ACTUARIAL STATISTICS

Volume II

CONSTRUCTION OF MORTALITY AND OTHER TABLES

BY

J. L. ANDERSON, B.A., F.I.A. and
J. B. DOW, M.A., F.F.A.

ACTUARIAL STATISTICS

COORDINATING EDITOR:
HARRY FREEMAN, M.A., F.I.A.

VOLUME I
STATISTICS & GRADUATION

by

H. TETLEY, M.A., F.I.A.

CAMBRIDGE
Published for the Institute of Actuaries and the
Faculty of Actuaries
AT THE UNIVERSITY PRESS
1953

PUBLISHED BY
THE SYNDICS OF THE CAMBRIDGE UNIVERSITY PRESS

London Office: Bentley House, N.W.1
American Branch: New York

Agents for Canada, India, and Pakistan: Macmillan

First Edition 1946
Second Edition 1950
Reprinted 1953

Printed in Great Britain at the University Press, Cambridge
(Brooke Crutchley, University Printer)

CONTENTS

STATISTICS

CHAPTER I

CONTINUOUS AND DISCONTINUOUS VARIABLES. GROUPED DATA

CHAPTER II

IMPORTANT FREQUENCY DISTRIBUTIONS

CHAPTER III

CORRELATION

CHAPTER IV

SAMPLING

GRADUATION

CHAPTER V

GENERAL CONSIDERATIONS AND TESTS

CHAPTER VI

THE GRAPHIC METHOD

CHAPTER VII

GRADUATION BY REFERENCE TO A STANDARD TABLE

CHAPTER VIII

GRADUATION BY A SUMMATION FORMULA

CHAPTER IX

GRADUATION BY MATHEMATICAL FORMULAE. MAKEHAM AND ALLIED CURVES

CHAPTER X

PIVOTAL VALUES AND OSCULATORY INTERPOLATION

EDITOR'S PREFACE

THIS BOOK was commenced in 1939. The actual writing was completed in 1941 and the first proofs were available during the next year. Owing however to war conditions and pressure of other work the editing took a considerable time and it was not until the summer of 1945 that there seemed to be any prospect of early publication.

Of the book itself I can only say that it supplies a long-felt want. Mr Tetley's knowledge of that part of actuarial statistics covered by this volume is far-reaching and as far as the subject-matter is concerned my editorial functions have been reduced to a minimum.

The task of editing the book has been a pleasurable one. On many occasions it has been the means of alleviating the monotony of war-time duties. Many hours which would otherwise have been weary have been agreeably spent both on reading the manuscript and correcting the proofs.

H. F.

AUTHOR'S PREFACE TO THE SECOND EDITION

Mistakes in the first edition have been corrected and the opportunity has been taken to re-write certain sections, particularly in Chapter IV.

In actuarial work the probabilities involved, such as those of death or falling sick, are usually very small, while the numbers exposed to risk are large. The Poisson distribution is therefore particularly appropriate in such work, and a demonstration of its derivation from the binomial has now been included in place of one of the developments of the normal approximation.

Although only the theory of large samples is considered it was felt that the notion of biased and unbiased estimates was sufficiently fundamental to require a brief explanation, and a section has been included in Chapter IV.

Table I of the Appendix based on the normal curve was formerly taken from Hardy's book, but in order to secure uniformity with the projected new text-books and Short Collection of Actuarial Tables, it has been replaced by the more usual type of table which will appear in the new publications.

The author is glad of this opportunity of thanking all those who have assisted him by helpful criticism and by drawing attention to mistakes in the earlier edition.

H. T.

AUTHOR'S PREFACE TO THE FIRST EDITION

As the title implies, no attempt has been made to produce a statistics text-book for general use. Many excellent works of this kind are already in existence (some are mentioned in the Bibliographics), but, until he has qualified, the actuary rarely has time at his disposal to make a thorough study of the subject of statistics. It is hoped that this book will give the reader a grasp of the fundamental ideas of statistics, which will not only enable him to examine critically the various tables with which he has to deal, but will help him to develop a 'statistical sense' so that he can take a lively and intelligent interest in developments outside, as well as inside, the actuarial world.

Some subjects have inevitably been dealt with in a rather superficial way and much interesting material has had to be excluded from the book, which may consequently present an unbalanced appearance to the professional statistician. Few readers will be content, therefore, to restrict their statistical studies to this volume, but the introduction which it provides should enable more ambitious works to be read with greater profit.

Mathematics has been given rather more prominence than is usual in statistical works, because it has been found that actuarial students find this form of treatment interesting and stimulating. The chapters on graduation reflect the theoretical complexity of the methods rather than their practical importance. Thus, graduation by a summation formula, because of its complicated analytical basis, has required much fuller treatment than the graphic method, which is more important from the practical point of view.

Chapter I should help the student to revise what he has learned in the Statistics chapters of *Mathematics for Actuarial Students* by extending to grouped data the technique he has already used for finding means, standard deviations and similar measures of location and dispersion. Chapter II deals with the two standard frequency distributions of most importance to actuaries, but the

Poisson distribution has been excluded both in the interests of brevity and because it proves rather a 'blind alley' in the present state of our knowledge. At the end of Chapter III some elementary ideas of non-linear regression and spurious correlation have been introduced because of their fundamental importance, although a satisfactory treatment of these subjects is quite outside the scope of the book.

Chapter IV is probably the most important because it is not until he understands the way in which sample results can be used that a student grasps the idea of 'statistical inference' which underlies all modern scientific methods.

It may seem somewhat illogical to explain tests of a graduation before the methods of graduation themselves, but Chapter V to a large extent develops from the sampling technique described in the previous chapter, while the process of graduation becomes more intelligible when the criteria of smoothness and goodness of fit are already appreciated. The outline of the χ^2 test is incomplete but the choice seemed to lie between dealing with it in this way and omitting all reference to it. A thorough treatment would involve a description of contingency tables and the multivariate normal frequency distribution, which would have greatly increased the scope of the whole book.

In each of the remaining chapters standard tables have been used as examples of the methods described, and it is hoped that the student may thus be saved a good deal of reference to original papers and memoranda which he has previously been obliged to consult.

The book has grown out of the lesson notes prepared by the tutors for the appropriate section of the examinations of the Institute of Actuaries and the Faculty of Actuaries. The author particularly wishes to acknowledge his indebtedness to Messrs A. T. Haynes and O. C. J. Klagge, who prepared the original set of notes when the Actuarial Tuition Service was inaugurated, and thus provided a framework, which with some modification of detail has remained virtually unchanged ever since. Mr H. W. Haycocks has also assisted greatly by informed criticism and many valuable suggestions.

Finally my grateful thanks are due to Mr H. Freeman, who has not only helped to ensure a logical development of the subject from the chapters in *Mathematics for Actuarial Students*, but has coordinated the two parts of this present book at a time when conditions made it impossible for the authors to meet. The book also owes much to his experience and care in reading and checking the proofs and in seeing them through the press.

H. T.

INTRODUCTION

THE IMPORTANCE OF STATISTICS TO THE ACTUARY

In studying statistics the student is usually handicapped by lack of practical background and is puzzled by the apparent uselessness of the science in actuarial work. It is hoped that the following remarks will be of assistance in clearing up his difficulties and giving a general survey of the scope of the book.

In our complex modern civilization we are constantly coming across enumerations and records of measurements, even if they are no more complicated than new business returns and details of claims. Owing to the limitations of the human mind some sort of classification and grouping is almost always essential in order to reduce them to comprehensible dimensions. Without a knowledge of statistics this classification and grouping may be carried out so that a misleading impression is conveyed, or, as more frequently happens, quite incorrect deductions may be made from data which have been collected and properly analysed.

Perhaps the most important aspect of statistics from the point of view of the actuary is that of reliability of results. For instance, a value of q_x derived from an "exposed to risk" of ten lives is, *ceteris paribus*, less reliable than one based on a hundred, and the Theory of Sampling dealt with in Chapter IV enables us to obtain a measure of this relative reliability.

This question is of particular importance in making a graduation and in considering the resulting values. The rates of mortality, sickness, retirement etc. derived from observations are not suitable for use until they have been graduated. This process may be said to be an attempt, by eliminating random errors from the observed data, to arrive at the true rates which would be obtained from ideal data of unlimited extent. The graduated rates may differ appreciably from the ungraduated rates and the question of reliability thus arises. Clearly a rate of mortality based on a thousand lives exposed to risk should be fairly close to the estimate of the true rate as shown by the graduated table, but if the exposed to risk is only fifty we should not be surprised to find the graduated rate differing considerably from the observed rate. To what extent are we justified in departing from the observed data and functions based on them in deriving the graduated values? To answer this question we again need to understand the Theory of Sampling.

Chapters I and II are concerned with general statistics and deal in somewhat greater detail with matters which the student will already have met in *Mathematics for Actuarial Students*. Chapter III on Correlation is included chiefly for the sake of completeness, as this subject is one more for the economist and the professional statistician than for the actuary.

CHAPTER I

CONTINUOUS AND DISCONTINUOUS VARIABLES. GROUPED DATA

1. Attributes and variables.

Before data derived from observations and measurements are of any practical use they have to be reduced to manageable proportions by classification and usually by grouping. The word "statistics" is in fact usually restricted to "arranged or classified facts". Such arrangement or classification can be made only with reference to some factor which varies from individual to individual and is capable of assessment. When this factor is capable of measurement, e.g. height, age, sum assured, etc., it is called a *variable* and may be continuous or discontinuous.

Continuous and discontinuous variables are constantly occurring in the mathematical work with which the student has previously had to deal and the words are used in the same sense in statistics. The phrase "discrete variation" is also fairly common instead of "discontinuous variation". In statistics relating to housing conditions the number of rooms is an important feature and is an example of discrete variation, since only integers are permissible.

When the factor used for classification is not capable of measurement it is called an *attribute*. Typical examples of attributes are nationality, class of policy, colour of eyes.

In this book we shall be concerned chiefly with variables, and when this theory has been mastered the student who is interested should find little difficulty in the statistics of attributes, although a different technique has to be developed.

2. Continuous and discontinuous variables.

When the variable is discontinuous the data are automatically divided into watertight compartments, although grouping may be used to reduce the statistics to more manageable proportions. For instance, if the results of a School Certificate examination were to be tabulated to show for a particular subject how many candidates obtained 0, 1, 2, ... up to 100 marks, the discontinuous variable (marks) would itself divide the data into 101 separate divisions which might be later aggregated into groups of five or ten marks together.* When the variable is continuous, however, the data are necessarily grouped, although this may not be obvious from the way in which they are published. For instance, an office will have tabulated the numbers of whole-life with profit policies on lives aged 18, 19, 20, ..., and although this may appear to be an example of discontinuous variation the lives are actually grouped, the commonest method being to combine all lives who have a birthday between the 1st July of one year and the 30th June of the next. Thus, those born on 1st July 1908 or 30th June 1909 or any intermediate date would be grouped and treated as aged 32 on 31st December 1940.

It is usual in classifying data involving a continuous variable to use values of this variable at equal intervals so that the frequencies in the various groups are comparable. The interval used for classification is termed the *class-interval*. The data so classified are said to form a *frequency distribution* with equal class intervals.

3. Histograms and frequency curves.

Where the variation is discrete we know the frequencies with which the values occur and can represent these frequencies by a series of points or by a series of ordinates as shown in Fig. 7, p. 256 of *Mathematics for Actuarial Students*, Part II. When the variable is continuous, however, the data are grouped and we know only the frequency with which values between x_1 and x_2, say, occur. An ordinate at the mid-point between x_1 and x_2 is not a very satisfactory way of representing the frequency, and the most usual way is to erect a rectangle with base from x_1 to x_2 and area proportionate to the frequency. If the class-interval is constant it is of course immaterial whether we regard the heights or the areas of the rectangles as representing the frequencies, but the method can be used with advantage when this is not the case.

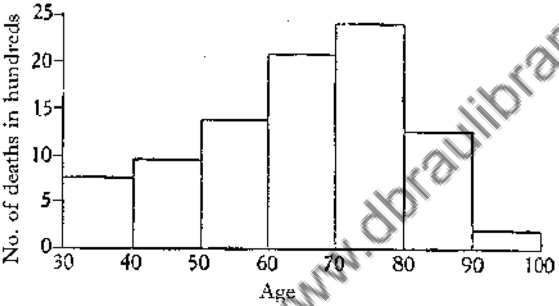
The series of rectangles is called a *histogram*.

Suppose, for instance, that we are given the following frequency distribution for the continuous variable 'age':

* In such cases a common convention is to include the frequency for the value x_1 but not for x_2 in the group "x_1-x_2". Similarly the frequency for the value x_2 is included in "x_2-x_3". Other conventions are frequently adopted.

Age last birthday	Number of deaths
30–39	7,408
40–49	9,482
50–59	13,953
60–69	20,865
70–79	23,990
80–89	12,714
90–99	1,269

These results could be represented by a histogram thus:



Ten years is a fairly wide class-interval to use and as a result the "steps" of the histogram are rather large, particularly above age 70, thus producing an irregular outline. If, instead, the class-interval had been unity, the process of drawing seventy rectangles would have been laborious; the outline would however have been much smoother, particularly over the range 70–90.

