_s d\bar{x}$. By this device of representing the sample by a point in multiple space we therefore now see that, from the probability (*a*), we can write

$$C\,\sigma_s^{n-2}\,e^{-\frac{[\Sigma(x_r'-\bar{x})^2+n(\bar{x}-m)^2]}{2\sigma^2}}\,d\sigma_s\,d\bar{x} \qquad \ldots(b)$$

where C is some constant, for the probability that the sample will have a mean lying between $\bar{x}$ and $\bar{x}+d\bar{x}$, and a standard deviation lying between σ_s and $\sigma_s+d\sigma_s$. This can at once be put as

$$C\left[e^{\frac{-n(\bar{x}-m)^2}{2\sigma^2}}\,d\bar{x}\right]\left[\sigma_s^{n-2}e^{-\frac{n\sigma_s^2}{2\sigma^2}}\,d\sigma_s\right] \qquad \ldots(c)$$

(1) *The Distribution of the Deviation $\bar{x}-m$.*

From the fact that $\bar{x}$ and σ_s are entirely separated in this expression it follows that *the distribution of the deviation $\bar{x}-m$ of the mean of a sample from the mean of the universe*, for any and all values of σ_s, is given by the normal form

$$C_M\,e^{-\frac{n(\bar{x}-m)^2}{2\sigma^2}}\,d\bar{x} \qquad \ldots(d)$$

where C_M is a constant.

C_M is obtained by making the area of the distribution curve unity, or $1 = C_M \int_{-\infty}^{+\infty} e^{-\frac{nx^2}{2\sigma^2}} dx$ where $x = \bar{x} - m$, which gives (cf. p. 209; B; 7) $C_M = \frac{\sqrt{n}}{\sigma\sqrt{2\pi}}$.

(2) *The Distribution of the Standard Deviation* σ_s

Similarly *the distribution of the standard deviation* σ_s *of a sample*, for any and all values of $\bar{x}$, is given by the second bracket of (c), namely,

$$C_S \, \sigma_s^{n-2} \, e^{-\frac{n\sigma_s^2}{2\sigma^2}} d\sigma_s \qquad \ldots .(e)$$

where C_S is a constant.

Here C_S may be found as follows (see p. 259; B; 29 for the Γ function):

Let $\frac{n\sigma_s^2}{2\sigma^2} = x$, so that $\frac{n\sigma_s}{\sigma^2} = \frac{dx}{d\sigma_s}$.

Then

$$1 = C_S \int_0^\infty \left(\frac{2\sigma^2 x}{n}\right)^{\frac{n-2}{2}} e^{-x} \left(\frac{\sigma^2}{n\sigma_s}\right) dx$$

$$= C_S \int_0^\infty \frac{(2x)^{\frac{n-2}{2}} (\sigma^2)^{\frac{n}{2}} e^{-x}}{n^{\frac{n-2}{2}} (2nx)^{\frac{1}{2}} \sigma} dx$$

$$= C_S \int_0^\infty \frac{(2x)^{\frac{n-3}{2}} \sigma^{n-1} e^{-x}}{n^{\frac{n-1}{2}}} dx = \frac{C_S 2^{\frac{n-3}{2}} \sigma^{n-1}}{n^{\frac{n-1}{2}}} \Gamma\left(\frac{n-1}{2}\right),$$

whence

$$C_S = \frac{n^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) 2^{\frac{n-3}{2}} \sigma^{n-1}}.$$

This distribution of σ_s was given first by Helmert in 1876 (see p. 165; A; 10). It may be of assistance to note that, since any σ is necessarily positive, the distribution (e) is, in fact (as was found also by "Student" in his method of approach—see p. 165; A; 10) a skew bell-shaped Pearson Type III curve (see p. 71) limited at 0 but extending theoretically to ∞.

(3) *The Distribution of "Student's z"* $= \dfrac{\bar{x}-m}{\sigma_s}$

In order now to find the distribution of "Student's" ratio, $\dfrac{\bar{x}-m}{\sigma_s}=z$, we may keep in mind the diagram of Figure 20. Expression (c) is the probability of a sample mean and standard deviation falling in the intervals $\bar{x}\pm\frac{1}{2}d\bar{x}$ and $\sigma_s\pm\frac{1}{2}d\sigma_s$. This same expression is also the probability that the "Student" ratio, z, and the standard deviation of the sample, σ_s, will fall in the intervals $z\pm\frac{1}{2}dz$ and $\sigma_s\pm\frac{1}{2}d\sigma_s$. To rewrite expression (c) in terms of σ_s, $d\sigma_s$, z, and dz, we replace $\bar{x}-m$ by $z\sigma_s$, and $d\bar{x}$ by $\sigma_s\,dz$, and thus obtain

$$\left[C_S\,\sigma_s^{n-2}\,e^{-\frac{n\sigma_s^2}{2\sigma^2}}d\sigma_s\right]\left[C_M e^{-\frac{nz^2\sigma_s^2}{2\sigma^2}}\sigma_s\,dz\right]$$

$$=(C_S C_M)\,e^{-\frac{n\sigma_s^2}{2\sigma^2}(1+z^2)}\,\sigma_s^{n-1}d\sigma_s\,dz \qquad \ldots(f)$$

This is the simultaneous distribution of σ_s and z. It is the probability of a sample falling within the elementary shaded area of Figure 20. All sample points within the wedge have the same value of z, to within the amount dz; hence to find the probability that "Student's" ratio falls within the interval $z\pm\frac{1}{2}dz$ for *all values of the sample standard deviation* σ_s, we must sum the probabilities expressed by (f) throughout the entire wedge, i.e., we must integrate (f), allowing σ_s to vary from 0 to ∞ while z and dz remain constant during the integration. Putting $\dfrac{n\sigma_s^2}{2\sigma^2}(1+z^2)=v^2$, it follows that

$$\int_0^\infty e^{-\frac{n\sigma_s^2}{2\sigma^2}(1+z^2)}\,\sigma_s^{n-1}\,d\sigma_s=\sigma^n\left[\frac{2}{n(1+z^2)}\right]^{\frac{n}{2}}\int_0^\infty e^{-v^2}v^{n-1}\,dv.$$

This latter integral (see p. 259; B; 29) is $\frac{1}{2}\,\Gamma\left(\dfrac{n}{2}\right)$. Inserting this value, therefore, and also those of the constants C_S and C_M, we obtain the distribution of "Student's" z as

$$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)}(1+z^2)^{-\frac{n}{2}}\,dz.$$

Another form of proof based on "characteristic functions" (first used by Lagrange and later systematically by Laplace—see P:*146*:240 and 264, and P:*22*:23) may be found in P:*146*:336-9 and P:*22*:47-8.

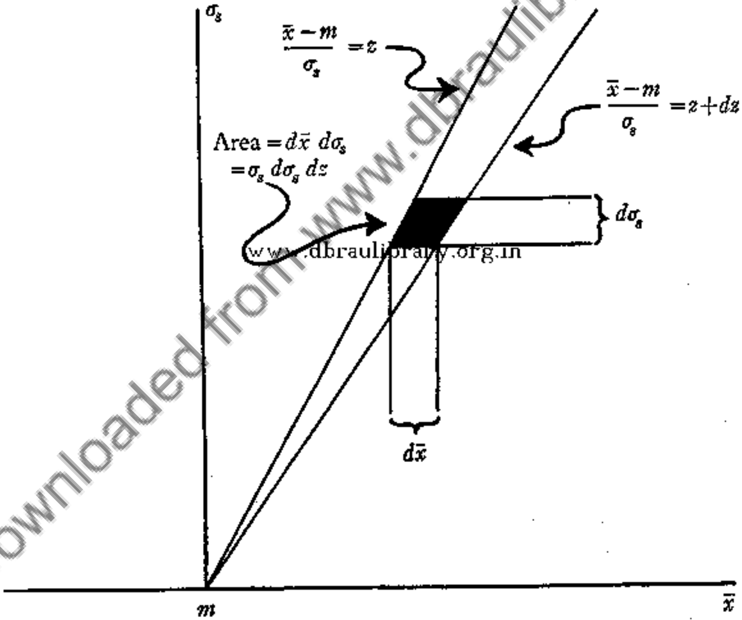

FIGURE 20.—Diagram showing the region within which "Student's" ratio z remains constant for varying sample mean $\bar{x}$ and standard deviation σ_s. All samples falling within the wedge have the same "Student's" ratio z to within the amount dz. "Student's" ratio is defined as $z=\frac{\bar{x}-m}{\sigma_s}$. When $\bar{x}$ varies by the amount $d\bar{x}$ while σ_s remains constant, z varies by the amount dz, which is related to $d\bar{x}$ by the equation $d\bar{x}=\sigma_s\,dz$.

The meaning of "Student's" distribution may also be made clearer by the use of "contours" (see p. 194; B; 3), as shown in P:*29*:140.

(4) *The Distribution of the Ratio* $\frac{{}_1\sigma_e}{{}_2\sigma_e}$

Suppose now that we have two estimates, ${}_1\sigma_e$ and ${}_2\sigma_e$, of the same σ, where ${}_1\sigma_e$ is based on a first sample of n_1 variates x'_r, and ${}_2\sigma_e$ is based on a second sample of n_2 variates x''_r, so that, by (42),

${}_1\sigma_e^2 = \frac{\sum_{r=1}^{r=n_1}(x'_r - \bar{x}_1)^2}{n_1 - 1}$ and ${}_2\sigma_e^2 = \frac{\sum_{r=1}^{r=n_2}(x''_r - \bar{x}_2)^2}{n_2 - 1}$, and the respective "degrees of freedom", say d_1 and d_2, are therefore $d_1 = n_1 - 1$ and $d_2 = n_2 - 1$.

Considering the first sample, we see at once that if we use the relation (42) between the sample variance, ${}_1\sigma_s^2$ say, and the estimate, ${}_1\sigma_e^2$, we have here approximately ${}_1\sigma_e^2 = \left(\frac{n_1}{n_1 - 1}\right){}_1\sigma_s^2$, whence

${}_1\sigma_s^2 = \left(\frac{n_1 - 1}{n_1}\right){}_1\sigma_e^2 = \left(\frac{d_1}{d_1 + 1}\right){}_1\sigma_e^2$, and $d\,{}_1\sigma_s = \left(\frac{d_1}{d_1 + 1}\right)^{\frac{1}{2}} d\,{}_1\sigma_e$. From (*e*) in (2) of this section B;13 we can therefore, using these relations, write down immediately, in terms of the degrees of freedom, d_1, and the estimate ${}_1\sigma_e$ (instead of in terms of n_1 and ${}_1\sigma_s$), that for the distribution of ${}_1\sigma_e$ the differential is given by

$$\frac{(d_1+1)^{\frac{d_1}{2}}}{\Gamma\left(\frac{d_1}{2}\right)2^{\frac{d_1-2}{2}}\sigma^{d_1}}\left(\frac{d_1}{d_1+1}\right)^{\frac{d_1-1}{2}} {}_1\sigma_e^{\,d_1-1}\, e^{-\frac{d_1\,{}_1\sigma_e^2}{2\sigma^2}}\left(\frac{d_1}{d_1+1}\right)^{\frac{1}{2}} d\,{}_1\sigma_e$$

$$= \frac{d_1^{\frac{d_1}{2}}}{\Gamma\left(\frac{d_1}{2}\right)2^{\frac{d_1-2}{2}}} e^{-\frac{d_1\,{}_1\sigma_e^2}{2\sigma^2}}\left(\frac{{}_1\sigma_e}{\sigma}\right)^{d_1-1}\frac{d\,{}_1\sigma_e}{\sigma} \qquad \ldots.(g)$$

As we are to investigate the distribution of $\frac{{}_1\sigma_e}{{}_2\sigma_e}$, let us denote this ratio by w; then ${}_1\sigma_e = w\,{}_2\sigma_e$, and for a given value of ${}_2\sigma_e$ the distribution of ${}_1\sigma_e$ becomes

$$\frac{d_1^{\frac{d_1}{2}}}{\Gamma\left(\frac{d_1}{2}\right)2^{\frac{d_1-2}{2}}}\, e^{-\frac{d_1 w^2\,{}_2\sigma_e^2}{2\sigma^2}}\,\frac{(w\,{}_2\sigma_e)^{d_1-1}}{\sigma^{d_1}}\,{}_2\sigma_e\,dw \qquad \ldots.(h)$$

The distribution of ${}_2\sigma_e$, however, is again in the form (g) with d_2 and ${}_2\sigma_e$ written for d_1 and ${}_1\sigma_e$. To obtain the distribution of w, therefore, since ${}_1\sigma_e$ and ${}_2\sigma_e$ are independent, we merely have to multiply that expression by (h) and integrate for values of ${}_2\sigma_e$ from 0 to ∞, and so find

$$dw\int_0^\infty \frac{d_1^{\frac{d_1}{2}}\;d_2^{\frac{d_2}{2}}}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)2^{\frac{d_1+d_2-4}{2}}}\, e^{-\frac{(d_1w^2+d_2)\,{}_2\sigma_e^2}{2\sigma^2}}\left(\frac{{}_2\sigma_e^{\,d_1+d_2-1}\;w^{d_1-1}}{\sigma^{d_1+d_2}}\right)d\,{}_2\sigma_e$$

$$=\frac{2d_1^{\frac{d_1}{2}}d_2^{\frac{d_2}{2}}\,\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\;\frac{w^{d_1-1}dw}{(d_1w^2+d_2)^{\frac{d_1+d_2}{2}}} \qquad \ldots.(i)$$

(5) *The Distribution of "Fisher's z"* $=\log_e\left(\frac{{}_1\sigma_e}{{}_2\sigma_e}\right)$.

It follows immediately that if now, as R. A. Fisher does, we write $z=\log_e w=\log_e\left(\frac{{}_1\sigma_e}{{}_2\sigma_e}\right)$, so that $dw=e^z dz$, the distribution of "Fisher's z" can be written down from (i) as

$$\frac{2d_1^{\frac{d_1}{2}}d_2^{\frac{d_2}{2}}\,\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\;\frac{e^{d_1z}dz}{(d_1e^{2z}+d_2)^{\frac{d_1+d_2}{2}}} \qquad \ldots.(j)$$

from which (47z) in the text emerges at once since

$$\frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}=\frac{1}{B\left(\frac{d_1}{2},\frac{d_2}{2}\right)}$$

as defined at p. 259; B;29.

B; 14. Derivation of the Multinomial Normal Law of Deviations

Commencing with the multinomial general term

$$\frac{N!}{f_1'! \ldots f_\nu'!} p_1^{f_1'} \ldots p_\nu^{f_\nu'} \qquad \ldots (49)$$

and replacing the factorials by their approximations according to Stirling's formula as in the deduction of the Normal Law (see p. 203; B; 5), we obtain

$$\frac{1}{(\sqrt{2\pi N})^{\nu-1}\sqrt{p_1 \ldots p_\nu}} \left(\frac{Np_1}{f_1'}\right)^{f_1'+\frac{1}{2}} \ldots \ldots \left(\frac{Np_\nu}{f_\nu'}\right)^{f_\nu'+\frac{1}{2}}.$$

Writing $\frac{f_r'-Np_r}{\sqrt{Np_r}} = j_r$, each factor $\left(\frac{Np_r}{f_r'}\right)^{f_r'+\frac{1}{2}}$

$$= \left(1+\frac{j_r}{\sqrt{Np_r}}\right)^{-(j_r\sqrt{Np_r}+Np_r+\frac{1}{2})}.$$

In order now to obtain an approximate expression for the product of all these ν factors when N is large we take the logarithm of the product, expand, collect terms as in the corresponding proof at p. 204; B; 5, and finds

$$\frac{1}{(\sqrt{2\pi N})^{\nu-1}\sqrt{p_1 \ldots p_\nu}} e^{-\frac{1}{2}\sum_{r=1}^{r=\nu} j_r^2} \qquad \ldots (50)$$

subject to a similar condition as the Normal Law, namely, that the approximation may not be satisfactory if Np_r is less than about 10 (see p. 267; C; 4, and p. 310; C; 14).

It will be noted that the conditions $f_1'+f_2'+\ldots+f_\nu'=N$ and $Np_1+Np_2+\ldots+Np_\nu=N$ mean that $p_1+p_2+\ldots+p_\nu=1$, since all the N cases must be disposed of into the ν cells.

B; 15. The Derivation and Characteristics of Poisson's Law of Small Numbers

We here seek to simplify the probability for $np+x$ successes and $nq-x$ failures in n trials, i.e., the fundamental formula

$$\frac{n!}{(np+x)!\,(nq-x)!} p^{np+x} q^{nq-x} \qquad \ldots (2)$$

under the special conditions of q being so small and yet n sufficiently large that $nq=m$, a small but finite number. [In this demonstration q is supposed to be small rather than p, since in those actuarial problems to which Poisson's formula is particularly applicable it is q_x—the rate of mortality at age x—rather than the complementary probability of survivorship, p_x, which usually is small, so that it may assist the actuarial student to think of q as being small.]

Writing for simplicity $nq-x=r$, (2) becomes

$$\frac{n!}{r!\,(n-r)!}q^r p^{n-r} \qquad \ldots.(a)$$

$$=\frac{n!}{r!\,(n-r)!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$

$$=\frac{n(n-1)\ldots(n-r+1)}{r!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r} \qquad \ldots.(b)$$

Now, since n is large, $\left(1-\dfrac{m}{n}\right)^n \doteq e^{-m}$ $\qquad \ldots.(c)$

and hence $\left(1-\dfrac{m}{n}\right)^{-r}=\left(1-\dfrac{m}{n}\right)^{-\frac{nr}{n}}\doteq e^{\frac{mr}{n}}\doteq 1$

if n is large compared with mr, so that (b) becomes

$$\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\left(1-\frac{r-1}{n}\right)\left(\frac{m^r}{r!}\right)e^{-m}.$$

Here, however, the product of the first $r-1$ brackets is a decreasing alternating series when expanded, namely, $1-\dfrac{r(r-1)}{2n}+\ldots$, and therefore lies between 1 and $1-\dfrac{r(r-1)}{2n}$, and consequently tends to 1 when $2n$ is large compared with $r(r-1)$. The expression therefore is, approximately,

$$\frac{m^r e^{-m}}{r!}\text{, the Poisson exponential} \qquad \ldots.(55)$$

Alternatively, the reduction may be effected by introducing Stirling's formula (9) for the factorials $n!$ and $(n-r)!$ in (a), thus obtaining

$$\frac{(n-r)^r e^{-r}}{r!\left(1-\frac{r}{n}\right)^{n+\frac{1}{2}}} q^r p^{n-r} \qquad \ldots (d)$$

But n being large, and r relatively small,

$$\left(1-\frac{r}{n}\right)^{n+\frac{1}{2}} \doteq \left(1-\frac{r}{n}\right)^n \doteq e^{-r}$$

by (c) above. Also, $p^{n-r}=(1-q)^{n-r} \doteq e^{-(n-r)q}$, as may be seen by expanding both expressions. Making these substitutions, (d) therefore becomes $\frac{(n-r)^r q^r e^{-(n-r)q}}{r!}$; and since $(n-r)q \doteq nq$ under the conditions assumed, we reach $\frac{(nq)^r e^{-nq}}{r!}=\frac{m^r e^{-m}}{r!}$ as before.

The mathematical analysis of the error involved in the approximation represented by the Poisson exponential is given in P:*146*:135-137.

The preceding proofs evolve Poisson's result by imposing special conditions upon the point binomial. The formula, however, may also be deduced by using Thiele's half-invariants (see p. 257; B; 28) and requiring that the half-invariants of orders higher than zero shall all be equal (see P:*36*:265).

In the foregoing deductions m, which was written for the expected number of happenings nq, has emerged as the single parameter by which the formula is determined. It will be noted, however, that, since (55) represents the probability of r occurrences in n trials, it is a discrete function which exists only for integral values $r=0, 1, 2, \ldots, n$, and consequently that the mean, standard deviation, etc., of the distribution should be determined in practice from that range only. The formula, moreover, is not a true probability distribution; if it were, the sum of the probabilities for all possible occurrences would be unity, i.e., $\sum_{r=0}^{r=n} \frac{m^r e^{-m}}{r!}$ would be 1, whereas in fact this expression is unity only if the upper limit of summation is extended from n to ∞.

It may, however, be shown as follows that the Mean $\doteq m$.

By the method adopted for the point binomial in the proof of (4) in the main text, and so taking the origin at the beginning of the range, the mean is

$$\sum_{r=0}^{r=n} e^{-m} r\left(\frac{m^r}{r!}\right) = me^{-m}\left[1+\frac{m}{1!}+\frac{m^2}{2!}+\ldots+\frac{m^{n-1}}{(n-1)!}\right].$$

Again here, if the limit of summation were extended from n to ∞, this would be $me^{-m}e^m = m$; when, however, n is merely large enough that m is finite, this result is only approximately true. Hence Mean $\doteqdot m$.

Similarly, by the method of proof for (5), (6), and (7) in the text hereof, the second moment about the origin is

$$\sum_{r=0}^{r=n} e^{-m}\left(r^2\frac{m^r}{r!}\right).$$

If, therefore, σ^2—the mean square deviation about the mean—be taken with reference to the approximate mean, m, we have

$$\sigma^2 \doteqdot \sum_{r=0}^{r=n} e^{-m}\left[(r-m)^2\frac{m^r}{r!}\right]$$

$$= \sum_{r=0}^{r=n} e^{-m}\left[\left(\frac{r^2m^r}{r!}\right) - 2m\left(\frac{rm^r}{r!}\right) + m^2\left(\frac{m^r}{r!}\right)\right].$$

The first term here is

$$e^{-m}\left[0+\frac{1^2m}{1!}+\frac{2^2m^2}{2!}+\frac{3^2m^3}{3!}+\ldots+\frac{n^2m^n}{n!}\right]$$

$$= me^{-m}\left[1+\frac{m}{1!}(1+1)+\frac{m^2}{2!}(1+2)+\ldots+\frac{m^{n-1}}{(n-1)!}(1+\overline{n-1})\right]$$

$$= me^{-m}\left[1+\frac{m}{1!}+\frac{m^2}{2!}+\ldots+\frac{m^{n-1}}{(n-1)!}\right]$$

$$+ m^2e^{-m}\left[1+\frac{m}{1!}+\frac{m^2}{2!}+\ldots+\frac{m^{n-2}}{(n-2)!}\right].$$

Now when n is large, as is here supposed, each of these brackets is not very different from e^m, so that this first term $\doteqdot m+m^2$. Similarly, as in the proof for Mean $\doteqdot m$, we may write

$$\sum_{r=0}^{r=n}\left(\frac{rm^r}{r!}\right) \doteqdot me^m, \text{ while } \sum_{r=0}^{r=n}\left(\frac{m^r}{r!}\right) \doteqdot e^m.$$

The whole expression for σ^2 therefore gives

$$\sigma^2 \doteq [(m+m^2)-2me^{-m}(me^m)+m^2e^{-m}(e^m)]=m.$$

Tables of Poisson's Exponential Function

A four-place table of $\frac{m^r e^{-m}}{r!}$ for values of m from .1 to 10 was published in H:*75* by Bortkiewicz in 1898. A six-place table for values of m from .1 to 15, and for r from 0 to 37, was computed by H. E. Soper in 1914 (Biometrika, X, 25), and is reprinted in P:*97*. In this latter volume of "Tables for Statisticians and Biometricians" are also to be found tables prepared by Lucy Whittaker to facilitate comparisons between the results of Poisson's function and the "normal" theory.

B; 16. Edgeworth's Generalized Law of Error

The main objective of Edgeworth's investigations of the "law of error" was to discover, from general *a priori* conditions, that true and unique law which could properly be held to represent the frequency distribution of a magnitude depending on a number of independently varying elements.

The problem was dealt with in his many papers by several different methods, of which Bowley gives an excellent account in H:*162*:29-35, 39-47, and 134-5. The derivation is there shown on the assumption that there are "m elemental frequency groups, such that the chance of drawing a magnitude from, say, the qth group is $\varphi(a_q+\xi_q)$, where φ is an unknown function and a_q the average of the group". Drawing then magnitudes ${}_p\xi_1, {}_p\xi_2, \ldots, {}_p\xi_m$ from the m groups to form an aggregate $\sum_{q=1}^{q=m}(a_q+{}_p\xi_q)$, or $A+{}_px$ where ${}_px=\sum_{q=1}^{q=m}({}_p\xi_q)$, and introducing certain postulates, he proceeded to determine the moments of the frequency distribution of the values of ${}_px$, and found them to be the same as the moments of his Generalized Law of Error (58). The method may be compared with that originally used by Laplace and Poisson, from which a form equivalent to (58) can be obtained as shown in P:*155*:168-172.

A simple algebraic approach results also from taking into account the terms neglected in the derivation of the Normal Law of Deviations, as shown on p. 205; B; 5.

B; 17. The Derivations of the Gram-Charlier and Poisson-Charlier Series

The first determination of the constants of the Gram-Charlier Type A series was effected by Gram (H:*59*) through the use of a least squares criterion that $\int_{-\infty}^{+\infty} \frac{1}{\varphi(x)} [y'_x - y_x]^2 \, dx$ shall be a minimum, where y'_x denotes the observed values (see P:*36*:203-6, and P:*116*:169-170). The weighting of the squares of the deviations with $\frac{1}{\varphi(x)}$ appears to have been adopted by Gram, without comment, on account perhaps of its algebraic convenience (see P:*116*:174).

Another derivation employed by Wicksell (H:*146*), which leads to the same result, is to develop the point binomial (3) by Laplace's method of generating functions (see P:*116*:156-161).

The values of the coefficients can also be obtained from the fact that the derivatives $\overline{\varphi}_n(x)$ of the normal function $\overline{\varphi}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, and the *Hermite polynomials*

$$H_n(x) = x^n - \frac{n(n-1)}{2} x^{n-2} + \frac{n(n-1)(n-2)(n-3)}{2.4} x^{n-4} - \ldots,$$

form a *biorthogonal system* satisfying the relations

$$\int_{-\infty}^{+\infty} \overline{\varphi}_n(x) H_m(x) dx = 0 \text{ when } m \neq n, \text{ and } \int_{-\infty}^{+\infty} \overline{\varphi}_n(x) H_n(x) dx = \frac{(-1)^n n!}{\sigma}$$

otherwise (see P:*116*:165-8, and P:*36*:199-202).

The determination of the Poisson-Charlier series (65) by Wicksell's development of the point binomial is given conveniently in P:*116*:161-4.

In connection with these methods of representing frequency distributions by means of series it may be well to emphasize that their practical success obviously must depend upon rapid rather than ultimate convergence. In actual application it is not desirable to proceed beyond the first three or four terms, so that it may be possible to describe a given frequency distribution by a few constants only (cf. P:*140*:39 et seq., and P:*114*:115).

B; 18. Transformation of the Variable

Suppose that the frequencies for a variable x do not accord with those of the Normal Curve (11), but that the corresponding frequencies for some function of x, say $f(x)=z$, do so. Then the relative frequency of the values between z and $z+dz$ may be written, from (11), as $\frac{1}{c\sqrt{\pi}}e^{-\frac{z^2}{c^2}}dz=F(z)dz$, say. Since $z=f(x)$, we have $\frac{dz}{dx}=f'(x)$, or $dz=f'(x)dx$, and the relative frequency of the values between z and $z+dz$, or $F(z)dz$, becomes

$$\frac{1}{c\sqrt{\pi}}f'(x)e^{-\frac{[f(x)]^2}{c^2}}dx.$$

This "transformed" frequency function, being expressible as $\varphi(x)dx$, now evidently represents a distribution of relative frequencies for the variable x.

As an example, if $z=\sqrt{x}$, so that $f(x)=\sqrt{x}$ and $f'(x)=\frac{1}{2\sqrt{x}}$, we can from the above write down immediately $\varphi(x)dx=\frac{1}{2c\sqrt{\pi x}}e^{-\frac{x}{c^2}}dx$; the transformed frequency function is consequently $\varphi(x)=\frac{1}{2c\sqrt{\pi x}}e^{-\frac{x}{c^2}}$.

Again, if $z=\frac{1}{\sqrt{2}}\log\left(\frac{x-a}{b}\right)=f(x)$, we have $f'(x)=\frac{1}{\sqrt{2}(x-a)}$, and $\varphi(x)dx=\frac{1}{c\sqrt{2\pi}(x-a)}e^{-\frac{1}{2c^2}\left[\log\left(\frac{x-a}{b}\right)\right]^2}dx$, which is the logarithmic frequency function (70) discussed in Chapter VII.

B; 19. The Fourth Degree Exponential as the System of Curves for which Moments up to μ_4 Provide the "Best" Method of Fitting

In Chapter VIII on the Fitting of Curves it is pointed out that the equations for the determination of the constants in the fitting process known as the Method of Least Squares (the theory of which is founded essentially on the assumption of normality in the errors of observation) will give, when those equations are duly "weighted" (so that the standard deviation is uniform throughout), results very similar to those of the "unweighted" equations of the Method of Moments (see p. 243; B; 23) in the case of the representation by an exponential form $f''_x = e^{a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}}$ of a frequency distribution for which the weights may be taken as approximately $\dfrac{1}{f''_x}$.

The same principle appears, in effect, in R. A. Fisher's conclusion (P:*37*:355) that the use of moments up to μ_4 (as in Pearson's system) is, on a criterion of "efficiency" implying the minimum variance, the "best" system of fitting when it is applied to a fourth degree exponential (72). Observing that the first two moments have "100% efficiency" for the Normal Curve, Fisher points out (loc. cit.) that for symmetrical curves of the Pearson type the method of moments "has an efficiency exceeding 80% only in the restricted region for which β_2 lies between the limits 2.65 and 3.42 and for which β_1 does not exceed 0.1". In then determining the system of curves for which moments up to μ_4 constitute the "best" method of fitting, he remarks that "if the frequency in the range dx be $y(x, \theta_1, \theta_2, \theta_3, \theta_4)dx$, then $\dfrac{\partial}{\partial \theta} \log y$ must involve x only as polynomials up to the fourth degree"; consequently $y = e^{-a^2(a_0 + a_1 x + a_2 x^2 + a_3 x^3 + x^4)}$ as in (72)—"the convergence of the probability integral requiring that the coefficient of x^4 should be negative, and the five quantities a, a_0, a_1, a_2, a_3 being connected by a single relation, representing the fact that the total probability is unity" (*ibid.*).

B; 20. The Genesis of the Verhulst-Pearl-Reed (the "Logistic") Curve of Population Growth

The conditions assumed in Verhulst's original deduction of the logistic form (102) with $A=0$ were that the proportionate rate of increase over time, t, of a population, P_t, growing in a restricted area, will tend to become less and less as the population becomes greater.

If the population were growing over time in a geometrical progression, so that $P_t = ar^t$, then the proportionate rate of growth, $\frac{1}{P_t}\left(\frac{dP_t}{dt}\right)$, would be a constant $\log_e r = m$, say. In order, however, that this rate of growth shall gradually decrease we may write, as the simplest assumption, $\frac{1}{P_t}\left(\frac{dP_t}{dt}\right) = m - nP_t$ where n is a constant. The solution of this differential equation then gives immediately the logistic expression in the form (102a) below (cf. P:*176*:4, 43, and 44). The fundamental principle of the curve is thus the assumption that the instantaneous proportionate rate of increase is a decreasing linear function of the population.

Formula (102) is a convenient form employed by Cramér (P:*23*:200).

Putting $A=0$ in (102), and multiplying numerator and denominator by the constant $e^{k\tau}=C$ we obtain $\frac{Be^{kt}}{C+e^{kt}}$, whence also

$$P_t = \frac{B}{1+Ce^{-kt}} = \frac{B'}{C'+e^{-kt}} \qquad \ldots (102a)$$

as alternative forms which are frequently used.

Another useful type, which is preferred by Yule (and is examined fully in P:*176*:5 and 46 et seq.), is

$$P = \frac{L}{1+e^{\frac{\beta-t}{\alpha}}} \qquad \ldots (102b)$$

where L is the limiting population P_∞, α determines the horizontal scale, and β is the time from zero to the point of inflexion.

The simplest expression (see P:*176*:5-6) results from choosing the scales so that L and a are unity and the point of inflexion is at zero time, so that

$$P_{t'} = \frac{1}{1+e^{-t'}} \qquad \ldots .(102c)$$

Exhaustive discussions of the hypotheses underlying the logistic may be found particularly in P:*176*, P:*96*:567 et seq., and P:*60*.

The mathematical relations between the summation of two logistic curves and the generalized form (103) are examined in P:*105*:742.

The various methods of fitting which have been employed in practice are indicated at p. 321; C; 19 and p. 327; C; 21.

B; 21. The Specification of the Parent Population from the Observed Values of a Sample

The problem of specifying the true, i.e., "parent", population from the observed values obtained by sampling is approached easily by using the Principle of Maximum Likelihood, which has already been noted on p. 39 (Chapter V). In order to give a simple illustration it will be of assistance to consider the problem of determining the parent population from which observed rates of mortality are drawn. Suppose, therefore, that the true rate of mortality at a particular age in the parent population is an unknown quantity, q, of which the best obtainable estimate is required so that the population may be specified thereby. Suppose, also, that E' persons, drawn as a sample from that parent population, have been observed, and that θ' of them have died, so that the observed rate of mortality, q', is $\frac{\theta'}{E'}$. Then, as in the derivation of (1) in Chapter III, the probability that out of E' exposed to risk there will be θ' deaths, when the true rate of mortality is q, is $({}^{E'}C_{\theta'})(q^{\theta'})(p^{E'-\theta'})$. Clearly the most probable hypothesis for the unknown q will be to assign to it that value which will make this expression a maximum. Taking logarithms, differentiating, and equating to zero, it is seen at once that the maximum value will be given when $q=\frac{\theta'}{E'}$. That is to say, the most probable estimate of the unknown q of the parent population will be provided by the observed value $\frac{\theta'}{E'}=q'$.

The preceding application of the principle of maximum likelihood can be extended immediately to the specification of a parent population which is characterized by more than one unknown parameter, as in the case of the multinomial distribution discussed in Chapter VI. For, as there, suppose the parent population consists of N persons distributed into ν cells with the unknown true relative frequencies $p_1, p_2, \ldots, p_\nu$; then, if a sample of N' persons is drawn and is observed to contain $f'_1, f'_2, \ldots, f'_\nu$ cases respectively attributable to the ν cells, what values for the true probabilities $p_1, p_2, \ldots, p_\nu$ will give the maximum likelihood to the sample drawn, i.e., what is the best estimate that can be made for those probabilities, by which the parent population can be specified? As in (48) and (49) the probability of drawing the particular sample observed is

$$\frac{N'!}{f'_1!f'_2!\ldots f'_\nu!}\,p_1^{f'_1}\,p_2^{f'_2}\ldots p_\nu^{f'_\nu}=P_s \text{ say} \qquad \ldots\text{(i)}$$

in which
$$p_1+p_2+\ldots+p_\nu=1 \qquad \ldots\text{(ii)}$$
and
$$f'_1+f'_2+\ldots+f'_\nu=N' \qquad \ldots\text{(iii)}$$

Now $\log P_s$ will be a maximum when P_s is a maximum; we therefore take the logarithm of (i), and differentiating and equating to zero find

$$\frac{f'_1}{p_1}\,\delta p_1+\frac{f'_2}{p_2}\,\delta p_2+\ldots+\frac{f'_\nu}{p_\nu}\,\delta p_\nu=0 \qquad \ldots\text{(iv)}$$

where for the variations $\delta p_1 \ldots \delta p_\nu$ in $p_1, \ldots, p_\nu$ we have, from (ii), the condition

$$\delta p_1+\delta p_2+\ldots+\delta p_\nu=0 \qquad \ldots\text{(v)}$$

From (iv) and (v) evidently $\frac{f'_1}{p_1}=\frac{f'_2}{p_2}=\ldots=\frac{f'_\nu}{p_\nu}=a$, say.