The data for this range for unit intervals are given below:

Age	Deaths	Age	Deaths
70	2434	80	2018
71	2468	81	1873
72	2490	82	1712
73	2496	83	1540
74	2487	84	1361
75	2459	85	1180
76	2412	86	1002
77	2343	87	830
78	2255	88	671
79	2146	89	527

These are actually the values of d_x in the H^M Table (Makeham Graduation), and a histogram based on these values would involve groups with unit class-interval, each group containing the deaths between one integral age and the next.

We can, however, go further and draw a continuous curve representing the limiting form of the histogram when the class-interval is indefinitely reduced. Such a curve (known as a *frequency curve*) can be used to find the frequency with which values between any limits (x_1 and x_2, say) will occur. This is represented by the area bounded by the curve, the x-axis and the ordinates at x_1 and x_2. If the ordinate of the curve is represented by $f(x)$ it is not correct to say that $f(x)$ represents the frequency with which the value x will occur. All we can say is that values between x and $x+\Delta x$ will occur approximately with frequency $f(x)\Delta x$ if Δx is very small.

It is important to realize that when the variable is continuous there is no such thing as the frequency with which a certain value x will occur exactly. Some range of values, however small, is always understood, although not always referred to explicitly. In a darts match we could record the frequency with which 17 was scored, but if we tried to record the frequency with which 1000 hens laid eggs weighing $2\frac{1}{4}$ oz. we should have to decide what range of values was to be included. We might decide to include all weights from $2\frac{1}{8}$ oz. to $2\frac{3}{8}$ oz. or from one-thousandth of an ounce below to one-thousandth of an ounce above $2\frac{1}{4}$ oz. However small the interval was made it would still exist and would be dependent on the degree of accuracy with which the weighing could be carried out.

In the above example $f(x)\Delta x$ gives the number of deaths occurring between ages x and $x+\Delta x$; $f(x)$ is therefore not a function of d_x but is of the form $l_x\mu_x$.

In practice, of course, we are usually able to obtain only the grouped data and may have to estimate as best we can the shape of the frequency curve which would be derived if the class-interval were indefinitely reduced and the numbers of groups accordingly increased. This problem will be dealt with later in this book, but it may be said at once that the method adopted is to start with a mathematical curve, the equation of which involves one or more constants, and to determine what values of these constants give the best "fit" to the given data. It should not be assumed, however, that a frequency curve to fit any frequency distribution can be arrived at by *a priori* reasoning from the data; a fairly good fit can usually be obtained, however, by empirical methods.

In dealing with grouped data it is usually sufficient to assume that the total frequency for any group is concentrated at the mid-point of that group, but in the following sections we shall deal with several instances for which such an assumption may not be sufficiently accurate.

4. Moments of a frequency distribution.

If we have a discontinuous or discrete variable which takes the values $x_1, x_2, x_3, \ldots, x_n$ with frequencies $f_1, f_2, \ldots, f_n$ (total frequency N), the rth moment about the origin is defined as

$$m_r = \frac{1}{N} \sum_{t=1}^{n} x_t^r f_t. \qquad \ldots\ldots(1)$$

If the origin is the *mean* of the distribution, we shall speak of "moments about the mean" and denote them by μ_r instead of m_r.

It is a simple matter to convert moments about one origin to moments about the mean ($x = M$).

If we denote by ξ_t the distance $x_t - M$, we have, using the above notation,

rth moment about the origin

$$= \frac{1}{N} \sum_{t=1}^{n} x_t^r f_t$$

$$= \frac{1}{N} \sum_{t=1}^{n} (\xi_t + M)^r f_t$$

$$= \frac{1}{N} \sum_{t=1}^{n} \{\xi_t^r + r_{(1)} \xi_t^{r-1} M + r_{(2)} \xi_t^{r-2} M^2 + \ldots\} f_t. \qquad \ldots\ldots(2)$$

But μ_r, the rth moment about the mean,

$$= \frac{1}{N} \sum_{t=1}^{n} \xi_t^r f_t.$$

Hence the above equation can be written

$$m_r = \mu_r + r_{(1)} M \mu_{r-1} + r_{(2)} M^2 \mu_{r-2} + \dots \qquad \dots\dots(3)$$

Similarly it can be shown that

$$\mu_r = m_r - r_{(1)} M m_{r-1} + r_{(2)} M^2 m_{r-2} - \dots . \qquad \dots\dots(4)$$

But it will be remembered that M, the mean, is defined by

$$\frac{1}{N}\{x_1 f_1 + x_2 f_2 + \dots + x_n f_n\} = m_1 = M \text{ above,}$$

and the standard deviation, σ, is defined by the equation

$$\sigma^2 = \frac{1}{N}\{\xi_1^2 f_1 + \xi_2^2 f_2 + \dots + \xi_n^2 f_n\} = \mu_2.$$

Hence, putting $r=1$ in equation (3), we have

$$m_1 = \mu_1 + M; \quad \text{giving } \mu_1 = 0,$$

i.e. the first moment about the mean is zero.

Similarly, putting $r=2$, we have

$$m_2 = \mu_2 + 2M\mu_1 + M^2$$
$$= \mu_2 + M^2, \quad \text{since } \mu_1 = 0;$$

or

$$\mu_2 = \sigma^2 = m_2 - M^2. \qquad \dots\dots(5)$$

Therefore, to find σ^2, we take the second moment about any convenient origin and subtract the square of the distance between this origin and the mean. This is of course the method with which the student is already familiar.

Again, putting $r=3$, we have from equation (4)

$$\mu_3 = m_3 - 3Mm_2 + 3M^2 m_1 - M^3$$
$$= m_3 - 3m_1 m_2 + 2m_1^3;$$

and

$$\mu_4 = m_4 - 4m_1 m_3 + 6m_1^2 m_2 - 3m_1^4.$$

In the same way we can find m's in terms of μ's from equation (3).

Note. The use of μ to denote moments about the mean is very common in statistical literature, and although in theoretical work confusion may arise with the force of mortality (and similarly m_r may be confused with the central rate of mortality) in numerical work this is unlikely to arise, as moments of higher order than the fourth are hardly ever met with.

For continuous variables we shall denote by $f(x)$ the ordinate to the frequency curve at the point x.

As stated above we can assume the frequency between x and $x+\Delta x$ to be approximately $f(x)\Delta x$.

Hence in the limit

$$m_r = \frac{1}{N}\int_0^n x^r f(x)\,dx, \qquad \text{......(6)}$$

where N = total frequency

$$= \int_0^n f(x)\,dx.$$

5. Sheppard's adjustments.

If we assume the total frequency of any group to be concentrated at the mid-point of that group, errors arise in the calculation of some of the moments. A method of making adjustments in certain circumstances is due to Dr W. F. Sheppard.

Let us consider a group for which the variable lies between

$$x_t - \frac{h}{2} \quad \text{and} \quad x_t + \frac{h}{2} \quad \text{(class-interval } h\text{)}.$$

Let the frequency curve stretch from $x=a$ to $x=b$ and, as before, let the ordinate at the point x be $f(x)$.

The *true* rth moment m_r is then

$$\frac{1}{N}\int_a^b x^r f(x)\,dx \quad (N = \text{total frequency}).$$

The approximate rth moment obtained by assuming the total frequency for each group to be concentrated at the mid-point is given by an expression of the form

$$m_r' = \frac{1}{N}\Sigma\, x_t^r \int_{-h/2}^{h/2} f(x_t+z)\,dz, \qquad \text{......(7)}$$

where the summation embraces all groups.

By Taylor's theorem,

$$f(x_t+z) = \left\{1 + zD + \frac{z^2D^2}{2!} + \frac{z^3D^3}{3!} + \dots\right\} f(x_t),$$

where $\qquad D \equiv \dfrac{d}{dx_t}.$

$$\therefore \int_{-h/2}^{h/2} f(x_t+z)\,dz = hf(x_t) + \frac{h^3}{24} f^{\text{ii}}(x_t) + \frac{h^5}{1920} f^{\text{iv}}(x_t) + \dots.$$

......(8)

Hence, from (7),

$$m'_r = \frac{1}{N}\left[h\,\Sigma x_t^r f(x_t) + \frac{h^3}{24}\Sigma x_t^r f^{\text{ii}}(x_t) + \frac{h^5}{1920}\Sigma x_t^r f^{\text{iv}}(x_t) + \dots\right]. \qquad \dots\dots(9)$$

It will be remembered that the Euler-Maclaurin expansion expresses $\int_a^b f(x)\,dx$ in terms of $\Sigma f(x)$ at intervals of h and terms involving the values of $f(x)$ and its derivatives at the limits a and b. If the frequency curve has contact of a high order with the x-axis at both ends—i.e. if $f(x)$ and its first few derivatives are zero at the limits $x=a$ and $x=b$—then

$$h\Sigma x_t^r f(x_t) \fallingdotseq \int_a^b x^r f(x)\,dx,$$

$$\frac{h^3}{24}\Sigma x_t^r f^{\text{ii}}(x_t) \fallingdotseq \frac{h^2}{24}\int_a^b x^r f^{\text{ii}}(x)\,dx.$$

$$\frac{h^5}{1920}\Sigma x_t^r f^{\text{iv}}(x_t) \fallingdotseq \frac{h^4}{1920}\int_a^b x^r f^{\text{iv}}(x)\,dx.$$

....

$$\therefore\; m'_r \fallingdotseq \frac{1}{N}\int_a^b x^r\left[f(x) + \frac{h^2}{24}f^{\text{ii}}(x) + \frac{h^4}{1920}f^{\text{iv}}(x) + \dots\right]dx.$$

The second term can be integrated by parts as follows:

$$\frac{1}{N}\int_a^b x^r \frac{h^2}{24} f^{\text{ii}}(x)\,dx = \frac{h^2}{24N}\Big[x^r f'(x)\Big]_a^b - \frac{rh^2}{24N}\int_a^b x^{r-1} f'(x)\,dx$$

$$= \frac{h^2}{24N}\Big[x^r f'(x) - r x^{r-1} f(x)\Big]_a^b + \frac{r(r-1)h^2}{24N}\int_a^b x^{r-2} f(x)\,dx.$$

Since $f(x)$ and its derivatives vanish at both limits this reduces to the last term.

Similarly, the term $\frac{1}{N}\int_a^b x^r \frac{h^4}{1920} f^{\text{iv}}(x)\,dx$ can be transformed into the term

$$\frac{r(r-1)(r-2)(r-3)h^4}{1920N}\int_a^b x^{r-4} f(x)\,dx.$$

In this way the expression for m'_r can be transformed to

$$m'_r = \frac{1}{N}\int_a^b \left[x^r + \frac{h^2}{24} r(r-1)x^{r-2} + \frac{h^4}{1920} r(r-1)(r-2)(r-3)x^{r-4} + \dots\right] f(x)\,dx$$

$$= m_r + \frac{h^2 r(r-1)}{24} m_{r-2} + \frac{h^4}{1920} r(r-1)(r-2)(r-3) m_{r-4} + \dots,$$

where the m's are true moments.

Hence

$$\left.\begin{aligned} m'_1 &= m_1 \\ m'_2 &= m_2 + \frac{h^2}{12} \\ m'_3 &= m_3 + \frac{h^2}{4} m_1 \\ m'_4 &= m_4 + \frac{h^2}{2} m_2 + \frac{h^4}{80} \end{aligned}\right\}.$$

From these we have successively that

$$\left.\begin{aligned} m_2 &= m'_2 - \frac{h^2}{12} \\ m_3 &= m'_3 - \frac{h^2}{4} m'_1 \quad (\text{since } m_1 = m'_1) \\ m_4 &= m'_4 - \frac{h^2}{2} m'_2 + \frac{7h^4}{240} \end{aligned}\right\}. \qquad \dots\dots(10)$$

For moments about the *mean*, μ_1 and μ_3 need no correction ($\mu_1 = \mu'_1 = 0$), so that:

$$\left.\begin{aligned} \mu_2 &= \mu'_2 - \frac{h^2}{12} \\ \mu_4 &= \mu'_4 - \frac{h^2}{2} \mu'_2 + \frac{7h^4}{240} \end{aligned}\right\}, \qquad \dots\dots(11)$$

where μ' represents a moment about the mean calculated on the assumption that the total frequency of each group is concentrated at the mid-point of that group.

The assumption made above that $f(x)$ and its derivatives vanish at a and b means that in practice Sheppard's adjustments should not be made unless the frequency dwindles to nothing at each end of the range considered and has negligible first and second differences as these ends are approached. (It should be borne in mind that we can usually only *estimate* the shape of the frequency curve from the given data.)

The most common adjustment is the deduction of $h^2/12$ in finding σ^2 $(=\mu_2)$ and this will be made in the example to be found later in the chapter. It is doubtful, however, whether the correction is worth making unless the data are fairly extensive (say, a total frequency of at least 1000), because errors of sampling would otherwise be large compared with errors of grouping.

6. Mode of grouped data.

Theoretically the only sound way of finding the mode of grouped data when the variable is continuous is to fit a frequency curve to the data and then find when the gradient dy/dx vanishes. Usually, however, the mode is not required to a degree of accuracy which would justify the considerable labour involved.

The work can be done graphically by drawing a histogram and replacing it by a smooth curve from which the mode can be found by inspection. Here again the work is fairly heavy, but the method has the great advantage that all the data are used.

(It should be remembered, however, that there may be more than one mode, each of which corresponds to a local maximum ordinate on the frequency curve.)

Of the short analytical methods perhaps the best is to fit a curve of the form $f(x)=a+bx+cx^2+\dots$ to the data in the immediate neighbourhood of the mode. This is usually fairly simple, but unfortunately only some of the data are used (see Example, para. 8).

7. Mean deviation of grouped data.

This section relates to the mean deviation which is the average of the *absolute* magnitudes of deviations from the arithmetic mean.

The group in which the mean lies presents difficulty when the mean deviation has to be calculated and this group is usually left until the last. For convenience we shall call it the special group.

The work is divided into three stages:

(1) Ignoring the special group the total in each group is assumed to be concentrated at the mid-point of the group and all distances are measured from the origin, which is, of course, chosen so as to reduce the arithmetical work.

(2) The result is adjusted so as to allow for all distances being measured from the mean (still ignoring the special group).

(3) The mean deviation of the special group about the mean is found and an appropriate adjustment made to the result of (2).

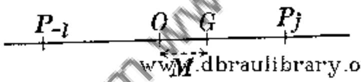
Denote the group frequencies by

$$u_{-r},\ u_{-r+1},\ \ldots,\ u_{-1},\ u_0,\ u_1,\ u_2,\ \ldots,\ u_{s-1},\ u_s$$

and the distances of the mid-points from the origin by

$$x_{-r},\ x_{-r+1},\ \ldots\ x_{-1},\ x_0,\ x_1,\ x_2,\ \ldots\ x_{s-1},\ x_s.$$

The products $u_{-r}x_{-r}$, $u_{-r+1}x_{-r+1}$, ... $u_{k-1}x_{k-1}$, $u_{k+1}x_{k+1}$ ... $u_s x_s$ (ignoring the special group term $u_k x_k$) are calculated and summed, treating all signs as positive.



In the illustration O represents the origin and G the position of the mean, distant M from the origin. P_j represents the mid-point of a group with frequency u_j lying to the right of O, while P_{-l} represents the mid-point of a group with frequency u_{-l} lying to the left of O.

$$OP_j = x_j \quad \text{and} \quad OP_{-l} = x_{-l}.$$

We have calculated $u_j x_j$, but to find the mean deviation about the mean we require

$$u_j \,.\, GP_j = u_j(x_j - M)$$
$$= u_j x_j - M u_j.$$

The same applies to all groups to the right of O, so that in finding the mean deviation about the mean the total frequency for these groups must be multiplied by M and subtracted from the previous result.

Similarly, the term $|u_{-l}x_{-l}|$ must be replaced by $u_{-l}\,.\,P_{-l}G$, i.e. by

$$|u_{-l}x_{-l}|+Mu_{-l}.$$

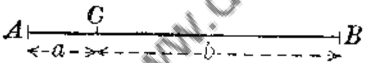
Dealing with all the groups to the left of O in the same way we see that we have to adjust the previous result by adding $M\times$(total frequency in the groups to the left of O).

Summing up, we may say that the process referred to as (2) above can be reduced to:

multiplying the total frequency in groups to the right of O by M and *subtracting* the product from the result of the first process; and

multiplying the total frequency in groups to the left of O by M and *adding* the product to the result of the first process.

The student should investigate the problem when G is to the left of O and is recommended, when in doubt, to make a drawing similar to the above to ensure that, in making the adjustment, the signs are correct.



Finally, the special group has to be dealt with.

As before, G represents the mean and AB the limits of the special group:

$$AG=a \quad \text{and} \quad GB=b.$$

If we assume that the group frequency u_k is evenly spread over the range, the frequency to the left of G will be $\dfrac{a}{a+b}u_k$ and the frequency to the right will be $\dfrac{b}{a+b}u_k$.

We can now assume that the first of these frequencies is concentrated at the mid-point of AG and the second at the mid-point of GB.

The special group thus contributes a term

$$\frac{a}{a+b}u_k\,.\,\frac{a}{2}+\frac{b}{a+b}u_k\,.\,\frac{b}{2}=\frac{a^2+b^2}{2(a+b)}u_k.$$

This is added to the previous result, and on dividing by the total frequency we obtain the mean deviation about the mean.

This seems a somewhat laborious process, but the numerical work is relatively simple, as will be seen in Example 2.

8. Illustrative examples.

The following example will serve to remind the student of what he has already learnt in *Mathematics for Actuarial Students*. It does not involve a continuous variable or grouped data.

Example 1.

The table below gives a frequency distribution of the scores returned in a veterans' golf competition for which there were 1000 competitors:

Score	Frequency	Score	Frequency
70	16	76	83
71	93	77	56
72	181	78	38
73	196	79	26
74	163	80	17
75	120	81	11

Determine the values of the mean, median, mode, quartile deviation, mean deviation and standard deviation.

To save arithmetic, measure deviations from the score 74 instead of from 0 or 70.

Value of variable x (1)	Frequency f_x (2)	Deviation from 74 $x-74$ (3)	$(x-74)\times f_x$ $(2)\times(3)$ (4)		$(x-74)^2\times f_x$ $(3)\times(4)$ (5)	Cumulative frequency Σf_x (6)
			−	+		
70	16	−4	64		256	16
71	93	−3	279		837	109
72	181	−2	362		724	290
73	196	−1	196		196	486
74	163	0	—		—	649
75	120	1		120	120	769
76	83	2		166	332	852
77	56	3		168	504	908
78	38	4		152	608	946
79	26	5		130	650	972
80	17	6		102	612	989
81	11	7		77	539	1000
Total	1000	—	−901	+915 = +14	5378	—

Mean $=74+\frac{14}{1000}=74\cdot014$.

Second moment about origin $(74)=\frac{5378}{1000}=5\cdot378$.

Distance of origin from mean $=\cdot014$.

$\therefore$ (standard deviation)$^2=5\cdot378-(\cdot014)^2$,

so that $$\sigma=2\cdot32.$$

Mean deviation from the mean. Total deviation from the origin (irrespective of sign) $=901+915=1816$. We must, however, measure from 74·014 instead of from 74.

All frequencies for values of x greater than 74 must therefore be multiplied by ·014 and *subtracted*, while all frequencies for values of x of 74 or less must be multiplied by ·014 and *added*.

We have $$1816+\cdot014(649-351)=1820.$$

$\therefore$ mean deviation from the mean $=\frac{1820}{1000}=1\cdot82$.

Median. As there are 1000 observations the median will lie between the 500th and the 501st. The nearest integral value of x satisfying this condition is 74, which may be taken as the median.

Lower quartile. A quarter of the total frequency is 250, but it cannot be said that one-quarter of the total frequency lies *below* the 250th observation (nor for that matter does three-quarters lie *above* the 251st observation). The lower quartile lies between the 250th and 251st observation and the nearest integral value of x satisfying this condition is 72.

Upper quartile. Similarly, the upper quartile separates the 750th and 751st observation and may be taken as 75.

Quartile deviation is therefore $\dfrac{75-72}{2}=1\cdot5$.

Mode. The value of x having the greatest frequency is 73.

Example 2.

If in the above example the frequency distribution relates, not to the scores but to the ages nearest birthday of the competitors, what alterations will there be in the values of the respective indices?

Here we are dealing with a continuous variable and the data are grouped. Thus the frequency 181 shown opposite $x=72$ really relates to ages from $71\frac{1}{2}$ to $72\frac{1}{2}$ (class-interval unity).

Mean. Assuming the frequencies to be concentrated at the mid-points of the groups we find that the previous calculations still hold good, and the mean is 74·014 as before (no Sheppard's adjustment is required to the first moment).

Standard deviation. Making the Sheppard's adjustment $(-\frac{1}{12})$, we have

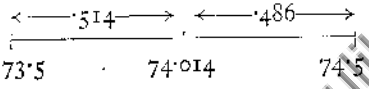
$$(\text{standard deviation})^2 = 5{\cdot}378 - ({\cdot}014)^2 - \tfrac{1}{12},$$
$$\therefore\ \sigma = 2{\cdot}30 \text{ years.}$$

Mean deviation from the mean. Ignoring the special group ($73\frac{1}{2}$–$74\frac{1}{2}$, or nearest age 74) we have the sum of the deviations from 74, irrespective of sign, 1816, as before.

In order to obtain the deviations from the mean (74·014) all frequencies in the groups for ages greater than $74\frac{1}{2}$ must be multiplied by ·014 and subtracted from the result. Similarly, all frequencies for ages less than $73\frac{1}{2}$ must be multiplied by ·014 and added.

This gives $1816 + {\cdot}014\,(486 - 351)$.

Finally, we must deal with the group $73\frac{1}{2}$–$74\frac{1}{2}$, with frequency 163.



Of this frequency ·514 × 163 are assumed to lie between 73·5 and the mean and ·486 × 163 between the mean and 74·5.

Each of these sub-groups may be assumed to be concentrated at the mid-point of its own range, thus giving a term

$$163\left\{{\cdot}514 \times \frac{{\cdot}514}{2} + {\cdot}486 \times \frac{{\cdot}486}{2}\right\}$$

to be added to the above.

Note. For the special group it is simpler to measure deviations not first from 74 but from the mean direct and afterwards to apply a correction.

The mean deviation therefore becomes

$$\tfrac{1}{1000}\left[1816 + {\cdot}014\,(486 - 351) + \tfrac{163}{2}({\cdot}514^2 + {\cdot}486^2)\right] = 1{\cdot}86 \text{ years.}$$

Median. As before, this is a value of x lying between the 500th and 501st observations. These lie in the group $73\frac{1}{2}$–$74\frac{1}{2}$, which includes 163 observations, while the total frequency for lower values of x is 486.

We assume that the 163 observations between $73\frac{1}{2}$ and $74\frac{1}{2}$ are evenly spread over the interval at distances $\frac{1}{326}, \frac{3}{326}, \frac{5}{326}, \ldots \frac{325}{326}$ from either end, i.e. they occur for values of x:

$$\left(73\tfrac{1}{2} + \frac{1}{326}\right),\quad \left(73\tfrac{1}{2} + \frac{3}{326}\right),\quad \ldots\quad \left(73\tfrac{1}{2} + \frac{2r-1}{326}\right),\quad \ldots\quad \left(73\tfrac{1}{2} + \frac{325}{326}\right).$$

The first of these observations is the 487th counting from the lowest values of x, the next is the 488th, and so on.

The 500th and 501st observations are the 14th and 15th *in that particular group* and correspond therefore to values of x of $73\frac{1}{2} + \frac{27}{326}$ and $73\frac{1}{2} + \frac{29}{326}$.

The median may be taken therefore as $73\frac{1}{2} + \frac{28}{326} = 73{\cdot}59$ years.

Lower quartile. 109 values lie below the group $71\frac{1}{2}$–$72\frac{1}{2}$, which itself includes 181 observations. These 181 observations are assumed to correspond to values of x of

$$\left(71\tfrac{1}{2}+\frac{1}{362}\right),\quad \left(71\tfrac{1}{2}+\frac{3}{362}\right),\quad \ldots\quad \left(71\tfrac{1}{2}+\frac{2r-1}{362}\right),\quad \ldots\quad \left(71\tfrac{1}{2}+\frac{361}{362}\right),$$

so that the 250th and 251st observations (141st and 142nd in that group) correspond to values of x:

$$71\tfrac{1}{2}+\tfrac{281}{362}\quad\text{and}\quad 71\tfrac{1}{2}+\tfrac{283}{362}.$$

The lower quartile may therefore be taken to be $71\frac{1}{2}+\frac{282}{362}=72{\cdot}28$ years.

Upper quartile. The 750th and 751st observations are the 101st and 102nd in the group $74\frac{1}{2}$–$75\frac{1}{2}$.

Hence the upper quartile is $74\frac{1}{2}+\frac{202}{240}=75{\cdot}34$ years.

Quartile deviation. $\frac{1}{2}(75{\cdot}34-72{\cdot}28)=1{\cdot}53$ years.

Mode. The student is recommended to try the graphic process of drawing the histogram, sketching in a smooth frequency curve and finding the mode by inspection.

An analytical method similar to the following is sometimes useful:

Assume that the frequency curve in the neighbourhood of the mode is of the form $y=a+bx+cx^2$.

Taking the origin at 73 we have the following equations from which to find a, b and c:

$$\int_{-1\frac{1}{2}}^{-\frac{1}{2}}(a+bx+cx^2)=181 \quad \text{(the frequency in the group } 71\tfrac{1}{2}\text{–}72\tfrac{1}{2}\text{);}$$

i.e.

$$a-b+\frac{13c}{12}=181.$$

Similarly

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y\,dx=a+\frac{c}{12}=196$$

and

$$\int_{\frac{1}{2}}^{1\frac{1}{2}} y\,dx=a+b+\frac{13c}{12}=163.$$

From these we obtain $b=-9$, $c=-24$.