Consequently $f'_1 = ap_1; \ldots ; f'_\nu = ap_\nu$; whence

$$\frac{f'_1}{p_1} = \ldots = \frac{f'_\nu}{p_\nu} = a = \frac{f'_1 + \ldots + f'_\nu}{p_1 + \ldots + p_\nu},$$ which by (ii) and (iii) $= N'$.

Finally, therefore, it follows that $p_1 = \frac{f'_1}{N'}; p_2 = \frac{f'_2}{N'}; \ldots ; p_\nu = \frac{f'_\nu}{N'}$.

That is to say, the assignment of the maximum value to the probability of drawing the sample actually observed means that the most probable estimate of the parent probabilities, p_r, will be provided by using the observed relative frequencies $\frac{f'_r}{N'}$ ($= p'_r$, say).

B; 22. The Classical Method of Approximation and Correction for the Least Squares Fitting of Transcendental Equations

If the curve to be fitted is $y''_x = f''(x; \alpha, \beta, \ldots)$, and approximate values $\alpha', \beta', \ldots$ have been found for the unknowns $\alpha, \beta, \ldots$, so that $\alpha = \alpha' + \delta\alpha$, $\beta = \beta' + \delta\beta, \ldots$, the problem becomes one of fitting the curve $y''_x = f''(x; \alpha' + \delta\alpha, \beta' + \delta\beta, \ldots)$, and hence of determining the corrections $\delta\alpha, \delta\beta, \ldots$, which therefore now are to be considered as the unknowns. Assuming that the approximate values $\alpha', \beta', \ldots$ are sufficiently close to $\alpha, \beta, \ldots$ that the squares and higher powers of $\delta\alpha, \delta\beta, \ldots$ may be neglected, y''_x can be expressed by Taylor's Theorem in the form

$$f''(x; \alpha', \beta', \ldots) + \delta\alpha\left[\frac{\partial}{\partial\alpha'} f''(x; \alpha', \beta', \ldots)\right] + \delta\beta\left[\frac{\partial}{\partial\beta'} f''(x; \alpha', \beta', \ldots)\right] + \ldots\ldots$$

Since this is to be fitted to the observed series of f'_x's, and the numerical values of $f''(x; \alpha', \beta', \ldots)$ are available by computation, while those of $\frac{\partial}{\partial\alpha'} f''(x; \alpha', \beta', \ldots), \ldots$, are also deducible by differentiation and calculation, the problem becomes simply one of fitting the linear function $k_\alpha(\delta\alpha) + k_\beta(\delta\beta) + \ldots$ to the numerical values of $[f'_x - f''(x; \alpha', \beta', \ldots)]$, where $k_\alpha, k_\beta, \ldots$ are constants of which the numerical values are known, and $\delta\alpha, \delta\beta, \ldots$ are the unknowns sought. The method of least squares can therefore be applied directly by the formation of linear "normal" equations in the form already stated and with the weights being employed in the manner previously given (cf. P:*51*:121 and H:*44*:169, in both of which, however, the definition and use of the weights in the formation of the normal equations must be watched carefully to avoid confusion—see p. 322; C; 20 here).

B; 23. The Approximate Equivalence of the Weighted Equations of Least Squares and the Unweighted Equations of Moments in the Case of an Exponential Function Representing a Frequency Distribution

In Chapter VIII the formulae, (111) to (113), were given for the weights of several ratios, such as $W\{q'_x\}$, etc., which are required in the fitting of any curve to such ratios by the method of least squares—these formulae being obtained from the mathematical model of E'_x exposed to risk at age x being subject to an observed rate of mortality $q'_x = \dfrac{\theta'_x}{E'_x}$, where $p'_x = 1 - q'_x$, and the corresponding true parent probabilities are q_x and p_x. Such ratios by age are not, of course, frequency distributions; they are simply curves depicting the progression of the ratios from age to age. The observed deaths, θ'_x, however, do form a frequency distribution. The weight, $W\{\theta'_x\}$, of each term, by (114), is then $\dfrac{1}{E'_x p_x q_x}$; at most ages, as in (115), this may often be taken as approximately $\dfrac{1}{E'_x q_x}$, which again will be given roughly by $\dfrac{1}{E'_x q''_x}$, that is, by the reciprocal of the graduated values of the frequency distribution. It therefore follows that in the fitting of a curve $f''_x = \varphi(x)$ to a *frequency distribution*, the weights may be taken as approximately $\dfrac{1}{f''_x}$ (cf. P:*137*:358, and P:*51*:129—remembering that in the latter $w_x = \sqrt{W_x} = \frac{1}{\sqrt{f''_x}}$ is used, as noted at p. 322; C; 20 here).

Now the exponential function $f''_x = e^{a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}}$ can represent approximately most frequency distributions. But here $\frac{\partial}{\partial a_i}(f''_x) = x^i f''_x$; and consequently, since $W_x \doteq \frac{1}{f''_x}$ as above, the *weighted* normal equations (116) of least squares becomes *approximately* $\Sigma\left[\frac{1}{f''_x}(f''_x - f'_x)(x^n f''_x)\right] = 0$, that is, $\Sigma[x^n(f''_x - f'_x)] = 0$ for $n = 0, 1, 2, \ldots$, which are the *unweighted* equations (119) of the method of moments.

We consequently see that *in the case of a frequency distribution*, for which the weights can be taken as approximately $\frac{1}{f''_x}$, and which can generally be represented roughly by an exponential $f''_x = e^{a_0 + a_1 x + \ldots + a_{n-1} x^{n-1}}$, the *weighted* equations of least squares may be expected to lead to practically the same results as the *unweighted* equations of the method of moments.

[It should be noted in the preceding demonstration that the approximate weights $\frac{1}{f''_x}$ involve, of course, the unknown parameters of which the values are being sought, but that their introduction into the argument at the stage of the differential form (116) results, through the subsequent cancellation of f''_x, in the elimination of the difficulty which would arise in the differentiations if $\frac{1}{f''_x}$ were inserted for W_x in (109). For the weights W_x in (109), being dependent only on the process of selection by which the observed f'''s are derived from the parent f's, are independent of the unknowns which are involved in the f'''s; they are therefore to be treated as constants in the differentiations of (109), and consequently so appear in (116). But if in (109) W_x could be taken as $\frac{1}{f''_x}$, it would evidently involve the unknowns, and the differentiations of (109) would immediately be complicated. The insertion of $\frac{1}{f''_x}$ as an *approximation* to W_x in the derived normal equations (116), however, will evidently lead to close results and is sufficiently justifiable under all the conditions assumed.]

B; 24. The Relation between the Criteria of Weighted Least Squares and Minimum-χ^2 in the Graduation of Mortality Tables

As shown by (109)—but dropping the limits of summation for convenience here—the method of least squares imposes the requirement that $\Sigma[W_x(f''_x - f'_x)^2]$ shall be a minimum, where f'_x are the observed values, f''_x the fitted (i.e., graduated) values, and the weights W_x are based on the true (i.e., parent) values and are thus independent of f'_x and also of the graduated values, f''_x, which are being sought.

By this method, therefore, a graduation of a series of observed deaths, θ'_x, by means of a fitted series θ''_x, requires that $\Sigma[W\{\theta'_x\}(\theta''_x - \theta'_x)^2]$ be minimized; and since by (114) the weight $W\{\theta'_x\}$ is $\frac{1}{E'_x p_x q_x}$, we see at once that the condition here imposed by the method of least squares is that $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p_x q_x}\right]$ must be a minimum.

Precisely the same result is reached, moreover, in a weighted least squares graduation of a series of observed rates of mortality, $q'_x\left(= \frac{\theta'_x}{E'_x}\right)$, where the graduated rates, q''_x, are applied to the observed E'_x in order to give a series of adjusted deaths $\theta''_x = E'_x q''_x$. For then the least squares expression is $\Sigma[W\{q'_x\}(q''_x - q'_x)^2]$; and since by (111) the weight $W\{q'_x\}$ is $\frac{E'_x}{p_x q_x}$, this expression to be minimized becomes $\Sigma\left[\frac{E'_x}{p_x q_x}(q''_x - q'_x)^2\right]$, which is here to be taken as $\Sigma\left[\frac{E'_x}{p_x q_x}\left(\frac{\theta''_x}{E'_x} - \frac{\theta'_x}{E'_x}\right)^2\right] = \Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p_x q_x}\right]$, being the same condition as before.

In the case of the minimum-χ^2 method the criterion, as explained in the statement of (121), is that $\sum\left[\frac{(f''_x - f'_x)^2}{f''_x}\right]$ shall be a minimum. Being obtained directly from the Multinomial Law of Deviations, which is based on a mathematical model of the various values of the observed series, f'_x, falling into cells so that the whole series is completely distributed, the minimum-χ^2 method is evidently applicable only to cases which can be viewed as frequency distributions. For the graduation of mortality data, therefore, we can contemplate either (*i*) a direct graduation of the observed deaths, θ'_x, which can be visualized as a frequency distribution, or (*ii*) a graduation of the rates of mortality, q'_x, in such a form that we are in fact dealing with a frequency distribution.

In (*i*)—*a direct graduation of* θ'_x—the criterion (121) becomes simply that $\sum\left[\frac{(\theta''_x - \theta'_x)^2}{\theta''_x}\right]$ must be a minimum. This interpretation of the χ^2 method is illustrated by the calculation of χ^2 for Weldon's dice data at p. 334; C; 25, and by the analogous example of a simple distribution of θ'_x which is also given at p. 339 thereof. The minimum-χ^2 method of Cramér and Wold, moreover, follows this approach—for their minimum-χ^2 graduation process is based on the principle of using (see P:*23*:172) $\sum\left[\frac{(\theta''_x - \theta'_x)^2}{\theta''_x}\right]$ where $\theta''_x = E'_x q''_x$.

In (*ii*)—*the graduation of rates of mortality,* q'_x, *but in such a manner that we are in fact dealing with a frequency distribution of deaths*—it is to be remembered that at each age the observed data are $E'_x q'_x = \theta'_x$ who die, and $E'_x p'_x = E'_x(1 - q'_x) = E'_x - \theta'_x$ who do not die, and that the graduation finds for them the fitted values $E'_x q''_x = \theta''_x$ who die, and $E'_x p''_x = E'_x(1 - q''_x) = E'_x - \theta''_x$ who do not die. At each age, therefore, the contribution to χ^2 will be $\frac{(\theta''_x - \theta'_x)^2}{\theta''_x} + \frac{[(E'_x - \theta''_x) - (E'_x - \theta'_x)]^2}{E'_x - \theta''_x}$, as in the binomial verification of the Multinomial Law (50). This expression is

$$(\theta''_x - \theta'_x)^2\left[\frac{1}{\theta''_x} + \frac{1}{E'_x - \theta''_x}\right] = (\theta''_x - \theta'_x)^2\left[\frac{1}{E'_x q''_x} + \frac{1}{E'_x p''_x}\right] = \frac{(\theta''_x - \theta'_x)^2}{E'_x p''_x q''_x},$$

since $p''_x + q''_x = 1$. The minimum-χ^2 principle in this case therefore requires that $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p''_x q''_x}\right]$ shall be a minimum. A numerical example of the calculation and use of this expression as the χ^2 test of goodness of fit is given at p. 341; C; 25.

From the preceding analyses it will be seen that the weighted least squares graduation of either the deaths, θ'_x, or the rates of mortality, q'_x, imposes the condition that $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p_x q_x}\right]$ shall be a minimum; the mininum-χ^2 method, on the other hand, requires that $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{\theta''_x}\right] = \Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x q''_x}\right]$ shall be a minimum for a graduation of θ'_x, and that $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p''_x q''_x}\right]$ shall be a minimum in a graduation of q'_x. In the least squares method the true values p_x and q_x must in practice be estimated by an approximate preliminary graduation (see p. 96); if these estimated values are reasonably close to p''_x and q''_x (as they should be by any proper method of determining them), the result of the weighted least squares process will evidently be very similar to that of minimizing $\Sigma\left[\frac{(\theta''_x - \theta'_x)^2}{E'_x p''_x q''_x}\right]$, which is the theoretical criterion for minimum-χ^2 in a graduation of q'_x. It is, however (cf. p. 102), difficult to minimize such an expression, for it involves the p''_x and q''_x, which are being sought, in the denominator; a practical device, therefore (as Cramér and Wold propose in their minimum-χ^2 graduation of θ'_x), would be to substitute the observed p'_x and q'_x, or to determine (as for least squares) approximate values for p''_x and q''_x by a preliminary graduation. On either of these approximate bases it is clear that the results by the weighted least squares and the minimum-χ^2 conditions will be very similar.

B; 25. The Frequency of Changes of Sign

Since the fact does not appear to be generally known, it may be of interest to record here that the first discussion and use of a method for examining the changes of sign in a mortality table graduation appears to have been given in 1876-1878 by De Forest, whose pioneer work on graduation by linear compounding constitutes one of the most remarkable accomplishments to be found in actuarial literature (see p. 284; C; 7, section (*xii*), and P:*166*). His first consideration of the problem was in H:*49*:29-35, on the following lines.

(*a*) In a *periodic series* (such as a sine curve), for which the first and last terms are consecutive, the probabilities that any particular 2, 3, 4, . . . signs for the deviations between the observed and graduated series will be alike are evidently $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, . . . , on the assumption that positive and negative deviations are equally likely to occur. The probability that any group of r consecutive signs will be the same is therefore $\frac{1}{2^{r-1}}$. Furthermore, the probability that any group of like signs will be isolated, so that the signs next preceding and following the group will be different, will be $(\frac{1}{2})(\frac{1}{2})$. The probability that any particular group of r signs will be alike and isolated is therefore $\frac{1}{2^{r+1}}$. If N

be the total number of signs in the periodic series, the expected number of isolated groups of r like signs is consequently $\frac{N}{2^{r+1}}$. Writing this as Np, where $p=\frac{1}{2^{r+1}}$, we see also, by formula (8), that the standard deviation of this number would be $\sqrt{Npq}$ on the assumption that the occurrences of the signs are independent (which, however, is open to question); and if, with De Forest (following the practice of his time), the "probable error" be adopted to indicate the limits of variation as in (21), we might say that in a periodic series of N terms the expected number of isolated groups of r like signs can be computed, with its probable error, as $\frac{N}{2^{r+1}} \pm \frac{.6745}{2^{r+1}}\sqrt{N(2^{r+1}-1)}$.

The expression $\frac{N}{2^{r+1}}$ may also be reached clearly as follows (see H:*115*:146): If in a series of numbers the conditions of selection are such that odd and even numbers are equally likely *a priori*, then the probability of any particular number being, for example, an even number is $\frac{1}{2}$. An isolated sequence of exactly r even numbers (i.e., the non-occurrence of change from even to odd) will occur through the appearance of r even numbers followed by an odd number (the appearance of the odd number being essential to terminate the sequence of the even numbers). The probability of a sequence of r even numbers is therefore $(\frac{1}{2})^r(\frac{1}{2}) = \frac{1}{2^{r+1}}$, and in a total of N numbers the expected number of sequences of r even numbers is $\frac{N}{2^{r+1}}$, as before. Visualizing now the occurrence of $+$ and $-$ signs in a series of differences between ungraduated and graduated rates of mortality, it follows that the infrequency of changes of sign to be anticipated as a result of a merely chance distribution of such changes (as might occur in an ideal graduation) may be measured by taking $\frac{N}{2^{r+1}}$ as the number of times a sequence of r plus or r minus signs may be expected to occur.

(*b*) De Forest also observed (H:*51*:111) that since with $r=1$ the expected number of sequences is $\frac{N}{4}$, and when $r=2$ it is $\frac{N}{8}$, so that the number of signs which occur singly or two alike is $\frac{N}{4} + 2\left(\frac{N}{8}\right) = \frac{N}{2}$, "we have this practical rule—that if a series has been well adjusted, the whole number of signs . . . which fall within groups of only one or two like signs each will probably be about equal to the whole number which fall within groups of more than two".

(*c*) From the preceding it follows that the average number of sequences of all orders is obtainable by summing $\frac{N}{2^{r+1}}$ for all possible values $r=1, 2, 3, \ldots$. The limit, as N approaches ∞, becomes $\frac{N}{2}\left[\frac{1}{2}+(\frac{1}{2})^2+\ldots\right] = \frac{N}{2}$.

For a *non-periodic series*, such as rates of mortality, De Forest pointed out (in H:*51*:111) that the preceding methods based on $\frac{N}{2^{r+1}}$, which strictly suppose that the series is periodic, will also give a close enough approximation provided that the first and last signs are treated as consecutive so that they will be considered as belonging to the same group if they are alike.

In order, however, to avoid that approximate method of dealing with a non-periodic series, De Forest also suggested (H:*49*:32) that the first and last signs be omitted—for they cannot strictly be included within any group, because they are not consecutive, and it is not known whether or not they are isolated (the next sign beyond being unknown). If we thus omit the first and last signs in a non-periodic series of N terms, it will be clear that the number of possible groups of r consecutive signs will be $N-r-1$ (for example, in 5 terms there is one middle group of 3; in 7 terms there is one middle group of 5, 2 groups of 4, 3 groups of 3, etc.). The expected number of isolated groups of r like signs is therefore $\frac{N-r-1}{2^{r+1}}$ $\Big($since the probability of any particular group of r signs being alike and isolated has already been established as $\frac{1}{2^{r+1}}\Big)$, with a corresponding probable error (if desired, and subject to the validity of assuming that the occurrences are independent) of $\pm\frac{.6745}{2^{r+1}}\sqrt{(N-r-1)(2^{r+1}-1)}$.

Furthermore, if the complete non-periodic series is to be examined, we must add to the preceding $\frac{N-r-1}{2^{r+1}}$ the two terms involving the first and last signs. Starting with the first we note that the probability of its being positive is $\frac{1}{2}$; the probability then of $r-1$ more positive signs (to make the sequence of r alike) is $\frac{1}{2^{r-1}}$; and the probability of the next sign being negative (to terminate the sequence of r) is $\frac{1}{2}$; consequently the probability of r isolated positive signs starting with the first is $\frac{1}{2}\left(\frac{1}{2^{r-1}}\right)\frac{1}{2}=\frac{1}{2^{r+1}}$.

The probability of r isolated negative signs starting with the first is similarly $\frac{1}{2^{r+1}}$. Consequently the total probability of r isolated positive or negative (i.e., like) signs starting with the first is $\left(\frac{1}{2^{r+1}}+\frac{1}{2^{r+1}}\right)=\frac{1}{2^r}$. By exactly the same reasoning it is evident that the probability of r isolated positive or negative (i.e., like) signs ending with the last is also $\frac{1}{2^r}$. The total expected number of isolated groups of r like (positive or negative) signs over the whole range of a non-periodic series of N terms (including the first and last) is therefore $\frac{N-r-1}{2^{r+1}}+\frac{1}{2^r}+\frac{1}{2^r}$ $=\frac{N-r+3}{2^{r+1}}$, as given by Seal in P:*125*:29.

More elaborate discussions of these principles are also available in De Forest's papers H:*54* and H:*55*:71.

B; 26. The Most Probable Value of the Mean Square Error of an Observation when there are ν Observation Equations, and k Unknowns have been Determined by the Method of Least Squares

The method of least squares is based on the assumption (see Chapter VIII, and p. 322; C; 20) that the observed values, f'_r, are drawn as a sample of the true values, f_r, and that the determination of the unknowns by the solution of the "normal equations" will give the best fitted values, f''_r. The true errors are thus $f_r-f'_r$; but after the fitting has been completed the *residuals*, $f''_r-f'_r=v_r$, say, will still remain. Furthermore, it will be remembered that when the observations, f'_r, are not all equally well drawn, the residuals become invested with "weights", $W_r=\frac{1}{\sigma_r^2}$, so that the whole process actually makes the sum of all the values of $W_rv_r^2$, or $(\sqrt{W_r}v_r)^2$, a minimum.

In order to simplify the proof to be now given (for which cf. P:*13*:100), let us write V_r for $\sqrt{W_r}v_r$—that is, in the usual language of the text-books, let us suppose that "each observation equation [$f''_r - f'_r = 0$, being $v_r = 0$] has been multiplied by the square root of its weight [$\sqrt{W_r}$], so that the residuals are all reduced to unit weight [V_r]". Then, on the assumption of the Normal Curve, the *a priori* probability of the whole series of reduced deviations $\sqrt{W_r}(f''_r - f'_r)$, or V_r, is $\left(\frac{1}{c\sqrt{\pi}}\right)^{\nu} e^{-\frac{\Sigma V^2}{c^2}}$ by (104), where the parameter c remains to be found.

Now the object of the fitting process is to determine the unknowns in the expression to be fitted, which we may suppose is $f''_x = f''(x; \alpha, \beta, \gamma, \ldots)$, so that the unknowns are $\alpha, \beta, \gamma, \ldots$, numbering, say, k. Since these unknowns are independent of each other, and might (if we adopt the Principle of Insufficient Reason — see p. 38, and p. 181; (*i*) of B; 1, and p. 222; B; 12) each have any values whatever between $-\infty$ and $+\infty$, the total probability of the system of residuals under consideration is

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\left(\frac{1}{c\sqrt{\pi}}\right)^{\nu} e^{-\frac{\Sigma V^2}{c^2}}\, d\alpha\, d\beta\, d\gamma \ldots, \quad \ldots (i)$$

where there are k integrations.

In performing the integration with respect to the first unknown, α, we may express the terms in ΣV^2 which involve α in the form $(A\alpha + B)^2$. Then

$$\int_{-\infty}^{+\infty}\left(\frac{1}{c\sqrt{\pi}}\right) e^{-\frac{\Sigma V^2}{c^2}}\, d\alpha = \left(\frac{1}{c\sqrt{\pi}}\right)\int_{-\infty}^{+\infty} e^{-\frac{(A\alpha+B)^2}{c^2}}\, d\alpha = \frac{1}{A}$$

by putting $\frac{A\alpha + B}{c} = t$ and using (*b*) of B; 7. The probability (*i*) with α thus eliminated becomes

$$\frac{1}{A}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\left(\frac{1}{c\sqrt{\pi}}\right)^{\nu-1} e^{-\frac{\Sigma V_\alpha^2}{c^2}}\, d\beta\, d\gamma \ldots,$$

where now $k-1$ integrations remain, and ΣV_α^2 is the quadratic function of $\beta, \gamma, \ldots$ which follows from ΣV^2 when α is chosen, by the method of least squares, to make ΣV^2 a minimum. Repeating the same argument for each of the k integrations we see that the final probability reduces to $K\left(\frac{1}{c}\right)^{\nu-k} e^{-\frac{\Sigma V_k^2}{c^2}}$, where K is a constant (absorbing the $\sqrt{\pi}$ factor in the denominator), and ΣV_k^2 is the value of ΣV^2 when all the k unknowns $\alpha, \beta, \gamma, \ldots$ are determined by least squares.

To find now the value of c for which this probability is a maximum, we take logarithms, differentiate with respect to c, and obtain $\frac{c^2}{2}$ or $\sigma^2 = \frac{\Sigma V_k^2}{\nu-k}$. That is to say, the mean square error of an observation of unit weight when there are ν observation equations and k unknowns is

$$\frac{\sum_{r=1}^{r=\nu} [W_r(f''_r - f'_r)^2]}{\nu-k} \qquad \ldots (ii)$$

where the unknowns in f'' have been determined by the method of least squares.

The preceding proof is simply an extension, to the case of k unknowns, of the demonstration of Bessel's formula (42), namely, $\sigma_e^2 \doteq \left(\frac{n}{n-1}\right)\sigma_s^2$, which is $\sigma^2 = \frac{\Sigma(x'_r - \bar{x})^2}{n-1}$ for the case of one unknown and uniform weights, as is shown in the proof of the latter form on the basis of the Principle of Insufficient Reason at p. 38. The proof just given, and the resulting formula, are consequently open to the criticisms which may be levelled against this principle (see p. 39 here, and the references there noted).

The alternative method of proof for Bessel's formula (42) which is given on p. 35 may likewise be extended to the case of k unknowns, as shown in P:*90*:80-82 (see also P:*155*:205 and 243-5).

The mean square error of "an observation", as it is contemplated in this formula (*ii*), is evidently the "most probable" or "presumptive" value (see p. 219; B; 11) of the mean square error of a fitted value which has been determined by the method of least squares. It is not the mean square error of any particular term of the fitted series; it is a hypothetical value, typifying the accuracy of the graduation as a whole, to which a "most probable" value has been assigned by maximizing the probability of actually obtaining the system of residuals resulting from the determination of the k unknowns from the ν observations by the method of least squares (cf. P:*124*:145, and P:*90*:137).

B; 27. Moments

For the point binomial $(q+p)^n$ it was shown, in the derivation of formula (4), that the mean is np, and in formula (5) that the average of the expected squares of the number of happenings is $np(np+q)$. The latter was obtained as the sum of the second powers of the variable each multiplied by the frequency; the former as the sum of the first powers multiplied by the frequencies. Consistently with these we could also take the sum of the 0th powers of the variable multiplied by the frequencies, namely, $q^n(0)^0+npq^{n-1}(1)^0+\ldots+p^n(n)^0=(q+p)^n=(1)^n=1$, showing, of course, that for the point binomial $(q+p)^n$, which represents a distribution of probabilities, the "total frequency" is 1. These are simply three cases, for $r=2$, 1, and 0, of a process of summing the rth powers of the variable multiplied by the appropriate frequencies.

If, therefore, instead of a point binomial probability distribution for which the total frequency must be 1, we have, in general, N cases altogether with frequencies $f_1, f_2, \ldots, f_n$ corresponding to values $x_1, x_2, \ldots, x_n$ of the variable, then, as for the above point binomial,

the total frequency, N, is $f_1(x_1)^0+f_2(x_2)^0+\ldots+f_n(x_n)^0$

$$=f_1+f_2+\ldots+f_n=\sum_{t=1}^{t=n} f_t(x_t)^0;$$

the first powers give $f_1(x_1)^1+f_2(x_2)^1+\ldots+f_n(x_n)^1=\sum_{t=1}^{t=n} f_t(x_t)^1$;

the second powers give $f_1(x_1)^2+f_2(x_2)^2+\ldots+f_n(x_n)^2=\sum_{t=1}^{t=n} f_t(x_t)^2$;

and generally

the rth powers give $f_1(x_1)^r+f_2(x_2)^r+\ldots+f_n(x_n)^r=\sum_{t=1}^{t=n} f_t(x_t)^r$.

If we reduce the values to a unit of frequency, and define the *rth moment* (per unit frequency) as the sum of the products of the frequencies per unit and the rth powers of the variable, we have

$$\text{Total Frequency (0th Moment)} = N$$

$$\text{and } r\text{th Moment} = \frac{1}{N}\sum_{t=1}^{t=n} f_t(x_t)^r = \mu_r', \text{ say.}$$

These moments are calculated with reference to the commencement of the range. It is often more convenient, however, to calculate them with reference to the arithmetic mean, which, it is to be noted, is the preceding first moment μ_1'. Clearly it is only necessary to measure the variable from μ_1' instead of from the commencement of the range; the rth moment about the mean is therefore

$$\frac{1}{N}\sum_{t=1}^{t=n} f_t(x_t-\mu_1')^r = \mu_r, \text{ say.} \qquad \ldots.(a)$$

Similarly, as for the changing of the origin of coordinates, the rth moment about any arbitrary origin X is $\frac{1}{N}\sum_{t=1}^{t=n} f_t(x_t-X)^r$.

Since the commencement of the range is, in relation to the mean, an arbitrary measure, (a) may be used to express the important relations between μ_r, the rth moment about the mean, and μ_r', the rth moment about any arbitrary origin. For, expanding the binomial in (a), and putting r successively $=0, 1, 2, 3, 4, \ldots$, we find immediately

$$\mu_0 = \mu_0' = 1$$
$$\mu_1 = 0$$
$$\mu_2 = \mu_2' - (\mu_1')^2$$
$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$$
$$\mu_4 = \mu_4' - 4\mu_1'\mu_3' + 6(\mu_1')^2\mu_2' - 3(\mu_1')^4.$$

The relation between the second moments, for example, is illustrated for the point binomial by formulae (4), (5), and (6). For we have above $\mu_2' = \mu_2 + (\mu_1')^2$ (which, from the definitions, means that the second moment about an arbitrary origin equals the second moment about the arithmetic mean plus the square of the arithmetic mean measured from the arbitrary origin); (4) means that $\mu_1' = np$, (5) that $\mu_2' = np(np+q)$, and (6) that $\mu_2 = npq$, which conform with $\mu_2' = \mu_2 + (\mu_1')^2$.

The same relation is used on p. 216; B; 9 in transferring the mean square deviation $np_tq_t (t=1, 2, \ldots, \nu)$ measured from the mean np_t, to np. For np_tq_t is the second moment about the mean np_t, and is to be transferred to the arbitrary origin np. Hence np_tq_t represents μ_2; by the relation $\mu_2' = \mu_2 + (\mu_1')^2$ we must add $(\mu_1')^2$, which is the square of the mean measured from the arbitrary origin np, or $(np_t - np)^2$, so that the second moment (the mean square deviation) measured from np instead of from np_t becomes $np_tq_t + (np_t - np)^2$.

In the preceding paragraphs it has been convenient to denote by μ a moment about the mean, and by μ' a moment about any arbitrary origin, in respect of a frequency distribution of ordinates. The same notation and relations evidently may be used in the case of a continuous curve, For if $y = f(x)$ represents the equation of the curve extending from $x = h$ to $x = k$, so that the area $= \int_h^k y\,dx$, then the average value of x (i.e., the mean μ_1') is $\int_h^k yx\,dx \div \int_h^k y\,dx$; generally the rth moment $\mu_r' = \int_h^k yx^r dx \div \int_h^k y\,dx$; correspondingly the rth moment about the mean is

$$\mu_r = \int_h^k y(x - \mu_1')^r dx \div \int_h^k y\,dx;$$

and the relations between μ and μ' are those already given.

Adjustments to Moments

An adjustment, however, obviously may be necessary when—as in the case of fitting a mathematical curve to a set of statistical data—we have to establish the relations between the moments obtained from the integration of a continuous curve on the one hand, and, on the other, the moments derived from (*a*) the given ordinates, or (*b*) areas, of the statistics themselves.

(*a*) In the case of a series of *ordinates* which evidently would finally vanish with high contact with the x-axis (i.e., asymptotically) at both ends of the range, it will be clear from the following formula that no adjustment would be required. But if the series at either end begins abruptly, an adjustment should be based on the appropriate quadrature formulae, with the assumption that $\int_{-\frac{1}{2}}^{+\frac{1}{2}} y\,dx \doteq y_0$; for example, for n given ordinates $y_0, \ldots, y_{n-1}$, and $\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y\,dx$ as the corresponding area of the curve, it is easy (see P:*32*:26-29) to establish the relation

$$\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y\,dx = 1.1220486(y_0+y_{n-1}) + .7588542(y_1+y_{n-2}) + 1.1578125(y_2+y_{n-3}) + .9612847(y_3+y_{n-4}) + (y_4+\ldots+y_{n-5}) \qquad \ldots.(i)$$

and by the right side of this expression to compute the corrected ordinates (for examples see P:*32*:28 and 35).

(*b*) When the statistics are given in groups (i.e., in a system of *areas*), and there is high contact at both ends, the same principle leads readily (see P:*32*:30, or P:*114*:93) to *Sheppard's Corrections* (H:*78*), by which μ_2, μ_3, and μ_4 are first computed directly from the data, and then

$$\mu_2 \text{ (corrected)} = \mu_2 - \frac{h^2}{12}$$

$$\mu_3 \text{ (corrected)} = \mu_3$$

$$\text{and } \mu_4 \text{ (corrected)} = \mu_4 - \tfrac{1}{2} h^2 \mu_2 + \frac{7}{240} h^4$$

where h is the width of the class interval. Numerical examples are shown accessibly in P:*177*:160, P:*32*:30, and P:*51*:55.

Instead of thus calculating from the grouped data and subsequently introducing Sheppard's corrections, Hardy has suggested a method of estimating the central ordinates of each group as the original numbers for each group less $\frac{1}{24}$ of their respective second central differences, and thence computing the moments directly without further adjustment. In the case of mortality data this process has the considerable advantage of giving values for the central ordinates which are useful in the calculation of the force of mortality, as well as providing a simple means of examining the nature of the curve from the differences of the logarithms of the ordinates (see P:*51*:57-59).

When there is not high contact at both ends, Sheppard's formulae are not applicable, and other methods must be followed.

The central ordinates can then be found as indicated in P:*32*:236, and adjusted by (*i*) herein. More elaborate corrections suggested by Pairman and Pearson, and by E. S. Martin, may also be applied (see P:*32*:231 and 236).

Hardy's Summation Method of Computing Moments

A very useful scheme for calculating moments by means of successive summations was suggested by Sir G. F. Hardy (P:*51*:59), and has been widely adopted in actuarial work. The method is explained clearly in schedule form in P:*32*:20 (see also P:*51*:59 and 124 for the mathematical relations between the summations and the moments). Other examples which are shown in detail may be found in P:*136*:3 and H:*106*:292 and 322. The abbreviated notation Σ, Σ^2, Σ^3, etc., is frequently used to denote the successive sums.

B; 28. Thiele's Half-Invariants

The remarkable system of symmetrical functions named by Thiele (H:*94*) "half-invariants" (called by some writers "semi-invariants" or "seminvariants") sometimes provides an elegant method of dealing with the sums of powers.

If $s_i = u_1^i + u_2^i + \ldots + u_n^i = \Sigma u_x^i$, where u_x $(x = 1, 2, 3, \ldots n)$ denotes an observed value, then the "half-invariants" $\lambda_1, \lambda_2, \lambda_3, \ldots$ are defined by the identity (with respect to v)

$$s_0 e^{\frac{\lambda_1 v}{1!} + \frac{\lambda_2 v^2}{2!} + \frac{\lambda_3 v^3}{3!} + \cdots} = s_0 + \frac{s_1 v}{1!} + \frac{s_2 v^2}{2!} + \frac{s_3 v^3}{3!} + \qquad \ldots\ldots (a)$$

$$= e^{u_1 v} + e^{u_2 v} + e^{u_3 v} + \ldots.$$

Differentiating (*a*) with respect to v we obtain

$$s_0 e^{\frac{\lambda_1 v}{1!} + \frac{\lambda_2 v^2}{2!} + \cdots} \left[\lambda_1 + \frac{\lambda_2 v}{1!} + \frac{\lambda_3 v^2}{2!} + \cdots\right] = s_1 + \frac{s_2 v}{1!} + \frac{s_3 v^2}{2!} + \ldots,$$

or, substituting by (*a*) for the first term,

$$\left[s_0 + \frac{s_1 v}{1!} + \frac{s_2 v^2}{2!} + \cdots\right]\left[\lambda_1 + \frac{\lambda_2 v}{1!} + \frac{\lambda_3 v^2}{2!} + \cdots\right]$$

$$= s_1 + \frac{s_2 v}{1!} + \frac{s_3 v^2}{2!} + \ldots,$$

from which, by equating coefficients of powers of v,

$$s_1=\lambda_1 s_0$$
$$s_2=\lambda_1 s_1+\lambda_2 s_0$$
$$s_3=\lambda_1 s_2+2\lambda_2 s_1+\lambda_3 s_0$$
$$s_4=\lambda_1 s_3+3\lambda_2 s_2+3\lambda_3 s_1+\lambda_4 s_0$$
$$\vdots$$

Hence $\lambda_1 = \left(\frac{s_1}{s_0}\right)$

$$\lambda_2 = \left(\frac{s_2}{s_0}\right)-\left(\frac{s_1}{s_0}\right)^2$$

$$\lambda_3 = \left(\frac{s_3}{s_0}\right)-3\left(\frac{s_2}{s_0}\right)\left(\frac{s_1}{s_0}\right)+2\left(\frac{s_1}{s_0}\right)^3$$

$$\lambda_3 = \left(\frac{s_4}{s_0}\right)-4\left(\frac{s_3}{s_0}\right)\left(\frac{s_1}{s_0}\right)-3\left(\frac{s_2}{s_0}\right)^2+12\left(\frac{s_2}{s_0}\right)\left(\frac{s_1}{s_0}\right)^2-6\left(\frac{s_1}{s_0}\right)^4$$
$$\vdots$$

by which the half-invariants are expressed in sums of powers.

The relation between the half-invariants and moments (see p. 253; B; 27) is likewise

$$e^{\frac{\lambda_1 v}{1!}+\frac{\lambda_2 v^2}{2!}+\cdots}=1+\frac{\mu_1' v}{1!}+\frac{\mu_2' v^2}{2!}+\ldots, \text{ whence}$$

$$\lambda_1 = \mu_1'$$
$$\lambda_2 = \mu_2'-(\mu_1')^2$$
$$\lambda_3 = \mu_3'-3\mu_2'\mu_1'+2(\mu_1')^3$$
$$\lambda_4 = \mu_4'-4\mu_3'\mu_1'-3(\mu_2')^2+12\mu_2'(\mu_1')^2-6(\mu_1')^4$$

or, for moments about the mean,

$$\lambda_1 = 0$$
$$\lambda_2 = \mu_2$$
$$\lambda_3 = \mu_3$$
$$\lambda_4 = \mu_4-3(\mu_2)^2.$$

The utility of these half-invariants in analysis is illustrated in numerous publications, such as P:*36*.

B; 29. The Gamma and Beta Functions

The *Complete Gamma Function* for any positive number, i.e., $n > 0$, is defined as

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1}\, dx.$$

In general, $\Gamma(n+1) = n\Gamma(n)$; and $\Gamma(0) = \infty$.

If n is a positive integer, $\Gamma(n+1) = n!$; and $\Gamma(1) = 1$.

Also, $\Gamma\left(\frac{n}{2}\right) = \left(\frac{n-2}{2}\right)!$ if n is even, and $= \frac{(n-2)(n-4)\ldots 1}{2^{\frac{n-1}{2}}}\sqrt{\pi}$ if n is odd; and $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

By the substitution $x = t^2$ it follows that

$$\int_0^\infty e^{-t^2} t^m\, dt \text{ (where } m > -1) = \tfrac{1}{2}\Gamma\left(\frac{1+m}{2}\right).$$

The proofs may be found in P:*21*:I, 250 and II, 323, or P:*32*:237.

The *Incomplete Gamma Function* is defined as

$$\Gamma_x(n+1) = \int_0^x e^{-x} x^n\, dx.$$

Tables of $\Gamma(n)$ or $\log \Gamma(n)$ are available in P:*97*, P:*47*, and elsewhere, and conveniently for actuaries in P:*32*:266.

Values of the "incomplete" function have been tabulated in P:*98*.

The *Complete Beta Function* of any two positive numbers, m and n, is defined by

$$B(m, n) = \int_0^1 x^{m-1}(1-x)^{n-1}\, dx,$$

and the *Incomplete Beta Function* by

$$B_x(m, n) = \int_0^x x^{m-1}(1-x)^{n-1}\, dx.$$

As shown, for example, in P:*21*:II, 337 or P:*32*:237,

$$B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}.$$

Values of the functions have been tabulated in P:*101*.

SECTION

C

APPLICATIONS

C; 1. The Relation between *a priori* "Deviations" and Observed "Statistical Frequencies"

In many presentations of this subject it is customary to illustrate the occurrence of a "deviation" by such experiments as those of tossing a supposedly "dynamically perfect" coin (for which the probability of head, say, falling is then known, *a priori*), and to picture, on the other hand, an "error"—that is to say, an error of observation occurring as a departure from a "true" value to be sought—as the observation of an expert marksman firing at a target with a perfect rifle on a windless day. Good as these modes of illustration are, however, it must be noted that their meaning must be carefully examined, for otherwise they may be over-simplified.

The matter of the coin cannot be put better than in the following words of Levy and Roth (P:*80*:28): "It is clear that with a given coin which is tossed by some mechanical process (beginning always with, say, the head upwards) it could be arranged that the result of each toss is always head or always tail; or, alternatively, that the ratio of the number of heads to the number of tails takes on a certain series of values within a specified range. [This] illustrates the fact . . . that in any physical process to which probability is to apply, there are three interlocked elements: (1) a 'population', P, in the above case, of head and tails; (2) a process of selection S (here a mode of tossing); and (3) a sample, s, drawn from P by the application of S. This process may be stated symbolically in the form $s=S(P)$".

With regard to the illustration of the marksman firing at a target—an instance where the *a priori* probabilities are not known—confusion will often be avoided if, with Herschel (H:*28*) and Ellis (H:*27*), we picture the problem not only as presenting a distribution of the shots around a "bull", but as implying also the inverted question which perhaps can be put more clearly in these terms: If the shots have been fired at a wafer which is afterwards removed, how are we to determine, from the distribution of the shot marks, the most probable position of the wafer?

The preceding viewpoint is essential to an understanding of the basic problem of Mathematical Statistics. The situation ordinarily encountered in practice does not usually concern simple text-book cases such as a bag containing 10 balls, indistinguishable except that 4 are black and 6 are white—where it is at once apparent that the *a priori* probabilities are known, namely, .4 and .6 as the probabilities of drawing a black, or a white, ball respectively, at the first attempt, or in each of a series of successive attempts when the ball is replaced after each trial. The problem met in practice generally presents the investigator with a very different situation, for example, a series of observations, such as the shots around the wafer, as an accomplished fact—the *a priori* probabilities, it is to be noted, not having been known; the investigator is then asked to estimate as closely as he can what the real or "true" position of the wafer was. The distinction cannot be emphasized too clearly.

C; 2. The Deviations in the Number of Occurrences, and in the Statistical Frequency

If s be written for the number, $np+x$, of actual occurrences, it will be seen that the observed statistical frequency is $\frac{s}{n}$ (as on p. 187; B; 2), and the deviation x in the number of occurrences is $s-np$. It may be emphasized at this point that the deviation to which a probability is being assigned by this analysis is the deviation between the actual and expected *numbers* of occurrences. It clearly involves, of course, in a similar manner but with a different scale of abscissae, the discrepancy $\left(\frac{s}{n}-p\right)$ between the statistical frequency, $\frac{s}{n}$, and the true *a priori* probability, p.

C; 3. The Symmetrical and Unsymmetrical Point Binomial

To give a simple example, if there are n independent trials ($n=10$), then if the chance of success in each trial is $\frac{1}{2}$ ($p=\frac{1}{2}$, $q=\frac{1}{2}$), the binomial distribution (3) gives for the probabilities of the happenings of 0, 1, 2, . . ., 10 successes the terms of the expansion

$$\frac{1}{2^{10}}\{1+10+45+120+210+252+210+120+45+10+1\},$$

and therefore also these terms multiplied by the number of trials, 10, for the distribution of the expected numbers of successes. These distributions are of course symmetrical, since $p=q$.

If, on the other hand, still with $n=10$ independent trials, the chance of success in each trial is .1 ($p=.1$, $q=.9$), the probabilities of the happenings of 0, 1, 2, . . ., 10 successes from (3) are the terms of $(.9+.1)^{10}$, and the distribution of the numbers of successes is again found by multiplying each term by the number of trials, 10. These distributions, however, are "skew", i.e., unsymmetrical, since $p\neq q$.

In both cases the terms when plotted form merely a series of points (of a symmetrical and unsymmetrical bell shape, respectively), since under the conditions of the problems here considered the quantities are obviously not capable of continuous variation, i.e., $np+x$ in (1) and (2) can take only the integral values 0, 1, 2, . . . 10.

As an illustration of the meaning of the above distributions in a practical case, it will be seen that if the true probability of death at age 75, say, were known to be .1, so that for each of a homogeneous group of 10 persons of that age the probability of death would be, invariably, .1, then the theoretical distribution of the number of deaths which would be expected to occur in a series of such homogeneous groups would be that given by the second expression above multiplied by 10.

C; 4. The Practical Applicability of the Normal and Skew-Normal Curves

The Normal Curve (10) is adopted very widely in practice, but the Skew-Normal form derived as (i) and (ii) in B; 5, and shown also in Chapter VII, is seldom used. It is particularly important for the actuary to appreciate the practical justification for, and the limitations of, this symmetrical "normal" expression, since, in his work concerning death rates, p and q are markedly dissimilar at nearly all ages—the probabilities of survivorship and mortality at most ages lying between about .995 and .75, and .005 and .25, respectively. The following examples may therefore be considered.

To take a case *when p and q are each near $\frac{1}{2}$*, suppose that $p=.44$ (and $q=.56$). Then, *when n is fairly large*, such as 10,000, and $x=+50$, say, the normal expression (10) gives .0048, and the skew term $e^{\frac{50(-.12)}{20000(.2464)}}=.9988$ does not affect the fourth decimal place.

Even when n is small, for example 50, (10) with, say, $x=+10$ gives .0020, and the "skew" term is .9525, which causes an alteration only of 1 in the fourth place. The excellence of the approximation afforded by the symmetrical form (10) will also be seen from the fact that in this latter case, as a simple instance, the true value of the original factorial form (2), to which (10) is an approximation, is .0021.

It is thus clear that the normal form gives exceedingly close results when p and q are each near to .5, even when n is small. This, of course, is to be expected, because the point binomial, of which (10) is an approximate representation, is but very slightly skew when p and q are nearly equal, whether n is large or small.

When p and q are not nearly equal to $\frac{1}{2}$, furthermore, the results are ordinarily very close, although in this case the value of n—the size of the "sample"—must be watched carefully.

If n is large—employing an example from mortality data (P:*174*:53)—we find that if at a certain age the "true" rate of mortality (here p, the probability that the event will happen) is known to be $\frac{1}{93}$ exactly $=.010753$ (so that $q=.989247$), and 41,385 lives were "exposed to risk", then the probability of a deviation of exactly $+30$ from the expected number of deaths, np, according to the normal expression (10), is .0068, and even in that case the effect of the skew term is only to change this to .0066.

If n is small, however, care must be taken in applying the Normal Curve. It will be noted, from the proof given at p. 204;

B; 5, that, in effect, n was supposed to be sufficiently large that $\frac{x}{2n}\left(\frac{1}{q}-\frac{1}{p}\right)$ could be neglected even if p is not nearly equal to q. Without entering into the somewhat complex mathematical considerations underlying the admissibility of this assumption under varying circumstances (for which see P:*146*:119-133), it will be evident that the use of the normal form (10) cannot be accepted with confidence when p (or q) is so small and n suf-

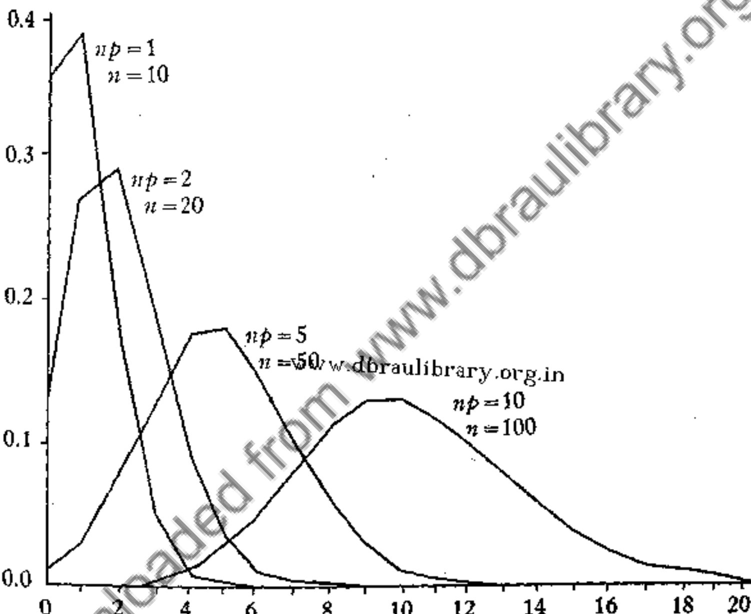

FIGURE 21.—Unsymmetrical Point Binomials for Small Values of np.

ficiently large that np (or nq) remains finite but small (see p. 230; B; 15 and p. 306; C; 14). This may be realized, also, from a consideration of Figure 21, showing the frequency polygons of an unsymmetrical point binomial $(q+p)^n$ such as $(.9+.1)^n$, for small values of np up to about 10, i.e., values of n up to about 100 (which might occur in small mortality groups).

It will be seen that, although the polygon rapidly assumes a shape very close to the normal as np increases, nevertheless for certain of the smaller values of np and n the lack of symmetry is very marked. Under such circumstances, therefore, it will be advisable to examine carefully the applicability of the Normal Curve in any particular case.

From the above examples, and from others shown in P;27: 182-4, it will be seen that the neglect of the skew term in the Skew-Normal form is usually unimportant. When the skewness is so marked that the Normal Curve is inappropriate it will, in fact, be preferable to use the Poisson exponential (55), which is developed later in Chapter VII and illustrated at p. 306; C; 14, rather than the Skew-Normal correction.

C; 5. The Finite Integration of the Normal Law of Deviations

Taking the mortality example of p. 266; C; 4 again, and hence supposing that n represents a number of lives, such as 41,385, exposed to risk at a certain age, and that the true rate of mortality is known *a priori*, or is given, as $\frac{1}{93}$ exactly, it follows that the probability of a deviation of *exactly* +30 is, by (10), approximately .0068. With the possibility of errors of different magnitudes, this probability of an error of a particular magnitude occurring is, of course, small, and, furthermore, is obviously not a matter of any special significance. The important question for investigation, in fact, is clearly not this small probability of some such specific deviation occurring, but is rather the probability that any deviation which may occur will be within certain limits, or that it will not exceed a certain amount. In this case the expected deaths are 445; let us therefore examine the probability that the actual deaths, instead of being the 445 expected, will lie between 415 and 475, i.e., that the deviation will not exceed 30 in either direction. Being $\sum_{x=-k}^{x=+k}\left(\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2npq}}\right)$, where $k=30$, the summation is ordinarily effected in practice, as explained at p. 207; B; 6, by taking either

$\frac{2}{\sqrt{\pi}}\int_0^{\frac{k}{c}} e^{-t^2}\,dt + y_k$, or $\frac{2}{\sqrt{\pi}}\int_0^{\frac{1}{c}(k+\frac{1}{2})} e^{-t^2}\,dt$, where $c=\sqrt{2npq}$, from tabulated values of the "probability integral" or "error function" $\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ (see p. 160; A; 5).

In the above illustrative case, $\frac{k+\frac{1}{2}}{c}$ where $c=\sqrt{2npq}$ is 1.0279, and $\frac{k}{c}=1.0111$. The alternative formulae therefore give from these tables $\frac{2}{\sqrt{\pi}}\int_0^{1.0111} e^{-t^2}dt + y_{30} = .854$, or $\frac{2}{\sqrt{\pi}}\int_0^{1.0279} e^{-t^2}dt = .854$. That is to say, it may be anticipated—neglecting the third decimal place—that in about 850 experiences out of 1,000 the actual deaths amongst 41,385 exposed to risk, when the true rate of mortality is $\frac{1}{93}$, will lie between 445+30 and 445−30; that is, in about 85 experiences out of 100 the deviation might range up to, but would not exceed, 30 in either direction. A deviation of 30 or less between the expected and the actual deaths would therefore not occasion surprise. Or, more precisely, the interpretation to be placed on this result would be that a deviation of 30 or less would not, as a practical matter, be significant, in the sense that it would not raise the presumption of the existence of a disturbing influence beyond that of merely chance fluctuations.

It may be useful to note also here that the complementary interpretation follows (as may similarly be demonstrated from the summations $\sum_{x=k}^{x=\infty} y_x + \sum_{x=-\infty}^{y=-k} y_x$), namely, that a deviation of 30 or more would be anticipated in only about 15 experiences in 100.

C; 6. Numerical Illustrations of the Mean Error, Mean Square Deviation, Standard Deviation, and Probable Error.

The application of these formulae may be illustrated as follows.

Using again, as a particular instance, the mortality case of C; 4 and C; 5, with $n=41{,}385$ exposed to risk and $q=\frac{1}{93}$ exactly $=.010753$ (so that $q=.989247$, and the expected deaths $np=445$), should find:

(a) *Mean Error, Irrespective of Sign*, by (20),

$$=.797885\sqrt{41385\left(\frac{1}{93}\right)\left(\frac{92}{93}\right)}=16.74 \text{ accurately,}$$

or $\frac{4}{5}\sqrt{41385\left(\frac{1}{93}\right)\left(\frac{92}{93}\right)}$ approximately $=16.78$ nearly.

That is to say, the average deviation (irrespective of sign) from the 445 expected deaths to be anticipated in an experience of 41,385 lives exposed to risk is 16.74 accurately, or approximately 16.78.

(b) *The Mean Square Deviation*, or *Variance*, being, by (7) and (14), $=npq$, is similarly 440.22.

(c) *The Standard Deviation*, σ, being $\sqrt{npq}$ by (8), accordingly $=20.98$.

It then follows, since the probabilities of deviations lying within the ranges $\pm\sigma$, $\pm 2\sigma$, $\pm 3\sigma$, . . . have been shown to be .6827, .9545, .9973, . . ., that in this case these latter probabilities are those of deviations within the ranges (to the nearest integer) ± 21, ± 42, ± 63, . . . from the mean (expected) number of deaths, 445. Or, to put the matter more specifically, the probability is, for instance, practically .9973 that in an experience of 41,385 lives, with a true mortality rate of $\frac{1}{93}$, the number of deaths will actually lie between $445-63$ and $445+63$, i.e., between 382 and 508.

If, as in a body of lives insured against the contingency of death, we view an excess number of deaths as unfavourable, the probability of actually experiencing $445+63=508$ deaths *or more* will be about .00135; while, on the other side, the probability of unfavourable experience in a group to whom annuities had been sold, to the extent of the deaths numbering only 382 *or less*, would be .00135 approximately. That is to say, insurance or annuity experiences unfavourable to this extent would be anticipated in only 13.5 (say 13 or 14) times in 10,000 trials, and thus would be very unlikely.

Experiences unfavourable to the extent of either $+2\sigma$ ($=42$) or more, or -2σ ($=-42$) or less, would similarly have probabilities of .0228, so that actual deaths of $445+42=487$ or more in an insurance experience, or $445-42=403$ or less in an annuity group, would be anticipated only 228 times in 10,000 trials, or only in slightly more than $2\frac{1}{4}\%$ of the experiences.

(*d*) For the ***Probable Error***, λ, which, by (21), $=.674489\sqrt{npq}$ $=14.2$, the numerical applications are made in exactly the same manner. Basing an illustration on $\pm\lambda$, for instance—since the use of λ itself rather than its multiples was its original concept—we see that the probability is .5, i.e., it is an even chance, that the actual deaths will lie (using integers) between $445-14$ and $445+14$, i.e., between 431 and 459. It is consequently also an even chance that they will exceed 459 or be less than 431. Considering, as before, the probabilities of positive and negative discrepancies separately, the probability is .25 of unfavourable deaths in an insurance experience to the extent of 459 actual deaths or more, while again in 25% of the experiences the unfavourable number of deaths in an annuitants' group would be 431 or less.

The solution of the inverse problem is also important, namely, *the determination of the number of trials necessary to secure a given probability* for a stated deviation between the actual and expected values. Suppose, for example, that we wish to find approximately how many trials, n, must be taken to secure a probability of .999 that $\frac{s}{n}$ will differ from p by ϵ or less.

From (11), the probability of a deviation in np of d or less in absolute magnitude is $\frac{2}{c\sqrt{\pi}}\int_0^d e^{-\frac{x^2}{c^2}}dx = \frac{2}{\sqrt{\pi}}\int_0^{\frac{d}{c}} e^{-t^2}dt = \text{Erf}\left(\frac{d}{c}\right)$;

and since the stated deviation of ϵ or less between $\frac{s}{n}$ and p means a deviation of ϵn or less between s and np, we have $d=\epsilon n$ and consequently must find (see p. 162; A; 5) the value of n for which

$\text{Erf}\left(\frac{\epsilon n}{c}\right) = .999$. But $.999 = \text{Erf}(2.327)$; hence we simply require the value of n for which $\frac{\epsilon n}{c} = 2.327$, where $c = \sqrt{2npq}$. To take a numerical case, let $p = \frac{1}{3}$, $q = \frac{2}{3}$, and $\epsilon = .01$; then $\frac{.01(\sqrt{n})}{\sqrt{2pq}} = 2.327$, whence $n = 24{,}066$, i.e., it would necessary to have (approximately) 24,066 trials in order to be satisfied, to the extent of a probability of .999, that the observed statistical frequency, $\frac{s}{n}$, would be within .01 of the true value, p.

This example is, of course, an illustration of the meaning of Bernoulli's Theorem (cf. p. 187; B; 2).

C; 7. The Mean Square Error, σ^2, of the Observed Values of Certain Actuarial Functions

The general formulae (23), (27), and (28) for the mean square error in a multiple, a linear compound, and a function, respectively, are frequently required in actuarial problems—particularly in connection with "graduation" (see Chapter VIII) and the "theory of risk". It is therefore important that the student should have available the expressions for the most usual of these cases. They are consequently assembled here.

It must be remembered that the analysis by which the general formulae have been deduced is usually admissible except when q (or p) is so small and yet n large enough that nq (or np) is less than about 10 (see p. 265; C; 4). The practical application of the formulae for the particular cases given below is consequently subject to the same limitation. In any doubtful cases the smallness of q (or p) and nq (or np) should therefore be examined critically in relation to the comments of p. 306; C; 14.

The employment of a descriptive notation will be especially convenient here in order to identify the instance under discussion. We shall therefore write $\sigma^2\{A\}$ to denote the mean square error of the *observed* quantity A, etc., and likewise $\sigma\{A\}$, $\eta\{A\}$, and

$\lambda\{A\}$, etc. Thus $\sigma^2\{q'_x\}$ will represent the mean square error of the *observed* value of the rate of mortality, q'_x; $\lambda\{q'_x\}$ will denote its "probable error"; and so forth.

(*i*) *The Survivors Observed at the End of a Year of Age*

If n persons comprising a homogeneous group, all of the same age x, are observed over the year of age x to $x+1$, and np'_x of them actually survive and so attain age $x+1$, it follows from (7) that the mean square error in this number of "successes" is np_xq_x, where p_x and q_x are the "true" (not the observed) probabilities of survival and death in the year of age x to $x+1$. The probable error, similarly, is $.6745\sqrt{np_xq_x}$ by (21). In the practical application of these expressions the "exposed to risk", E'_x, is of course often shown instead of n.

This formula for the probable error has been stated and illustrated in H:*34*:187-8. If, for example (employing the data which are, in fact, there used), n is taken as 7,943 and the observed survivors at the end of the year as 7,847, with $p_x=.9879$ and $q_x=.0121$, we find this probable error to be 6.57, which means that it is an even chance that the survivors at the end of the year (out of the 7,943 entering upon that year) will lie between 7,847+6.57 and 7,847−6.57, i.e., between 7,853.57 and 7,840.43, or (using integers) between 7,854 and 7,840.

It is to be noted carefully that the formulae should be applied and interpreted, as here shown, for the purpose of examining the mean square error, probable error, etc., in isolated observed values over a single year of age when the true p_x and q_x are known or can be assumed.

Since the preceding statistical example is stated in H:*34*:188 on the basis of certain numerical values from a hypothetical "life-table", it may therefore be advisable to warn the student that the formulae used therein cannot be applied directly to find the mean square error, etc., of the hypothetical life-table function l'_x (the number of survivors at age x from the l'_{x-1}, l'_{x-2}, . . . at earlier ages) when that l'_x is constructed from a series of inter-related observations at earlier ages. For if an observed l'_x is thus built up from an arbitrary radix, say l_a at age a, by successive multiplication by the observed probabilities of survival $p'_a, p'_{a+1}, \ldots$, we have $l'_x = l_a p'_a p'_{a+1} \ldots p'_{x-1}$. The mean square error in this function would then be found approximately by means of (28) as $\sum_{t=a}^{t=x-1}\left[\left(\frac{\partial l_x}{\partial p_t}\right)^2 \sigma^2\{p'_t\}\right] = \sum_{t=a}^{t=x-1}\left[\left(\frac{l_x}{p_t}\right)^2\left(\frac{p_t q_t}{E'_t}\right)\right]$ by (*v*) on p. 275, $= l_x^2 \sum_{t=a}^{t=x-1}\left(\frac{q_t}{p_t E'_t}\right)$, where again the undashed symbols refer to the "true" values.

(*ii*) *The Deaths Observed during a Year of Age,* θ'_x

By the same reasoning as that in the first paragraph of (*i*), it follows at once that if n, or E'_x, persons aged x in a homogeneous group are observed from age x to $x+1$, and if θ'_x $(= nq'_x = E'_x q'_x)$ of them die, then from (7) the mean square error in these deaths is again np_xq_x $(= E'_x p_x q_x)$, where p_x and q_x are the "true" values; and the probable error, by (21), is $.6745\sqrt{np_xq_x}$ $(= .6745\sqrt{E'_x p_x q_x})$.

As in (*i*), these formulae refer to an isolated observation of a single year of age, and do not apply to the hypothetical life-table deaths, d_x, where d_x is derived from a series of inter-related observations at earlier ages.

(*iii*) *The Observed Probability of Death,* $q'_x = \dfrac{\theta'_x}{E'_x}$

Using a similar argument for this case also, it will be seen immediately from (14) and (22), or (23), that $\sigma^2\{q'_x\} = \dfrac{p_x q_x}{E'_x}$.

(*iv*) *The Observed Central Death Rate,* $m'_x = \dfrac{\theta'_x}{E'_{x+\frac{1}{2}}}$

Since m'_x is defined as a death rate operating upon an observed exposed to risk $E'_{x+\frac{1}{2}}$ in the middle of the year of age, and producing during that year the observed deaths θ'_x, it follows, as in (*iii*), that $\sigma^2\{m'_x\} = \dfrac{m_x(1-m_x)}{E'_{x+\frac{1}{2}}}$ where m_x is the "true" value. This formula has been used in H:*44*:164.

An alternative approximate formula, $\sigma^2\{m'_x\} \doteq \dfrac{(m_x)^2}{E'_x q_x}$, has been derived in P:*51*:100, and it is there also suggested that in practice this may be modified to give $\sigma\{m'_x\} \doteq \frac{(m'_x)}{\sqrt{\theta'_x}}$ as a rough approximation for the standard deviation. For, since $E'_{x+\frac{1}{2}} \doteq E'_x\left(1-\frac{q'_x}{2}\right)$ and $m_x = \frac{d_x}{l_x - \frac{1}{2}d_x} = q_x\left(1-\frac{q_x}{2}\right)^{-1}$, we see that the expression in the preceding paragraph may be written

$$\sigma^2\{m'_x\} \doteq \frac{m_x(1-m_x)}{E'_x\left(1-\frac{q'_x}{2}\right)} \doteq \frac{m_x(1-m_x)}{E'_x\left(1-\frac{q_x}{2}\right)} \doteq \frac{(m_x)^2(1-m_x)}{E'_x q_x} \doteq \frac{(m_x)^2}{E'_x q_x}.$$

Consequently $\sigma\{m'_x\} \doteq \frac{m_x}{\sqrt{E'_x q_x}} \doteq \frac{m'_x}{\sqrt{E'_x q'_x}} \doteq \frac{m'_x}{\sqrt{\theta'_x}}$, being simply, and conveniently, the rate divided by the square root of the deaths. It will be realized, however, from the nature of the approximations involved, that these formulae should not be relied upon for anything more than rough indications.

(v) The Observed Probability of Survival, p'_x

By exactly the same principle as that used in *(iii)*,

$$\sigma^2\{p'_x\} = \frac{p_x q_x}{E'_x}.$$

(vi) The Observed $\operatorname{colog}_e p'_x$

This can be obtained at once, approximately, from (28). For in that formula $F = f(F_1) = f(p_x)$; $\left(\frac{\partial F}{\partial F_1}\right)^2 = \left[\frac{d(\operatorname{colog} p_x)}{dp_x}\right]^2 = \left(\frac{1}{p_x}\right)^2$; $\sigma_1^2 = \sigma^2\{p'_x\} = \frac{p_x q_x}{E'_x}$ by *(v)* here.

Hence $\sigma^2\{\operatorname{colog} p'_x\} \doteq \left(\frac{1}{p_x}\right)^2\left(\frac{p_x q_x}{E'_x}\right) = \frac{q_x}{p_x E'_x}$ (cf. H:*43*:318 and P:*138*:281-2).

A method of deducing this expression from $\sigma^2\{q'_x\}$, without the use of (28), is given in H:*108*:256. The same formula may also be obtained from first principles as shown in H:*70*:393.

(vii) The Ratio of Actual to Expected Deaths

Since the "expected deaths" are entirely independent of the rate of mortality, q'_x, actually experienced, this is a simple case of (23) where the multiplier $\kappa = \dfrac{1}{\text{Expected Deaths}} = \dfrac{1}{E'_x q_x}$,

and $\sigma^2\{\text{Actual Deaths}\} = E'_x p_x q_x$ by (7). Hence

$$\sigma^2\left\{\frac{\text{Actual Deaths}}{\text{Expected Deaths}}\right\} = \sigma^2\left\{\frac{E'_x q'_x}{E'_x q_x}\right\} \text{ is } \left(\frac{1}{E'_x q_x}\right)^2 (E'_x p_x q_x) = \frac{p_x}{E'_x q_x}.$$

The student should note that in some practical applications this formula has been modified. It is somewhat common practice to assume that p_x may be taken as unity without sensible loss of accuracy. This, however, should not be done without careful examination—see p. 293; C; 10 here, and (for example) J.I.A., LXVIII, 62.

If it is justifiable,

$\dfrac{p_x}{E'_x q_x}$ becomes simply $\dfrac{1}{E'_x q_x} = \dfrac{1}{\text{Expected Deaths}}$, as used in P:*5*.

If, in addition, it should be possible to assume that q'_x and q_x would not differ markedly, then

$\dfrac{1}{E'_x q_x}$ becomes $\dfrac{1}{E'_x q'_x} = \dfrac{1}{\text{Actual Deaths}}$ (cf. J.I.A., LXVIII, 62).

In dealing with a special category of lives—such as those with some particular medical impairment, or in a hazardous occupation—which may be subject to mortalities differing markedly from any standard table, it will be evident that the observed probabilities of death, q'_x, of the special category may themselves provide a much closer approximation to the "true" probabilities of the special category than will be afforded by the values, say q^s_x, of the standard table with which the observed probabilities might be compared. Under these circumstances, therefore, it will be better to take $\sigma^2\{\theta'_x\} \doteq E'_x p'_x q'_x$ rather than $E'_x p^s_x q^s_x$.

Then $\sigma^2\left\{\dfrac{\theta'_x}{E'_x q^s_x}\right\} \doteq \dfrac{E'_x p'_x q'_x}{(E'_x q^s_x)^2}$, so that $\sigma\left\{\dfrac{\theta'_x}{E'_x q^s_x}\right\}$, or

$\sigma\left\{\dfrac{\text{Actual Deaths}}{\text{Expected Deaths}}\right\} \doteq \dfrac{\sqrt{E'_x p'_x q'_x}}{E'_x q^s_x} \doteq \dfrac{\sqrt{E'_x q'_x}}{E'_x q^s_x} = \dfrac{M.R.}{\sqrt{\theta'_x}}$ where the "mortality ratio", *M.R.*, is the ratio of actual to expected deaths, i.e., $\dfrac{\theta'_x}{E'_x q^s_x}$. This method is used, for example, in P:*6*:12.

(viii) The Rate of Mortality by Amounts

The preceding formulae for $\sigma^2\{q'_x\}$, etc., are based on the customary definition of the probability of death in the year of age x to $x+1$ (the "rate of mortality", as it is usually called), being the probability that a life aged x will die between ages x and $x+1$. In the practical estimation of the monetary losses to be anticipated in plans of insurance, however, the measure of "success" or "failure" may obviously be indicated more closely in some cases by the amounts of money involved than by the mere enumeration of the lives without regard to the various monetary burdens which they represent. Under such circumstances it becomes important to consider the mean square error of the so-called "rate of mortality by amounts", i.e., the probability that a random monetary unit (say \$1.00) exposed to risk will become a claim by death in the year of age x to $x+1$. The necessary formulae for this case have been deduced by the following reasoning in P:*19*. [In the notation which is adopted here, a takes the place of Cody's s in that paper, and the experienced values are, as throughout this book, distinguished by dashes.]

Since at age x the various lives exposed, which hitherto have been assumed to be a homogeneous group of individuals, will now be associated with varying amounts of the monetary unit, it will be seen that the total "exposed to risk" E'_x must be defined as $\Sigma({}^aE'_x)$, where ${}^aE'_x$ is the exposure in the "amount class" a when the amounts at risk in the year of age x to $x+1$ are classified by the varying amounts of the monetary unit, and where the Σ covers all the values of a. Similarly the actual losses which emerge upon death, totalling θ'_x in the year of age, will be $\Sigma({}^a\theta'_x)$. From these definitions it follows that in the amount class a the observed probability of loss in the year of age x to $x+1$ will be $\frac{{}^a\theta'_x}{{}^aE'_x} = {}^aq'_x$ say, and that for all amount classes in that year of age the observed probability will be $\frac{\Sigma({}^a\theta'_x)}{\Sigma({}^aE'_x)} = \frac{\theta'_x}{E'_x} = q'_x$ say. Correspondingly, if the "true" probability in the amount class a be written ${}^aq_x$, the expected losses in that class will be ${}^aE'_x\,{}^aq_x = {}^a\theta_x$ say; and if the "true" probability for all the amount classes combined be denoted by q_x, then the expected losses in all amount classes will be $E'_x q_x = \theta_x$ say. Furthermore, we must evidently have the relation $\theta_x = \Sigma({}^a\theta_x)$, consistently with $\theta'_x = \Sigma({}^a\theta'_x)$ for the observed data.