The mode (the value of x for the maximum ordinate) is given by

$$\frac{dy}{dx}=0,\quad\text{i.e. } b+2cx=0.$$

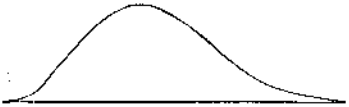
$\therefore\ x=-\dfrac{b}{2c}=-\dfrac{9}{48}$, referred to 73 as origin.

The mode is therefore $\quad 73-\frac{9}{48}=72{\cdot}81$ years.

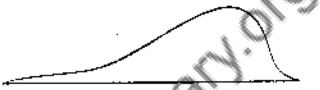
Note: all the values calculated are *values of x*.

9. Skewness.

If we imagine a frequency distribution represented by a smooth frequency curve, the measures already demonstrated will tell us a great deal about the curve. We know, for instance, where its highest point occurs (mode), whether it is like a steep-sided peak (small standard deviation) or a broad plateau (large standard deviation). There is a further characteristic in which we are sometimes interested, namely, its lack of symmetry or "skewness".



Positive skewness.



Negative skewness.

In a symmetrical curve the mean, median and mode all coincide, and the extent to which they fail to do so gives rise to one well-known measure of skewness, viz.

$$\frac{\text{mean}-\text{mode}}{\text{standard deviation}}=\frac{3\,(\text{mean}-\text{median})}{\text{standard deviation}}\ \text{(approx.)}. \qquad \ldots\ldots(12)$$

For reasons which will be discussed later, this approximate measure should be used only if the skewness is relatively small.

A second measure sometimes used is

$$\frac{(\text{upper quartile}-\text{median})-(\text{median}-\text{lower quartile})}{\frac{1}{2}(\text{upper quartile}-\text{lower quartile})}. \qquad \ldots\ldots(13)$$

This is a very cumbersome measure and difficult to calculate. Perhaps the most convenient measure is

$$\frac{\sqrt[3]{\mu_3}}{\sigma}, \qquad \ldots\ldots(14)$$

where μ_3 is the third moment about the mean. This expression has the great advantage that it is susceptible of arithmetical or algebraical calculation. It will be observed that when the hump of a curve occurs at low values of x the skewness is positive, while a hump to the right gives negative skewness. (A symmetrical curve gives of course a zero result.)

10. King's formula.

Before we leave grouped data we must mention two important formulae which will be needed later in the book. They are commonly known as "King's Formula" and "Hardy's Formula".

Let u_x be a function whose fourth and higher differences are negligible.

Let

$$\begin{aligned} w_{-1} &= u_{-3m-1}+u_{-3m}+\dots &&+u_{-m-1},\\ w_0 &= u_{-m}+u_{-m+1}+\dots+u_{-1}+u_0+u_1+\dots+u_{m-1}+u_m,\\ w_1 &= u_{m+1}+u_{m+2}+\dots &&+u_{3m+1}. \end{aligned}$$

w_{-1}, w_0 and w_1 are therefore three consecutive groups each of $2m+1$ values.

We wish to express u_0, the middle term of the middle group, in terms of w_{-1}, w_0 and w_1.

Stirling's formula is

$$u_x = u_0+\frac{x}{2}(\Delta u_0+\Delta u_{-1})+\frac{x^2}{2}\Delta^2 u_{-1}+\frac{x(x^2-1)}{12}(\Delta^3 u_{-1}+\Delta^3 u_{-2})+\dots.$$

(*Mathematics for Actuarial Students*, Part II, p. 64.)

Summing from $-m$ to $+m$:

$$\begin{aligned} w_0 &= (2m+1)u_0+\Delta^2 u_{-1}(1^2+2^2+\dots+m^2)\\ &= (2m+1)u_0+\frac{m(m+1)(2m+1)}{6}\Delta^2 u_{-1}. \qquad \dots\dots(15) \end{aligned}$$

Summing from $-3m-1$ to $3m+1$:

$$w_{-1}+w_0+w_1 = (6m+3)u_0+\frac{(3m+1)(3m+2)(6m+3)}{6}\Delta^2 u_{-1}. \qquad \dots\dots(16)$$

Hence $$w_{-1}-2w_0+w_1=(2m+1)^3\Delta^2 u_{-1}.$$

Substituting for $\Delta^2 u_{-1}$ in (15), we obtain:

$$w_0=(2m+1)u_0+\frac{m(m+1)}{6(2m+1)^2}(w_{-1}-2w_0+w_1)$$

or

$$u_0=\frac{1}{2m+1}\left[w_0-\frac{m(m+1)}{6(2m+1)^2}(w_{-1}-2w_0+w_1)\right]. \qquad \dots\dots(17)$$

Putting $2m+1=n$, this may be written in the more usual form

$$u_0=\frac{1}{n}\left[w_0-\frac{(n^2-1)}{24n^2}\Delta^2 w_{-1}\right], \qquad \ldots\ldots(18)$$

where Δ now denotes the differencing of *group totals* and not individual u's.

When $n=5$, we have

$$\begin{aligned} u_0 &= \tfrac{1}{5}\left[w_0-\tfrac{1}{25}\Delta^2 w_{-1}\right] \\ &= \cdot 2w_0-\cdot 008\Delta^2 w_{-1}, \qquad \ldots\ldots(19) \end{aligned}$$

the usual form of King's formula. The more general expression (18) should, however, be remembered.

So far we have considered only odd values of n.

If n is even ($=2r$ say), let

$$w_0=u_{-\frac{2r-1}{2}}+u_{-\frac{2r-3}{2}}+\ldots+u_{-\frac{3}{2}}+u_{-\frac{1}{2}}+u_{\frac{1}{2}}+u_{\frac{3}{2}}+\ldots+u_{\frac{2r-3}{2}}+u_{\frac{2r-1}{2}}.$$

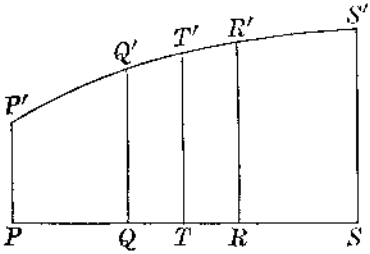
Proceeding as before, and remembering that $2m+1=n$, we have

$$\begin{aligned} w_0 &= 2ru_0+\frac{r(2r-1)(2r+1)}{12}\Delta^2 u_{-1} \\ &= nu_0+\frac{n(n^2-1)}{24}\Delta^2 u_{-1}, \text{ corresponding to (15).} \end{aligned}$$

Formula (18) then follows as before, but it will be seen that $\frac{1}{n}\left[w_0-\frac{n^2-1}{24n^2}\Delta^2 w_{-1}\right]$ no longer gives the central term of the middle group (there is no central term when n is even) but the value of u for an argument half-way between those of the two central terms. For instance, if the data are given for groups 40–43, 44–47 and 48–51 the application of the formula would give $u_{45\frac{1}{2}}$.

11. Hardy's formula.

This formula can be applied only to continuous functions. Unlike King's formula therefore it cannot be applied to functions such as E_x (the "initial" exposed to risk), which is discontinuous at the end of every year of experience. On the other hand, E_x^c, the exposed to risk in central form, is to all intents and purposes continuous. It is true that it increases or decreases by whole numbers (or, in certain formulae, by fractions), but entrants and exits are allowed for as they occur and there is no sudden "jump" as with E_x. Hardy's formula can therefore be applied to E_x^c.



$$\text{Let } w_{-1}=\int_{-3n/2}^{-n/2} f(x)\,dx, \quad \text{represented by the area } PQQ'P',$$

$$w_0=\int_{-n/2}^{n/2} f(x)\,dx, \quad \text{,, \quad ,,} \quad QRR'Q',$$

$$\text{and} \quad w_1=\int_{n/2}^{3n/2} f(x)\,dx, \quad \text{,, \quad ,,} \quad RSS'R'.$$

Hardy's formula gives the central ordinate u_0 in terms of the three areas, i.e. $f(0)$ in terms of w_{-1}, w_0 and w_1, assuming fourth and higher differences to be negligible.

Let $$f(x)=a+bx+cx^2+dx^3.$$

Then $$w_0=\int_{-n/2}^{n/2} f(x)\,dx=na+\frac{n^3}{12}c, \qquad \ldots\ldots(20)$$

$$w_{-1}+w_0+w_1=\int_{-3n/2}^{3n/2} f(x)\,dx=3na+\frac{9n^3}{4}c$$

and $$\Delta^2 w_{-1}=w_{-1}-2w_0+w_1=2n^3c, \qquad \ldots\ldots(21)$$

where Δ is the operator for differencing grouped values.

Hence, from (20), $\quad w_0=na+\frac{1}{24}\Delta^2 w_{-1}$,

and a, the central ordinate $=\frac{1}{n}\left[w_0-\frac{1}{24}\Delta^2 w_{-1}\right]$. $\qquad \ldots\ldots(22)$

This is Hardy's formula.

12. Application to exposed to risk and deaths.

It is important that the student should realize the essential difference between King's and Hardy's formulae.

King's formula is a finite difference formula enabling the central term to be found when only three group totals are given.

If, for example, we are given ΣE_x for groups 40–44, 45–49 and 50–54, i.e. $E_{40}+E_{41}+E_{42}+E_{43}+E_{44}$, etc., the formula gives a value for E_{47}, the central term of the middle group. Similarly, if the functions given were ΣE_x^c or $\Sigma\theta_x$ (the number of deaths observed), the formula would give E_{47}^c or θ_{47} as the case may be.

Hardy's formula cannot be applied to E_x, which is discontinuous. In applying it to E_x^c we must regard $\sum\limits_{45}^{49} E_x^c$ not as

$$E_{45}^c+E_{46}^c+E_{47}^c+E_{48}^c+E_{49}^c$$

but as an integral.

$$E_x^c=\int_0^1 P_{x+t}\,dt,$$

where P_{x+t} denotes the number exposed to risk at exact age $x+t$.

Hence Hardy's formula gives $P_{47\frac{1}{2}}$ (not P_{47}).

If we have the corresponding deaths (θ_x) similarly grouped it is difficult at first to see what function is given by Hardy's formula.

Actually the function is $P_{x+t}\mu_{x+t}$ for the central point, because

$$\int_A^B P_{x+t}\,\mu_{x+t}\,dt = \text{deaths occurring between ages } A \text{ and } B.$$

Hence Hardy's formula applied to $\Sigma\theta_x$ gives $P_{47\frac{1}{2}}\mu_{47\frac{1}{2}}$, and since $P_{47\frac{1}{2}}$ has been found from ΣE_x^c we can arrive at $\mu_{47\frac{1}{2}}$ by division. This will be dealt with later in Chapter X.

Example 8.

Given $\sum\limits_{1}^{7} u_x = 3865$, $\sum\limits_{8}^{14} u_x = 2618$ and $\sum\limits_{15}^{21} u_x = 1885$, find u_{11}, assuming that fourth and higher differences are negligible.

Denoting the groups by w_{-1}, w_0 and w_1, respectively,

$$\Delta^2 w_{-1} = 514$$

and

$$u_{11}=\frac{1}{7}\left[2618-\frac{7^2-1}{24\times 7^2}514\right]$$

$$=371 \text{ almost exactly.}$$

Actually the data are the values of $100\overset{\circ}{e}_x$ from ages 75 to 95 inclusive in E.L. No. 8, where $100\overset{\circ}{e}_{85}$ (corresponding to u_{11}) is 372.

Example 4.

If we were given $\sum_{1}^{6} u_x = 3401$, $\sum_{7}^{12} u_x = 2434$, $\sum_{13}^{18} u_x = 1778$, i.e. an *even* number of terms in each group, the application of King's formula would give

$$u_{9\frac{1}{2}} = \frac{1}{6}\left[2434 - \frac{6^2-1}{24\times 6^2}311\right]$$

$$= 404 \text{ (to nearest integer).}$$

It will be noted that $u_{9\frac{1}{2}}$ is centrally situated in the group 7–12.

By third difference interpolation, using the tabulated values of u_8, u_9, u_{10} and u_{11} (i.e. $100\overset{\circ}{e}_{82}$, etc., in E.L. No. 8), the value of $u_{9\frac{1}{2}}$ is found to be 404, so that on this occasion the group formula gives a good result.

13. Weighted mean.

This term sometimes leads the student to think that he is dealing with a further measure of statistical average, whereas all means are, in a sense, weighted means.

Thus, in the standard expression for the mean, $\frac{\Sigma x f_x}{\Sigma f_x}$, the observed values of x are weighted with the frequencies with which they occur; e.g. a batting average is obtained by weighting the scores with the frequencies of their occurrence and dividing by the total frequency (number of innings completed). The phrase *weighted mean* is usually used in statistical books when the actual frequencies are not available and have to be estimated. Provided that the values of the variable are not greatly unequal and that the weights used are not wide of the mark, the value thus obtained will usually be very close to the true mean which would have been obtained by using the actual frequencies.