In this model there are a units in the amount class a for every unit which would be tabulated if the data were compiled on the basis of lives only. We can therefore write ${}^a\theta'_x = a({}^a d'_x)$ where ${}^a d'_x$ denotes the actual deaths (as distinct from the monetary losses) in the amount class a, and similarly ${}^aE'_x = a({}^a l'_x)$ where ${}^a l'_x$ denotes the lives exposed to risk. Consequently $\sigma^2\{{}^a\theta'_x\} = \sigma^2\{a({}^a d'_x)\} = a^2\sigma^2\{{}^a d'_x\}$ by (23). Moreover, by section (*ii*) of this Appendix C; 7, we shall have $\sigma^2\{{}^a d'_x\} = {}^a l'_x\ {}^a p_x\ {}^a q_x$, so that $\sigma^2\{{}^a\theta'_x\} = a^2({}^a l'_x\ {}^a p_x\ {}^a q_x)$. Proceeding now to the consideration of $\sigma^2\{\theta'_x\}$ for all the groups, it follows at once from these relations and (27) that $\sigma^2\{\theta'_x\} = \sigma^2\{\Sigma({}^a\theta'_x)\} = \Sigma[a^2\ {}^a l'_x\ {}^a p_x\ {}^a q_x]$ $= \Sigma[a\ {}^aE'_x\ {}^a p_x\ {}^a q_x]$; and from this result we see that $\sigma^2\{q'_x\}$, or $\sigma^2\left\{\dfrac{\theta'_x}{E'_x}\right\}$, for all the amount classes, which by (23) is $\dfrac{1}{(E'_x)^2}\sigma^2\{\theta'_x\}$, becomes $\dfrac{1}{(E'_x)^2}\Sigma[a\ {}^aE'_x\ {}^a p_x\ {}^a q_x]$.

In order to get this into a form related to the usual $\sigma^2\{q'_x\} = \dfrac{p_x q_x}{E'_x}$ as stated in (*iii*) here for the probability of death when only lives (not amounts) were under consideration, we note that, since ${}^a q_x = q_x + (q^a_x - q_x)$ and ${}^a p_x = p_x - (q^a_x - q_x)$, the expression just given can be put as

$$\frac{1}{(E'_x)^2}\Sigma[a\ {}^aE'_x p_x q_x] + (p_x - q_x)\Sigma[a\ {}^aE'_x(q^a_x - q_x)] - \Sigma[a\ {}^aE'_x(q^a_x - q_x)^2]$$

$$= \frac{1}{(E'_x)^2}\left\{\Sigma[a\ {}^aE'_x p_x q_x] + \Sigma[a\ {}^aE'_x(q^a_x - q_x)(p_x - q^a_x)]\right\}$$

$$= \frac{1}{(E'_x)^2}\Sigma[a\ {}^aE'_x p_x q_x]\left\{1 + \frac{\Sigma[a\ {}^aE'_x(q^a_x - q_x)(p_x - q^a_x)]}{\Sigma[a\ {}^aE'_x p_x q_x]}\right\}$$

$$= R^2_x\left(\frac{p_x q_x}{l'_x}\right)$$

$$\text{where } R^2_x = \frac{l'_x\Sigma[a\ {}^aE'_x]}{(E'_x)^2}\left\{1 + \frac{\Sigma\left[a\ {}^aE'_x\left(\dfrac{q^a_x}{q_x} - 1\right)\left(1 - \dfrac{q^a_x}{p_x}\right)\right]}{\Sigma[a\ {}^aE'_x]}\right\}.$$

and $l'_x = \Sigma^a l'_x$, namely, the total number of lives exposed to risk.

This $\frac{p_x q_x}{l'_x}$, in accordance with section (*iii*) here, is the mean square error of the rate of mortality observed in a group of l'_x lives exposed to risk, in which the "true" rate of mortality, q_x, is $\frac{\theta_x}{E'_x}$ as previously defined; and R^2_x represents the ratio in which that mean square error, based on lives, is increased when the investigation follows a classification by amounts.

A numerical illustration may be found in P:*19*:72.

(*ix*) *The Ratio of Actual to Expected Losses by Amounts*

In section (*vii*) an approximate formula for the standard deviation of the ratio of the actual to expected deaths by lives was found to be $\sigma\left\{\frac{\theta'_x}{E'_x q^s_x}\right\} \doteq \frac{M.R.}{\sqrt{\theta'_x}}$, where the "mortality ratio", $M.R. = \frac{\theta'_x}{E'_x q^s_x}$, and q^s_x is the rate of mortality according to the

standard table used in computing the expected deaths. When the ratio of actual to expected losses is to be taken by amounts (not by lives), the development in section (*viii*) evidently must be used instead, with the formula $\sigma^2\{q'_x\} = R^2_x\left(\frac{p_x q_x}{l'_x}\right)$ as the basis, where p_x and q_x are the "true" probabilities (based on amounts) for all the amount classes combined. In dealing with a special category of lives, moreover, as pointed out in section (*vii*), it will be preferable to use the observed probabilities, rather than those of the standard table, as an estimate of the "true" probabilities.

Now, corresponding to $\sigma^2\{q'_x\} = R^2_x\left(\frac{p_x q_x}{l'_x}\right)$ we have $\sigma^2\{\theta'_x\} = \sigma^2\{E'_x q'_x\} = (E'_x)^2\sigma^2\{q'_x\} = (E'_x R_x)^2\left(\frac{p_x q_x}{l'_x}\right)$, which here is to be taken approximately as $(E'_x R_x)^2\left(\frac{p'_x q'_x}{l'_x}\right)$. Hence also

$\sigma^2\left\{\frac{\theta'_x}{E'_x q^s_x}\right\} \doteq \left(\frac{1}{E'_x q^s_x}\right)^2 (E'_x R_x)^2 \left(\frac{p'_x q'_x}{l'_x}\right) = \left(\frac{R^2_x}{l'_x}\right)\frac{p'_x q'_x}{(q^s_x)^2}$; and dropping the p'_x, as in (*vii*), since $p'_x \doteq 1$, this is $\doteq \left(\frac{R^2_x}{l'_x}\right)\frac{q'_x}{(q^s_x)^2} = R^2_x\left(\frac{E'_x}{l'_x}\right)\frac{\theta'_x}{(E'_x q^s_x)^2}$ $= R^2_x\left(\frac{E'_x}{l'_x}\right)\frac{(M.R.)^2}{\theta'_x}$. The corresponding approximate formula for the standard deviation, σ, is therefore $R_x\sqrt{\frac{E'_x}{l'_x}}\left(\frac{M.R.}{\sqrt{\theta'_x}}\right)$, or $R_x \frac{M.R.}{\sqrt{\frac{l'_x}{E'_x}\theta'_x}}$ as stated in P:*19*:72.

(*x*) *The Observed Expectation of Life*, e'_x

A convenient approximate formula for $\sigma^2\{e'_x\}$ may be obtained easily by writing

$$e'_x = \frac{l'_{x+1}+l'_{x+2}+\ldots}{l'_x} = p'_x + {}_2p'_x + \ldots$$
$$= (1-q'_x) + (1-q'_x)(1-q'_{x+1}) + \ldots,$$

and applying (28) thereto since the values of q' are independent. Dropping the primes for the preliminary formula which follows, we see that

$$\frac{\partial e_x}{\partial q_{x+n}} = \frac{\partial}{\partial q_{x+n}}\Big[(p_x + {}_2p_x + \ldots + {}_np_x) + \{(p_x p_{x+1}\ldots p_{x+n}) + (p_x p_{x+1}\ldots p_{x+n}p_{x+n+1}) + \ldots\}\Big]$$
$$= \frac{\partial}{\partial q_{x+n}}\Big[(p_x + {}_2p_x + \ldots + {}_np_x) + (1-q_{x+n})\{(p_x p_{x+1}\ldots p_{x+n-1}) + (p_x p_{x+1}\ldots p_{x+n-1}p_{x+n+1}) + \ldots\}\Big]$$
$$= \frac{(-1)}{p_{x+n}}\Big[{}_{n+1}p_x + {}_{n+2}p_x + \ldots\Big]$$
$$= \frac{(-1)}{p_{x+n}}\left(\frac{l_{x+n}}{l_x}\right)e_{x+n}, \text{ since the } q\text{'s are independent}$$

so that $\frac{dp_{x+t}}{dq_{x+n}} = -\frac{dq_{x+t}}{dq_{x+n}} = -1$ when $t=n$, and is 0 otherwise.

Using this result in (28), therefore, and remembering that, by (*iii*) here, $\sigma^2\{q'_{x+n}\} = \dfrac{p_{x+n}q_{x+n}}{E'_{x+n}}$, we find immediately ($\omega$ being the limiting age)

$$\sigma^2\{e'_x\} = \sum_{n=0}^{n=\omega}\left(\frac{l_{x+n}\, e_{x+n}}{p_{x+n}l_x}\right)^2\left(\frac{p_{x+n}q_{x+n}}{E'_{x+n}}\right)$$
$$= \left(\frac{1}{l_x}\right)^2 \sum_{n=0}^{n=\omega}\left[\frac{q_{x+n}(l_{x+n}e_{x+n})^2}{p_{x+n}E'_{x+n}}\right]$$

which corresponds to Steffensen's formula for the life annuity in (*xi*) below.

A comparable formula for $\sigma^2\{\mathring{e}'_x\}$, with a discussion of the effect of using groups of ages, is to be found in P:*158*.

(*xi*) *The Immediate Life Annuity*, a'_x

The first examination of the mean square error of a'_x was undertaken by G. F. Hardy in P:*51*:99, where he deduced an approximate formula only, and gave numerical illustrations. It may be useful to note his conclusion that "the standard deviation in the values of the 3% annuities in an aggregate experience such as the $O^{M(5)}$ [the British Offices' Life Tables, 1863-93, Males, excluding the first 5 years of insurance] is about one-fiftieth of a year's purchase from about 30 to 65 years of age". He also pointed out that in "an experience including about 1000 deaths distributed approximately as in the $O^{M(5)}$ data, the deduced annuity values between ages 30 and 60 would on the average be uncertain to about $\pm .20$, or from 1% to $1\frac{1}{2}$% of their values".

A closer approximation was next given by J. F. Steffensen (P:*138*:281), who used the method of proof based on (28) as in (*x*) above, and thus reached the following formula corresponding to that already given for the expectation of life (which is, of course, the immediate life annuity at zero rate of interest):

$$\sigma^2\{a'_x\} = \left(\frac{1}{D_x}\right)^2 \sum_{n=0}^{n=\omega}\left[\frac{q_{x+n}N^2_{x+n}}{p_{x+n}E'_{x+n}}\right].$$

This expression shows clearly that if the exposed to risk, E', be multiplied at each age by a constant factor, k, then $\sigma\{a'_x\}$

will be divided by $\sqrt{k}$. For example, the total exposed in the O^M experience was 7,659,454, and in the H^M [Institute of Actuaries' Life Tables, Healthy Males, 1863] was 1,199,093; if, therefore, the same proportionate distribution could be assumed over the various ages, $\sigma\{a_x'\}$ for the O^M experience should be about .4 of that for the H^M—an indication with which the figures at decennial ages "agree very fairly except at the highest ages" (P:*138*:282).

Finally, T. Tinner (P:*142*:305) deduced from first principles an exact expression in the form

$$\sigma^2\{a_x'\} = \sum_{t=1}^{t=\omega}\left[v^{2t}{}_tp_x^2(1+2a_{x+t})\left\{\left(1+\frac{q_x}{E_x'p_x}\right)\left(1+\frac{q_{x+1}}{E_{x+1}'p_{x+1}}\right)\cdots\left(1+\frac{q_{x+t-1}}{E_{x+t-1}'p_{x+t-1}}\right)-1\right\}\right]$$

and demonstrated its relation to Hardy's and Steffensen's approximations. Tinner's investigations showed that the ratio of the exact standard deviation to the value thereof derived from the approximate expressions is generally much less than a maximum of about 1.04, so that, "as it would never be necessary to compute the standard deviation with minute accuracy, it is clear that the approximate expression gives a result close enough for all practical purposes".

(*xii*) *"Linear Compounding" in the Theory of Graduation*

Formula (27), which shows the relation of the mean square error of any linear compound of a number of independent quantities to the mean square errors in the quantities themselves, is of very special importance to actuaries, since it is used for certain basic criteria and comparative tests in the theory of graduation (see Chapter VIII). Its customary method of application for those purposes, and the manner in which it affords a measure by which the graduating power of different "linear compound" graduation formulae may be assessed, are explained in the author's paper P:*166*, and in P:*167*:111, to which actuarial students must be referred for details. It will be sufficient here to give (from that paper) the following statement of the principles involved.

"Graduation" by "linear compounding" concerns the replacement of an observed value u'_r (of a "true" value u_r) by a linear compound, v_r, of u'_r and terms $u'_{r+1}, u'_{r+2}, \ldots u'_{r-1}, u'_{r-2}, \ldots$ on either side of it, on the assumption that differences of u beyond a certain order j may be neglected. Any such graduation formula, therefore, will be of the form

$$v_r = l_r u'_r + (l_{r+1}u'_{r+1} + l_{r-1}u'_{r-1}) + (l_{r+2}u'_{r+2} + l_{r-2}u'_{r-2}) + \ldots \\ \ldots + (l_{r+n}u'_{r+n} + l_{r-n}u'_{r-n}) \quad \ldots (a)$$

for a range of $2n+1$ terms; and where the l's are symmetrical, so that $l_{r+t} = l_{r-t} = b_t$, say, this becomes

$$v_r = b_0 u'_r + b_1 u'_{r\pm1} + b_2 u'_{r\pm2} + \ldots + b_n u'_{r\pm n} \quad \ldots (b)$$

where $u'_{r\pm t}$ is written for $u'_{r+t} + u'_{r-t}$.

Now if the "error" in the observed u'_r be e_r, so that $u'_r = u_r + e_r$, it may generally be assumed, in dealing with series of observed data such as rates of mortality, etc., that the e's are independent and that the mean square error of each is (say) e^2. It follows then, by (27), that the mean square error of v_r in (a) is $(\Sigma l_r^2)e^2$. The process of "linear compounding" expressed in the general form (a) therefore replaces the original u'_r, with its mean square error e^2, by the "graduated" v_r with mean square error $(\Sigma l_r^2)e^2$; that is to say, the graduation will have reduced the mean square error of u'_r to $\frac{(\Sigma l_r^2)e^2}{e^2}$, or Σl_r^2, of its original amount. If a particular linear compound formula (i.e., a formula of given range and with certain values of the l's) has been determined in some manner, this "reduction of error", Σl_r^2, effected in the observed u'_r may consequently be taken as a measure of the "accuracy" of the graduation. Furthermore, clearly we can set out, by making Σl_r^2 a minimum, to find the l's (for a given range) which will give the greatest possible reduction of mean square error; the resulting linear compound formula will then be the "best" formula (for that range) on the particular assumptions made.

The same principle can be applied to investigate the effect of any linear compound graduation upon the mean square error in the differences of u'_r, instead of in u'_r itself. Since in practice differences beyond a certain order j may usually be assumed (as stated above) to be negligible, several cases arise according to the value assigned to j, and depending also upon whether the mean square error of $\Delta^{j+1}v$, or of $\Delta^{j}v$, or of $\Delta^{j-1}v$, . . . is considered. The reductions thus effected in the mean square error of the differences may be taken as criteria of "smoothness", in contrast to the "accuracy" of v itself; and again, by determining the l's to give the greatest possible reductions, the formulae may be evolved which will be the "best" expressions under the conditions assumed.

The formulae which secure the greatest possible reduction in the mean square error of v itself are, in fact, identical with those which result from "fitting" a polynomial $v_x = A + Bx + Cx^2 + \ldots + Jx^j$ by the Method of Least Squares (see P:*166*:100, and p. 132 here). Their first complete treatment was given in 1871 by Erastus L. De Forest, after preliminary consideration by Schiaparelli (H:*36*).

For actuarial purposes, however, when usually $j=3$, the important cases are those giving the greatest possible reduction in the mean square error of $\Delta^4 v$ or of $\Delta^3 v$. The former were investigated originally in a most valuable and elegant series of papers by De Forest as early as 1873—De Forest's work (which for many years remained unknown) thus antedating in its conception many other later contributions which still are often credited with priority. The case of $\Delta^3 v$ was indicated first by G. F. Hardy, then stated completely by W. F. Sheppard, and later restated independently in different form by R. Henderson and again by J. R. Larus. The history and details of those investigations are set out in P:*166*, with an appraisal of De Forest's fundamentally important contributions. A summary is given in section VII of Chapter XI here.

(*xiii*) *The "Theory of Risk"*

Although it does not seem advisable in this study to examine in detail the somewhat extensive mathematical and practical considerations involved in the so-called "Theory of Risk" attaching to the "grant" or sale of life annuities and insurances, it may be well to indicate briefly the nature of the problem, the manner in which it utilizes the concepts here under discussion, and the appropriate bibliography.

It is important, firstly, to realize that the matter is essentially distinct from that dealt with in reaching such formulae as those for $\sigma^2\{a'_x\}$ in (*xi*). The latter concern the "errors" which arise as a result of limited data (the observed exposed to risk and deaths), and thus seek to measure the effect of such errors upon the computed annuity value a_x. The formulae of (*xi*) accordingly represent the mean square error in a'_x, i.e., the average of the squares of the deviations expected in the observed value of a'_x as deduced from a given set of exposed to risk (E') and deaths ($E'q'$). They consequently provide a means of testing the permissible range of variation in the annuity values so observed, in accordance with the principles summarized on p. 21, and illustrated at p. 269; C; 6. For example, if we have a known and widely accepted mortality experience, such as the O^M, from which the "true" values of q and p could be taken, and if some other body of data had yielded at age x the observed annuity value a'_x, then a deviation of $\pm 3\sigma\{a'_x\}$ or more would be very unlikely as a result of merely chance variations, and its actual occurrence would indicate the existence in that other body of data of some definite cause leading to so great a variation from the O^M experience.

The "Theory of Risk", on the other hand, considers the deviations encountered in the estimated value of an annuity (or an insurance) which may be supposed to have been issued on a specified life, or lives. The distinction between the two problems is well drawn by Steffensen's remark (P:*138*:280) that the formulae such as those for $\sigma^2\{a'_x\}$ in (*xi*) concern the *origin* of the tabulated values, whereas the "theory of risk" examines their *application*.

The first treatment of the fundamental formulae required was published by Bremiker (H:*41*:286) in 1871. A very convenient restatement of Bremiker's reasoning and conclusions was given by G. F. Hardy (P:*51*:104) in 1909, and the question has again been examined on the same lines within the last few years (see P:*103* and P:*89*). It will be sufficient here to give the essence of the method, as shown by Hardy, for the case of the continuous life annuity, $\bar{a}_x$.

Suppose that an annuity is granted to a person aged x at a price $\bar{a}_x$—the annuity payments being assumed to be made continuously throughout the year, and interest being correspondingly based on the force of interest $\delta(=\log_e(1+i)=-\log_e v)$. Then if the annuitant should die at the end of time t, the deviation (as at the date of entry) from the mean value $\bar{a}_x$ for which the annuity was sold to the specified individual would be $\bar{a}_{\overline{t|}}-\bar{a}_x$; the probability of the annuitant so dying at time t is ${}_tp_x\mu_{x+t}$ $=-\dfrac{d}{dt}({}_tp_x)$; and therefore the mean square deviation, being the sum, for all values of t, of the squares of the deviations multiplied by the frequency in each case, is $\int_0^\infty(\bar{a}_{\overline{t|}}-\bar{a}_x)^2{}_tp_x\mu_{x+t}dt$.

Now $\bar{a}_{\overline{t|}}-\bar{a}_x=\left(\dfrac{1-v^t}{\delta}\right)-\left(\dfrac{1-\bar{A}_x}{\delta}\right)=\dfrac{\bar{A}_x-v^t}{\delta}$, so that the integral becomes $\dfrac{(\bar{A}_x)^2}{\delta^2}\int_0^\infty {}_tp_x\mu_{x+t}dt-\dfrac{2\bar{A}_x}{\delta^2}\int_0^\infty v^t{}_tp_x\mu_{x+t}dt+\dfrac{1}{\delta^2}\int_0^\infty v^{2t}{}_tp_x\mu_{x+t}dt$ $=\dfrac{(\bar{A}_x)^2-2\bar{A}_x(\bar{A}_x)+\bar{A}'_x}{\delta^2}=\dfrac{\bar{A}'_x-(\bar{A}_x)^2}{\delta^2}$ where $\bar{A}'_x$ is computed at a rate of interest corresponding to $v'=v^2$ (that is, $\delta'=2\delta$, or at a rate $j=(1+i)^2-1$). We thus find that *the standard deviation of the distribution for which $\bar{a}_x$ is the mean value* is $\dfrac{1}{\delta}\sqrt{\bar{A}'_x-(\bar{A}_x)^2}$. Illustrative numerical values are given in H:*41*:289, P:*103*:243-4, and P:*89*:71, which show that the standard deviation ranges generally from about 5 at the younger ages to about 2 or 3 at age 80, and represents an increasing proportion of the mean value $\bar{a}_x$.

The comparable formulae for the standard deviations of the distributions for which the mean values are other functions such as $\bar{A}_x$, $\mathring{e}_x$, a_x, the terminal reserves, etc., are indicated easily in P:*51*:105, and are proved also, with numerical interpretations, in P:*103* and P:*89*.

The formulae of the type just considered in effect examine the mean square errors and standard deviations over the *entire lifetime* subsequent to the age at entry x. They are consequently to be distinguished carefully (cf. P:*89*:78 and 602) from those referring only to *each year separately*, which form the basis of the discussions of the "risk" problem in numerous other papers (such as P:*113*, P:*117*:296, and P:*53*:40). In this latter case suppose that each of n persons effected t years ago a policy of 1 at age x; then the standard deviation in the deaths for the year of age $x+t$ to $x+t+1$ (being the policy year $t+1$) is, by (8), $\sqrt{np_{x+t}q_{x+t}}$; the net amount at risk on each policy at the beginning of the year is $v(1-{}_{t+1}V_x)$, where ${}_{t+1}V_x$ is the terminal reserve at the end of the year; and consequently the standard deviation in the amounts at risk for the year is

$$v(1-{}_{t+1}V_x)\sqrt{np_{x+t}q_{x+t}}={}_tR_x \text{ (say)} \ldots\ldots\ldots\ldots (c)$$

The obvious generalizations of this basic formula for the cases when the amounts at risk are not the same for every individual, or the age varies, are discussed in P:*53* and P:*117*:296.

The relation between this formula for a single year and the corresponding expression for the entire duration (see also P:*89*:602) is given by Hattendorff's Theorem (H:*38*), which states that if R^2 is the mean square error in the amounts at risk for the entire duration, and ${}_tR_x^2$ by (c) above is that for the $(t+1)$th year, then $R^2=\sum_{t=0}^{t=\infty} {}_tE'_x\, {}_tR_x^2$ where ${}_tE'_x=v^{2t}{}_tp_x$ and is therefore a pure endowment calculated at a rate of interest corresponding to $v'=v^2$ (that is, $\delta'=2\delta$). A demonstration of this theorem, by Steffensen, and two other proofs due to N. P. Bertelsen, are given in P:*139*:8.

C; 8. The Standard Deviation of the Arithmetic Mean

Particularly in view of certain later discussions in Chapter V, it will be well for the student to be quite clear as to the meaning of this result. A single quantity is supposed to be under observation; r independent and unbiassed observations are made; the arithmetic mean of the results is taken; and it is assumed that σ^2 is known.

For example, imagine a large homogeneous group of n persons (all of the same age, sex, habitat, etc.), who may properly—within practicable limits—be considered to represent a "random sample" (see Chapter V); and suppose that the number of deaths (within a year, say) is the "single observed quantity". Then if s_1 deaths actually occur (giving an observed statistical frequency of deaths of $\frac{s_1}{n}$), the mean square error, by (7), is npq, where p and q are the "true" probabilities of survival and death. Now if another similar group, of precisely the same size n, is again observed, and s_2 deaths are found (with an observed statistical frequency of $\frac{s_2}{n}$), the mean square error is again npq. Suppose, therefore, that in r such similar groups, each of size n, the deaths observed quite independently are $s_1, s_2, \ldots, s_r$. The mean square error for every "observation" will be npq; the arithmetic mean of the r independent determinations is $\frac{s_1+s_2+\ldots+s_r}{r}$; then by (30) the mean square error of that result is $\frac{\sigma^2}{r}=\frac{npq}{r}$. That is, $\sigma\{A.M.\}=\frac{\sigma}{\sqrt{r}}$.

C; 9. Numerical Applications of σ^2 {Ratios} to Actuarial Data

(1) Suppose that in a homogeneous sample group of 1000 men, 65% are eligible to qualify for admission under the rules of a specified pension plan; what are the limits of variation thence to be anticipated in the "universe" of men from which the sample was drawn?

By formula (8), $\sigma=\sqrt{1000pq}$, where p is the true percentage of eligibles in the "universe" or "parent population". Although p is thus not known, pq cannot exceed $(\frac{1}{2})(\frac{1}{2})$, or $\frac{1}{4}$, so that σ cannot exceed $\sqrt{1000(\frac{1}{4})}$, or 16. Using now the principle that $\pm 3\sigma$ embraces over 99% of the deviations (see p. 20), it follows that the limits of variation may be put at $\pm 3(16)=\pm 48$, which is $\pm 4.8\%$ of the sample of 1000. Hence the percentage of eligibles in the parent population would be between $65\% \pm 4.8\%$, i.e., between 60.2% and 69.8%.

In view of the meaning of Bernoulli's Theorem concerning the approach of the observed statistical frequency to the true probability as n increases (see p. 187; B; 2), an approximation of a somewhat closer kind would be obtainable in such a case as this, where p and q are unknown, by taking the observed statistical frequency, .65, as an estimate of p. Without at the moment discussing the justifiability of this substitution except to remark upon its clearly reasonable nature (see p. 292; C; 10), we should then find that $\sigma \doteq \sqrt{1000(.65)(.35)} = 15.08$, so that $\pm 3\sigma \doteq \pm 45.24$, and the percentage of eligibles in the parent population would lie between 65%±4.5%, i.e., 60.5% and 69.5%, which differs only slightly from that previously obtained.

(2) Suppose that in the above example it were known (or could be confidently assumed) that 65% is the "true" value of the percentage of eligibility, and that a comparable random sample of 1500 cases had been taken some years later, which showed 1100 eligibles. Can this be attributed to chance fluctuations alone? In a group of 1500 the expected number of eligibles $= 1500(.65) = 975$; $\sigma = \sqrt{1500(.65)(.35)} = 18.47$; the $\pm 3\sigma$ limits are therefore ± 55.41, so that in a sample of 1500 the eligibles would lie between 975 ± 55.41, i.e., between say 920 and 1030, if the fluctuations were due solely to chance. The occurrence of so many as 1100 eligibles therefore suggests that some specific cause must be sought as the reason for so great a variation.

(3) Again assuming 65% as the "true" percentage of eligibility, suppose that the 1500 sample showed 1100 eligibles (i.e., 73%), while a second sample of 2000 gave 1200 eligibles (i.e., only 60%). Can the difference shown have arisen solely from chance variations? Here formula (31) is directly applicable, giving σ^2 for the difference between the observed proportions (73% and 60%) of $(.65)(.35)\left(\frac{1}{1500} + \frac{1}{2000}\right)$, from which $\sigma = .01629$. Hence $\pm 3\sigma$ for the difference is $\pm .048$. The difference between the proportions is .13, which lies far outside the $\pm 3\sigma$ limits; consequently it is to be concluded that some influence other than chance fluctuations has operated to cause the change between the 60% and 73% in the two samples.

While this rapid method of applying the $\pm 3\sigma$ rule is very generally adopted, and indicates the conclusion clearly in such a case as this, the probability itself can of course be calculated directly from the tables of the probability integral very easily. For it has been shown in (25) that the Normal Curve is applicable to the case of the sum of two independent quantities, and it follows from its extension to (27) that the normal law also holds in the more general linear compound case, of which this problem is an instance. Now (see p. 162; A; 5) the probability, under the normal law, of an "error" of $\pm d$ or less is $\text{Erf}\left(\frac{d}{\sigma\sqrt{2}}\right)$, and the probability of an error as much as, or more than, d is therefore $1-\text{Erf}\left(\frac{d}{\sigma\sqrt{2}}\right)$. Here $d=.13$, and $\sigma=.01629$, whence $\frac{d}{\sigma\sqrt{2}}=5.64$. The probability that the chance distribution of errors of the normal law would give an error so great numerically as .13 when $\sigma=.01629$ is therefore $1-\text{Erf}(5.64)$, or practically zero. That is to say, so great a difference as that between the observed 73% and 60% cannot have occurred by chance—some specific cause must have influenced the change.

(4) Since the above procedure may, with proper reservations, be applied to the mortality rates of different samples or groups, it may be well to give another example (from P:*16*:270, slightly modified).

Suppose that at a certain age in a particular district one group of 224,728 males exhibits a probability of death (q'_x) of .00486, while another group of 244,906 males gives q'_x as .00420 (the data having been collected with as much regard as is practicable for the homogeneity requirements of simple sampling); it is known that the "true" rate of mortality for males of that age in that district can be taken as .00453; does the difference of .00066 in the death rates of the two groups indicate a real difference in the mortality, or might it have arisen from purely chance variations? By formula (31), σ^2 for the difference in the two observed rates of mortality is $(.00453)(.99547)\left(\frac{1}{224728}+\frac{1}{244906}\right)$, whence $\sigma=.000196$. Consequently the probability of a difference (an "error") as great as .00066 is $1-\text{Erf}\left(\frac{d}{\sigma\sqrt{2}}\right)$ where $d=.00066$ and $\sigma=.000196$, giving $1-\text{Erf}(2.38)=1-.9992=.0008$, so that the probability is almost negligible that the difference observed can have been due to chance, i.e., there is a real difference in the mortalities of the two groups.

Using the $\pm 3\sigma$ rule, we see that $\pm 3\sigma=\pm .00059$; the observed difference of .00066 again lies outside that range, and hence is not attributable to mere chance. In this case, however, the indication given by the probability itself is clearer on account of the order of magnitude of the figures involved.

In examples (3) and (4) it has been supposed that the "true" p and q are definitely known. If this is not so, however (and in actuarial problems it usually is not so), it would be necessary to form an estimate of their values by reference to previous experience, or to use the evidence of the data alone by resorting to formula (32) as an approximation (see p. 292; C; 10).

C; 10. On Certain Approximations for npq when the "True" p and q are not known

Throughout this study it has been emphasized that the primary objective has been to explore, mathematically and statistically, the deviations which may be expected to occur, by pure chance, between an observed statistical frequency, $\frac{s}{n}$, of s successes in n trials, and the true *a priori* probability, p. The fundamental thought underlying Bernoulli's theorem is the approach of $\frac{s}{n}$ to p as n is increased indefinitely.

Now it has been seen that the various formulae involving the probabilities of these deviations require that the true probabilities, p and q, be known. The fundamental mean square error, for example, of the $np=s$ successes in n trials, as given by npq in formula (7), means that in a series of independent but similar observations, based successively on $n_1, n_2, \ldots, n_\nu$ cases, the respective values of σ^2 would be n_1pq, n_2pq, ..., $n_\nu pq$—for the basic conditions are unchanged, so that p and q remain the same.

In some problems these true values of p and q may be known. But if they are not, it is obviously necessary in many cases, if a numerical solution is to be reached at all, to form some estimate of their values. If there is no outside source of information which can be used, it may be necessary to employ a value based on the observed values as an estimate of the true p—a procedure which clearly would become less and less of an approximation as n increases.

In the practical application of these methods to actuarial problems the true p and q are seldom known. If, consequently, it becomes essential to base an estimate on the observed values, the immediately obvious method would be to assume that, if n is large, the observed statistical frequency may be employed in each case, so that the above series would be taken as $n_1p'_1q'_1$, $n_2p'_2q'_2$, . . . , $n_\nu p'_\nu q'_\nu$ where $p'_y = \frac{s_y}{n_y}$ $(y=1, 2, \ldots, \nu)$ (cf. P:*115*). This method of approximation, however, must be used with considerable care. Yule and Kendall express the following view: "Precisely how large n must be for this approximation to be valid it is not easy to say. Samples of 1000 are almost certainly large enough, and we may often apply the foregoing procedure to much smaller samples, say of 100. For samples below that figure it is well to examine carefully the circumstances of any given case and to proceed with caution" (P:*177*:355).

When we are dealing with more than one sample it is clear that the estimate of p would be based on a larger body of data if, instead of taking $p'_y = \frac{s_y}{n_y}$ $(y=1, 2, \ldots, \nu)$ from each separate sample, the observed values were all combined as $\frac{s_1+s_2+\ldots+s_\nu}{n_1+n_2+\ldots+n_\nu}$. This is, of course, the method shown by formula (32) for two samples only.

In the case of mortality statistics, for which at many ages q is small, it is frequently suggested (as in P:*115*) that an approximation may be used by first substituting the observed values $p_1', p_2', \ldots, p_\nu'$ in each case for p (and correspondingly for q), as noted above, and then taking $p_1' = p_2' = \ldots = p_\nu' \doteq 1$. If this is legitimate, then the values of σ^2 for each of the ν independent samples of $n_1, n_2, \ldots, n_\nu$, which strictly are n_1pq, n_2pq, $\ldots, n_\nu pq$, become first $n_1p_1'q_1'$, $n_2p_2'q_2'$, $\ldots$, $n_\nu p_\nu' q_\nu'$, and then $n_1q_1', n_2q_2', \ldots, n_\nu q_\nu'$. The σ^2 of their sum, by (27), would be $n_1q_1' + n_2q_2' + \ldots + n_\nu q_\nu'$. But $n_y q_y' = d_y'$, where d_y' represents the observed deaths of sample y, so that $\Sigma n_y q_y'$ is the total of all the observed deaths. The approximation is thus reached that σ^2 for the deaths in such a series of mortality statistics might be taken as merely the total of the actual deaths—a conclusion which the student will often find stated in the form $\sigma \doteq \sqrt{\text{Actual Deaths}}$.

The fact that this approximation must be applied with circumspection can hardly require emphasis. From its nature it evidently results from several assumptions which may, under particular circumstances, be far from close approximations. For, like the original npq of (7), its basis is questionable if q is so small (i.e., p near enough to 1) that nq is small with a large n; the substitution of the individual sample values $p_1', p_2', \ldots, p_\nu'$ (and $q_1', q_2', \ldots, q_\nu'$) for the true p (and q) may be unjustifiable if $n_1, n_2, \ldots, n_\nu$ are not large; the assumption that $p_1' = p_2' = \ldots = p_\nu'$ must be examined carefully; and the further supposition that each p_y' (or even $\sqrt{p_y'}$) can be taken as 1 also imagines q_y' to be so small that again the validity of the whole basis may be questioned unless each $n_y q_y'$ is perhaps 10 or over when n is not also small. These are the theoretical reservations; in any practical case they should be given full consideration, and the complete formula resulting from the use of a "true" p (and q) for each $n_1, n_2, \ldots, n_\nu$ should be tested against the $\sigma \doteq \sqrt{\text{actual deaths}}$ approximation before the latter should be adopted.

C; 11. Applications of the Lexis Theory

Experimental Verifications

Examples of actual drawings of cards, and balls from urns, in conformity with these Bernoulli, Poisson, and Lexis methods of sampling are given in P:*36*:136-145 and P:*114*:84-7, which show close agreement between the results derived from the actual experiments and the theoretical values by the formulae.

Hypothetical Illustration of the Three Types

The following very simple hypothetical example, constructed by Forsyth (P:*44*:195), illustrates the practical application of the method. Suppose that the numbers of deaths in 10 districts in each of three countries (the populations of each district being assumed to be 1,000), are as given below:

Actual Deaths per 1,000 Population (D_r) *in 10 Districts* (r) *in Three Countries* (A, B, *and* C)

District (r)	*Country* (A)	*Country* (B)	*Country* (C)
1	17	18	24
2	16	17	22
3	11	10	20
4	12	11	19
5	13	9	18
6	14	12	10
7	15	19	9
8	16	16	8
9	14	10	6
10	12	18	4
Totals	140	140	140

The average number of deaths in all three countries is 14 per district. Assuming for the moment that we may here use the observed instead of the unknown true values (which strictly should be used in conformity with the analyses in B; 9) we therefore have $np = 14$; $p = .014$; $q = .986$; and $\sigma_B = \sqrt{14(.986)} = 3.72$, which would be the standard deviation if the probability of death were constant throughout.

The actual values of σ, however, are now computed directly from the data. For Country (C), for example, we have

$\lvert D_r - np \rvert$	$\lvert D_r - np \rvert^2$
10	100
8	64
6	36
5	25
4	16
4	16
5	25
6	36
8	64
10	100
	482, whence $\sigma^2 = \frac{482}{10} = 48.2$,
	or $\sigma = 6.94$.

We thus find by actual calculation from the data that

For Country (A), $\sigma = 1.90$, and $L = \frac{1.90}{3.72} = .51$

For Country (B), $\sigma = 3.74$, and $L = \frac{3.74}{3.72} = 1.01$

For Country (C), $\sigma = 6.94$, and $L = \frac{6.94}{3.72} = 1.87$

Remembering that here $\sigma_B = 3.72$, and that an actual value less than σ_B indicates variation of the Poisson type within the sample, while a greater value denotes variation of the Lexis type from sample to sample, these results suggest the conclusion, on the Lexis theory, that greater variation occurred within each district of Country (A) than from district to district; that variation was about the same throughout (B); and that there was greater variation from district to district in (C) than within each district.

This example was originally given, and is here reproduced, only as a simple illustration of the Lexis method of approach. Students should note carefully, however, that the observed data are used therein to give estimates of the "true" values which strictly are required by the analysis of B; 9; rigorously, therefore, Bessel's correction (42) should be introduced in such cases by multiplying (41) and its results by $\frac{\nu-1}{\nu}$, as indicated in P:*63*:508, or by adopting an even more refined adjustment due to Tschuprow which is demonstrated and illustrated in P:*146*:216-222.

(a) *Practical Illustration of Bernoulli Type*

An application of the Lexis theory under conditions which were found to be Bernoullian has been given by the Russian Jastremsky (P:*66*) in connection with the Austro-Hungarian mortality investigation of 1876-1900 (P:*34*:135). If $q_{[x-t]+t}$ denotes the "select" rate of mortality at attained age x for the year of duration t since entry, and $q_x^{(5)}$ the "ultimate" rate at attained age x based on the data of 5 or more years of duration since entry, the ratio $\frac{q_{[x-t]+t}}{q_x^{(5)}}$ for values of t from 0 to 4 will indicate the effect of duration upon the rate of mortality at attained age x—being <1 if the effects of "selection" at entry are still apparent, and equalling 1 if the mortality is no longer dependent on duration. If now, for each value of t from 0 to 4, this ratio is computed for the various values of the attained age x, five series are obtained (one for each t from 0 to 4), each of which will be constant as x varies if the ratio is independent of x, but will vary as x varies if the ratio depends on x. In order to examine whether the series of ratios (for, say, any one of the values of t) could be treated as being independent of x, Jastremsky in effect computed σ (about its average value for all ages) from the actual series of ratios, and also the Bernoullian σ_B on the hypothesis of independence, so that the Lexis ratio $L=\frac{\sigma}{\sigma_B}$ would test the admissibility of that hypothesis. For endowment assurances, for examples, the values of L for $t=0$, 1, 2, 3, and 4 were found to be 1.01, .96, 1.05, .98, and .91, and were thus all close to 1. The inference to be drawn therefore was that the Bernoullian hypothesis was plausible for each value of t, so that the observed fluctuations in the ratios by age could be attributed to chance alone, and $\frac{q_{[x-t]+t}}{q_x^{(5)}}$ for each value of t might reasonably be treated as being independent of the attained age x (see P:*36*:166).

(b) Practical Illustration of Poisson Type

The Poisson type of variation from group to group within the "universe" is encountered directly in dealing with mortality statistics. Yule and Kendall, for example (P:*177*:366) consider the following as an illustration: In a population of n persons, all of one sex and one age, with a rate of mortality of 12 per 1,000 (being .012) throughout—thus conforming with the Bernoullian conditions of simple sampling—σ_B in the death rate, being $\sqrt{\frac{pq}{n}}$ by (36), is $\sqrt{\frac{(.012)(.988)}{n}} = \frac{108.9}{1000\sqrt{n}}$. If, however, a population (still composed of one sex only) had the same average (crude) rate of mortality (.012) throughout, but now were to include various age groups with different rates of mortality, such as about .064 in infancy, decreasing to about .0025 in childhood, and thence continuously increasing until old age, the standard deviation, σ_p, of the death rate about its mean within such a population would be about .024; hence, from (37), σ_P, the Poisson type standard deviation, would be

$$\frac{1}{\sqrt{n}}\sqrt{(.012)(.988)-(.024)^2} = \frac{106}{1000\sqrt{n}}.$$

The small difference between this value and the standard deviation of simple sampling, namely, $\frac{108.9}{1000\sqrt{n}}$, consequently indicates that the effect of the variation among the individuals within the population of a country under such practical mortality conditions is not likely to be serious.

Similar calculations may be made to test the importance of non-homogeneity in any group of lives to which a single mortality rate is applied, or from which such a mortality rate is computed. It is interesting to note that non-homogeneity actually decreases the standard deviation in comparison with that of a homogeneous group.

(c) *Practical Illustration of Lexis Type*

Since usually the populations will not really all be equal (as was assumed in the demonstration of the formulae), it is desirable in practice to employ a technique which will give due weight to such variations in size. A mathematical examination of the corrections theoretically necessary, with a numerical example, has been given by Arne Fisher in P:*36*:157-160. For practical purposes in connection with mortality data, however, it will generally be sufficient to use the groups of varying size as if they were all of the same size, except that due allowance should be given for the varying size in determining the basic rates of mortality involved. The point may be illustrated from the following example given by Rietz (P:*114*:89) for the rate of mortality in the first year of age in nine states of the United States:

State Number	Exposed to Risk	Death Rate per 1000
1	50,707	70
2	33,370	85
3	57,915	78
4	35,392	68
5	53,658	77
6	51,452	66
7	51,832	74
8	41,656	78
9	54,472	79
	430,454	

The simple arithmetic mean of the nine death rates is 75 per 1,000. If they had all arisen from groups of equal size their σ^2, computed directly by using only column (3) above, would be

$$\frac{1}{9}[(70-75)^2+(85-75)^2+\ldots+(79-75)^2]=32.67, \text{ whence } \sigma=5.72.$$

One obvious interpretation of the assumption that they had arisen from groups of equal size would be to suppose that in each state the same death rates per 1,000 would have been shown if in fact the populations had all been equal—in which case the equal populations could clearly be taken as merely the average $\frac{430,454}{9}=47,828$. Under these circumstances the Bernoullian σ_B for the rate of mortality, being $\sqrt{\frac{pq}{n}}$ by (36), would have been $\sqrt{\frac{(.075)(.925)}{47828}}=.00120$ per person $=1.20$ per 1,000. The Lexis Ratio is therefore $\frac{5.72}{1.20}=4.77$, and Charlier's Coefficient of Disturbancy is $\frac{100\sqrt{(5.72)^2-(1.20)^2}}{75}=7.45$. The data are thus shown to be "hypernormal", of the Lexis type—indicating that the rates of mortality vary significantly from state to state.

It is clearly preferable, however, to give due weight to the varying sizes of the populations, which may be done as follows. The weighted average death rate per 1,000, which allows properly for the variations in size, is

$$\frac{70(50707)+85(33370)+\dots+79(54472)}{430454}=74.864.$$

We now compute, from the data, the σ^2 for the rate with reference to this mean instead of to the simple average 75 previously used—which, by the rules given at p. 254; B; 27 is

$$\frac{50707(70-75)^2+33370(85-75)^2+\ .\ .\ .\ +54472(79-75)^2}{430454}\text{ less}$$

$(75-74.864)^2$, whence $\sigma=5.40$. But if the conditions had been Bernoullian, groups of equal size, viz., 47828 for each state as before, would have shown the weighted average death rate of 74.864; under such circumstances σ_B would have been

$$\sqrt{\frac{(.074864)(.925136)}{47828}}=1.203\text{ per }1,000.$$

The Lexis Ratio is consequently $\frac{5.40}{1.203}=4.49$, and Charlier's Coefficient of Disturbancy is again positive. The significant variation from state to state, which was found by the first simpler method, is confirmed.

In these examples, again (as noted on p. 295), a correction of Bessel's type should strictly be introduced—cf. the similar illustrations, with Bessel's correction, in H:*163*(English translation): 217-220.

Many other examples of the analysis of birth, death, and marriage rates for various localities or periods by this procedure are to be found in the literature (P:*116*:153; P:*36*:151-165; and P:*27*:320 and 330). The actuary, however, with an innate distrust of supposed homogeneity, will of course in practice examine closely every such series not only by locality and period, but also according to age, sex, occupation, and any other characteristics which may have influenced the data. He will not defer detailed analysis until he is warned by the theory of the Lexis ratio. That theory, nevertheless, while thus of limited applicability in the actuary's practicing equipment, is fundamentally important in the development of the underlying theories of Mathematical Statistics.

C; 12. The Standard Error of the Mean

(1) From the results of Weldon's experiment in casting 5 or 6 in 26,306 throws of dice, as tabulated at p. 334; C; 25, it may easily be calculated that the mean of the distribution on the hypothesis of unbiassed dice is 4, and for the observed series is 4.0524. The standard deviation computed from the observed frequencies is $\sqrt{2.69826}$. The standard error of the mean (which is based on the 26,306 values) is therefore

$$\frac{\sqrt{2.69826}}{\sqrt{26306}} = .01013\,.$$

This result may now be used, instead of the χ^2 method of C; 25, to test the admissibility of the hypothesis that the dice were unbiassed. For the deviation between the hypothetical and observed means is .0524; this, however, is over 5 times the standard error of the mean (.01013); and as it consequently exceeds the limit of twice the standard error which is generally adopted (in conformity with the principles of Chapter III) as a reasonable test, and even exceeds 3 times the standard error which is usually regarded as a conclusive test, the inference to be drawn is that the observed values are not compatible with the hypothesis that the dice were unbiassed.

(2) The use of the rule that, on the assumption of normality, ± 3 times the standard error will define the limits within which a sample value of a parameter may be expected to lie may be illustrated by considering a sample distribution of the heights of 1,000 men, with an arithmetic mean of 5′7″ and a standard deviation (computed from the sample) of 2″, from which the standard error of the mean is seen to be $\frac{2}{\sqrt{1000}} = .063$. Then it is practically certain that the true mean will lie within the range $67 \pm 3(.063)$ inches, i.e., between 67.19″ and 66.81″.

(3) As another illustration Elderton suggests, in P:*32*:191, that the formula for the standard error of the mean could be applied similarly to examine the average profit from various classes of business for a series of years, and thence to determine whether some particular average profit should be attributed to chance alone. The standard error of the profits in the various years would be computed as the standard deviation of the observed series, divided by the square root of the number of years.

C; 13. Illustrations of "Student's" Distribution

(1) The following example of the application of formula (45) is given by Deming and Birge in P:*29*:138. Suppose that 10 equally reliable readings on a micrometer (so taken that the assumptions of a normal universe and randomness are satisfied) show the values 1.078, 1.080, 1.071, 1.076, 1.081, 1.077, 1.075, 1.073, 1.079, and 1.070, with mean $\bar{x} = 1.0760$ and standard deviation $\sigma_s = .00355$; examine the hypothesis that the true mean, m, of the parent population from which the sample was drawn is 1.0740. Here $\bar{x} - m = .002$, and "Student's" $z = \frac{\bar{x} - m}{\sigma_s} = .563$;

$n=10$; and P_z in (45) (which is read easily from Nekrassoff's nomograph) is .13. That is to say, in about 1 in 8 samples of 10 we should expect to find z in absolute value as large as or larger than .563 (and in about 1 in 16 samples z would be as large as or larger than +.563). On the assumption that $\sigma_s=.00355$ is not unusual, the inference to be drawn would therefore be that there is no strong reason to reject the hypothesis that the population mean, m, is 1.0740.

As pointed out in the text on p. 46 here, it is important to realize that the validity of this conclusion depends on the supposition that the value of σ_s, namely, .00355, is not unusual. This reservation is overlooked so frequently, and it bears so vitally upon the inferences to be deduced, that space may be taken here to illustrate its meaning further, with extracts from the admirable presentation by Deming and Birge in P:*29*:138-9.

Let us recall, then, that in the above example nothing whatsoever is stated to be known with respect to the σ of the universe from which the sample of 10 readings came; the position therefore is that "without some knowledge concerning σ the only thing we can do is to postulate that the sample was not extraordinary" (loc. cit., 138), and proceed as already shown. If we knew, however, from previous comparable observations, that σ may be supposed to be very close to .0040, for example, it will be apparent that the reasonableness of $\sigma_s=.00355$ in comparison with $\sigma \doteq .0040$ can be examined if we have the law of distribution of σ_s. This distribution is available in (*e*) at p. 225; B; 13; and for samples of 10 from a normal universe with $\sigma=.0040$, the average standard deviation can be placed at $.0040\times.9227=.0037$ from Table I in P:*29*:128. This is evidently so close to the observed $\sigma_s=.00355$ that the latter can be accepted as being not unusual. With such knowledge that σ_s is not unusual, the "Student" test, as shown here in (1), obviously could be applied with confidence. Under these circumstances, however, the classical normal theory can also be employed, for σ is known closely; in this example, for instance, the normal theory would use $\dfrac{\bar{x}-m}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}=\dfrac{.002}{\left(\dfrac{.004}{\sqrt{10}}\right)}=1.58$, and from the chart in P:*29*:134 it is seen immediately that the probability here of drawing a sample of 10 with an absolute difference in the mean as great as, or greater than, the postulated .002 is .114—"which means that there is about 1 chance in 9 that $|\bar{x}-m| \geqq .002$, or that there is about 1 chance in 18 that $(\bar{x}-m) \geqq +.002$" (loc. cit., 139). This test and the "Student" test (as previously applied in (1) here) "therefore concur, as they will when σ_s is not extraordinary" (loc. cit.).

To see in another way how dependent the "Student" test is on σ_s being not unusual, suppose now that in the universe $\sigma=.0025$, so that the observed $\sigma_s=.00355$ seems to be unusually large (for, as shown in P:*29*:139 from the distribution of σ_s again, in samples of 10 with $\sigma=.0025$ we should expect to get σ_s as large as or larger than .00355 only about 17 times in 1,000 trials). The "Student" test, with its ignorance of σ, still gives $P_z=.13$, as in (1) here, and so does not suggest rejecting the hypothesis that the population mean, m, is 1.0740. But with σ now known to be .0025, the normal theory is available, and a calculation similar to that in the preceding paragraph shows that the probability of then drawing a sample of 10 with an absolute difference in the mean as great as, or greater than, the postulated .002 is only the much lower value .0114—which certainly would suggest rejecting the hypothesis, in contradiction of the "Student" inference. This disagreement "shows how misleading the latter would be if used alone; the trouble comes, of course, from the fact that σ_s is now exceptionally high" (loc. cit.).

(2) The same type of investigation can be made with respect to the heights, weights, etc., of individuals (see, for instance, P:*177*:440) when we know that the universe may be supposed to be normal, that randomness exists, and that σ_s is not unusual.

(3) Another illustration frequently quoted (e.g., in P:*110*:583 and P:*43*:127) was given by "Student" himself in the following statement of additional hours of sleep induced in the same patients by the use of two different drugs:

Patient	Drug No. 1	Drug No. 2	Difference (No. 2 − No. 1)
1	+0.7	+1.9	+1.2
2	−1.6	+0.8	+2.4
3	−0.2	+1.1	+1.3
4	−1.2	+0.1	+1.3
5	−0.1	−0.1	0.0
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4
Mean ($\bar{x}$)	+0.75	+2.33	+1.58
Standard Deviation (σ_s)	1.70	1.90	1.17

Here again let it be assumed that the necessary conditions are satisfied regarding normality of the universe, randomness of the samples, and the values of σ_s being not unusual. Then:

(*a*) If it were desired to examine the probability that drug No. 1 will cause an increase of sleep, it would be necessary to take $\bar{x}-m$, being the deviation of the mean of the sample from the specified mean of the universe, as $.75-0$; then

$$z=\frac{\bar{x}-m}{\sigma_s}=\frac{.75-0}{1.70}=.44;\ t=z\sqrt{n-1}=.44(3)=1.32;$$

and from the tables the corresponding probability is .888.

(*b*) To obtain the probability that drug No. 2 is more effective than No. 1 we similarly take $z=\dfrac{1.58-0}{1.17}=1.35$, whence $t=(1.35)(3)=4.05$, and the probability is .999. Otherwise, from the tables in Fisher's form, it is seen that for $n-1=9$ only 1 value in 100 will exceed 3.250 by chance (P:*43*:127 and 177). On either reading, therefore, the probability from the tables of the observed or a larger positive difference appearing by chance alone is very small, so that the difference between the results of the two drugs is undoubtedly significant.

With regard to the practical utility of these (and similar) examples, it must of course be realized that a number of additional and comparable experiments would be necessary, in order to test the stability of the indications, before the inference suggested by the single experiment could properly be made the basis for any future action (here the future prescription of either drug).

(4) Example 3(*b*) just given exhibits the effects of two different actions (i.e., the administration of drugs No. 1 and No. 2) upon the same $n(=10)$ individuals, and, from the evidence so tabulated, tests the significance of the difference (+1.58) between the two means. In the same way we could take two comparable sets of individuals, with $n(=10$, say) different persons in each set, then tabulate the effects of drug No. 1 upon set No. 1 and of drug No. 2 upon set No. 2, calculate therefrom $\bar{x}$ and σ_s for the differences so shown by the two samples (as in the last column of the table in example (3) above), and proceed as in example 3(*b*). Illustrations of this method may be found conveniently in P:*177*:441 (example 23.2), and in the first calculation at P:*43*:133.

(5) As pointed out in the text, however, the use of different individuals in the two sets will obviously decrease the reliability of the results obtained therefrom as in (4); some advantage may consequently then be gained by using R. A. Fisher's extension of the "Student" principle for the case of two independent sets (cf. P:*43*:132-3). Furthermore, if there is no correspondence at all between the members of the two sets, Fisher's extension for the case of two independent sets will clearly be the only applicable method.

As an illustration, take again the data of example (3), and suppose that the results of drugs No. 1 and No. 2 were secured from two sets of patients so distinct as to require the assumption of complete independence. Then (as shown in P:*43*:130, ex. 20) $\bar{x}_1-\bar{x}_2=+1.58$ as before; but now we calculate $\sigma_e^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)=.7210$ by taking $\sigma_e^2=\frac{n_{1\,1}\sigma_s^2+n_{2\,2}\sigma_s^2}{n_1+n_2-2}$ with $n_1=10$,

$n_2 = 10$, ${}_1\sigma_s = 1.90$, and ${}_2\sigma_s = 1.70$; hence $t = \dfrac{1.58}{\sqrt{.721}} = +1.861$; and entering the table of t for $d = 10 + 10 - 2 = 18$ we find that the probability lies between .1 and .05, which cannot be viewed as significant. This conclusion, based on the supposition that the sets of patients were quite distinct, affords an interesting contrast with the conclusion of significance found by the method of example 3(*b*) on the supposition that the patients were identical. As Fisher points out (P:*43*:131), it provides a good illustration of "the value of design in small scale experiments, and that the efficacy of such design is capable of statistical measurement".

Another similar example may be found in P:*43*:133. In P:*118*:153-4 two comparisons are also shown between the results of the method here discussed and the classical normal theory.

(6) A simple illustration of the use of Fisher's z-distribution (47z) in testing the difference between two sample variances, ${}_1\sigma_s$ and ${}_2\sigma_s$, is given by Rider in P:*112*:118, where ${}_1\sigma_s^2 = 9.6$ for a sample of $n_1 = 7$ variates is to be compared with ${}_2\sigma_s^2 = 4.8$ for $n_2 = 9$. To calculate $z = \log_e\left(\dfrac{{}_1\sigma_e}{{}_2\sigma_e}\right)$ we remember that ${}_1\sigma_e^2 = \left(\dfrac{n_1}{n_1 - 1}\right){}_1\sigma_s^2 = \left(\dfrac{7}{6}\right)9.6 = 11.2$, and that similarly ${}_2\sigma_e^2 = \left(\dfrac{9}{8}\right)4.8 = 5.4$, so that $z = \frac{1}{2}\log_e\left(\dfrac{11.2}{5.4}\right) = .3648$. From Fisher's tables (P:*43*:251) we see that, since the degrees of freedom are $d_1 = 6$ and $d_2 = 8$, the "5% point" is .6378; the value .3648 for z is thus well inside that point; and the inference to be drawn therefore is that, on that test (and subject to the reservation concerning ${}_1\sigma_s^2$ and ${}_2\sigma_s^2$ being not unusual), ${}_1\sigma_s^2$ would not be regarded as significantly greater than ${}_2\sigma_s^2$.

C; 14. The Applicability of the Poisson Exponential

The striking ability of the Poisson function $y_r \doteq \dfrac{m^r e^{-m}}{r!}$ to depict markedly skew point binomials will be seen from the following examples (given in P:*36*:267, in P:*117*:299, and P:*36*:267 respectively), where the symmetrical Normal Curve $y_x = \frac{1}{c\sqrt{\pi}} e^{-\frac{x^2}{c^2}}$ would clearly give a very poor representation.

$10{,}000(.001+.999)^{100}$, i.e., $q = .001$ and $nq = .1$

Number of Occurrences	Point Binomial	Poisson Exponential
0	9048	9048
1	906	905
2	45	45
3	1	2
4	0	0

$10{,}000(.008+.992)^{250}$, i.e., $q = .008$ and $nq = 2$

Number of Occurrences	Point Binomial	Poisson Exponential
0	1343	1353
1	2707	2707
2	2718	2707
3	1812	1805
4	902	902
5	358	361
6	118	120
7	33	34
8	8	9
9	2	2
10	0	0

10,000(.05+.95)[100], i.e., $q=.05$ and $nq=5$

Number of Occurrences	Point Binomial	Poisson Exponential
0	59	67
1	312	337
2	812	842
3	1396	1404
4	1781	1755
5	1800	1755
6	1500	1462
7	1060	1044
8	649	653
9	349	363
10	167	181
11	72	82
12	28	34
13	3	13
14	1	5

It will interest actuaries to realize that the importance in practical statistics of the Poisson function $y_r \doteq \frac{m^r e^{-m}}{r!}$ for the probability of r occurrences of a rare event in n trials, where m is computed as the mean, was first illustrated by Bortkiewicz from the deaths in the Prussian army from an unusual cause, namely, the kicks of horses (see p. 166; A; 11). The data covered 10 army corps for 20 years, i.e., 200 observations, and from 122 deaths gave .61 as m, from which (by tables of the Poisson function—see p. 234; B; 15) the comparison between the observed and calculated values was remarkably close, as follows:

Number of Deaths per Annum	Observed Frequency of Occurrence	Theoretical Poisson Frequency
0	109	108.67
1	65	66.29
2	22	20.22
3	3	4.11
4	1	.63
5	0	.08
6	0	.01

The preceding illustrations demonstrate the closeness of the approximation to the ordinates themselves in certain types of cases. In considering, however, whether the Poisson function is to be preferred to the Normal Curve as a means of examining the probability of deviations within a certain range or exceeding a given amount, it will be well here to examine further some examples first given by H. L. Rietz (P:*117*:299 and P:*115*).

For the point binomial $(.008+.992)^{250}$ already noted—which might represent a group of 250 persons of the same age for each of whom the rate of mortality is .008—we have $q=.008$, $nq=2$, and $\sigma=\sqrt{npq}=1.4085$.

If the group of persons were annuitants, a number of deaths smaller than the expected 2 would be unfavourable. Using the binomial and Poisson values previously calculated, and from tables of the "probability integral" (see p. 161; A; 5), the probabilities of deviations not exceeding 2 in defect (i.e., of only 1 or 0 deaths instead of 2) are

For the point binomial.........	.4050
By the Poisson exponential.....	.4060
By the normal curve (area).....	.3233

The result by the normal curve is therefore about 20% too low.

On the other hand, if an insurance experience were under consideration, for which an excess of deaths would constitute an unfavourable event, the probabilities of deviations not exceeding 2 in excess are

For the point binomial.........	.2714
By the Poisson exponential.....	.2707
By the normal curve (area).....	.3233

Here the normal curve gives a result about 19% too high.

When, however, we examine the probabilities in relation to the criterion suggested for the normal curve that a deviation of 3σ or more is very improbable, the skewness of the distribution again has a marked effect. Since $3\sigma=4.2$, and a deviation of $+3\sigma$ would therefore imply 6.2 deaths, we may examine the probabilities of 7 or more deaths actually occurring, thus:

For the point binomial.........	.0043
By the Poisson exponential.....	.0045
By the normal curve (area).....	.0007

In this case the normal value is about 84% too low—a serious discrepancy on the wrong side.

If, therefore, such results—for small q (or p) and nq (or np)—were to be used as a basis for determining the risk of unfavourable deviations from the expected mortality, it would be essential to apply the normal theory with great care, and with particular reference to the degree of skewness and to the question whether positive or negative deviations of given extent would constitute, in practice, an unfavourable outlook (cf. P:*117*:299-303).

Although it thus becomes clear that the Poisson exponential may well, in certain cases, give much more reliable conclusions than the normal theory—especially when the investigations only involve either positive or negative deviations—it must be noted that the advantages of Poisson's formula are not usually so marked when both positive and negative deviations are examined together, as is often sufficient in the practical consideration of sampling errors.

As an example, in the somewhat extreme case just employed, the probabilities of deviations not exceeding ± 2, taken together, would be

For the point binomial	.6764
By the Poisson exponential	.6767
By the normal curve (area)	.6466

Here the normal curve gives an indication only about 4% too low, in comparison with 20% and 19% respectively for the negative and positive deviations.

Another similar illustration is shown by Rietz in P:*115*, which again indicates that the normal curve gives a reasonable representation when positive and negative deviations are considered together, even though the approximation is not good when one side (either positive or negative) is taken alone.

It will accordingly be realized from these examples that the applicability of the "normal" theory should be tested in any practical case where q (or p) is small but n large enough that nq (or np) is about 10 or less (cf. the diagram on p. 267; C; 4).

C; 15. The Practical Applicability of Edgeworth's Generalized Law of Error

The ability of Edgeworth's curves to represent distributions which are not very skew has been amply demonstrated by examples which may be found in his original papers, in H:*107*:329-334 (where in (52) $c=1.683$ and $j=.0728$), in H:*87*:39 (where $c=3.623$ and $j=.06$), and in P:*32*:134-7. They will generally be applicable to the same types of data which can be graduated by the Gram-Charlier Type A series. Under conditions of marked skewness, however, Pearson's system or the Poisson-Charlier Type B curve may be expected to give more satisfactory graduations.

It should be noted that theoretically $j=\frac{\mu_3}{c^3}$, so that the fitting might be performed by the method of moments (see Chapter VIII, and H:*107*:330-4 for a numerical example). In practice, however, that method does not always give good results (see H:*87*:40). Edgeworth consequently devoted much effort to the development of alternatives—particularly (*i*) a "method of percentiles", (*ii*) a process based on the condition that the values of the constants should minimize the improbability of the observations arising at random from a curve of the given form, and (*iii*) his "method of translation"—of which an account is given in H:*162*:55-78 and 85.

C; 16. Applications of the Gram-Charlier Type A and Poisson-Charlier Type B Series

An example of using successively one, two, and three terms of Type A is given in P:*114*:117-8. Several graduations by the Type A and the Type B series are discussed by Elderton in P:*32*:134-140, together with valuable comparisons of the results which can be obtained alternatively by Edgeworth's or Pearson's systems. A very detailed examination of the practical utility of both Types A and B, and the technique of fitting, is available in P:*36*:215 et seq.

Although Type A certainly can be made to assume a very skew shape—as is shown at P:*36*:226-232 for a distribution of six values only—it would appear to be generally conceded that it is not of great practical service in such cases, and that then either the method of logarithmic transformation with Type A, or the adoption of Type B, is likely to be preferable (see especially P:*36*:235-260, and P:*32*:131-140).

An objection which has been levelled against Type A is that the series expansion sometimes gives rise to negative frequencies in dealing with fairly skew distributions (P:*32*:140, and P:*140*:48). The advocates of the Gram-Poisson-Charlier school, however, reply that in reality this is a matter of little practical importance, because in any event the observations at the ends of the distribution are very small (P:*36*:217).

C; 17. The Practical Use of Pearson's Frequency Curves in Actuarial Work

So far at least as readers of English are concerned, the Pearsonian system of frequency curves far overshadows Edgeworth's and the Scandinavian methods as an essentially practical means of representing statistical frequency distributions which occur in practice. Not unnaturally, controversy has at times surrounded the extensive applications which have been made by Pearson and his followers, and some supporters of the Scandinavian school have advanced the claim that the Gram-Charlier or Poisson-Charlier series will deal more adequately with difficult cases (cf. P:*36*:184-5, 216, and 232-4). The comparatively simple manner, however, in which the appropriate Pearson curve can first be selected by the use of the "criterion" (68), and then fitted by the "method of moments" (see p. 97 here), has led to the wide acceptance of the system in many fields, and in actuarial work has stimulated several interesting developments which are of particular importance to the actuarial student.

The Direct Representation of Actuarial Data by Pearson's Curves

The Type I curve is useful as a means of representing a wide variety of actuarial distributions. Being related to the skew

www.dbraulibrary.org.in

point binomial (and so being, in fact, capable of expressing the terms of the point binomial very closely even when n is as small as 6—see P:*51*:46), the curve in its skew bell-shaped form often resembles the numbers exposed to risk in a mortality or similar experience (see H:*79*:Diagram 1, P:*51*:47 and 55-57, and P:*32*:62). In its bell-shaped form it may also be employed to describe the actual deaths (see H:*79*:Diagram 1, and H:*89*) or the "entrants" (see H:*45*:diagram, and P:*51*:6 and 47) in a mortality experience, and such diverse functions as the number of marriages and the rates of marriage, or disability retirements, the average number of children, and the cost of their pensions, in a pension fund (P:*51*:47), or the distributions of sums assured or premiums by age groups (H:*181*:35). It was also used extensively in H:*122*:322 to represent the age distributions of the populations of India. A numerical example of the J-shaped form may be found in P:*32*:125-6, while the U-shaped curve is noted in P:*32*:112, and the twisted J-shape in P:*32*:111.

The second main type, No. IV, is not particularly useful in actuarial work, since it is unlimited in both directions, and the diminishing rate at which the values decrease at the ends is not encountered often except in dealing with a function like the "rate of withdrawal". Numerical examples are shown in P:*32*:68 and 137, and P:*20*:7-11, and its applicability for representing the distributions of sums assured or premiums is indicated in H:*181*:35. The work involved in calculating the ordinates, which previously had always made Type IV the most troublesome of all the Pearson curves to fit, has now been greatly facilitated by the publication of the tables noted at P:*20*.

The skew bell-shaped form of No. VI—the third main type—has been illustrated in P:*32*:77 as a representation of the number of "entrants" in a mortality experience, and is suggested in H:*181*:35 for sums assured or premiums by age. In general, however, its utility is reduced by the fact that its range is limited in one direction only (cf. P:*51*:49).

Coming now to the Transition Types, the Normal Curve has been used very widely, of course, in much of the basic theory with which this study deals. Its utility (despite its theoretically unlimited range) as an approximation to the terms of the symmetrical point binomial is illustrated numerically in P:*51*:44, and its fitting to certain actuarial data (sums assured with bonuses, and reserves, grouped by office years of birth) by the method of moments is shown in P:*32*:81. It may be employed also to represent approximately a hypothetical (though not the actual) exposed to risk in the special method of determining the Makeham constant c which is described on p. 318 here. Another ingenious method is illustrated by Sir G. F. Hardy in P:*51*:91-98 (see also H:*122*:389), which uses the Normal Curve in the representation of an asymmetrical series, consisting of only a few groups, by taking the proportionate numbers of exposed to risk, deaths, living, etc., above age t as $\frac{1}{\sqrt{\pi}}\int_{-\infty}^{z} e^{-t^2}\,dt$, and then adjusting the values of z on the assumption of constant third or fifth differences.

The transition Type II in its bell-shaped form, and with its advantage (unlike the Normal Curve) of being limited in both directions, is also closely related to the symmetrical point binomial and to the Normal Curve, as may be seen from the examples in P:*32*:87-89 and 134, and from P:*51*:42-43 and 44. Being of limited range, its values naturally decrease at the ends more rapidly than those of the Normal Curve. Its utility for representing the distributions of sums assured or premiums is shown in H:*181*:35. It should be noted that it assumes a symmetrical U-shape when m is negative.

The transition Type VII, being again symmetrical, is of rather limited application to actuarial data, though again it is suggested for the distributions of sums assured or premiums in H:*181*:35. The curve differs from Type II in that its values decrease at the ends more slowly than the Normal Curve.

The utility of the skew transition Type III is often limited by its tendency to run into a geometrical progression. An example of its application is given in P:*32*:93. It may be useful, however, for the hypothetical representation of the exposed to risk in the indirect method of graduating rates of mortality, marriage, etc., which is discussed below.

The skew transition Type V also has a restricted applicability to actuarial statistics for the same reason as that noted for Type IV, namely, the diminishing rate at which the values decrease at the ends.

The remaining Types VIII to XII are not bell-shaped, and might sometimes be useful, as indicated by Elderton in P:*32*: 104-112, if it should be desired to represent such data as maturities of endowment assurances (Type VIII), the exposed to risk by duration (Type IX), withdrawals (Type XI), or withdrawals in select tables (Type XII).