Occasionally weighted means arise in another sense. Sometimes weights are applied to individual observations to allow for some element of relative importance other than numerical frequency. Weighted means of this kind are to be regarded as indicators or indices of some condition rather than as averages.

14. Index numbers.

Economists often wish to have a single measure of the combined results of many factors operating together. Index numbers are commonly used for this purpose. They are a special type of weighted mean of which perhaps the best known is the "Cost of Living Index", which reflects the effect of changes in price from year to year on a fixed "basket" of goods bought by working-class housewives.

A convenient year was taken as a base and the figure for that year taken arbitrarily as 100. For many years the "Cost of Living Index" was based on prices in 1914, but a more up-to-date base year is now desirable and the necessary data have been collected. The present (1949) index is only an interim arrangement.

Another example with which the student will meet in investment work is the "Actuaries Investment Index", which gives a measure of how shares in certain broad groups are changing in value from time to time.

One of the best examples in actuarial work, however, is provided by the Comparative Mortality Figure (C.M.F.) used after the 1921 Census to compare the mortality in different occupations with that in the country as a whole and for comparing different occupations among themselves.

The C.M.F. may be represented by the formula

$$1000\frac{\Sigma P^s_x m^a_x}{\Sigma P^s_x m^s_x},$$

where P^s_x = the number in the specified age-group x of the standard population,

m^a_x = central death-rate for the age-group x of the occupation specified,

m^s_x = central death-rate for the age-group x of the standard population,

and the summation extends over ages 20–65.

The standard population was based on the number of occupied and retired civilian males between the ages of 20 and 65 enumerated at the 1921 Census. The census numbers were scaled down so that the number of deaths expected between the limiting ages according to the rates of mortality m^s_x was 1000.

By the use of these reduced populations the expression for the C.M.F. reduces to $\Sigma P^s_x m^a_x$. If the result is less than 1000 the mortality according to this index is lighter than the average; if it is more than 1000 the mortality is relatively heavy.

Any index number is open to the objection that it may convey misleading impressions in exceptional circumstances, and against the C.M.F. it may be urged that "normal" weight is thereby attached to values of m_x^a which may be based on very scanty data and may be quite unreliable in consequence.

Index numbers cannot be expected to convey all the information given by the data they are intended to summarize, but they are nevertheless useful in reducing a mass of classified data to manageable proportions for purposes of comparison.

A change of base year affects the relative size of all the indices already calculated and for this reason a mean akin to the geometric mean has often been recommended.

This would involve terms such as

$$y_1^{x_1},\ y_2^{x_2},\ y_3^{x_3},\ \dots \quad \text{instead of} \quad x_1y_1,\ x_2y_2,\ \dots.$$

An objection is that if one of the y's should happen to vanish in any given year (a by no means impossible occurrence) the whole index also vanishes. One advantage of this form of average is, however, that a change of base year does not affect the relative values of previously calculated indices. (The Actuaries Investment Index involves this type of geometric mean.)

BIBLIOGRAPHY

Mathematics for Actuarial Students. Part II, Chapters xi and xii. H. Freeman. Camb. Univ. Press.

An Introduction to the Theory of Statistics. Chapters 6–9. G. Udny Yule and M. G. Kendall. London, 1948.

EXAMPLES 1

1. An office has analysed its new business figures over a number of years for endowment assurances with profits. The following table shows the distribution according to age next birthday at entry:

Age next birthday at entry	No. of policies	Age next birthday at entry	No. of policies
15–19	30	45–49	270
20–24	200	50–54	180
25–29	450	55–59	80
30–34	420	60–64	20
35–39	400	65–69	5
40–44	350	70 and over	Nil

Assuming that the exact age is on the average ·35 yr. less than the age next birthday, calculate the mode, standard deviation, mean deviation and quartile deviation. Apply any tests you know to check approximately the last two of these values and comment on the results.

2. The following table gives a frequency distribution of ages of bridegrooms:

Age of bridegroom	Frequency	Age of bridegroom	Frequency
15–	15	54–	83
18–	550	57–	55
21–	3050	60–	40
24–	3653	63–	32
27–	2825	66–	24
30–	1674	69–	16
33–	1028	72–	11
36–	714	75–	6
39–	466	78–	4
42–	312	81–	1
45–	238	84–	1
48–	181	—	—
51–	110		

Calculate the mean, median, mode, quartile deviation, mean deviation and standard deviation. Apply the approximate relationships connecting the values of these indices to check your results.

3. The frequency distribution of a measurable characteristic x varying between 0 and 2 may be represented by the following expressions:

The frequencies are proportional to x^3 for values of x between 0 and 1 and to $(2-x)^3$ for values of x between 1 and 2.

Find by separate calculation in each case the values of the mean deviation, standard deviation and probable error of the distribution.

4. A frequency curve fitting an observed distribution is given by the two equations

$$\left.\begin{aligned} x &= a + a\sin\theta \\ y &= a - a\cos\theta \end{aligned}\right\},$$

where θ varies between $-\pi/2$ and $\pi/2$. y is the frequency with which the value x occurs.

Calculate the mean, median, mode, standard deviation, mean deviation, probable error and a measure of skewness. Draw the frequency curve.

If the observed distribution relates to the number of hours' sunshine recorded at Greenwich on each 21st of March over a period of 50 years, state what units a will represent in relation to x and in relation to y.

5. The following table shows the number of deaths (in thousands) among the male population in England and Wales in the years 1930-32:

Age at death	No. of deaths	Age at death	No. of deaths	Age at death	No. of deaths
0–	97	35–	18	70–	84
5–	12	40–	24	75–	72
10–	7	45–	33	80–	45
15–	13	50–	44	85–	20
20–	17	55–	57	90–	5
25–	16	60–	68	95–	1
30–	16	65–	81	100–	0
				Total	730

Calculate the mean age at death, the standard deviation of the age at death and the coefficients of skewness by formulae (12) and (14). Can you suggest a reason for the difference between the two measures of skewness?

6. The following table shows the distribution in groups, according to sum assured and rate of premium per cent, of a number of policies effected with an insurance company:

Central sum assured in group, £	Central rate of premium in group						Total no. of policies	Average rate of premium £
	10*s.*	£1. 10*s.*	£2. 10*s.*	£3. 10*s.*	£4. 10*s.*	£5. 10*s.*		
50	1	2	6	6	2	1	18	3·00
150	2	1	3	8	3	0	17	3·03
250	0	1	3	4	1	1	10	3·30
350	0	3	6	7	5	1	22	3·27
450	2	2	4	8	4	1	21	3·12
550	1	1	3	7	0	0	12	2·83
Total no. of policies	6	10	25	40	15	4	100	—
Average sum assured, £	300	300	282	310	290	275	—	—

Criticize the following observations regarding the figures and calculate any statistical measures you consider necessary in dealing with the questions mentioned:

"The average values of the sums assured shown in the last line of the table vary between £275 and £310—i.e. a range of £35 in relation to an average amount of about £300—while the average values of the rate of premium given in the last column vary between 2·83 and 3·30—i.e. a range of ·47 in relation to about 3·00. It is clear from these figures that the sum assured under the policies is on the whole a much more stable quantity than the rate of premium per cent under the policies. Further, the arithmetic mean of the first three values of the average sum assured in the last line of the table is £294 as compared with a figure of about £292 for the second three values, while the first three premiums in the last column have an average value of 3·11 as compared with a corresponding average of about 3·07 in respect of the second three. Both of these results show clearly that the larger sums assured are associated with the smaller rates of premium and vice versa."

7. Find the mode of the following data derived from Friendly Society records (*a*) by a graphic process, (*b*) by an analytical process.

Age at commencement of illness	No. of claims	Age at commencement of illness	No. of claims
16–18	27	40–50	172
18–22	48	50–55	60
22–25	40	55–60	52
25–30	64	60–65	40
30–40	135	Over 65	Nil

8. From the following table of the exposed to risk in central form and the corresponding deaths per annum, find the values of μ_x and q_x for ages 32, 37, 42 and 47. How would you find the same functions for ages 27 and 52?

Age group	Exposed to risk in central form	No. of deaths per annum
25–30	7,300	44
30–35	10,500	74
35–40	14,700	118
40–45	15,400	140
45–50	14,000	154
50–55	12,300	160

9. The following table shows the frequencies with which values of a continuous variable were observed to lie within the ranges shown. Find the fourth moment about the mean, using Sheppard's adjustments.

Value of variable	Frequency	Value of variable	Frequency
0–	41	·6–	500
·1–	152	·7–	411
·2–	235	·8–	260
·3–	318	·9–	72
·4–	470	Over 1	Nil
·5–	560		

CHAPTER II

IMPORTANT FREQUENCY DISTRIBUTIONS

1. The binomial frequency distribution.

For reasons which will be apparent when we come to consider Mortality Tables we shall denote the probability of a *success* by q and the probability of a *failure* by p.

If we have n independent trials with the probability of success at each trial q, we know that the probability of n successes is q^n. Similarly, the probability of $n-1$ successes is $nq^{n-1}p$ and, generally, that of r successes ${}^nC_rq^rp^{n-r}$. In fact, the probabilities of 0, 1, 2, ... n successes are the successive terms in the expansion of $(p+q)^n$.

In statistics we are interested in frequencies rather than probabilities as such and we usually imagine that the n trials are repeated N times. Then

the expected *frequency* with which n successes will be obtained is Nq^n,

,, ,, $n-1$,, ,, $Nnq^{n-1}p$,

....

,, ,, r ,, ,, $N{}^nC_rq^rp^{n-r}$,

....

,, ,, 0 ,, ,, Np^n.

Obviously the actual frequencies in a sequence of N repetitions of n trials will all be integers.

The theoretical frequencies shown above will not, in general, be integers, but it is to be expected that the actual frequencies will not differ greatly from them.

Before we proceed to calculate the various statistical constants of the theoretical distribution it is as well to examine exactly what assumptions have been made: the important practical example of mortality data will serve as a test case.

If we were to consider n lives for each of whom the probability of dying within one year was q, the probability of r deaths would be ${}^nC_rq^rp^{n-r}$. Again, if there were N groups each of n lives, as before, we should expect to find that r people had died in $N\,{}^nC_r q^r p^{n-r}$ groups and the frequency distribution would be as follows:

No. of deaths	0	1	2	...	r	...	n
No. of groups with above no. of deaths	Np^n	$Nnqp^{n-1}$	$N\,{}^nC_2q^2p^{n-2}$	...	$N\,{}^nC_rq^rp^{n-r}$	...	Nq^n

Incidentally we sometimes omit the N and refer to the probabilities as "proportionate frequencies", i.e. the proportion of trials in which, say, r successes would be obtained out of n.

We have assumed that all the N groups are exactly alike and that each of the n lives in each group has the same chance of dying within a year. In practice data rarely relate to the number of deaths because of the difficulties of ensuring that each life is counted once only in the exposed to risk and the deaths; it is usual to base an investigation on the number of policies, with certain adjustments with which we are not at present concerned. Thus a man who dies having four policies in force will probably give rise to four claims, and these will be reckoned as four separate deaths. In this way the inclusion of duplicates upsets the assumptions made above, and such factors as epidemics, wars, local environment, etc. may mean that the chance of one man dying is not independent of the chance of another man dying.

Statistical methods should therefore be used with discretion in dealing with such data and results obtained by them should not be interpreted too dogmatically.

2. Mean and standard deviation of the binomial distribution.

In the general case of class-interval h we shall have the values $0, h, 2h, 3h, \dots nh$ occurring with frequencies $Np^n, Nnp^{n-1}q, \dots Nq^n$.

The total frequency $= N(p+q)^n = N$.

The mean is therefore

$$\begin{aligned}\frac{1}{N}[0Np^n + hNnp^{n-1}q + 2hN\,{}^nC_2p^{n-2}q^2 + \dots + nhNq^n]\\ = nqh[p^{n-1} + (n-1)p^{n-2}q + \dots + q^{n-1}]\\ = nqh[p+q]^{n-1}\\ = nqh. \qquad \dots\dots(1)\end{aligned}$$

The second moment about the origin (m_2) is

$$\frac{1}{N}[0^2Np^n+h^2Nnp^{n-1}q+4h^2N\,{}^nC_2p^{n-2}q^2+\dots+n^2h^2Nq^n]$$

$$=nqh^2\left[p^{n-1}+2(n-1)p^{n-2}q+3\frac{(n-1)(n-2)}{2!}p^{n-3}q^2+\dots+nq^{n-1}\right]. \qquad \text{......(2)}$$

The expression in brackets is the first moment about -1 of the distribution $(p+q)^{n-1}$. From (1) we know that the first moment about -1 is $(n-1)q+1$, and the expression (2) therefore reduces to

$$nqh^2[(n-1)q+1].$$

The second moment about the mean (μ_2) is derived by subtracting the square of the mean from m_2. It is therefore

$$nqh^2[(n-1)q+1]-n^2q^2h^2=nq(1-q)h^2$$

$$=npqh^2. \qquad \text{......(3)}$$

Hence $$\sigma=h\sqrt{npq}. \qquad \text{......(4)}$$

The results (1) and (4) are very important.