The Compounding of Frequency Curves for Actuarial Data

It may be of interest to note that some cases arise in practice which obviously cannot be dealt with by any single frequency curve, but evidently might be represented by the addition of several such curves. The actual deaths in certain experiences, or the values of d_x in the hypothetical community of a "life table", for example, when tabulated for all ages from infancy to old age, might begin at a high figure (say 100,000 for ages 0-5), decrease sharply during the early years (to about 6,000 for ages 10-15), then increase (steadily, or with a point of inflexion at about age 25 or 30) to a maximum (of perhaps 70,000 for ages 73-78), and finally decline gradually to zero at about age 100. In such a distribution there would be at least two prominent maxima (in infancy and at ages 73-78), with perhaps a third slightly marked at the point of inflexion about age 25 or 30. A J-shaped curve therefore might be used for the sharply descending portion in the earliest years, with a skew bell-shaped type thereafter, and perhaps a third small bell-shaped curve (symmetrical or skew) to provide for the point of inflexion. This method of "compounding" frequency curves is illustrated, for the case of the English Life Table No. 4, by Karl Pearson's analysis of the life-table function $l_x\mu_x$ (i.e., the numbers dying per annum at the moment of attaining age x) into five superimposed frequency curves representing old age, middle life, youth, childhood, and infancy respectively (see H:*74*, H:*77*, and P:*102*:49).

Another example of compounding (though of a rather different kind) is to be found in Howell's addition of two Normal Curves to a Makeham graduation in H:*131*:198.

[The splitting of the distribution of actual deaths into a series of frequency curves has been advocated by Arne Fisher (H:*130*, and H:*137*) as part of a process for determining, from proportionate death ratios by causes of death, the rates of mortality of an experience without any information concerning the "exposed to risk" from which the deaths arose. A summary of the method is stated in P:*167*:86 (footnote), where it is pointed out, however, that the procedure is necessarily unsafe.]

The Indirect Graduation of Rates of Mortality, Sickness, Marriage, Withdrawal, etc., by Pearson's Frequency Curves

Rates of mortality by age, from infancy to the limit of life, usually take the form of a contorted U-shaped curve (see Figure 19, p. 82), and cannot ordinarily be represented by a single frequency curve. Nor can sickness rates be so depicted. Rates of marriage according to age, on the other hand, and of withdrawal by age or duration, do assume a form much like the skew bell-shape of Type III. The fit resulting from a direct frequency-curve graduation of such rates is generally poor, however, mainly because weights are given to the values at the end of life equal to those assigned at the other ages where the "exposed to risk" are much more numerous, with the result that those end values exercise a disproportionate influence on the results (see, for example, columns (4) and (5) of P:*32*:117).

It is consequently advisable to employ a method which will avoid this difficulty by giving due weight to the varying magnitudes of the exposed to risk upon which the rates are based. Since in many cases both the exposed to risk and the deaths (or marriages, etc.) assume after the earliest ages the shape of a frequency curve like Type I, an obvious procedure would be to fit a Type I curve separately to each, and thence to take the ratios of the graduated deaths to the graduated exposed as the finally graduated rates of mortality. An example may be found in a paper by Elderton, H:*89* (see also H:*181*:4-5).

Such a process, however, suffers from the defect that the two separate graduations fail to make allowance for the fact that the fluctuations in the observed exposed to risk and deaths are not independent. The method can therefore be improved—as seems to have been suggested first by Calderon (H:*79*:164 et seq., and H:*89*:521)—by "adopting a formula for the deaths whereby the rate of mortality would be incorporated into the expression for the death curve, and connecting it with the exposed curve, so that the two were graduated together rather than separately". On this principle, accordingly, we should first represent the exposed to risk by a suitable frequency curve; then we compute the deaths (so that they will correspond with the frequency-curve representation of the exposed) by multiplying the graduated exposed by the original rates of mortality; next graduate these recomputed deaths by a frequency curve; and lastly take the ratios of these graduated deaths to the graduated exposed as the finally graduated rates of mortality. Since this representation of the exposed by a frequency curve is intended to provide, in effect, only a series of approximate weights, it will be sufficient to adopt merely a hypothetical (rather than a fitted) curve for that purpose, such as even a Normal Curve for which the ordinates are readily available.

The first published experiments with this simplified technique were given by Elderton in H:*83*, and two examples are shown in P:*32*:116 and 118 for the graduation of marriage rates and death rates—the recomputed marriages being graduated by Type III, and the recomputed deaths by Type I.

Illustrations are also available in H:*181* of using colog p_x or ${}_np_x$ instead of the rate of mortality q_x, and more elaborate discussions of the underlying theories may be found in H:*148* and H:*181*. In H:*137*:234 an illustration is also shown in which the exposed are represented by the Normal Curve and the recomputed deaths are graduated by the Gram-Charlier Type A series instead of by one of Pearson's curves. In H:*181*:33 the principle here under discussion is again applied with the Normal Curve for the exposed, but with the survivors (instead of the recomputed deaths), obtained by multiplying the hypothetical exposed by ${}_{10}p_x$ (instead of q_x), being graduated alternatively by the Gram-Charlier Type A series and Pearson's Type I curve for purposes of comparison.

The Determination of Makeham's Constant c by means of Frequency Curves

The principle just described is of special value in the calculation of the constant c in Makeham's formula for the force of mortality, namely, $\mu_x = A + Bc^x = A + Be^{x\lambda}$ where $\lambda = \log_e c$.

Calderon was the first (H:*79*:164 and 192) to publish the suggestion, as the result of work with G. F. Hardy on the graduations of the British Offices' Annuitants' Experience, 1863-93 (loc. cit., 169 and 191). His original idea in that paper was to use a point binomial for the exposed to risk, which should be arranged in not more than five or six broad groups (of at least 10 ages in each group) in order to give a reasonable fit (see P:*51*: 67 and 134). The expression for the "recomputed deaths", being then a point binomial $(\alpha+\beta)^n$ multiplied by the force of mortality, evidently assumes the form $(\alpha+\beta)^n(A+Bc^x)$; the moment relations are $\theta_0 = AE_0 + BE_0'$ for the total, and similarly for the first and second moments, where θ denotes the moments of the recomputed deaths, E those for the exposed, and E' those of the exposed multiplied by c^x. From the resulting expressions the value of c may be deduced easily, as shown in H:*79*:164-5 and P:*59*:88-90.

A development of Calderon's point binomial method was next employed by Elderton in H:*83* (see also H:*181*:4), and is now to be found conveniently at P:*32*:120-3. The improvement uses the Normal Curve as a hypothetical representation of the exposed, and is easily applied since the ordinates of that curve are readily available in tables such as P:*97*. The necessary formulae are demonstrated clearly in P:*32*:120. The excellent results of a numerical application to the $O^{NM(5)}$ mortality table are shown in P:*32*:122, and another illustration is given in P:*33*:256.

Pearson's Type III curve has likewise been suggested by Hardy for the hypothetical exposed, since in combination with $\mu_x = A + Be^{x\lambda}$ it also leads to a simple expression from which $\lambda(=\log_e c)$ can be found. The formulae are proved in P:*51*:134, and are stated again in P:*59*:87-88. The ordinates of the Type III curve are now available (cf. P:*33*:256, footnote) in P:*120*:63.

[It may further be noted here that Hardy has given, in P:*51*:135, the necessary formulae for this method when the exposed are represented by the form $y = kx^{\alpha}(1-x)^{\beta}$, where x represents a proportionate part of the range of the curve so that it varies between 0 and 1, as in (75) here.]

C; 18. The Principle of "Uniform Seniority"

The particular advantage of Makeham's formula (83) for actuarial purposes lies in its possession of the valuable property of "uniform seniority", by which an annuity on n joint lives of any ages may be computed by the substitution of the same number of lives, all of the same age, in accordance with the relation

$$\mu_{x+t} + \mu_{y+t} + \ldots = n\mu_{w+t} \text{ where } c^w = \frac{1}{n}(c^x + c^y + \ldots).$$

Makeham's second expression, (84), permits the use of a modified and less convenient uniform seniority method, involving a special rate of interest for the substituted annuity at equal ages (see H:*187*:535).

Hardy pointed out that formula (95) also preserves the principle in a modified form, and in P:*85*:413 and P:*94*:542 its application in the case of the modified double and triple geometric expressions (96) and (97) is examined.

The name "uniform seniority" arises from the fact that the calculation of the equivalent equal ages, w, is made, in effect, by an addition to the youngest age. It has been observed, however (P:*85*:413 and P:*94*:547), that its use may be facilitated if it is applied as a method of "uniform juniority", i.e., deduction from the oldest rather than addition to the youngest age.

C; 19. The Application and Fitting of the Verhulst-Pearl-Reed (the "Logistic") Curve of Population Growth

Very complete statistical illustrations of the "logistic" curve have been worked out by several investigators—sometimes in order to examine its acceptability as a "law" of population growth, on other occasions to illustrate its dangers as an instrument of prediction, and again to investigate the difficulties of fitting such a transcendental form.

Pearl and Reed have been in the forefront amongst those who would claim some measure of universality for the method. Summarizing in P:*96*:637, for example, the results of a number of their papers, Pearl shows the logistic (sometimes in its original symmetrical form, and sometimes as the sum of two such curves or in the generalized form (103) to give effect to cycles of growth) for 16 different countries, and concludes that it "does in fact describe the known (or, in the case of the world, estimated) population growth with great precision and fidelity", so that "this evidence makes it probable that the curve is at least a first approximation to a descriptive law of population growth". It is of interest, also, to note that Pearl's rediscovery of this form of curve was supported by observations of the growing numbers of the fruit-fly *Drosophila melanogaster* enclosed for breeding and observation in the "limited universe" of a milk-bottle, which followed the curve quite closely—a type of evidence which prompted Sir Athelstane Baines to observe (P:*176*:61) that "so far as the habits of the *Drosophila* are concerned, I must confess that I sympathize with Dickens' Eugene Wrayburn [the calm but briefless barrister in 'Our Mutual Friend'] who, when taunted with the example of the ant and the bee, 'protested, on principle, as a biped' ".

Other valuable statistical examinations are shown in P:*176*:7-23 by Yule, and in P:*105* by Reed and Pearl. In P:*124*:162 Schultz has given further illustrations, and has emphasized the importance of the standard errors of the forecast values (see also P:*60*:311). In connection with the use of the logistic for long-range prediction, his conclusion, indeed, may be quoted: "There is no necessary relation between the goodness of fit of a curve to past observations and its reliability for forecasting purposes. A curve may fit the data for the past 100 years with a high degree of accuracy, and yet fail to predict the situation for the next year or so".

Another extensive series of examples for 29 populations is provided by Wilson and Puffer in P:*159*, which draws attention to cases in which the logistic gives eventually an infinite population.

An excellent contribution, also, is the paper (P:*23*) by Cramér and Wold. In addition to a very useful examination of the characteristics and methods of fitting the logistic, their general conclusion (loc. cit., 203) should be noted: "The method gives good results in cases when the data are not too few in number and show an evident trend of the logistic type. In other cases, the advantage of fitting the logistic curve to statistical data seems to us somewhat doubtful".

Methods of Fitting the Logistic

The difficulties of fitting the curve have been investigated exhaustively.

(1) Verhulst, and Pearl and Reed, have used simply 3, or 5, equidistant ordinates (see P:*176*:49; P:*96*:576; and P:*27*:242).

(2) Yule has suggested summing the reciprocals in 3 successive groups, and thus equating the harmonic means of the actual and calculated populations in each group (P:*176*:7 and 51).

(3) The same author also investigated the fitting of the form (102b) in B; 20 by finding L and α from the fitting of a straight line to the proportionate increases over successive intervals of time by the method of least squares, and then determining β from sums of the reciprocals of the given populations (P:*176*:52; see also particularly P:*60*:293).

(4) Will has illustrated a method based on finite differences (see p. 172; A; 17) in H:*170*:177.

(5) The application of the method of least squares is discussed at p. 327; C; 21.

C; 20. The Meaning and Use of the "Weights" in forming the "Normal Equations" in the Method of Least Squares

Although the "weight" is, in accordance with its definition by (107) and (108), the reciprocal of the square of the probable error (to adopt here the use of λ, rather than σ or c, in accordance with the phraseology usually associated with the classical explanations of the method of least squares), it should be observed that it obviously may also be interpreted as expressing "the number of observations of weight unity of which it is the equivalent". For, in forming the "normal equations" (110), if the "observation equation", when $x=1$ and the weight is W_1, were merely written down W_1 times, each time with weight unity, and if in general the observation equation for $x=s$ is merely written down W_s times, each time with unit weight, it will be clear that we obtain exactly the same scheme of equations as if when $x=1$ the observation equation is merely multiplied by W_1 and the observation equation when $x=s$ is multiplied by W_s; and the partial differential coefficients, when equated to 0, will give exactly the set of normal equations (110).

The student must here be warned to watch for an alternative mode of stating the rule of formation of the normal equations, since otherwise it may cause confusion. The condition of least squares, by (109), is that $\Sigma[W_x(f''_x-f'_x)^2]$ must be a minimum, where by (108)—and again adopting here the classical use of the probable error—$W_x=\frac{1}{\lambda_x^2}$. This, however, can also be written that $\Sigma\left[(\sqrt{W_x}f''_x-\sqrt{W_x}f'_x)^2\right]$ is to be a minimum, where $\sqrt{W_x}=\frac{1}{\lambda_x}$, which means that $\Sigma\left[(w_xf''_x-w_xf'_x)^2\right]$ must be minimized where $w_x=\frac{1}{\lambda_x}$. The partial differentiation with regard to the several unknowns in the case of $f''_x=\alpha+\beta x+\gamma x^2+\ldots$ then leads to the following (omitting the common factor 2):

$$\left.\begin{aligned}\Sigma[w_x\{w_x(\alpha+\beta x+\gamma x^2+\ldots\;)-w_xf'_x\}]&=0\\ \Sigma[xw_x\{w_x(\alpha+\beta x+\gamma x^2+\ldots)-w_xf'_x\}]&=0\\ \Sigma[x^2w_x\{w_x(\alpha+\beta x+\gamma x^2+\ldots)-w_xf'_x\}]&=0\end{aligned}\right\}\ldots.(110a)$$
$$\vdots$$
etc.

These are of course the same as the normal equations (110), since $w_x^2 = W_x$. In order that they may be written down in the form (110a), however, they evidently require the following slight re-wording of the verbal rule previously given (in Chapter VIII): "Set down the 'observation equation' $(\alpha+\beta x+\gamma x^2+\ldots)-f'_x=0$ for each value of x, and prepare it by multiplying it by the *square root of its weight*, i.e., by $w_x=\sqrt{W_x}=\frac{1}{\lambda_x}$. Form the normal equation for the unknown α by then multiplying each observation equation, *thus prepared*, by the coefficient of α in that *prepared* equation, and adding the results; similarly form the normal equation for β by multiplying each *prepared* observation equation by the coefficient of β in that equation, and adding the results; and so on". It will be noted that in the verbal rule just given each observation equation itself is first multiplied by the *square root* of the weight, so that $\sqrt{W_x}(=w_x)$ becomes part of the coefficient of the unknown by which the equations are multiplied in forming the normal equations. An example of this procedure is given in H:*44*:164.

It is important for the student to grasp thoroughly the difference between the systems (110) and (110a), and also to understand that in both systems the "weight" is $W_x=\frac{1}{\lambda_x^2}$, so that $w_x=\frac{1}{\lambda_x}$ is the square root of the weight. Actuarial students, in particular, should note that—contrary to the above accepted definitions of the method of least squares—in P:*51*:119 and 129, and P:*174*:46, the weight is defined as $\frac{1}{\lambda_x}$, so that in reality the term "weight" is there used for w_x instead of for the usual W_x.

C; 21. The Practical Application of the Method of Least Squares

The Systematic Solution of Linear Normal Equations

A large proportion of the literature and text-book descriptions of the method of least squares concerns itself with the systematic solution of the linear "normal equations", and the controls of the necessary calculations. As the equations to be solved constitute a linear simultaneous set, the solutions can be expressed very conveniently by means of determinants (see P:*155*:231). In the explanations two special notations are generally employed: (*a*) The term "residual", or "residual error", with the symbol v_x, is used for the difference between the fitted and observed values, $f''_x - f'_x$; and (*b*) following Gauss (cf. P:*90*:49), square brackets are employed to denote summation, in the form $[aa] = a_1^2 + a_2^2 + \ldots + a_\nu^2$, $[ab] = a_1b_1 + a_2b_2 + \ldots + a_\nu b_\nu$, etc. It may also be noted that in many of the classical texts the unknowns are denoted by $x, y, z, \ldots$.

Since these expositions of systematic methods are so complete (with many numerical examples) and are so readily available, it will be adequate here to give the following references only:

(1) For the "method of determinants" see P:*155*:231 and P:*13*:99.

(2) Gauss's "method of substitution" (H:*17*:Supplementum) is fully explained in P:*155*:234, P:*13*:90, and P:*90*:175.

(3) "Doolittle's method" (H:*56*), which is a modification of Gauss's, is described in P:*13*:96 and P:*28*:104 et seq.

(4) For the case when the number of unknowns is large, the "method of successive approximation (iteration)" developed by Gauss, Seidel, and Jacobi is given in P:*155*:255.

(5) Other variations are noted in P:*155*:236 and P.*28*:106.

(6) An electrical machine has recently been devised for the automatic solution of large sets of linear simultaneous equations (H:*179*).

Some of these processes are particularly convenient on account of the manner in which they facilitate the calculation of the "standard errors" of the unknown parameters as well as the parameters themselves (for examples, see P:*155*:209 et seq., P:*13*:103, P:*124*:144 et seq., and P:*28*).

The Determination of the Constants in Makeham's Formula (83)

One of the main curve-fitting problems with which actuaries are concerned is the determination of the constants in Makeham's formula (83). For the purposes of this discussion (83) will be written $y_x = A + Bc^x$ where y_x represents any one of the functions m_x, μ_x, $\mu_{x+\frac{1}{2}}$, or colog p_x, and the unknowns A, B, and c may therefore be supposed to stand for the Makeham constants in any of the formulae for those several functions. [This Makeham constant c, which is used here in conformity with universal practice, of course has nothing to do with the c of the Normal Curve (11) and of the weight W_x defined by (107).] Then, by (109), the method of least squares requires that

$$\Sigma[W_x\{(A+Bc^x)-f'_x\}^2] \text{ must be a minimum} \quad \ldots . (128)$$

and the "normal equations" resulting from the partial differentiations with regard to A, B, and c are

$$\left.\begin{aligned} A(\Sigma W_x)+B(\Sigma c^x W_x)-\Sigma f'_x W_x &= 0 \\ A(\Sigma c^x W_x)+B(\Sigma c^{2x} W_x)-\Sigma f'_x c^x W_x &= 0 \\ A(\Sigma x c^x W_x)+B(\Sigma x c^{2x} W_x)-\Sigma f'_x x c^x W_x &= 0 \end{aligned}\right\} \quad \ldots\ldots (129)$$

These equations are not linear with regard to the Makeham constant c; it is therefore necessary either to adopt the method of approximation stated on p. 94 and at p. 241; B; 22, or to employ some other device in order to reach linear equations for solution.

The classical method of approximation seems to have been illustrated first by Chandler in H:*44*:169. Later Karup also gave a thorough application (H:*66*).

Since the non-linear nature of the normal equations is due to the manner in which c is involved, and as $\log_{10}c$ is often found to lie between about .038 and .04, another obvious procedure is to adopt a succession of trial values for c, and to determine the corresponding values of A and B from the first two of the equations (129). In order, however, to reach the absolute minimum for (128) it would be necessary to adopt some criterion for the degree of approximation secured, as observed by Rosmanith (H:*104*:329).

A very valuable method which is less onerous than the preceding has been proposed by Steffensen (P:*135*). He eliminates A and B from (129) by determinants, and obtains a single equation in c, namely

$$D \equiv \begin{vmatrix} \Sigma W_x & \Sigma c^x W_x & \Sigma f'_x W_x \\ \Sigma c^x W_x & \Sigma c^{2x} W_x & \Sigma f'_x c^x W_x \\ \Sigma x c^x W_x & \Sigma x c^{2x} W_x & \Sigma f'_x x c^x W_x \end{vmatrix} = 0 \quad \ldots . (130)$$

Since $\log c = .04$ is usually a close approximation, the numerical solution of this equation can generally be effected (to 5 figures for c) from 3 trial values for $\log c$ and subsequent interpolation to find the value for which $D=0$. With c thus determined, A and B follow from any two of the equations (129). An example is given in P:*135*, and the excellent results obtained are shown by the comparisons in P:*33*:253.

The "weights" in all the preceding methods will be taken in accordance with formulae (111), (112), and (113), or the other particular applications to mortality functions discussed at p. 272; C: 7. The weight of the observed m'_x as $\dfrac{E'_{x+\frac{1}{2}}}{m_x(1-m_x)}$ was thus used in Chandler's graduation in H:*44*:164, or may be taken as $\dfrac{E'_x q_x}{(m_x)^2}$ as suggested in P:*51*:100. The weight of the observed colog p'_x, being $\dfrac{E'_x p_x}{q_x}$, was used in H:*108*:256, and in Steffensen's least squares graduation of P:*135*. Although it is not usually required for a Makeham graduation, the weight of the observed q'_x would similarly be $\dfrac{E'_x}{p_x q_x}$.

It must also be remembered, as pointed out in the text, that if the series f''_x is to be fitted on the assumption of uniform weights, but is itself transformed by a device such as the taking of its logarithm, then that logarithmic series $\log f''_x$ must have weights $(f_x)^2$ assigned to it. This would be of importance if the fitting of such generalized Makeham forms for μ_x as (86)-(88) were to be undertaken (as suggested in P:*144*:287) by fitting $\log \mu_x$ by the method of least squares.

The fitting of Makeham's formula (83) in its l_x or $\log l_x$ form evidently involves, with any method, more awkward expressions than those which emerge from m_x, $\mu_{x+\frac{1}{2}}$, or colog p_x. It may be noted, nevertheless, that the fitting of $\log l_x$, or $\Sigma \log l_x$ in decennial groups, by least squares (but in each case without weights) has been discussed in H:*120*:231.

The Determination of the Constants of the Logistic Curve

The fitting by the method of least squares of the transcendental "logistic" curve has been discussed in several forms. If it be assumed in (102) that the enumerated populations P_t—here f'_x—are equally well determined, so that the weights of the several observations can be taken as uniform, the least squares condition requires that

$$\Sigma_x\left[\frac{A+Be^{k(x-\tau)}}{1+e^{k(x-\tau)}}-f'_x\right]^2 \text{ must be a minimum} \ldots . (102d)$$

The necessary differentiations with respect to the unknowns here lead to non-linear "normal equations". The classical method of approximation for such a case (as stated on p. 94 and at p. 241; B; 22) by which preliminary values are first found and corrections then determined by the method of least squares, is set out fully by Schultz in P:*124*:164. The problem has also been examined in a valuable paper by Wilson and Puffer (P:*159*; see also P:*65*:108), where it is observed that disappointing results may arise from the neglect of terms of the second order in the method of approximation as ordinarily used.

As in the case of Steffensen's method for dealing with the non-linear normal equations of the Makeham function, a useful least squares technique, which involves a comparatively small amount of work, has been evolved by Cramér and Wold (P:*23*:201). In (102), crude approximate values of the asymptotes A and B are first determined by graphical inspection of the data. The approximate position of the inflexion point, τ, in this symmetrical curve then follows from $P_\tau = \dfrac{A+B}{2}$. For k we have, from (102) by differentiation, $\dfrac{dP_t}{dt} = \dfrac{k(P_t-A)(B-P_t)}{B-A}$, whence

$\left(\frac{dP_t}{dt}\right)_{t=\tau}=k\left(\frac{B-A}{4}\right)$. Next regarding τ and k as fixed, they determine the equations of condition for A and B to satisfy the usual unweighted least squares criterion. Using then the fixed τ, but different values of k near its approximate value, they find by interpolation the value of k for the fixed τ which produces the minimum. Repeating the process for a number (generally only 2) further values of τ, it is possible to find with sufficient accuracy the value of τ for the required absolute minimum, and thence k, B, and A by interpolation. A numerical example is shown, and tables are supplied to facilitate the process.

Another method of introducing the least squares principle, with the logistic in the form (102*a*), was used by Pearl and Reed (P:*96*:579 et al.; P:*124*:163; and P:*27*:245). First determining from three equidistant points an approximate value for the exponent of e, they multiply through by the denominator of the expression to be minimized, and then take the partial derivatives therefrom in order to make the unweighted sum of the squares of the resulting residuals a minimum. This process, however—even though it may give good results—is not the true least squares procedure. The minimizing of (102*d*) is based on the assumption of uniform weights for P_t; but the multiplication through by the denominator changes the system entirely, and introduces implicitly a series of different weights, so that—as Schultz observes (P:*124*:164)—"in fact it is difficult to give meaning to the residuals which they are minimizing" (cf. also p. 96 here).

C; 22. Practical Applications of the Unweighted Method of Moments in Actuarial Work

The relationship between the method of least squares and the method of moments which is discussed in the text has an important bearing upon the extent to which the method of moments should be applied in practice—as is very commonly done—in its unweighted form.

The excellent results obtainable by the unweighted method of moments in the fitting of curves, such as Pearson's, to frequency distributions are to be anticipated from the principles brought out in that discussion. The application of the unweighted moment equations in the fitting of curves to series (such as μ_x or colog p_x) which cannot be viewed as frequency distributions should, however, be undertaken with some caution; the entire omission of the weights clearly may lead to a system of fitting essentially different from that of strict least squares, while certain alterations in the method of stating the basic equations may result, in effect, in the use of approximate weights which are in reasonable conformity with those of least squares. These considerations are of particular importance in connection with some of the procedures by which it is said that the "method of moments" (without any reference to whether weights are or are not used) has been applied in the fitting of Makeham's formula (83).

One example of the use of unweighted moments is the application of the moment principle directly to the graduation by Makeham's form colog $p_x = \alpha + \beta c^x$, without any introduction of weights, in H:*92*:11, and the valuable discussions and detailed illustrations of that procedure in H:*135* and P:*136* (see also P:*162*:547). The mode of computation followed is that of successive summation, so that the equations to be solved for α, β, and c are the equations for Σcolog p_x, Σ^2colog p_x, and Σ^3colog p_x over the range of ages selected (cf. p. 257; B; 27, and see H:*135*:656 and P:*136*:3-6 for the precise development of the equations in this case). Since the method of course leads (just as does the method of least squares) to an awkward equation for the determination of c, it should be noted that Steffensen (P:*136*:7) performs the solution for c by elimination with determinants and subsequent interpolation from trial values (cf. his similar method for least squares stated at equation (130) on p. 326; C; 21 here). Trachtenberg (H:*135*) also gives a useful table by which the solution by trial of the equation for c may be accomplished when the range of ages is from 20 to 79 inclusive. Steffensen's careful numerical application of the method to the Danish $D^{M(5)}$ experience—which follows Makeham's formula closely (see P:*136*:2)—shows that the results are not very satisfactory (P:*136*:16), and that the failure can be attributed to the manner in which this direct application of the method of moments to colog p_x ignores the weights of the observations.

The unsatisfactory results which are thus to be anticipated from colog $p_x = a + \beta c^x$ when the constants are obtained directly from the unweighted moment equations are greatly improved—as again is to be expected from the principles stated in the text—when the equation just given is applied in a form which in effect deals, by unweighted moments, with a frequency distribution instead of with a series of rates. For colog $p_x \doteq \mu_{x+\frac{1}{2}} = \dfrac{\theta_x}{E_{x+\frac{1}{2}}}$; if for this, by (83), we use the Makeham form $A + Bc^{x+\frac{1}{2}}$, we may write $E_{x+\frac{1}{2}}(A + Bc^{x+\frac{1}{2}}) = \theta_x$; we are then dealing with a frequency distribution, of deaths θ_x, instead of a series of rates colog p_x, and (in accordance with the principles explained in the text) may use unweighted moments as being likely to give results close to those obtainable from colog p_x by strictly weighted least squares. Sir G. F. Hardy made extensive use of this principle—performing the calculations by his method of successive summations as the equivalent of the method of moments, and employing trial values of c so that only the first and second summations are required in order to determine A and B from two linear equations for each such value of c (see P:*51*:65 and H:*106*:292). [In some applications Hardy also subsequently introduced an adjustment to correct for the approximate nature of the relation colog $p_x \doteq \mu_{x+\frac{1}{2}}$, so that the constants a and β for colog p_x could be derived from the graduation as well as A and B in $\mu_{x+\frac{1}{2}}$ (see H:*95*:502 and H:*106*:293).] In the case of the $D^{M(5)}$ experience (where, as already noted, unweighted moments applied directly to colog p_x gave unsatisfactory results) this procedure of Hardy (which transforms the process into an application of unweighted moments to a frequency distribution) showed, as would be expected, results very close to those of strictly weighted least squares (see methods Nos. 1 and 2 in P:*33*:253).

The methods of the preceding paragraphs contemplate the use of Makeham's formula in its colog p_x or $\mu_{x+\frac{1}{2}}$ form. The expression $\log l_x = \log k + x \log s + c^x \log g$ will evidently be less easily handled; it may be noted, nevertheless, that the application of unweighted moments by integration in this case was worked out fully by Karl Pearson in H:*86*:298, and has been used by Glover (H:*84*) and Thompson (H:*120*:226), while a simpler method involving summation instead of integration is given at P:*59*:95.

C; 23. The Practical Application of the Minimum-χ^2 Method to Actuarial Data

An interesting illustration of the principle of minimizing χ^2 in accordance with condition (122) is shown by Cramér and Wold, in P:*23*:173, for the case of a Makeham graduation. Remembering that the method is applicable particularly when the data are in the form of a frequency distribution, and that the observed deaths, θ'_x, may be so considered, it will evidently be advisable to make, in effect, a graduation of θ'_x—for a ratio such as $\mu_{x+\frac{1}{2}}=\dfrac{\theta_x}{E_{x+\frac{1}{2}}}$, which by Makeham's formula (83) may be taken in the form $A+Bc^{x+\frac{1}{2}}$, is not a frequency distribution. We should therefore deal with the frequency distribution of deaths, $\theta_x=E_{x+\frac{1}{2}}(A+Bc^{x+\frac{1}{2}})$; in (122), consequently, we put $f'_r=\theta'_x$ and $f''_r=E'_{x+\frac{1}{2}}(A+Bc^{x+\frac{1}{2}})$; and the expression to be minimized becomes $\Sigma\left[\dfrac{(AE'_{x+\frac{1}{2}}+BE'_{x+\frac{1}{2}}c^{x+\frac{1}{2}}-\theta'_x)^2}{\theta'_x}\right]$, where A, B, and c are the unknowns.

Since here, as in the method of least squares, the differentiation with regard to c leads to a non-linear equation for solution, it will be advisable to adopt certain trial values for c, and thence to determine the best value by interpolation (cf. Steffensen's method, p. 326; C; 21, and P:*23*:176). For each such value of c, the differentiations with regard to A and B when equated to zero give easily (dropping the multiplier 2)

$$A\Sigma\left[\frac{(E'_{x+\frac{1}{2}})^2}{\theta'_x}\right]+B\Sigma\left[\frac{c^{x+\frac{1}{2}}(E'_{x+\frac{1}{2}})^2}{\theta'_x}\right]=\Sigma[E'_{x+\frac{1}{2}}]$$

and $$A\Sigma\left[\frac{c^{x+\frac{1}{2}}(E'_{x+\frac{1}{2}})^2}{\theta'_x}\right]+B\Sigma\left[\frac{(c^{x+\frac{1}{2}}E'_{x+\frac{1}{2}})^2}{\theta'_x}\right]=\Sigma[c^{x+\frac{1}{2}}E'_{x+\frac{1}{2}}]$$

as shown, in slightly different notation, in P:*23*:173.

The numerical illustrations given by Cramér and Wold indicate, as would be expected (cf. P:*137*:358, quoted on p. 99 here), that the values of c, A, and B obtained by this method are insignificantly different from the corresponding equations of the unweighted method of moments, when the latter also is applied (see p. 330; C; 22) to graduate the frequency distribution of deaths in the form $\theta_x = E_{x+\frac{1}{2}}(A + Bc^{x+\frac{1}{2}})$.

C; 24. Applications of the Tests of Goodness of Fit based on (i) Changes of Sign, and (ii) Standard Deviations

(*i*) The following example of the formula $\dfrac{N-r-1}{2^{r+1}}$ and its probable error (if it may be assumed that the occurrences are independent) for a non-periodic series with the first and last signs omitted (as given at p. 249; B; 25), was shown by De Forest (H:*49*:34) in order to test the deviations between certain observed and graduated rates of mortality over a range of 70 ages:

Number of Like Signs in Sequence $= r$	Expected Sequences $= \dfrac{N-r-1}{2^{r+1}} \pm p.e.$	Observed Sequences
1	17.0 ± 2.4	26
2	8.4 ± 1.8	10
3	4.1 ± 1.3	6
4	2.0 ± 0.9	1
5	1.0 ± 0.7	0
6	.5 ± 0.5	0
7	.2 ± 0.3	0

Here the excess in the observed sequences for $r=1$, 2, and 3, and the deficiencies thereafter, indicate that the graduated values follow the observations rather too closely, i.e., the original series has not been smoothed quite enough. This may be seen also from the fact that, even if adjustment were made in respect of the first and last signs in order to deal with the series as a periodic one, the signs falling within groups of 1 or 2 would total about 36, and are thus much greater than—instead of being approximately equal to—the 7 in groups of more than 2. The sequences of all orders, being 43, are similarly too large in comparison with the limit $\frac{N}{2}=35$ even if the series were periodic.

Two examples of the similar method based on $\frac{N}{2^{r+1}}$ (but without knowledge of De Forest's work) may be found in H:*69*:331 and P:*135*:5 (see also a reference thereto in P:*125*:28).

(*ii*) In applying the test for a permissible graduation based on $\pm 3\sigma$ or $\pm 2\sigma$ ($\doteq \pm 4\lambda$ or $\pm 3\lambda$), or the test for a satisfactory graduation based on $\frac{4}{5}\sigma$, it is necessary only to compute $\sigma\{q'_x\} \doteq \sqrt{\frac{p''_x q''_x}{E'_x}}$ at each age (where E'_x denotes the exposed to risk, q'_x the ungraduated rate of mortality, and q''_x the graduated rate—see par (*iii*); C; 7), and then to compare the actual deviation $(q''_x - q'_x)$ with 3σ or 2σ, or with $\frac{4}{5}\sigma$, as the case may be. A numerical example of the 2σ method is shown at H:*160*:12 and 25.

Alternatively, of course, the comparisons may be shown for the deaths instead of for the rates of mortality. The standard deviation to be used is then $\sigma\{\theta'_x\} \doteq \sqrt{E'_x p''_x q''_x}$ by formula (8) and par. (*ii*); C; 7, and the actual deviations to be compared with 3σ or 2σ, or with $\frac{4}{5}\sigma$, are $(\theta''_x - \theta'_x) = (E'_x q''_x - \theta'_x)$. Illustrations of the $\frac{4}{5}\sigma$ test are given in H:*90*:128 and 155. The examination is there made in 5-year age groups. It should be noted, however, that this customary method of using age groups may give misleading conclusions through the cancellation of positive and negative deviations within a group (cf. P:*125*:6-7); it is therefore advisable in practice to make the comparisons age by age.

C; 25. Illustrations of the χ^2 Test

(1) *The Binomial; Probabilities Known a priori; One Constraint—Weldon's Dice Data*

A classical example which brings out clearly the technique of the χ^2 test, and also its relation to the simpler probable error methods, is afforded by the results of Weldon's experiment in throwing 12 dice together, 26,306 times, and observing the number at each throw which turned up 5 or 6. With one unbiassed die the chance of 5 or 6 at a single throw is $\frac{1}{3}$; the theoretical frequencies when 12 unbiassed dice are thrown together 26,306 times are therefore the terms of the binomial $26306(\frac{1}{3}+\frac{2}{3})^{12}$, which are shown in column (2) of Table A. Weldon, however, found that the distribution which occurred was that shown in column (3). Knowing thus the "true" distribution to be expected from unbiassed dice, and having a series actually observed, the question which immediately arises is whether the observed series fits the theoretical series within such deviations as may be attributable to chance alone, or, on the other hand, whether it is incompatible with this hypothesis so that some cause may have operated to produce the results—as, for example, that all the dice may not have been unbiassed.

TABLE A

Number of Dice with 5 or 6 r	Theoretical Frequency, f_r Unbiassed Dice ($p=.\dot{3}$)	Observed Frequency, f'_r	$\frac{(f'_r-f_r)^2}{f_r}$
(1)	(2)	(3)	(4)
0	202.75	185	1.554
1	1216.50	1149	3.745
2	3345.37	3265	1.931
3	5575.61	5475	1.815
4	6272.56	6114	4.008
5	5018.05	5194	6.169
6	2927.20	3067	6.677
7	1254.51	1331	4.664
8	392.04	403	.306
9	87.12	105	3.670
10	13.07	14 }	
11	1.19	4 }	.952
12	.05	0 }	
	26306.02	26306	$\chi_0^2=35.491$

In applying the χ^2 test to this enquiry we find χ_0^2 as in column (4)—the last two groups being merged with that for $r=10$ to give frequencies not less than 10 throughout. There are thus 11 groups; one constraint has been imposed by the equality of the totals, as in the fundamental derivation of (50) and (52), so that there are $11-1=10$ "degrees of freedom"; and from the tables of P and χ_0^2 it is seen that for 10 degrees of freedom P lies well below .01—being, in fact, about .0004—when $\chi_0^2=35.491$. It is therefore evident that the "fit" is very poor, i.e., the observed series is not compatible with the hypothesis that the true frequencies would be those of column (2) resulting from unbiassed dice, since we should expect to get a set of deviations giving a value of χ^2 as large as or larger than that actually observed only about 4 times in 10,000 trials. The inferences to be drawn would therefore be (separately or in conjunction) that the dice were biassed, or that the method of throwing them was faulty (an unlikely circumstance, since special care was taken), or that a most improbable event actually occurred.

(2) *The Binomial; Probabilities Estimated from the Data; Two Constraints—Weldon's Dice Data*

Suppose now that we are presented merely with the observed frequencies of column (3), without any prior knowledge of the true probabilities as reflected in the frequencies of column (2). Or, as an alternative but really equivalent viewpoint, suppose that, being presented only with the observed frequencies, we are asked to examine whether they are consistent with the hypothesis that the true probability of throwing 5 or 6 with a particular die is not the unbiassed estimate $\frac{1}{3}$ assumed above, but is instead *a value to be estimated from the data*. If we adopt as the best estimate the weighted mean $\dfrac{\Sigma r f_r'}{12(26306)} = \dfrac{106602}{315672} =$ .3376986, the problem becomes that of comparing the "goodness of fit" of the observed series shown again in column (2) of Table B, with the theoretical values resulting from the distribution $26306(.337699+.662301)^{12}$, which are shown in column (3). The value of χ_0^2 is then found, as in column (4), to be 8.179.

TABLE B

Number of Dice with 5 or 6 r	Observed Frequency, f'_r	Estimated Frequency, f_r ($p=.3376986$)	$\frac{(f'_r-f_r)^2}{f_r}$
(1)	(2)	(3)	(4)
0	185	187.38	.030
1	1149	1146.51	.005
2	3265	3215.24	.770
3	5475	5464.70	.019
4	6114	6269.35	3.849
5	5194	5114.65	1.231
6	3067	3042.54	.197
7	1331	1329.73	.001
8	403	423.76	1.017
9	105	96.03	.838
10	14	14.69	
11	4	1.36	.222
12	0	.06	
	26306	26306.00	$\chi_0^2=8.179$

The last three frequencies being grouped, as before, there are still 11 groups. It is to be noted especially, however, that there is now an additional "constraint"—making 2 in all—because p has been made the same in the observed and theoretical series, in addition to the equality of the totals. The degrees of freedom are consequently only $11-2=9$. From the tables it is seen that P is .52 when $d=9$ and $\chi_0^2=8.179$. The "fit" is therefore good, i.e., there is, on this test, no reason to doubt the hypothesis that the probability for each die of throwing 5 or 6 was actually the biassed figure .3376986.

(3) *Frequency Curves—Number of Constraints*

It will be useful in connection with the concept of "degrees of freedom" to point out that the observed frequencies, f'_r, of data such as Weldon's could be represented, or "graduated", by some analytical function, and that if then it were desired to examine the goodness of fit by the χ^2 test, it would be accomplished by computing χ_0^2 as in the preceding illustrations with, however, due allowance for the proper number of constraints. Thus if 11 groups were still used, and Poisson's formula were fitted, there would be $11-2=9$ degrees of freedom, since in determining the graduated f_r'' two constraints result from equating the total number and the mean; the Normal Curve, fitted from the total, mean, and standard deviation, imposes 3 constraints; Pearson's Type III is fitted by using the total, the mean, μ_2, and μ_3, so that there are 4 constraints; Pearson's Main Types I, IV, and VI require μ_4 as well, thus increasing the number of constraints to 5; and the Gram-Charlier series similarly employs the total and the first four moments, with 5 constraints again.

(4) *The Graduation of Mortality Statistics; Makeham's Formula; Number of Constraints when c is Assumed*

The problem of graduating mortality statistics does not ordinarily present itself to the actuary as a mere graduation of a series of observed deaths. The data usually appear in the form of a column of exposed to risk, E_x', at each age x—which have been obtained by observing the exposures of a group of lives with due allowances for the fluctuations caused by new entrants and exitants—and another column of the corresponding observed deaths, θ_x', at each age. This means that for each age x there are, in effect, E_x' cases, of which θ_x' are observed to die, and $E_x'-\theta_x'$ do not die. A graduation of the observed rate of mortality $q_x'\left(=\dfrac{\theta_x'}{E_x'}\right)$, or a function such as $\text{colog}_{10}\, p_x'$ or $\mu_{x+\frac{1}{2}}'$, is then generally made. If it is next desired to test the fit obtained, the customary procedure is to compute the expected deaths, say θ_x'', by multiplying the original E_x' cases by the graduated rate of mortality, say q_x'', and then to compare θ_x'' with the observed θ_x' of the data. That is to say, the original data in reality consist (taking two age groups only as an example) of observations in the form set out in Table C, and a graduation might have produced the adjusted figures of Table D if it had been performed so that the total expected deaths, θ_x'', equal the total actual deaths, θ'_x. This form of presentation is known as a *Contingency Table*; in this particular case we have a "2×2" contingency table, since there are (to the left and above the double lines) 2 rows and 2 columns, giving 4 "cells" in all—a "fourfold" table.

TABLE C—OBSERVED VALUES

Age Group	θ'_x die	$E'_x-\theta'_x$ do not die	Total E'_x
20-24	15	1,734	1,749
25-29	113	15,876	15,989
Total....	128	17,610	17,738

TABLE D—GRADUATED VALUES

Age Group	θ''_x die	$E'_x-\theta''_x$ do not die	Total E'_x
20-24	12	1,737	1,749
25-29	116	15,873	15,989
Total....	128	17,610	17,738

Now in such a table the E'_x is fixed, i.e., the totals of the rows are fixed; and this is necessarily so in both the tables. Similarly when, as here, the graduation has made the total expected deaths (θ''_x) equal to the total of the actual deaths (θ'_x), it follows that the totals of the columns are also equal in the two tables. Under such circumstances, in a fourfold table, the fixing of one cell fixes them all—for example, if θ''_x for the 20-24 group were to have emerged as 30, all the rest of the table could be written down at once from the fixed values of the totals of the two rows and the first column. Since the value in only one cell can here be assigned at will, it follows that there is only one "degree of freedom". By similar reasoning it is easy to see, in general, that in a $p\times q$ contingency table, with pq cells, the frequencies in the first $p-1$ columns and $q-1$ rows can be determined at will, and the remainder follow automatically from the totals, so that there are $(p-1)(q-1)$ degrees of freedom.

Let us now consider a more extended hypothetical example (taken from H:*90*:155, with slight adjustment for illustrative purposes here, in order to make the totals of the actual and expected deaths precisely equal). Suppose, therefore, that the data consist of the exposed to risk, E'_x, and actual deaths, θ'_x, of columns (2) and (3) of Table E, and that we wish to apply the χ^2 method to test the fit of the graduated values stated in the second half of that table.

Now if the deaths had been obtained by some graduation process which operated directly upon the observed deaths, θ'_x, alone, to produce simply the graduated values of column (6), and in so doing merely imposed the condition that the totals of θ'_x and θ''_x should be equal, the problem would be analogous to

TABLE E

Age, x	Observed Data			Graduated		
	Exposed to Risk, E'_x	Deaths, θ'_x	$E'_x-\theta'_x$, who do not die	Exposed to Risk, E''_x	Deaths, θ''_x	$E'_x-\theta''_x$, who do not die
(1)	(2)	(3)	(4)	(5)	(6)	(7)
20-24	1,749	15	1,734	1,749	12	1,737
25-29	15,989	113	15,876	15,989	116	15,873
30-34	107,629	864	106,765	107,629	861	106,768
35-39	346,276	3,119	343,157	346,276	3,136	343,140
40-44	588,003	6,461	581,542	588,003	6,300	581,703
45-49	728,094	9,761	718,333	728,094	9,698	718,396
50-54	757,987	13,071	744,916	757,987	13,183	744,804
55-59	701,051	16,521	684,530	701,051	16,636	684,415
60-64	590,761	19,628	571,133	590,761	19,815	570,946
65-69	442,842	21,428	421,414	442,842	21,527	421,315
70-74	286,647	20,731	265,916	286,647	20,505	266,142
75-79	151,977	16,160	135,817	151,977	16,105	135,872
80-84	62,595	10,003	52,592	62,595	9,796	52,799
85-89	18,059	3,946	14,113	18,059	4,131	13,928

that already considered for Weldon's dice data, and there would be one constraint, giving, with 14 groups, $14-1=13$ degrees of freedom. Then χ_0^2 would be found as $\Sigma\left[\frac{(\theta_x'-\theta_x'')^2}{\theta_x''}\right]=24.7582$, as in column (3) of Table F. P corresponding to $\chi_0^2=24.7582$ for 13 degrees of freedom is .027. This is below .05, and not much above .02; there is consequently strong reason to conclude that there is a real discrepancy between the observed deaths of column (3) and the values shown in column (6), if they had been obtained by direct graduation.

TABLE F

Age x (1)	$(\theta_x'-\theta_x'')$ (2)	$\frac{(\theta_x'-\theta_x'')^2}{\theta_x''}$ (3)	$(E_x'-\theta_x')-(E_x'-\theta_x'')$ (4)	$\frac{[\text{Col. (4)}]^2}{(E_x'-\theta_x'')}$ (5)
20-24	3	.7500	−3	.0052
25-29	− 3	.0776	3	.0006
30-34	3	.0104	−3	.0001
35-39	−17	.0922	17	.0008
40-44	161	4.1144	−161	.0446
45-49	63	.4093	−63	.0055
50-54	−112	.9515	112	.0168
55-59	−115	.7950	115	.0193
60-64	−187	1.7648	187	.0612
65-69	−99	.4553	99	.0233
70-74	226	2.4909	−226	.1919
75-79	55	.1878	−55	.0223
80-84	207	4.3741	−207	.8115
85-89	−185	8.2849	185	2.4573
Total	+718 −718	$\chi_0^2=24.7582$		$\chi_0^2=3.6604$

Suppose now, however, that the graduated θ_x'', instead of being obtained as above by direct adjustment of the observed deaths only, have been computed from a graduation of some mortality ratio derived from the data of Table E, such as $q_x'\left(=\frac{\theta_x'}{E_x'}\right)$ or $\mu_{x+\frac{1}{2}}'\left(=\frac{\theta_x'}{E_x'-\frac{1}{2}\theta_x'}\right)$, in which both θ_x' and E_x' are involved. Then the position is that the observed data of both cols. (3) and (4) of Table E have contributed to the graduation, and have emerged from the process as cols. (6) and (7) of that table. The observed data, in fact, constitute a contingency table with 2 columns, θ'_x and $E'_x-\theta'_x$, and 14 rows, and we desire to test whether the corresponding 2×14 graduated values in the second half of the table represent a good "fit" by the χ^2 test. In so doing the number of constraints imposed by the method of graduation employed must be taken into account. If, as on the previous supposition, the method provided only for an equality in the totals, the "degrees of freedom" for such a table would be $(p-1)(q-1)$ where $p=2$ and $q=14$, or 13 still. But suppose now—in order to deal with a case which occurs prominently in actuarial practice—that the graduation had been made by fitting Makeham's formula colog $p_x=\alpha+\beta c^x$, and that the fitting had been performed by first choosing c arbitrarily (or assuming it from prior knowledge of other data) as $\log^{-1}.039$, and by imposing an equality in the total frequencies, as before, and now also in the first moments. The arbitrary selection of $\log c$ as .039 means that c has not been determined from the data, and thus has not in any way imposed a constraint, so that no degree of freedom should be deducted; the equality of totals is allowed for in $(p-1)(q-1)=13$; but now an additional constraint results from the use of the first moments, so that the degrees of freedom become $13-1=12$. From a slightly different viewpoint, there are 14 age groups; the formula actually fitted is $\alpha+\beta c^x$, where c is assumed arbitrarily, and two constants α and β are actually determined from the data by moments; there is consequently no constraint in respect of c, but 2 for α and β; the degrees of freedom are therefore $14-(0+2)=12$. The calculation of χ_0^2 must in this case, since both θ'_x and $E'_x-\theta'_x$ are involved, include the $E'_x-\theta'_x$ elements, as shown in cols. (4) and (5) of Table F, which gives χ_0^2 for both the θ'_x and $E'_x-\theta'_x$ series as $24.7582+3.6604=28.4186$. These calculations, which give $\sum\frac{(\theta'_x-\theta''_x)^2}{\theta''_x}+\sum\frac{[(E'_x-\theta'_x)-(E'_x-\theta''_x)]^2}{E'_x-\theta''_x}$, may of course also be made in the form $\sum\left[\frac{(\theta'_x-\theta''_x)^2}{E'_x p''_x q''_x}\right]$ in accordance with the analysis on

p. 245, and as used in P:*125*:35. Here P for the 12 degrees of freedom is .0056; and the conclusion to be drawn from so small a value of P is that there is a real discrepancy, not attributable to chance alone, between the observed and graduated tables, i.e., that the fit of Makeham's formula is, on this criterion, not satisfactory, since deviations as large as or larger than those which have here produced 28.4186 for χ_0^2 would occur from random sampling alone in only about 6 out of 1,000 trials.

This case offers an interesting example of the desirability of proceeding beyond a routine application of the χ^2 test in examining the general goodness of fit of a curve to extensive mortality data. The test here indicates a need for care and further analysis—not necessarily rejection of the graduation. For it must be remembered, as noted at (*iii*) on p. 113, that when the data, as here, are very large, P is often small even though the fit appears to be good. In practice, therefore, an examination should next be made of the general manner in which the deviations between the actual and expected deaths are balanced. Thus it will be seen from the following Table G that the features of the data are well reproduced over broader age groups.

TABLE G

Age Group	Actual Deaths, θ'_x	Expected Deaths, θ''_x	Deviation, $\theta'_x - \theta''_x$
20-39	4,111	4,125	−14
40-59	45,814	45,817	− 3
60-79	77,947	77,952	− 5
80-89	13,949	13,927	+22

The smallness of the deviations over these broad ranges, and the satisfactory agreement which would also be found between the ungraduated and graduated annuity values, would amply justify the acceptance of the graduation in practice—particularly where the facilities of Makeham's formula in the calculation of joint-life annuities might be important.

In the illustration just given it was assumed that c was chosen arbitrarily, without any reference to the data; and under such circumstances it is quite clear that there is no constraint, and consequently that no degree of freedom should be removed, in respect of that constant. If, however, c had been adjusted to the data by inspection, in order to give effect to a realization that a value derived from experiences with other material could not be assumed, the interpretation of the degrees of freedom is not so clear—for it cannot then be felt that there is no constraint at all with regard to c, although it is evident that one degree of freedom should not be deducted, and that the effect upon χ^2 of the adjustment by inspection cannot be measured.

C; 26. Illustrations of the Concept of "Confidence" or "Fiducial" Limits

The charts of Clopper and Pearson (P:*18*:410-411) provide the confidence limits for values of n up to 1,000 with great facility. An example which they give for a small sample illustrates the procedure clearly: Out of 30 individuals, selected at random from a population, 8 are observed to die, so that $\frac{s}{n} = p' = \frac{8}{30} = .267$; within what limits may p be expected to lie? If we are satisfied to accept a risk of error of not more than 1 in 20, or 5%, so that the prediction is to be based on a *confidence coefficient* of .95, we read at once from the .95 chart, for $\frac{s}{n} = .267$, that the lower curve thereon for $n = 30$ gives $p = .12$, and the upper curve gives $p = .46$; that is to say, p, which has been observed in this sample to be .267, may be expected in the long run to lie between the limits .12 and .46. If, on the other hand, we desire to have greater confidence in the prediction, and should only be satisfied with a risk of error of 1 in 100, the chart for the .99 confidence coefficient would be used, from which the limits are seen to be .09 and .52. By similar reasoning the graphs can be employed to determine the size of the sample necessary in order to attain a stated degree of accuracy in estimation (see P:*18*:411-413).

When the smallness of $\frac{s}{n} = p'$ or of q' (<about .03), and of np'

or nq' (< about 10), warns of the doubtful applicability of the normal distribution, and suggests instead the use of the Poisson exponential, Ricker's table in P:*108*:354 may be used conveniently for confidence coefficients of .95 or .99. For the purposes of these confidence limits it may be noted that, in comparison with the approximate condition just stated ($p'<.03$, $np'<10$) for the use of Poisson's distribution, Ricker considers that the Poisson values should be used when $\frac{s}{n} \not> .01$, and may even be preferable up to .05. The close agreement of the Poisson and binomial results for this last value is shown by the example of 5 occurrences in 100 trials, so that $\frac{s}{n}=.05$ and $np'=5$; for if we desire a confidence of 99%, Ricker's Poisson table gives p lying between .01 and .14, whereas Clopper and Pearson's binomial chart indicates limits of .02 and .13.

The use of the preceding charts and tables may be compared with the example at pp. 271-2; C; 6, for large values of n on the basis of the probability integral of the normal curve.

C; 27. Applications of Correlation Theory to Actuarial Problems

A completely worked numerical discussion of the correlation between the mean ages at maturity and the unexpired terms of endowment assurances is shown by Elderton in P:*32*:142-155, 194, and 210-220. Correlation between stature of father and stature of son is noted in P:*177*:199, 211, 238, and 246; correlation between births in a certain district and the proportion of male births per thousand of all births in England and Wales is shown at p. 213, loc. cit.; and the correlation between infantile mortality under 1 year of age and the general rate of mortality at all ages in England and Wales is examined at pp. 292-294 of the same volume. In P:*43*:179-190 the calculations are given for correlation between the statures of fathers and daughters.

Another interesting application is the construction—discussed in P:*31*—of the unknown values of, say, ${}^{I}q_{a}^{s}$ at a particular age

a in the calendar year z for community I, from the known value ${}^{II}q_a^z$ for community II, by calculating the correlation coefficient and the regression equation between the known series ${}^{I}q_a^{z-1}$, ${}^{I}q_a^{z-2}$, ... and the corresponding known series ${}^{II}q_a^{z-1}$, ${}^{II}q_a^{z-2}$, ... for earlier calendar years.

The multiple correlation method has been applied by E. C. Snow (H:*119*; see also P:*167*:61) in the calculation of post-censal population estimates, on the assumption that the increase of population between two censuses may be expressed as a linear function of two or three different variables such as (a) the increase of births during the period over those of the preceding intercensal period, and the similar increase in the deaths, and in the marriages, or (b) the natural increase (i.e., births less deaths), and the increase in the number of inhabited houses, or (c) the increase in the inhabited houses and the increase in rateable values.

BIBLIOGRAPHY

The following abbreviations are used:

J.I.A.	Journal of the Institute of Actuaries (Great Britain).
T.F.A.	Transactions of the Faculty of Actuaries (Scotland).
T.A.S.A.	Transactions of the Actuarial Society of America.
R.A.I.A.	Record of the American Institute of Actuaries.
J.R.S.S.	Journal of the Royal Statistical Society (Great Britain).
J.A.S.A.	Journal of the American Statistical Association.
Ann. Math. Stat.	Annals of Mathematical Statistics.
Phil. Mag.	Philosophical Magazine (London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science).
Phil. Trans.	Philosophical Transactions, Royal Society of London.
Camb. Phil. Trans.	Cambridge Philosophical Transactions.

"H" LIST
OF
PUBLICATIONS OF HISTORICAL SIGNIFICANCE

1657

1. HUYGENS, Christiaan: "De Ratiociniis in Ludo Aleae" (English translations by J. Arbuthnot and W. Browne, London, 1692 and 1714).

1685

2. WALLIS, John: "Algebra".

1708

3. MONTMORT, P. de: "Essai d'Analyse sur les Jeux de Hasard".

1713

4. BERNOULLI, Jacques (James): "Ars Conjectandi". [Edited and published after his death by his nephew Nicholas. A German translation was published in Ostwald's "Klassiker" series in Leipzig in 1899.]

1718

5. DE MOIVRE, Abraham: "The Doctrine of Chances".

1730

6. STIRLING, James: "Methodus Differentialis".

1733

7. DE MOIVRE, Abraham: "Approximatio ad Summam Terminorum Binomii $\overline{a+b}|^{n}$ in Seriem Expansi" (Second Supplement, 1733, to "Miscellanea Analytica", 1730).

1739

8. HUME, David: "Treatise on Human Nature", and "An Enquiry concerning Human Understanding".

1757

9. SIMPSON, Thomas: "Miscellaneous Tracts on some Curious and very Interesting Subjects in Mechanics, Physical Astronomy, and Speculative Mathematics".

1763

10. BAYES, Thomas: "An Essay towards Solving a Problem in the Doctrine of Chances" (*Phil. Trans.*, LIII, 370).

1798

11. MALTHUS, T. R.: "An Essay on the Principle of Population".

1805

12. LEGENDRE, A. M.: "Sur la Méthode des Moindres Carrés" (Appendix to "Nouvelles Méthodes pour la Détermination des Orbites des Comètes"—English translation in H:*164*:576).

1808

13. **Adrain**, Robert: "Research concerning the Probabilities of the Errors which Happen in Making Observations" (*The Analyst, Philadelphia*, I, 93).

1809

14. **Gauss**, Carl Friedrich: "Theoria Motus Corporum Cœlestium" (translated into English by C. H. Davis, 1857).

1812

15. **Laplace**, Pierre Simon, Marquis de: "Théorie Analytique des Probabilités".

1814

16. **Laplace**, Pierre Simon, Marquis de: "Essai Philosophique sur les Probabilités" (English translation, by Truescott and Emory, published in 1902).

1823

17. **Gauss**, Carl Friedrich: "Theoria Combinationis Observationum Erroribus Minimis Obnoxiae" (French translation entitled "Méthode des Moindres Carrés; Mémoires sur la Combinaison des Observations" published in Paris by Bertrand in 1855).

1825

18. **Gompertz**, Benjamin: "On the Nature of the Function Expressive of the Law of Human Mortality" (*Phil. Trans.*, 1825, 513).

1827

19. **Poisson**, S. D.: "Connaissances des Temps" (1827 and 1832).
20. Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, précédées des Règles Générales du Calcul des Probabilités".

1834

21. **Encke**, Johann Franz: "Uber die Methode der kleinsten Quadrate" (Berliner Astronomisches Jahrbuch).

1837

22. **Hagen**, G.: "Grundzuge der Wahrscheinlichkeitsrechnung".

1838

23. **Verhulst**, P. F.: "Notice sur la Loi que la Population Suite dans son Accroissement" (Correspondance Mathématique et Physique Publiée par A. Quetelet—Tome X, 113).

1843

24. **Ellis**, R. L.: "On the Foundations of the Theory of Probabilities" (*Camb. Phil. Trans.*, VIII) [reprinted in "Mathematical and Other Writings", 1863].

1844

25. **Ellis**, R. L.: "On the Method of Least Squares" (*Camb. Phil. Trans.*, VIII, 204).

1845

26. **Verhulst**, P. F.: "Recherches Mathématiques sur la Loi d'Accroissement de la Population" (*Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Bruxelles*, XVIII, 1).

1850

27. Ellis, R. L.: "Remarks on an Alleged Proof of the Method of Least Squares Contained in a Late Number of the Edinburgh Review" (*Phil. Mag.*, XXXVII, 321).

28. Herschel, Sir John F. W.: "On the Theory of Probabilities" (*Edinburgh Review*, July, 1850—reprinted largely in *J.I.A.* XV, 179).

1853

29. Bienaymé, J.: "Considerations à l'Appui de la Découverte de Laplace sur la Loi de Probabilité dans la Méthode des Moindres Carrés" (*Comptes Rendus des Séances de l'Academie des Sciences*, XXXVII, 309; reprinted in *Liouville's Journal de Mathématique*, Second Series, XII, 158).

1854

30. Ellis, R. L.: "Remarks on the Fundamental Principles of the Theory of Probabilities" (*Camb. Phil. Trans.*, IX) [reprinted in "Mathematical and Other Writings", 1863].

1860

31. Makeham, W. M.: "On the Law of Mortality, and the Construction of Annuity Tables" (*J.I.A.* VIII, 301).

1861

32. Airy, Sir George Biddell: "On the Algebraical and Numerical Theory of Errors of Observations, and the Combination of Observations" (second edition, 1875).

1865

33. Todhunter, Isaac: "History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace".

34. Wittstein, Theodor: "On Mathematical Statistics and its Application to Political Economy and Insurance—Translated by T. B. Sprague" (*J.I.A.* XVII, 178, 355, and 417).

1866

35. Venn, J.: "The Logic of Chance".

1867

36. Schiaparelli, G.: "Sul Modo di Ricavare la Vera Espressione delle Leggi della Natura dalle Curve Empiriche".

37. Tchebycheff, P. L.: "Des Valeurs Moyennes" (*Liouville's Journal de Mathématique*, Second Series, XII, 177).

1868

38. Hattendorff, K.: "Das Risiko bei der Lebensversicherung" (*Rundschau der Versicherungen*, E. A. Massius, Jahrgang XVIII, 434).

39. Lazarus, Wilhelm: "On Some Problems in the Theory of Probabilities" (*J.I.A.* XV, 245).

1870

40. Woolhouse, W. S. B.: "Explanation of a New Method of Adjusting Mortality Tables, with Some Observations upon Mr. Makeham's Modification of Gompertz's Theory" (*J.I.A.* XV, 389).

1871

41. Bremiker, C.: "On the Risk Attaching to the Grant of Life Assurances —Translated by T. B. Sprague" (*J.I.A.* XVI, 216 and 285).
42. Crofton, Morgan W.: "On the Proof of the Law of Errors of Observation (*Phil. Trans.*, CLX, 175).
43. Thiele, T. N.: "On a Mathematical Formula to Express the Rate of Mortality throughout the Whole of Life—Translated by T. B. Sprague" (*J.I.A.* XVI, 313).

1872

44. Chandler, S. C., Jr.: "On the Construction of a Graduated Table of Mortality from a Limited Experience" (*J.I.A.* XVII, 161).
45. " " "On the Law of the Ages at which Life Insurances are Effected" (*J.I.A.* XVII, 56).
46. Glaisher, J. W. L.: "On the Law of Facility of Errors of Observation, and on the Method of Least Squares" (*Memoirs of the Royal Astronomical Society*, XXXIX, 75).
47. Makeham, W. M.: "On the Laws of Sickness and Invalidism, and their Relation to the Law of Mortality" (*J.I.A.* XVI, 408).

1873

48. De Forest, Erastus L.: "Additions to a Memoir on Methods of Interpolation Applicable to the Graduation of Irregular Series" (Annual Report of the Smithsonian Institution for 1873, 319).

1876

49. De Forest, Erastus L.: "Interpolation and Adjustment of Series".
50. Helmert, F. R.: "Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit" (*Astronomische Nachrichten*, LXXXVIII, 122).

1877

51. De Forest, Erastus L.: "On Adjustment Formulas" (*The Analyst, Des Moines*, IV, 79 and 107).
52. Lexis, W.: "Zur Theorie der Massenerscheinungen in der menschlichen Gesellschaft".
53. Merriman, Mansfield: "A List of Writings Relating to the Method of Least Squares, with Historical and Critical Notes" (*Transactions of the Connecticut Academy*, IV).

1878

54. De Forest, Erastus L.: "On the Grouping of Signs of Residuals" (*The Analyst, Des Moines*, V, 1).
55. " " "On Repeated Adjustments and on Signs of Residuals" (*The Analyst, Des Moines*, V, 65).
56. Doolittle, M. H.: "U.S. Coast and Geodetic Survey, Report for 1878" (Appendix 8, 115).
57. Gyldén, H.: "Försäkringsföreningens Tidskrift".

1879

58. Balfour, Arthur James (afterwards Earl Balfour): "A Defence of Philosophic Doubt—being an Essay on the Foundations of Belief".

59. Gram, J. P.: "Om Raekkendviklinger, bestemte ved Hjaelp af de mindste Kvadraters Methode" (reprinted in German in *Journal für Mathematik*, XCIV (1894), 41).

60. Kummel, C. H.: "Reduction of Observation Equations which Obtain more than One Observed Quantity" (*The Analyst, Des Moines*, VI, 97).

61. Sprague, T. B.: "On the Construction and Use of a Series of Select Mortality Tables, to be Employed in Combination with the Institute $H^{M (5)}$ Table" (*J.I.A.* XXI, 229).

1880

62. King, George, and Hardy, G. F.: "On the Practical Application of Mr. Makeham's Formula to the Graduation of Mortality Tables" (*J.I.A.* XXII, 191).

1882

63. Ackland, Thomas G.: "On the Graduation of Mortality Tables" (*J.I.A.* XXIII, 352).

64. De Forest, Erastus L.: "On an Unsymmetrical Probability Curve" (*The Analyst, Des Moines*, IX, 135 and 161).

1883

65. Wittstein, Theodor: "The Mathematical Law of Mortality" (*J.I.A.* XXIV, 153 and XXXIII, 399).

1884

66. Karup, Johannes: "Die Ausgleichung der Sterblichkeitserfahrungen der Gothaer Bank nach der Gompertz-Makeham'schen Sterblichkeitsformel" (*Rundschau der Versicherungen*, XXXIV) [see also *J.I.A.* XXIV, 70].

1886

67. Von Kries, J.: "Die Principien der Wahrscheinlichkeit".

1888

68. Hardy, George F.: "Friendly Societies" (*J.I.A.* XXVII, 245).

1889

69. Makeham, W. M.: "On the Further Development of Gompertz's Law" (*J.I.A.* XXVIII, 152, 185, and 316).

1890

70. Makeham, W. M.: "Demonstration of a Formula Relating to the Theory of Errors" (*J.I.A.* XXVIII, 393).

1892

71. Pizzetti, P.: "I Fondamenti Matematici per la Critica dei Risultati Sperimentali" (*Atti della Regia Università di Genova*, XI, 113).

1893

72. Quiquet, Albert: "Représentation Algébrique des Tables de Survie" (*Bulletin de l'Institut des Actuaires Français*, III).

1894

73. HARDY, G. F.: Discussion of "The Application of Makeham's Modification of Gompertz's Expression for the Law of Mortality" (*J.I.A.* XXXI, 359).

1897

74. PEARSON, Karl: "The Chances of Death, and Other Studies in Evolution".

1898

75. BORTKIEWICZ, L. VON: "Das Gesetz der kleinen Zahlen".
76. HARDY, G. F.: "Mortality Experience of Assured Lives and Annuitants in France" (*J.I.A.* XXXIII, 485).
77. " " "Mortality Tables and Frequency Curves" (*J.I.A.* XXXIII, 530).
78. SHEPPARD, W. F.: "On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged according to Equidistant Divisions of a Scale" (*Proceedings of the London Mathematical Society*, XXIX, 353).

1900

79. CALDERON, H. P.: "Some Notes on Makeham's Formula for the Force of Mortality" (*J.I.A.* XXXV, 157).
80. PEARSON, Karl: "On the Criterion that a Given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling" (*Phil. Mag.*, XL (Series 5), 157).

1901

81. GRAM, J. P.: "La Science Actuarielle en Danemark", with Summary in English (*Troisième Congrès d'Actuaires, Documents*, 770).

1902

82. EDGEWORTH, F. Y.: "Law of Error" (Encyclopaedia Britannica, new volumes of 9th edition, XXVIII).
83. ELDERTON, W. Palin: "Graduation and Analysis of a Sickness Table" (*Biometrika*, II, 260 and 503, and III, 52).
84. GLOVER, J. W.: "A Graduation of the American Experience Table of Mortality to Makeham's Formula by the Method of Moments" (*T.A.S.A.* VII, 339).
85. KING, George: "Institute of Actuaries' Text Book, Part II, Life Contingencies" (Second Edition).
86. PEARSON, Karl: "On the Systematic Fitting of Curves to Observations and Measurements" (*Biometrika*, I, 265).

1903

87. BOWLEY, A. L.: "The Measurement of Groups and Series" (A Course of Lectures delivered at the Institute of Actuaries, London).
88. CHATHAM, James: "On the Graduation of the British Offices Annuity Experience (1863-93) by the Graphic Method" (*J.I.A.* XXXVII, 526).
89. ELDERTON, W. P.: "Temporary Assurances" (*J.I.A.* XXXVII, 501).
90. INSTITUTE OF ACTUARIES, AND FACULTY OF ACTUARIES IN SCOTLAND: "British Offices Life Tables, 1893; An Account of the Principles and Methods Adopted in the Compilation of the Data, the Graduation of the Experience, etc.".
91. KAPTEYN, J. C.: "Skew Frequency Curves in Biology and Statistics" (2 papers, in 1903 and 1916).