Example 1.

A throw of 5 with two dice being counted as a success, we have $q=\frac{1}{9}$, where q is the probability of success.

Hence, if four pairs of dice are thrown, the chances of four, three, two, one and no successes are the successive terms in the expansion of

$$(\tfrac{1}{9}+\tfrac{8}{9})^4.$$

If the four pairs of dice are thrown 9^4 times the frequencies of 0–4 successes would be as follows, if the theoretical probabilities were realized:

No. of successes	4	3	2	1	0
Frequency	1	32	384	2048	4096

The mean of these is $\frac{1}{6561}[4+(3\times 32)+(2\times 384)+(1\times 2048)]=\frac{4}{9}$, as it should be according to formula (1).

Similarly, the standard deviation will be found to be $\sqrt{\frac{32}{81}}$, which is

$$\sqrt{npq}, \quad \text{where} \quad n=4,\ q=\tfrac{1}{9},\ p=\tfrac{8}{9}.$$

3. The normal curve of error.

In its simplest form the equation of this curve is

$$Y = e^{-X^2}.$$

Any other form can be reduced to this merely by a change of scale and a change of origin.

The most common form of the equation is $y = y_0 e^{-\frac{x^2}{2\sigma^2}}$.

This is derived from the simple equation above by putting

$$Y = \frac{y}{y_0} \quad \text{and} \quad X = \frac{x}{\sigma\sqrt{2}},$$

where y_0 and σ are constants, the significance of which will be considered later.

There is in fact only one normal curve, a fact which makes it of the first importance in statistical theory.

The curve $y = y_0 e^{-\frac{x^2}{2\sigma^2}}$ is clearly symmetrical about the y-axis and approaches the x-axis asymptotically as $x \to \pm\infty$; the total area corresponding to the total frequency of the distribution represented by the curve

$$= y_0 \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx = 2y_0 \int_0^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx.$$

Putting $\frac{x^2}{2\sigma^2} = t$ the integral becomes

$$\sqrt{2}y_0\sigma \int_0^{\infty} e^{-t} t^{-\frac{1}{2}} dt = \sqrt{2}y_0\sigma\Gamma(\tfrac{1}{2})$$

$$= \sqrt{2\pi}y_0\sigma.$$

[The proof that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ is outside the scope of this book.]

Hence, if we denote as usual the total frequency by N, the equation of the curve is

$$y = \frac{N}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \qquad \ldots\ldots(5)$$

So far we have regarded σ merely as a constant in the equation. We shall now show that it is in fact the standard deviation of the distribution. As the curve is symmetrical the mean is clearly zero, as are the mode and median.

The square of the standard deviation (S.D.) is

$$\frac{1}{N}\int_{-\infty}^{\infty} x^2 \frac{N}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{\sqrt{2}}{\sqrt{\pi}\sigma}\int_0^\infty x^2 e^{-\frac{x^2}{2\sigma^2}}\,dx.$$

Integrate by parts, taking x as the first part and $xe^{-\frac{x^2}{2\sigma^2}}$ as the second.

$$\text{Square of S.D.} = \frac{\sqrt{2}}{\sqrt{\pi}\sigma}\left[x\left(-\sigma^2 e^{-\frac{x^2}{2\sigma^2}}\right)\right]_0^\infty + \frac{\sqrt{2}\sigma}{\sqrt{\pi}}\int_0^\infty e^{-\frac{x^2}{2\sigma^2}}\,dx$$

$$=\sigma^2,$$

since the first bracket vanishes at both limits and

$$\int_0^\infty e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{\sigma\sqrt{\pi}}{\sqrt{2}}, \text{ as shown above.}$$

4. Standard tables.

When the total frequency N and σ are both unity the equation of the curve is

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Extensive sets of tables have been published for the curve in this form. The most important are those which give:

(i) the ordinate y for values of x which are at very close intervals when x is small and y changing fairly rapidly, and for values of x at less frequent intervals, when x is large and y is changing only slowly;

(ii) values of $\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$ for different positive values of z.

This function is usually denoted by $\frac{1}{2}(1+\alpha_z)$ and represents the area of the normal curve lying to the left of the ordinate $x=z$.

Since the area to the left of the origin is $\frac{1}{2}$ (the total area being unity) it follows that the area bounded by the curve, the axis of x and the ordinates $x=0$, $x=z$ is $\frac{1}{2}\alpha_z$.

Hence α_z represents the area of the curve lying between the ordinates $x=\pm z$ and can readily be obtained from the tabulated values of $\int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$ by doubling and subtracting unity from the result.

Full tables of the ordinate $\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ and of the area $\frac{1}{2}(1+\alpha)$ are given in Sheppard's *Tables of Area and Ordinate in terms of Abscissa.* In using them for the general Normal Distribution, x/σ must be taken as a new variable x' (say) for entering the table and the ordinate so obtained must be multiplied by N/σ, while the area must be multiplied by N.

Table I in the Appendix can also be used as shown in Example 2. The student will find this table reproduced in *A Short Collection of Actuarial Tables* for use in the examinations.

The importance of this table will be appreciated when sampling is discussed (Chapter IV), but the following will at once be obvious.

Taking the general equation $y=\frac{N}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$, the probability that an observation taken at random lies within k, say, of the mean is clearly the area of the curve lying between the ordinates $x=-k$ and $x=k$ divided by the total frequency: this probability is therefore

$$\frac{2}{\sqrt{2\pi}\sigma}\int_0^k e^{-\frac{x^2}{2\sigma^2}}dx. \qquad \ldots\ldots(6)$$

By taking x/σ as a new variable the values of these probabilities can be read off (or interpolated where necessary) from the prepared tables. The following values are important:

k	Probability
$\cdot 6745\sigma$	·5000
σ	·6827
2σ	·9545
3σ	·9973

Thus we see that about 95½ per cent of the total area lies between $x=-2\sigma$ and $x=+2\sigma$, while no less than 99·73 per cent lies between $x=-3\sigma$ and $x=+3\sigma$.

Expressed differently, this means that in a normal frequency distribution about 95½ per cent of the observations lie within a distance 2σ of the mean while about 99·7 per cent lie within a distance 3σ of the mean.

5. Probable error.

The first entry in the table relates to the *probable error* (*Mathematics for Actuarial Students*, Part II, Chap. XII, para. 11).

If we consider a general frequency curve $y=f(x)$, the probable error (P) is given by the equation

$$\int_{M-P}^{M+P} f(x)\,dx = \frac{1}{2}\int_a^b f(x)\,dx, \qquad \ldots\ldots(7)$$

where M is the mean and the frequency curve stretches from $x=a$ to $x=b$. Expressed in words this means that half the total frequency occurs for values of x between (mean $-P$) and (mean $+P$). An observed value of x selected at random is as likely to fall within the range $M-P$ to $M+P$ as it is to lie outside that range.

The probable error is rarely used nowadays; for the normal curve it is approximately equal to $\cdot 6745\sigma$.

6. Mean deviation of the normal distribution.

Since the mean and median are zero, the mean deviation

$$= \frac{2}{N}\int_0^\infty \frac{N}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} x\,dx$$

$$= -\frac{2\sigma}{\sqrt{2\pi}}\left[e^{-\frac{x^2}{2\sigma^2}}\right]_0^\infty$$

$$= \frac{\sqrt{2}\sigma}{\sqrt{\pi}}$$

$$= \cdot 7979\sigma \text{ approx.}$$

This value for the mean deviation is also widely used in the form

$$\text{mean deviation} = \tfrac{4}{5} \text{ standard deviation}, \qquad \ldots\ldots(8)$$

but may be wide of the mark if the distribution considered is not approximately normal.

Example 2.

A normal distribution of a continuous variable has a mean of 13·4 and a standard deviation of 2·5. Find the probability that a value selected at random lies between the values 11·8 and 15·0.

Taking the mean as origin the limits become $-1\cdot 6$ and $1\cdot 6$. Hence the probability required is, by (6) above,

$$\frac{2}{\sqrt{2\pi}\times 2\cdot 5}\int_0^{1\cdot 6} e^{-\frac{x^2}{2\times 2\cdot 5^2}}\,dx.$$

To make use of the tables in the Appendix we proceed as follows:

Let $$x=2\cdot 5x', \quad dx=2\cdot 5dx'.$$

When $x=1\cdot 6$, $$x'=1\cdot 6/2\cdot 5=\cdot 64.$$

The required probability is therefore $\frac{2}{\sqrt{\pi}}\int_0^{\cdot 64} e^{-\frac{x'^2}{2}}\,dx'$, which by use of the tables is found to be $\cdot 478$ approx.

More generally, the substitution $x=\sigma x'$ reduces the integral

$$\frac{2}{\sqrt{2\pi}\sigma}\int_0^{k} e^{-\frac{x^2}{2\sigma^2}}\,dx$$

to $$\frac{2}{\sqrt{\pi}}\int_0^{k/\sigma} e^{-\frac{x'^2}{2}}\,dx' \qquad \ldots\ldots(9)$$

and the tables can be used directly.

Example 3.

The following table (A) gives the distribution of heights of 1000 men. Find the normal curve representing the same total frequency and with the same mean and the same standard deviation. Draw the curve and the histogram.

Compare the value of the interquartile range obtained from the statistics with that of the corresponding normal distribution.

Table A

Stature in inches	No. of men within these limits of stature	Stature in inches	No. of men within these limits of stature
61·5–62·5	4·0	69·5–70·5	138·5
62·5–63·5	19·0	70·5–71·5	108·0
63·5–64·5	24·5	71·5–72·5	53·5
64·5–65·5	40·5	72·5–73·5	47·5
65·5–66·5	84·5	73·5–74·5	21·0
66·5–67·5	123·5	74·5–75·5	12·0
67·5–68·5	139·0	75·5–76·5	5·0
68·5–69·5	179·0	76·5–77·5	·5

Table B gives values of the function $y' = e^{-\frac{x'^2}{2}}$.

Table B

x'	$y' = e^{-\frac{x'^2}{2}}$	$\log y'$	x'	$y' = e^{-\frac{x'^2}{2}}$	$\log y'$
0	1·00000	0	2·6	·03405	$\bar{2}\cdot53209$
0·2	·98020	$\bar{1}\cdot99131$	2·8	·01984	$\bar{2}\cdot29757$
0·4	·92312	$\bar{1}\cdot96526$	3·0	·01111	$\bar{2}\cdot04567$
0·6	·83527	$\bar{1}\cdot92183$	3·2	·00598	$\bar{3}\cdot77641$
0·8	·72615	$\bar{1}\cdot86103$	3·4	·00309	$\bar{3}\cdot48978$
1·0	·60653	$\bar{1}\cdot78285$	3·6	·00153	$\bar{3}\cdot18577$
1·2	·48675	$\bar{1}\cdot68731$	3·8	·00073	$\bar{4}\cdot86439$
1·4	·37531	$\bar{1}\cdot57439$	4·0	·00034	$\bar{4}\cdot52564$
.1·6	·27804	$\bar{1}\cdot44410$	4·2	·00015	$\bar{4}\cdot16952$
1·8	·19790	$\bar{1}\cdot29644$	4·4	·00006	$\bar{5}\cdot79603$
2·0	·13534	$\bar{1}\cdot13141$	4·6	·00003	$\bar{5}\cdot40516$
2·2	·08892	$\bar{2}\cdot94901$	4·8	·00001	$\bar{6}\cdot99693$
2·4	·05614	$\bar{2}\cdot74923$	5·0	—	$\bar{6}\cdot57132$

There are two points to note about Table A. First, the frequencies ending in ·5 are probably caused by men whose height (to the degree of accuracy adopted) coincided with the limit of a group, e.g. 65·5 inches. It is customary in such circumstances to allot ·5 to each of the two groups adjoining. Thus three men of height 65·5 inches would be counted as 1·5 in the group 64·5–65·5 and 1·5 in the group 65·5–66·5.

The second point to notice is that the distribution is roughly symmetrical and an attempt to fit a normal curve seems likely to be fairly successful in view of the run of the data.

The general equation of the normal curve is

$$y = \frac{N}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}},$$

referred to the ordinate through the mean as axis of y.

Hence we need to find the mean (so as to fix the axes), the total frequency N and the standard deviation σ.

Assuming that the total frequency of each group is concentrated at the mid-point, we proceed thus (see table on p. 40):

69 is taken as an arbitrary origin. Referred to this origin the mean is

$$-\tfrac{145}{1000} \text{ inches.}$$

$$\therefore \ \text{mean} = 69 - \cdot145 = 68\cdot855 \text{ inches.}$$

Stature in inches	(1) − 69	Frequency f_x	(2) × (3)		(2) × (4)	Σ (3)
(1)	(2)	(3)	(4)		(5)	(6)
			−	+		
62	−7	4	28		196	4
63	−6	19	114		684	23
64	−5	24·5	122·5		612·5	47·5
65	−4	40·5	162		648	88
66	−3	84·5	253·5		760·5	172·5
67	−2	123·5	247		494	296
68	−1	139	139		139	435
69	0	179	0		0	614
70	1	138·5		138·5	138·5	752·5
71	2	108		216	432	860·5
72	3	53·5		160·5	481·5	914
73	4	47·5		190	760	961·5
74	5	21		105	525	982·5
75	6	12		72	432	994·5
76	7	5		35	245	999·5
77	8	·5		4	32	1000
Total	—	1000	−1066	+921	6580	—
			= −145			

m_2 (second moment about the origin) $= \frac{6580}{1000}$, ignoring Sheppard's adjustment.

$\therefore$ μ_2 (second moment about the mean) $= 6{\cdot}580 - (-{\cdot}145)^2$, ignoring Sheppard's adjustment, $= 6{\cdot}559$.

Subtracting $\frac{1}{12}$ of the class-interval (Sheppard's adjustment) from this value for μ_2 we obtain 6·476 and

$$\sigma = \sqrt{6{\cdot}476} = 2{\cdot}54 \text{ inches.}$$

The lower quartile separates the 250th and the 251st observations and from the last column we see that it lies in the range 66·5–67·5, which includes 123·5 observations. Hence the lower quartile

$$= 66{\cdot}5 + \frac{77{\cdot}5}{123{\cdot}5} = 67{\cdot}13 \text{ inches.}$$

Similarly, the upper quartile, which separates the 750th and the 751st observations,

$$= 69{\cdot}5 + \frac{136}{138{\cdot}5} = 70{\cdot}48 \text{ inches,}$$

or alternatively $= 70{\cdot}5 - \frac{2{\cdot}5}{138{\cdot}5} = 70{\cdot}48$ inches as before.

$\therefore$ interquartile range $= 3{\cdot}35$ inches.

It will be remembered that by definition one-quarter of the total frequency occurs for values of the variable between the lower quartile and the median and also between the median and the upper quartile.

The difference between the lower quartile and the median is equal to the difference between the median and the upper quartile only in a symmetrical curve.

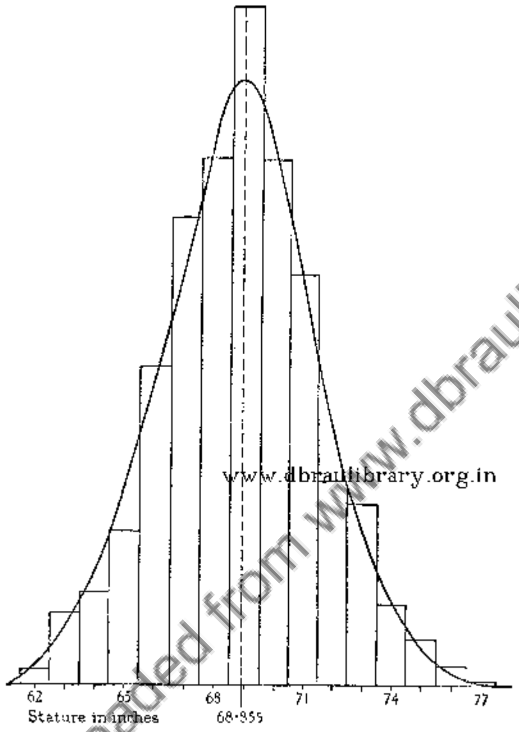


Fig. 1

If we measure from the mean a distance equal to P, the probable error, in both directions we enclose half the total frequency.

In a symmetrical curve such as the one considered, the median and mean coincide and it will be seen that the interquartile range is equal to $2P = 2 \times \cdot 6745\sigma$ in a normal curve.

$$\text{Hence the interquartile range} = 2 \times \cdot 6745 \times 2 \cdot 54$$
$$= 3 \cdot 43,$$

as compared with 3·35 obtained above from the data.

This is one indication that the normal curve is not a perfect fit.

We proceed to plot the curve and the histogram representing the given data.

The equation of the curve is

$$y=\frac{1000}{\sqrt{2\pi}\times 2{\cdot}54}e^{-\frac{x^2}{2(2{\cdot}54)^2}}.$$

To put this in the form $y'=e^{-\frac{x'^2}{2}}$, so that the given table of values can be used, we let

$$y=\frac{1000}{\sqrt{2\pi}\times 2{\cdot}54}y',$$

i.e. $\log y = 3-\tfrac{1}{2}\log 2\pi - \log 2{\cdot}54+\log y'$

and $x=2{\cdot}54x'$.

$\log\pi = {\cdot}4971496$, $\log 2 = {\cdot}3010300$ and $\log 2{\cdot}54 = {\cdot}4048337$.

$$\therefore\ \log y=\log y'+2{\cdot}19608.$$

The following table is constructed from Table B by adding this constant to the given values:

x	$\log y$	y	x	$\log y$	y
0	2·19608	157·06	6·604	·72817	5·35
·508	2·18739	153·95	7·112	·49365	3·12
1·016	2·16134	144·99	7·620	·24175	1·74
1·524	2·11701	130·93	8·128	$\bar{1}$·97249	·94
2·032	2·05711	114·05	8·636	$\bar{1}$·68586	·49
2·540	1·97893	95·26	9·144	$\bar{1}$·38185	·24
3·048	1·88339	76·45	9·652	$\bar{1}$·06047	·11
3·556	1·77047	58·95	10·160	$\bar{2}$·72172	·06
4·064	1·64018	43·67	10·668	$\bar{2}$·36560	·02
4·572	1·49252	31·08	11·176	$\bar{3}$·99211	·01
5·080	1·32749	21·26	11·681	$\bar{3}$·60124	·00
5·588	1·14509	13·97	12·192	$\bar{3}$·19301	·00
6·096	·94531	8·82	12·700	$\bar{4}$·76740	·00

In order to illustrate the rapid tailing off of the values of y more significant figures have been kept than can be used in drawing the graph.

The curve and the histogram are shown in Fig. 1.

7. Approximation of the binomial distribution to the normal curve.

The binomial distribution may be represented by a curve drawn through $n+1$ points corresponding to the $n+1$ terms of the expansion. This diagram will be symmetrical if $p=q$, and if n is large will not be very different from the normal curve.

We shall now show that as $n\to\infty$ the distribution does in fact approximate to the normal distribution. Incidentally, the series of points will at the same time approximate to a continuous curve.

As n is to tend to infinity we can, without loss of generality, assume that $n=2k$. As $p=q=\frac{1}{2}$, the binomial distribution can be written

$$N(\tfrac{1}{2}+\tfrac{1}{2})^{2k}.$$

The central term $=N\dfrac{2k!}{k!\,k!}(\tfrac{1}{2})^{2k}=y_0$.

To arrive at the equation of the normal curve this term must correspond to $x=0$, and if we imagine the points representing successive terms of the expansion to be at intervals of h, the term corresponding to $x=\pm rh$ will be

$$N\frac{2k!}{(k-r)!\,(k+r)!}(\tfrac{1}{2})^{2k}=y_{rh}.$$

Hence
$$\frac{y_{rh}}{y_0}=\frac{k(k-1)(k-2)\dots(k-r+1)}{(k+r)(k+r-1)\dots(k+1)}$$
$$=\frac{1\left(1-\frac{1}{k}\right)\left(1-\frac{2}{k}\right)\dots\left(1-\frac{r-1}{k}\right)}{\left(1+\frac{1}{k}\right)\left(1+\frac{2}{k}\right)\dots\left(1+\frac{r}{k}\right)}.$$

$\therefore$ provided $r<k$
$$\log y_{rh}-\log y_0=\log\left(1-\frac{1}{k}\right)+\log\left(1-\frac{2}{k}\right)+\dots+\log\left(1-\frac{r-1}{k}\right)$$
$$-\log\left(1+\frac{1}{k}\right)-\log\left(1+\frac{2}{k}\right)-\dots-\log\left(1+\frac{r}{k}\right)$$
$$=-\frac{2}{k}(1+2+3+\dots+\overline{r-1})-\frac{r}{k}-\text{terms}$$
involving second and higher powers of $1/k$
$$=-\frac{r^2}{k}\text{ approximately.}$$

$$\therefore\; y_{rh}=y_0e^{-\frac{r^2}{k}}. \qquad \dots\dots(10)$$

To derive a continuous curve the interval h between the points must tend to zero. In that event rh becomes the abscissa x.

Thus
$$\frac{r^2}{k}=\frac{x^2}{kh^2}.$$

We know that for the binomial distribution $\sigma^2=npqh^2$.

As $n=2k$ and $p=q=\frac{1}{2}$,

$$\sigma^2=2k\,.\,\tfrac{1}{2}\,.\,\tfrac{1}{2}h^2 \quad \text{and} \quad \frac{r^2}{k}=\frac{x^2}{2\sigma^2}.$$

Substituting in (10), we obtain

$$y=y_0e^{-\frac{x^2}{2\sigma^2}}. \qquad \ldots\ldots(11)$$

8. The Poisson Distribution.

In deriving the normal curve as an approximation to the binomial it was assumed that $p=q$ and $n\rightarrow\infty$. In practical work this means that p and q should not be very dissimilar and that n should be large.

In some fields of statistics, q is often very small indeed although n is so large that the product nq is appreciable. For instance, the chance of a given employee being involved in an accident in a factory (q) is usually very small but the number employed (n) is so large that nq, the chance of an accident occurring to someone is quite an important consideration. Similarly, the probability of any given person dying within a year is very small indeed except at high ages but the number exposed to the risk of death in most investigations into mortality is very large.

In such circumstances the binomial distribution can be represented approximately as follows:

Let $nq=m$ and assume that $n\rightarrow\infty$ and $q\rightarrow 0$, m remaining finite.

The probability of r successes in n trials $=\dfrac{n!}{r!(n-r)!}q^rp^{n-r}$ which may be written

$$\frac{m^r}{r!}\left(1-\frac{m}{n}\right)^n\frac{n!}{(n-r)!\,n^r(1-m/n)^r}. \qquad \ldots\ldots(12)$$

To simplify this expression we make use of Stirling's approximation:

$$n!\doteqdot\sqrt{2\pi}\,e^{-n}n^{n+\frac{1}{2}}$$

Substituting for the terms involving factorials we obtain:

$$\frac{n!}{(n-r)!\,n^r(1-m/n)^r}\doteqdot\frac{e^{-n}n^{n+\frac{1}{2}}}{e^{-n+r}(n-r)^{n-r+\frac{1}{2}}n^r(1-m/n)^r}$$

i.e.

$$\doteqdot\frac{1}{e^r(1-r/n)^{n-r+\frac{1}{2}}(1-m/n)^r}. \quad \ldots\ldots(13)$$

As $n\to\infty$ $\left(1-\frac{r}{n}\right)^{n-r+\frac{1}{2}}\to e^{-r}$ and $(1-m/n)^r\to 1$.

Hence (13)→1 and (12) tends to the form:

$$\frac{m^r}{r!}(1-m/n)^n \quad \text{i.e.} \quad \frac{m^r}{r!}e^{-m}. \qquad \text{......(14)}$$

Thus the probabilities of 0, 1, 2 ... r successes become in the limit the successive terms of the series:

$$e^{-m}\left\{1+\frac{m}{1!}+\frac{m^2}{2!}+\dots\frac{m^r}{r!}+\dots\right\}. \qquad \text{......(15)}$$

The distribution represented by the successive terms of (15) is called the *Poisson distribution.*

It will be seen that the total probability for all values of r is unity. It is left to the student to prove that the mean is m and the standard deviation $\sqrt{m}$.

9. The use of the normal approximation.

It may seem at first that the normal curve will not in general be a very close approximation to the binomial distribution. In this connection it will be realized:

1. That the binomial distribution is a series of points approximating to a continuous curve only when the number of terms is indefinitely increased.
2. That the binomial distribution is finite, while the normal curve approaches infinity in either direction.
3. That, unless $p=q$, the binomial distribution is skew, while the normal curve is symmetrical.

In actual practice the approximation in the neighbourhood of the mean is much better than these considerations would suggest.

If we measure a distance $K\sigma$ from the mean in both directions a balance of errors is obtained and the frequency enclosed by the normal curve may be a satisfactory approximation to the frequency in the binomial distribution within the same limits.

This is very important, because it means that the sets of tables prepared for the normal curve may sometimes be used for a binomial distribution. Thus from the table on p. 36 we often assume that about two-thirds of the total frequency occurs for values of x lying between $M-\sigma$ and $M+\sigma$, where M is the mean and σ the standard deviation, while about 90 per cent occurs for values of x lying between $M-2\sigma$ and $M+2\sigma$. The student will, however, realize that unless n is fairly large or p is nearly equal to q these assumptions may be far from the truth.

Example 4.

The following table gives the first 13 terms in the expansion of

$$10{,}000(\cdot 95+\cdot 05)^{100},$$

the figure of 10,000 being introduced to eliminate decimals.

Value of variable x	Frequency	Accumulated frequency	Value of variable x	Frequency	Accumulated frequency
0	59	59	8	649	9,368
1	311	370	9	349	9,717
2	812	1,182	10	167	9,884
3	1,396	2,578	11	72	9,956
4	1,781	4,359	12	28	9,984
5	1,800	6,159			
6	1,500	7,659	13–100 inclusive	16	10,000
7	1,060	8,719			

The distribution is clearly very skew.