92. Lidstone, George J.: "Further Remarks on the Valuation of Endowment Assurances in Groups" (*J.I.A.* XXXVIII, 1).
93. Pearson, Karl (Editorial): "On the Probable Errors of Frequency Constants" (Part 1, 1903, *Biometrika*, II, 273; Part 2, 1913, *ibid.* IX, 1; Part 3, 1920, *ibid.* XIII, 113).
94. Thiele, T. N.: "Theory of Observations" (an abridged English edition of the original "Almindelig Iagttagelseslaere", Copenhagen, 1884) [reprinted in *Ann. Math. Stat.*, II (1931), 165].

1904

95. Hardy. George Francis: "The British Offices Life Tables, 1893; Memorandum on the Graduation of the Whole-Life Without-Profit Mortality Table, Male Lives" (*J.I.A.* XXXVIII, 501).
96. Quiquet, Albert: "Sur l'Emploi Simultané de Lois de Survie Distinctes" (*Proceedings of the Fourth International Congress of Actuaries*, I, 382).
97. Spencer, John: "On the Graduation of the Rates of Sickness and Mortality Presented by the Experience of the Manchester Unity of Oddfellows during the Period 1893-97" (*J.I.A.* XXXVIII, 334).
98. Todhunter, R.: Review of "British Offices Life Tables, 1893; An Account of the Principles and Methods, etc." (*J.I.A.* XXXVIII, 356).

1905

99. Cantelli, F. P.: "Sull' Adattamento delle Curve ad una Serie di Misure o di Osservazioni" (reviewed in *J.I.A.* XXXIX, 376).
100. Charlier, C. V. L.: "Über das Fehlergesetz" (*Arkiv för Matematik, Astronomi och Fysik*, II, No. 8, 1); "Die zweite Form des Fehlergesetzes" (*ibid.*); "Researches into the Theory of Probability" (*Meddelanden Lunds Astronomiska Observatorium*, 1906); "Vorlesungen über die Grundzüge der Matematischen Statistik" (2nd edition, 1920); "A New Form of the Frequency Function" (*Meddelanden Lunds Astronomiska Observatorium*, 1928).
101. Elderton, W. P.: Review of "On Fitting Curves to a Series of Measurements or Observations" (*J.I.A.* XXXIX, 376).

1906

102. Bruns, H.: "Wahrscheinlichkeitsrechnung und Kollektivmasslehre".
103. Henderson, Robert: "A Practical Interpolation Formula" (*T.A.S.A.* IX, 211).
104. Rosmanith, G.: "Die verschiedenen Methoden der Anwendung der Gompertz-Makehamschen Formel" (*Proceedings, Fifth International Congress of Actuaries*, II, 317).
105. Steffensen, J. F.: "Notes on the Practical Graduation of Life Insurance Tables" (*Proceedings, Fifth International Congress of Actuaries*, II, 247).

1907

106. Ackland, Thomas G.: "Notes on the British Offices Life Annuity Tables (1893)" (*T.F.A.* III, 285).
107. Bowley, A. L.: "Elements of Statistics" (Third Edition).
108. Jensen, Christian: "Mortality Table for Female Beneficiaries in Survivorship Annuities" (*T.A.S.A.* X, 253).
109. King, George: "Notes on Summation Formulas of Graduation, with Certain New Formulas for Consideration" (*J.I.A.* XLI, 530).
110. " " "On the Error Introduced into Mortality Tables by Summation Formulas of Graduation" (*J.I.A.* XLI, 54).

111. LIDSTONE, G. J.: "On the Rationale of Formulas for Graduation by Summation" (*J.I.A.* XLI, 348, and XLII, 106).

112. SPENCER, John: "Some Illustrations of the Employment of Summation Formulas in the Graduation of Mortality Tables" (*J.I.A.* XLI, 361).

113. TODHUNTER, R.: Review of "Frequency Curves and Correlation" (*J.I.A.* XLI, 443).

1908

114. BUCHANAN, James: "Osculatory Interpolation by Central Differences, with an Application to Life Table Construction" (*J.I.A.* XLII, 369).

115. CZUBER, E.: "Wahrscheinlichkeitsrechnung".

116. MARKOFF, A. A.: "Calculus of Probability" (in Russian; German edition, "Wahrscheinlichkeitsrechnung", published in 1912; 4th Russian edition, 1924; English edition by Professor F. M. Weida in preparation).

117. "STUDENT": "The Probable Error of a Mean" (*Biometrika*, VI, 1).

1909

118. TODHUNTER, R.: Review of "The Theory of the Construction of Tables of Mortality, etc." by G. F. Hardy (*J.I.A.* XLIII, 471).

1911

119. SNOW, E. C.: "The Application of the Method of Multiple Correlation to the Estimation of Post-Censal Populations" (*J.R.S.S.* LXXIV, 575).

120. THOMPSON, J. S.: "A Determination of the Constants in Makeham's Formula by the Method of Least Squares" (*T.A.S.A.* XII, 225).

1912

121. ACTUARIAL ADVISORY COMMITTEE: "Life Tables adopted in Computing Reserve Values—National Insurance Act, Part I, Health Insurance" (Cmd. 6907—extracts printed in *J.I.A.* XLVII, 548).

1913

122. ACKLAND, T. G.: "On the Estimated Age Distribution of the Indian Population, as Recorded at the Census of 1911, and the Estimated Rates of Mortality Deduced from a Comparison of the Census Returns for 1901 and 1911" (*J.I.A.* XLVII, 315).

123. SHOVELTON, S. T.: "On the Graduation of Mortality Tables by Interpolation" (*J.I.A.* XLVII, 285).

1915

124. ACKLAND, T. G.: "On Osculatory Interpolation, where the Given Values of the Function are at Unequal Intervals" (*J.I.A.* XLIX, 369).

1916

125. JÖRGENSEN, N. R.: "Undersögelser over Frequensflader og Korrelation".

1917

126. KNIBBS, G. H.: "The Mathematical Theory of Population, of its Character and Fluctuations, and of the Factors which Influence them".

127. STEFFENSEN, J. F.: "On a Formula Facilitating the Application of Select Mortality for All Durations" (*Svenska Aktuarieföreningens Tidskrift*, 1917).

1918

128. PEARSON, Karl: "On Generalized Tchebycheff Theorems in the Mathematical Theory of Statistics" (*Biometrika*, XII, 284).

1919

129. CARVER, H. C.: "On the Graduation of Frequency Distributions" (*Proceedings of the Casualty Actuarial and Statistical Society of America*, VI, 52).
130. FISHER, Arne: "Note on the Construction of Mortality Tables by Means of Compound Frequency Curves" (*Proceedings of the Casualty Actuarial and Statistical Society of America*, VI).

1920

131. HOWELL, Valentine: "Two Graduations of the American-Canadian Mortality Experience" (*T.A.S.A.* XXI, 178).
132. LIDSTONE, G. J.: "An Elementary Demonstration of Stirling's Approximate Formula for the Value of Factorial n" (*J.I.A.* LII, 102, 161, and 574).
133. PEARL, Raymond, and REED, Lowell J.: "On the Rate of Growth of the Population of the United States since 1790 and its Mathematical Representation" (*Proceedings of the National Academy of Sciences*, VI, 275).
134. STEWART, R. Meldrum: "Adjustment of Observations" (*Phil. Mag.*, (6) XL, 217).
135. TRACHTENBERG, H. L.: "The Relation of Life Tables to the Makeham Law" (*J.R.S.S.* LXXXIII, 656).

1922

136. CAMP, B. H.: "A New Generalization of Tchebycheff's Statistical Inequality" (*Bulletin of the American Mathematical Society*, XXVIII, 427).
137. FISHER, Arne: "An Elementary Treatise on Frequency Curves and their Application in the Analysis of Death Curves and Life Tables".
138. FISHER, R. A.: "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P" (*J.R.S.S.* LXXXV, 87).
139. RIETZ, H. L.: "Frequency Distributions Obtained by Certain Transformations of Normally Distributed Variates" (*Annals of Mathematics*, XXIII, 292).
140. YULE, G. UDNY: "On the Application of the χ^2 Method to Association and Contingency Tables, with Experimental Illustrations" (*J.R.S.S.* LXXXV, 95).

1923

141. ELSTON, J. S.: "Survey of Mathematical Formulas that have been Used to Express a Law of Mortality" (*R.A.I.A.* XII, 66).
142. INSTITUTE OF ACTUARIES, AND FACULTY OF ACTUARIES IN SCOTLAND: "Mortality of Life Annuitants, 1900-1920".
143. UHLER, H. S.: "Method of Least Squares and Curve Fitting" (*Journal of the Optical Society of America and Review of Scientific Instruments*, VII, 1043).
144. WHITTAKER, E. T.: "On a New Method of Graduation" (*Proceedings of the Edinburgh Mathematical Society*, XLI, 63).
145. " " "On the Theory of Graduation" (*Proceedings of the Royal Society of Edinburgh*, XLIV, 77).
146. WICKSELL, S. D.: "Contributions to the Analytical Theory of Sampling" (*Arkiv för Matematik, Astronomi och Fysik*, XVII, 1).

1924

147. **Brownlee**, John: "Some Experiments to Test the Theory of Goodness of Fit" (*J.R.S.S.* LXXXVII, 76).
148. **Elderton**, W. Palin: "Mathematical Law of Mortality—A Suggestion" (*Proceedings of the International Mathematical Congress, Toronto*, II, 867).
149. **Reilly**, J. F.: "Certain Generalizations of Osculatory Interpolation" (*R.A.I.A.* XIII, 4).
150. **Romanovsky**, V.: "Generalization of Some Types of the Frequency Curves of Professor Pearson" (*Biometrika*, XVI, 106).
151. **Watson**, Sir Alfred W., and **Weatherill**, H.: "Mortality Experience of Government Life Annuitants, 1900-1920" (a résumé appears in *J.I.A.* LV, 144).

1925

152. **Aitken**, A. C.: "On the Theory of Graduation" (*Proceedings of the Royal Society of Edinburgh*, XLVI, 36).
153. **Buchanan**, J.: "Notes on Graduation" (*T.F.A.* X, 289).
154. **Reilly**, J. F.: "On Lidstone's Demonstration of the Osculatory Interpolation Formula" (*R.A.I.A.* XIV, 12).
155. **Traversi**, A. T.: "Note on the Construction of the [Life] Tables" (Government Statistician's General Report on the Census of 1921, New Zealand—Section XXII).
156. **Tschuprow**, A. A.: "Grundbegriffe und Grundprobleme der Korrelationstheorie".

1926

157. **Buchanan**, James: "The Theory of Selection; its History and Development" (*T.F.A.* XI, 43).
158. **Fisher**, R. A.: "Applications of 'Student's' Distribution" (*Metron*, V, 90).
159. **Reilly**, J. F.: "Osculatory Interpolation with Unequal Intervals" (*R.A.I.A.* XV, 34).

1927

160. **Keffer**, Ralph: "Group Sickness and Accident Insurance" (*T.A.S.A.* XXVIII, 5).
161. **Reitz**, H. L.: "On Certain Properties of Frequency Distributions of the Powers and Roots of Variates of a Given Distribution" (*Proceedings of the National Academy of Sciences*, XII, 817.

1928

162. **Bowley**, A. L.: "F. Y. Edgeworth's Contributions to Mathematical Statistics".
163. **Mises**, R. **Von**: "Wahrscheinlichkeit, Statistik, und Wahrheit" (First edition, 1928; second edition, 1936; translated into English by J. Neyman, D. Sholl, and E. Rabinowitsch as "Probability, Statistics, and Truth", 1939)
164. **Smith**, David Eugene: "A Source Book in Mathematics".

1930

165. **Berkson**, J.: "Bayes' Theorem" (*Ann. Math. Stat.*, I, 42).
166. **Molina**, E. C.: "The Theory of Probability; Some Comments on Laplace's Théorie Analytique" (*Bulletin of the American Mathematical Society*, XXXVI, 369; reprinted by Bell Telephone Laboratories, New York).
167. **Mouzon**, E. D. Jr.: "Equimodal Frequency Distributions" (*Ann. Math. Stat.*, I, 137).

168. Nekrassoff, V. A.: "Nomography in Applications of Statistics" (*Metron*, VIII, 95).

169. Steffensen, J. F.: "Infantile Mortality from an Actuarial Point of View" (*Skandinavisk Aktuarietidskrift*, 1930).

170. Will, Harry S.: "On Fitting Curves to Observational Series by the Method of Differences" (*Ann. Math. Stat.*, I, 159).

1931

171. Macaulay, Frederick R.: "The Smoothing of Time Series".

172. Mises, R. von: "Wahrscheinlichkeitsrechnung".

173. Molina, E. C.: "Bayes' Theorem; An Expository Presentation" (*Ann. Math. Stat.*, II, 23).

174. Pearson, Karl: "Historical Note on the Distribution of the Standard Deviation of Samples of any Size drawn from an Indefinitely Large Normal Parent Population" (*Biometrika*, XXIII, 416).

175. Rietz, H. L.: "On Certain Properties of Frequency Distributions Obtained by a Linear Fractional Transformation of the Variates of a Given Distribution" (*Ann. Math. Stat.*, II, 38).

176. Winfrey, Robley, and Kurtz, Edwin B.: "Life Characteristics of Physical Property" (Iowa Engineering Experiment Station, Iowa State College, Bulletin 103).

1932

177. Haldy, M.: "Influence des Variations de l'Invalidité sur les Réserves Mathématiques" (*Mitteilungen der Vereinigung sweiserischer Versicherungsmathematiker*, XXVII).

1933

178. Kolmogoroff, A.: "Grundbegriffe der Wahrscheinlichkeitsrechnung".

179. Mallock, R. R.: "An Electrical Calculating Machine" (*Proceedings of the Royal Society of London*, Series A, CXL, 457).

180. O'Toole, A. L.: "On the System of Curves for which the Method of Moments is the Best Method of Fitting" (*Ann. Math. Stat.*, IV, 1); "A Method of Determining the Constants in the Bimodal Fourth Degree Exponential Function" (*ibid.* IV, 79).

1934

181. Elderton, W. Palin: "An Approximate Law of Survivorship, and Other Notes on the Use of Frequency Curves in Actuarial Statistics" (*J.I.A.* LXV, 1).

182. Hansmann, G. H.: "On Certain Non-Normal Symmetrical Frequency Distributions" (*Biometrika*, XXVI, 129).

183. Schweer, W.: "Eine Graphische Methode zur Berechnung der Konstanten der Makeham-Gompertzchen Formel" (*Transactions of the Tenth International Congress of Actuaries*, V, 292, with English summary).

1935

184. Winfrey, Robley: "Statistical Analyses of Industrial Property Retirements" (Iowa Engineering Experiment Station, Iowa State College, Bulletin 125).

1936

185. Harper, Floyd S.: "An Actuarial Study of Infant Mortality" (*Skandinavisk Aktuarietidskrift*, 1936).

1937

186. Copeland, A. H.: "Consistency of the Conditions Determining Kollektivs" (*Transactions of the American Mathematical Society*, XLII, 333).

187. Lidstone, G. J.: "Some Properties of Makeham's Second Law of Mortality, and the Double and Triple Geometric Laws" (*J.I.A.* LXVIII, 535).

1938

188. Pearson, E. S.: "Karl Pearson; An Appreciation of some Aspects of his Life and Work".

1939

189. McMullen, L. and Pearson, E. S.: "William Sealey Gosset, 1876-1937" (*Biometrika*, XXX, 205).

1940

190. Deming, W. Edwards: "Facsimiles of Two Papers by Bayes, with Commentaries".

"P" LIST
OF
PUBLICATIONS OF PRESENT VALUE

1. AITKEN, A. C.: "The Accurate Solution of the Difference Equation Involved in Whittaker's Method of Graduation, and its Practical Application" (*T.F.A.* XI, 31); 1926.
2. " " "On the Graduation of Data by the Orthogonal Polynomials of Least Squares" (*Proceedings of the Royal Society of Edinburgh*, LIII, 54); 1933.
3. " " "Statistical Mathematics" (Oliver and Boyd, Edinburgh and London); 1939.
4. ALLEN, R. G. D.: "Mathematical Analysis for Economists" (Macmillan and Co., London); 1938.
5. ASSOCIATION OF LIFE INSURANCE MEDICAL DIRECTORS AND ACTUARIAL SOCIETY OF AMERICA: "Occupation Study; Report of the Joint Committee on Mortality"; 1929.
6. ASSOCIATION OF LIFE INSURANCE MEDICAL DIRECTORS AND ACTUARIAL SOCIETY OF AMERICA: "Medical Impairment Study; Report of the Joint Committee on Mortality"; 1931.
7. BARTLETT, M. S.: "The Effect of Non-Normality on the *t*-Distribution" (*Proceedings, Cambridge Philosophical Society*, XXXI, 223); 1935.
8. BATEN, W. D.: "Elementary Mathematical Statistics" (John Wiley and Sons, New York, and Chapman and Hall, London); 1938.
9. BERGSTRESSER, R. L.: Discussion of "The Whittaker-Henderson Graduation Formula A" (*T.A.S.A.* XXXIX, 52); 1938.
10. BERKSON, Joseph: "Some Difficulties of Interpretation Encountered in the Application of the Chi-Square Test" (*J.A.S.A.* XXXIII, 526); 1938.
11. BIRGE, Raymond T., and SHEA, John D.: "A Rapid Method for Calculating the Least Squares Solution of a Polynomial of Any Degree" (*University of California Publications in Mathematics*, II, No. 5); 1927.
12. BOWERMAN, Walter G.: "Henderson's Mechanico-Graphic Method of Graduation" (*T.A.S.A.* XXXVIII, 7); 1937.
13. BRUNT, David: "The Combination of Observations" (Second Edition; Cambridge University Press, Cambridge and London); 1931 (First Edition, 1917).
14. BUCHANAN, James: "Recent Developments of Osculatory Interpolation, with Applications to the Construction of National and Other Life Tables" (*T.F.A.* XII, 117); 1929.
15. " " Discussion of "Graduation by the General Formulae of Osculatory Interpolation" (*T.F.A.* XIV, 208); 1933.
16. CAMP, B. H.: "The Mathematical Part of Elementary Statistics" (D. C. Heath and Co., New York); 1931.
17. CAMP, Kingsland: Discussion of "Henderson's Mechanico-Graphic Method of Graduation" (*T.A.S.A.* XXXVIII, 512); 1937.
18. CLOPPER, C. J., and PEARSON, E. S.: "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial" (*Biometrika*, XXVI, 404); 1934.

19. Cody, Donald D.: "The Standard Deviation in the Rate of Mortality by Amounts" (*T.A.S.A.* XLII, 69); 1941.
20. Comrie, L. J.: "Tables of $\tan^{-1}x$ and $\log\ (1+x^2)$" (Tracts for Computers, No. XXIII; Department of Statistics, University of London); 1938.
21. Courant, R.: "Differential and Integral Calculus" (Two volumes, New Revised Edition; Nordemann Publishing Company, New York); 1937.
22. Cramér, Harald: "Random Variables and Probability Distributions" (Cambridge Tracts in Mathematics and Mathematical Physics, No. 36) [See also review in *J.I.A.* LXIX, 193]; 1937.
23. Cramér, H., and Wold, H.: "Mortality Variations in Sweden; a Study in Graduation and Forecasting" (*Skandinavisk Aktuarietidskrift* 1935, 161); 1935.
24. Davidson, Andrew R., and Reid, A. R.: "A New Type of Summation Graduation Formulae Related to Whittaker's Analytical Formula" (*T.F.A.* XI, 1); 1926.
25. Davis, H. T.: "Polynomial Approximation by the Method of Least Squares" (*Ann. Math. Stat.*, IV, 155); 1933.
26. Davis, Mervyn: Discussion of "A New Method of Graduation" (*T.A.S.A.* XXV, 300); 1924.
27. Davis, H. T., and Nelson, W. F. C.: "Elements of Statistics" (Principia Press, Bloomington, Indiana); 1935.
28. Deming, W. Edwards: "Some Notes on Least Squares" (Graduate School, Department of Agriculture, Washington, D.C.); 1938.
29. Deming, W. Edwards, and Birge, Raymond T.: "On the Statistical Theory of Errors" [Reprinted from *Reviews of Modern Physics*, VI, 119, with Additional Notes dated 1937 and 1938] (Graduate School, Department of Agriculture, Washington, D.C.); 1938.
30. Dodd, E. L.: Review of Mises' "Wahrscheinlichkeit, Statistik, und Wahrheit" (*J.A.S.A.* XXXI, 758); 1936.
31. Dublin, L. I., Lotka, A. J., and Spiegelman, M.: "The Construction of Life Tables by Correlation" (*Metron*, XII, 121); 1935.
32. Elderton, W. Palin: "Frequency Curves and Correlation" (Third Edition; Cambridge University Press, Cambridge and London); 1938 (First Edition, 1906).
33. Elderton, W.P., and Rowland, S. J.: "Graduation by Makeham's Hypothesis" (*J.I.A.* L, 251); 1917.
34. Elston, James S., and Others: "Sources and Characteristics of the Principal Mortality Tables" [Actuarial Studies, No. 1, Revised Edition] (Actuarial Society of America, New York); 1932.
35. Fieller, E. C.: "Recent Advances in Probability Theory—The Law of Large Numbers and other Investigations in Probability" (*J.R.S.S.* XCIX, 715); 1936.
36. Fisher, Arne: "The Mathematical Theory of Probabilities" (Second Edition; Macmillan Company, New York); 1923.
37. Fisher, R. A.: "On the Mathematical Foundations of Theoretical Statistics" (*Phil. Trans.*, Series A, CCXXII, 309); 1922.
38. " " "On a Distribution Yielding the Error Functions of Several Well Known Statistics" (*Proceedings of the International Mathematical Congress*, Toronto, II, 805); 1924.
39. " " "Applications of 'Student's' Distribution" (*Metron*, V, No. 3, 90); 1925.

40. FISHER, R. A.: "Moments and Product Moments of Sampling Distributions" (*Proceedings of the London Mathematical Society*, Series 2, XXX, 199); 1929.

41. " " "The Concepts of Inverse Probability and Fiducial Probability Referring to Unknown Parameters" (*Proceedings, Royal Society of London*, Series A, CXXXIX, 343); 1933; and "Probability, Likelihood, and Quantity of Information in the Logic of Uncertain Inference (*ibid.* CXLVI, 1); 1934.

42. " " "The Logic of Inductive Inference" (*J.R.S.S.* XCVIII, 39); 1935.

43. " " "Statistical Methods for Research Workers" (Seventh Edition; Oliver and Boyd, Edinburgh and London); 1938 (First published 1925).

44. FORSYTH, C. H.: "Simple Derivation for the Formulas for the Dispersion of Statistical Series" (*American Mathematical Monthly*, XXXI); 1924.

45. GARWOOD, F.: "Fiducial Limits for the Poisson Distribution" (*Biometrika*, XXVIII, 437); 1936.

46. GERHARD, F. Bruce: "A Graphic Method of Applying Makeham's Formula to Mortality Experience" (*T.A.S.A.* XXIV, 398); 1923.

47. GLOVER, James W.: "Tables of Applied Mathematics, Finance, Insurance, and Statistics" (George Wahr, Ann Arbor, Michigan); 1923.

48. GOULDEN, C. H.: "Methods of Statistical Analysis" (John Wiley and Sons, New York, and Chapman and Hall, London); 1939.

49. GREENWOOD, Major: "English Death Rates, Past, Present, and Future" (*J.R.S.S.* XCIX, 674); 1936.

50. HARDY, G. F.: "Graduation Formulas" (*J.I.A.* XXXII, 371); 1896.

51. " " "The Theory of the Construction of Tables of Mortality and of Similar Statistical Tables in Use by the Actuary" (Institute of Actuaries, London); 1909.

52. HENDERSON, R.: "Frequency Curves and Moments" (*T.A.S.A.* VIII, 30; reprinted in *J.I.A.* XLI, 429); 1904.

53. " " "Note on Limit of Risk" (*T.A.S.A.* IX, 40); 1905.

54. " " "Central Difference Interpolation" (*T.A.S.A.* XXII, 175); 1921.

55. " " "The Logical Basis of the Theory of Probabilities" (*T.A.S.A.* XXIII, 305); 1922.

56. " " "A New Method of Graduation" (*T.A.S.A.* XXV, 29); 1924.

57. " " "Further Remarks on Graduation" (*T.A.S.A.* XXVI, 52); 1925.

58. " " Discussion of "The Whittaker-Henderson Graduation Formula A" (*T.A.S.A.* XXXIX, 50); 1938.

59. " " "The Mathematical Theory of Graduation" [Actuarial Studies, No. 4] (Actuarial Society of America, New York); 1939 (First Edition, 1919).

60. HOTELLING, Harold: "Differential Equations Subject to Error, and Population Estimates" (*J.A.S.A.* XXII, 283); 1927.

61. IMMERWAHR, George: Discussion of "Henderson's Mechanico-Graphic Method of Graduation" (*T.A.S.A.* XXXVIII, 516); 1937.

62. IRWIN, J. O.: "Mathematical Theorems in the Analysis of Variance" (*J.R.S.S.* XCIV, 284); 1931.

63. " " "Recent Advances in Mathematical Statistics (1931)" (*J.R.S.S.* XCV, 498); 1932.

106. REID, A. R., and DOW, J. B.: "Graduation by the General Formulae of Osculatory Interpolation" (*T.F.A.*. XIV, 185); 1933.
107. RHODES, E. C.: "On the Generalized Law of Error" (*J.R.S.S.* LXXXVIII, 576); 1925.
108. RICKER, William E.: "The Concept of Confidence or Fiducial Limits Applied to the Poisson Frequency Distribution" (*J.A.S.A.* XXXII, 349); 1937.
109. RIDER, Paul R.: "A Generalized Law of Error" (*J.A.S.A.* XIX, 217); 1924.
110. " " "A Survey of the Theory of Small Samples" (*Annals of Mathematics*, 2nd Series, XXXI, 577); 1930.
111. " " "A Note on Small Sample Theory" (*J.A.S.A.* XXVI, 172); 1931.
112. " " "An Introduction to Modern Statistical Methods" (John Wiley and Sons, New York); 1939.
113. RIETZ, H. L.: "On the Mathematical Theory of Risk and Landré's Theory of the Maximum" (*R.A.I.A.* II, 1); 1913.
114. " " "Handbook of Mathematical Statistics" (Houghton Mifflin Co., Boston and New York); 1924.
115. " " "On Certain Applications of Mathematical Statistics to Actuarial Data" (*R.A.I.A.* XIII, 214); 1924.
116. " " "Mathematical Statistics" (Carus Mathematical Monographs, No. 3; Open Court Publishing Co., Chicago); 1927.
117. " " "On the Risk Problem from a Mathematical Point of View" (*Transactions, Ninth International Congress of Actuaries*, II, 294); 1930.
118. " " "Comments on Applications of Recently Developed Theory of Small Samples" (*J.A.S.A.* XXVI, 150); 1931.

119. RUTHERFORD, C. D.: "An Annuity Table Complying with the Requirements of the New Canadian Valuation Standard" (*T.A.S.A.* XXVIII, 54); 1927.
120. SALVOSA, L. R.: "Tables of Pearson's Type III Function" (*Ann. Math. Stat.*, I, Nos. 2 and 3); 1930.
121. SASULY, Max: "Trend Analysis of Statistics" (Brookings Institution, Washington, D.C.); 1934.
122. SCARBOROUGH, James B.: "Numerical Mathematical Analysis" (Johns Hopkins Press, Baltimore; Oxford University Press, London); 1930.
123. SCHULTZ, Henry: Discussion on "Applications of the Theory of Error to the Interpretation of Trends" (*J.A.S.A.* XXIV, 86); 1929.
124. " " "The Standard Error of a Forecast from a Curve" (*J.A.S.A.* XXV, 139); 1930.
125. SEAL, H. L.: "Tests of a Mortality Table Graduation" (*J.I.A.* LXXI, 5); 1940.
126. SHANNON, S.: "An Alternative Method of Solution of Certain Fundamental Problems in the Individual Theory of Risk" (*R.A.I.A.* XXVII, 372); 1938.
127. " " Discussion of "A Rapid Method of Graduating Select Mortality Tables" (*R.A.I.A.* XXVII, 245); 1938.
128. SHEPPARD, W. F.: "The Relation between Probability and Statistics" (*T.F.A.* XII, 25); 1928.
129. " " "The Probability Integral" [Vol. VII of the British Association Mathematical Tables] (Cambridge University Press, Cambridge and London); 1939.

130. SHEWHART, Walter A.: "Statistical Method from the Viewpoint of Quality Control" [Edited by W. Edwards Deming] (Graduate School, Department of Agriculture, Washington, D.C.); 1939.
131. SNEDECOR, George W.: "Statistical Methods Applied to Experiments in Agriculture and Biology" (Collegiate Press, Ames, Iowa); 1938.
132. SOPER, H. E.: "Frequency Arrays" (Cambridge University Press, Cambridge and London); 1922.
133. SPOERL, Charles A.: "Actuarial Note: Henderson's Graduation Formula B" (*T.A.S.A.* XXXII, 60); 1931.
134. " " "The Whittaker-Henderson Graduation Formula A" (*T.A.S.A.* XXXVIII, 403); 1937.
135. STEFFENSEN, J. F.: "On the Fitting of Makeham's Curve to Mortality Observations" (*Proceedings of the Fifth International Congress of Mathematicians*, Cambridge, II, 389); 1912.
136. " " "On the Graduation of Mortality Tables by G. F. Hardy's Modification of the Method of Moments" (*Svenska Aktuarieforeningens Tidskrift*, I); 1915.
137. " " "On the Relation between the Method of Moments and the Method of Least Squares" (*J.I.A.* XLIX, 355); 1915.
138. " " "On Certain Inequalities and Methods of Approximation" (*J.I.A.* LI, 274); 1919.
139. " " "On Hattendorff's Theorem in the Theory of Risk" (*Skandinavisk Aktuarietidskrift*, I-II, 1); 1929.
140. " " "Some Recent Researches in the Theory of Statistics and Actuarial Science" (Institute of Actuaries, London; Cambridge University Press, London); 1930.
141. STEVENS, W. L. and IRWIN, J. O.: "Recent Advances in Mathematical Statistics (1934)" (*J.R.S.S.* XCIX, 752); 1936.
142. TINNER, Thomas: "On the Average Value and the Standard Deviation of a Life Annuity Based on a Given Experience" (*J.I.A.* LVI, 301); 1925.
143. TIPPETT, L. H. C.: "The Methods of Statistics" (Williams and Norgate, London); 1931 (Second Edition, 1937).
144. TRACHTENBERG, H. L.: "The Wider Application of the Gompertz Law of Mortality" (*J.R.S.S.* LXXXVII, 278); 1924.
145. UNSIGNED: Review of "Actuarial Studies Nos. 1 and 4 of the Actuarial Society" (*J.I.A.* LI, 367); 1919.
146. USPENSKY, J. V.: "Introduction to Mathematical Probability" (McGraw-Hill Book Co., New York); 1937.
147. VAUGHAN, HUBERT: "Summation Formulas of Graduation with a Special Type of Operator" (*J.I.A.* LXIV, 428); 1933.
148. " " "Further Enquiries into the Summation Method of Graduation" (*J.I.A.* LXVI, 463); 1935.
149. WALKER, Helen M.: "Studies in the History of Statistical Method" (Williams and Wilkins, Baltimore); 1929.
150. WATSON, A. D.: "Summation Formulae with Second Difference Errors" (*J.I.A.* L, 259); 1917.
151. " " "Summation Formulae Introducing Second Difference Errors, and Formulae Correct to Third Differences Derived Therefrom" (*J.I.A.* L, 305); 1917.
152. " " Discussion of "A Guide to Graphic Graduation" (*T.A.S.A.* XXXVIII, 521); 1937.

153. WELLS, E. H.: "A Rapid Method of Graduating Select Mortality Tables" (*R.A.I.A.* XXVI, 551); 1937.

154. WHITTAKER, E. T.: "On Some Disputed Questions of Probability" (*T.F.A.* VIII, 163); 1920.

155. WHITTAKER, E. T. and ROBINSON, G.: "The Calculus of Observations" (Blackie and Son, London); 1924.

156. WILKS, S. S.: "The Theory of Statistical Inference" (Princeton University Press, Princeton, N.J.); 1937.

157. WILSON, E. B.: "First and Second Laws of Errors" (*J.A.S.A.* XVIII, 841); 1923.

158. " " "The Standard Deviation of Sampling for Life Expectancy" (*J.A.S.A.* XXXIII, 705); 1938.

159. WILSON, E. B. and PUFFER, Ruth: "Least Squares and Laws of Population Growth" (*Proceedings of the American Academy of Arts and Sciences*, LXVIII, 285); 1933.

160. WISHART, J.: "Notes on Frequency Constants" (*J.I.A.* LXII, 174); 1931.

161. WOLFENDEN, Hugh H.: Discussion of "Note on Mean Population" (*T.A.S.A.* XX, 218 and 296); 1919.

162. " " Discussion of "Two Graduations of the American Canadian Mortality Experience" (*T.A.S.A.* XXI, 546); 1920.

163. " " "On the Determination of the Rates of Mortality at Infantile Ages from Statistics of the General Population" (*T.A.S.A.* XXIV, 126); 1923.

164. " " "On the Methods of Comparing the Mortalities of Two or More Communities, and the Standardization of Death Rates" (*J.R.S.S.* LXXXVI, 399); 1923.

165. " " Review of "U.S. Abridged Life Tables, 1919-1920" (*T.A.S.A.* XXV, 145); 1924.

166. " " "On the Development of Formulae for Graduation by Linear Compounding, with Special Reference to the Work of Erastus L. DeForest" (*T.A.S.A.* XXVI, 81); 1925.

167. " " "Population Statistics and their Compilation" [Actuarial Studies, No. 3] (Actuarial Society of America, New York); 1925.

168. " " Review of Report by H. G. W. Meikle on the "Age Distribution and Rates of Mortality Deduced from the Indian Census Returns of 1921 and Previous Enumerations" (*T.A.S.A.* XXVII, 467); 1926.

169. " " Review of "The Life Tables, 1926, of the Association of German Life Assurance Companies" (*T.A.S.A.* XXVIII, 135); 1927.

170. " " Review of "Census of England and Wales, 1921: Life Tables, etc." (*T.A.S.A.* XXIX, 327); 1928.

171. " " Discussion of "Canadian Life Tables from Census Returns" (*T.A.S.A.* XXXV, 281); 1934.

172. " " Discussion of "Errors and Bias in the Reporting of Ages in Census Data" (*T.A.S.A.* XLII, 78); 1941.

173. WOODWARD, R. S.: "Probability and Theory of Errors" (John Wiley and Sons, New York); 1906.
174. WOOLHOUSE, W. S. B.: "On the Philosophy of Statistics" (*J.I.A.* XVII, 37); 1872.
175. YUAN, Pae-Tsi: "On the Logarithmic Frequency Distribution and the Semi-Logarithmic Correlation Surface" (*Ann. Math. Stat.*, IV, 30); 1933.
176. YULE, G. Udny: "The Growth of Population, and the Factors which Control It" (*J.R.S.S.* LXXXVIII, 1); 1925.
177. YULE, G. Udny, and KENDALL, M. G.: "An Introduction to the Theory of Statistics" (Eleventh Edition; Charles Griffin and Co., London); 1937.

INDEX

INDEX