In the previous notation $N=10{,}000$, $n=100$, $p=\cdot 95$, $q=\cdot 05$.

The mean $=nq=5$.

The standard deviation $=\sqrt{npq}=10\sqrt{\cdot 0475}=2\cdot 18$.

The limits $M-\sigma$ to $M+\sigma$ become 2·82 to 7·18 and include 75·37 per cent of the total frequency.

The limits $M-2\sigma$ to $M+2\sigma$ become ·64 to 9·36 and include 96·58 per cent of the total frequency.

If we used a Poisson distribution as an approximation we should take m, the Poisson parameter, equal to the observed mean viz. 5 and the standard deviation would be $\sqrt{5}$ or 2·24 as compared with the more accurate value of 2·18.

Example 5.

An assurance company is about to assure a group of 10,000 lives all aged x for a sum of £100 payable on each death within one year. The company wishes to charge the minimum single premium which will ensure that its probability of suffering a loss on the whole transaction will not be greater than one-fourth.

Calculate the approximate single premium per cent which should be charged to each member, ignoring interest and expenses, and assuming that for a life aged x the probability of death within one year is ·01.

Discuss briefly the effect on the problem of each of the following variations:

(i) the sum assured being £200 instead of £100;
(ii) the number of lives being 100 instead of 10,000;
(iii) the specified probability of loss being some figure other than one-fourth.

Calculate the approximate single premium per cent in cases (i) and (ii).

The probabilities of 0, 1, 2, ... deaths occurring are the successive terms in the expansion of $(·99+·01)^{10,000}$.

The terms are proportionate frequencies and are exactly similar to the frequencies we have been discussing except that $N=1$.

The mean number of deaths is therefore nq or 100 and the standard deviation $\sqrt{npq}$ or 10 very nearly.

Since n is large we can assume that the distribution is nearly normal, especially as only an approximate result is required. P, the probable error, is, in that event, $·6745\sigma$

$$=6·745$$

If M is the mean, half the total frequency lies between $M-P$ and $M+P$, or *in a symmetrical distribution* a quarter of the total frequency occurs for values of x greater than $M+P$. Applied to our proportionate frequencies this means that the total of the frequencies for deaths in excess of $100+6·745$ is only one-quarter, i.e. if we base our premium on the assumption that 107 deaths will occur the premium will prove inadequate in less than one case in four.

The required premium is therefore £10,700 or £1. 1*s*. 5*d*. per cent, say.

The effects of the modifications set out in the question are as follows:

(i) This has no effect on the premium per cent since the amount of the claim is the same for each death. If the sum assured were not the same for all the lives the problem would be more complicated. This aspect will be discussed later. (Chapter III, Ex. 3.)

(ii) The expression we now have to consider is $(·99+·01)^{100}$ and the mean of the distribution formed by the terms of the expansion is unity.

The standard deviation is $\sqrt{100\times·99\times·01}=·995$.

We can no longer assume that the distribution approximates to the normal over most of its range, because n is now relatively small. We can now, however, expand the binomial to obtain:

Chance of no deaths occurring $= (\cdot 99)^{100}$ $= \cdot 366$ approx.

,, one death ,, $= 100(\cdot 99)^{99}(\cdot 01)$ $= \cdot 370$,,

,, two deaths ,, $= \frac{100 \times 99}{2}(\cdot 99)^{98}(\cdot 01)^{2} = \cdot 185$,,

Hence the chance of more than *one* death is

$$1 - \cdot 366 - \cdot 370 = \cdot 264,$$

while the chance of more than *two* deaths is

$$1 - \cdot 366 - \cdot 370 - \cdot 185 = \cdot 079.$$

A premium based on one death occurring out of a hundred would be inadequate in more than one case in four and we must allow for two deaths by charging a premium of £2 per cent.

(iii) The probability of one-fourth enables us to use the probable error when 10,000 lives are involved. For any other probability, say one-fifth, we might refer to prepared tables based on the normal curve. Such a table shows, for instance, that the probability of an observation lying outside the ordinates $x = \pm \cdot 8\sigma$ (the mean being the origin) is ·42371, while the probability that it lies outside the ordinates $x = \pm \cdot 9\sigma$ is ·36812.

Since the curve is symmetrical we deduce that the probability of a value being greater than $\cdot 8\sigma$ is half of $\cdot 42371 = \cdot 21185$, while the probability that it is greater than $\cdot 9\sigma$ is ·18406.

By interpolation the chance that it is greater than $\cdot 84\sigma$ is ·2, or one-fifth.

In other words, if we were to charge a premium which could be expected to prove inadequate on only one occasion in five we should allow for a number of deaths equal to the mean $+ \cdot 84\sigma$.

For 10,000 assured we therefore allow for $100 + \cdot 84\sigma \times 10$ deaths, or, say, 109 deaths, by charging a premium of £1·09 or £1. 1*s*. 10*d*. per cent.

For 100 deaths the normal curve is not a sufficiently good approximation, and unless the probability of loss were small the labour involved in evaluating successive terms of the expansion would be heavy.

Actually the premium of £2 per cent calculated in (ii) would cover two deaths and the probability that it would prove inadequate is only ·079, i.e. less than one-twelfth.

BIBLIOGRAPHY

The Theory of the Construction of Tables of Mortality, etc. G. F. HARDY. London, 1909.

Methods of Statistical Analysis, chap. 3. C. H. GOULDEN. New York, 1939.

An Introduction to the Theory of Statistics, chap. 10. G. UDNY YULE and M. G. KENDALL. London, 1948.

"Tests of a Mortality Table Graduation." H. L. SEAL. *J.I.A.* Vol. LXXI.

EXAMPLES 2

1. In 100 investigations into the mortality of 500 lives all aged 70, the number of deaths occurring in one year were as follows:

Investigation no.	No. of deaths	Investigation no.	No. of deaths	Investigation no.	No. of deaths	Investigation no.	No. of deaths
1	20	26	10	51	16	76	15
2	24	27	11	52	19	77	16
3	19	28	10	53	21	78	13
4	20	29	9	54	15	79	17
5	7	30	12	55	16	80	16
6	8	31	9	56	12	81	15
7	25	32	13	57	13	82	16
8	14	33	15	58	17	83	13
9	10	34	16	59	19	84	15
10	8	35	14	60	15	85	17
11	9	36	16	61	16	86	12
12	7	37	18	62	15	87	17
13	8	38	13	63	12	88	20
14	18	39	17	64	14	89	22
15	16	40	19	65	17	90	15
16	12	41	10	66	18	91	14
17	13	42	14	67	11	92	11
18	11	43	16	68	15	93	16
19	9	44	19	69	17	94	17
20	18	45	22	70	21	95	12
21	14	46	15	71	13	96	15
22	16	47	13	72	15	97	16
23	11	48	18	73	14	98	19
24	21	49	13	74	18	99	23
25	12	50	15	75	20	100	16

Calculate

(*a*) The mean number of deaths.

(*b*) The mean rate of mortality $q_{70} = \dfrac{\text{No. of deaths in one year}}{\text{No. of lives investigated}}$.

(*c*) The standard deviation of the number of deaths.

Would you modify your method of calculation of item (*b*) if the number of lives in each investigation had not been the same?

If the frequencies of the numbers of deaths in these 100 investigations could have been represented exactly by the binomial distribution $N(p+q)^n$, having a mean equal to the observed mean, calculate:

(*a*) The standard deviation of the rate of mortality.

(*b*) The number of investigations in which exactly 15 deaths would have been recorded.

2. A large issue of Bonds of £100 each is redeemable through the operation of a sinking fund, by annual drawings at par, the proportion redeemable at the next drawing (which is to take place shortly) being 1 per cent of the total outstanding. The current market price of the Bonds is £110, so that an immediate loss of £10 arises in respect of each Bond drawn for repayment.

A holder of 5000 Bonds, whilst anticipating a loss of £500 on the above basis at the next drawing, desires to effect an indemnity policy to cover him in respect of *excess* losses arising if the proportion of his holding drawn for repayment exceeds the anticipated 1 per cent, e.g. if 51 of his Bonds are drawn, he requires indemnity to the extent of £10, and so on. Calculate approximately the net premium required.

3. A normal distribution has a mean of 7·52 and a standard deviation of 2·38. Using Table I in the Appendix calculate the probabilities

(i) that a value selected at random is greater than 12·00;

(ii) that a value selected at random lies between the limits 6·00 and 9·00. (Note that these are not equidistant from the mean.)

4. A large transport organization decided to increase all its passenger fares on 1st January 1941 by 10 per cent. You are asked to estimate the passenger receipts for 1941 on the assumption that the volume of traffic remained unchanged. The 1940 figures were:

(*a*) Passenger receipts £1,250,000.

(*b*) Fare, $1\frac{1}{2}d.$ per mile.

(*c*) Average mileage per journey 50.

In calculating the revised fares, fractions of a penny are to be taken as one penny. How would your estimate vary if fares were to be calculated to the nearest penny, halfpennies being taken as one penny?

CHAPTER III

CORRELATION

1. Hitherto we have considered only a single variable and the frequencies with which it occurs. When we investigate two variables x and y and the frequencies with which pairs of values are associated we meet the phenomenon of correlation.

Suppose, for instance, that we tabulate the height to the nearest inch of a number of fathers and their eldest sons. We might have a table similar to the following:

Table I

Height of father (nearest inch)	Height of son (nearest inch)	Height of father (nearest inch)	Height of son (nearest inch)
63	65	69	67
64	62	69	69
65	67	69	73
65	70	70	68
66	64	70	69
66	66	70	74
66	71	71	67
67	70	71	70
67	66	72	70
68	68	72	73
68	70	73	70
68	72	74	74

This is an example of the simplest type of correlation table, which consists merely of a list of observed values of x and the corresponding values of y. These values need not necessarily be arranged according to a definite scheme, and if there happened to be two observations for which a given value of x was associated with a given value of y they would appear as two separate items in the list. For instance, in Table I we might have had two fathers of height 70 inches to the nearest inch shown as having sons of height 68 inches to the nearest inch. This would be represented by two separate entries in the table.

In the more general type of table the data are extensive and have to be grouped so as to show the frequencies with which values of x within a given range are associated with values of y within the various ranges adopted in grouping. For instance, Table II shows for a given year how maximum and minimum temperatures were associated. As frequencies have to be shown as well as values of the two variables, a double-entry table is used.

Table II

Maximum temperatures in degrees Fahrenheit	Minimum temperatures in degrees Fahrenheit					
	Below 31	31–39	39–47	47–55	55–63	Over 63
Below 45	10	30	5	—	—	—
45–55	10	50	40	10	—	—
55–65	—	10	50	40	5	—
65–75	—	—	10	30	10	—
75–85	—	—	5	15	25	—
Over 85	—	—	—	—	5	5

We usually require to know to what extent one variable varies with another. The extremes of these comparable variations are important:

(*a*) If large values of x tend to be associated with large values of y there is said to be positive correlation.

(*b*) If large values of x tend to be associated with small values of y and small values of x tend to be associated with large values of y there is said to be negative correlation.

The figures in Table II suggest positive correlation.

The assessment of the magnitude of correlation requires considerable analysis and usually involves the calculation of an index known as the coefficient of correlation. This will be dealt with in para. 4, but before we proceed to the analytical details there are one or two general principles which should always be borne in mind. The application of them to a given set of data may even indicate that analytical work is unnecessary and liable to produce misleading results.

In the first place it is important to see that the pairs of observations have some definite link quite apart from the association which it is desired to measure. For instance, in Table I the link is that of father and son, while in Table II the pairs of readings relate to the same day of the year.

Secondly, although the coefficient of correlation will nearly always have to be found, it should be remembered that unless some hypothesis is made concerning the mathematical form of the population the parameters of which are to be estimated from the given data as set out in tabular form (as for instance in Table II), these actually give more information than any single index can do. Careful examination of the data will often, therefore, give useful information which is submerged in subsequent analytical work. An example of this will be given later.

Finally, it must never be assumed that correlation implies causation. Because x and y show a marked tendency to vary together it must on no account be inferred that a change in x will *cause* a change in y.

An example will make this clear. An investigation of cases of sunstroke in a series of years and the amount of home-grown wheat in those years would probably show a marked degree of correlation between the two factors, although neither could be said to influence the other in any way. The true explanation would almost certainly be that a hot summer tends to produce a bumper harvest and many cases of sunstroke. Two variables which are correlated are, in fact, very often both affected by a common cause, or combination of causes, but only rarely is one caused directly by the other.

2. Scatter-diagrams.

A method of representing the given data which naturally suggests itself is to plot on squared paper the various associated values of x and y, thus producing what is known as a *scatter-diagram*.

For instance, the data given in Table II could be represented by the following scatter-diagram (Fig. 2) if the observations were assumed to be concentrated at the mid-points of the intervals and if the group "below 31" were taken as 23–31 and the groups at the other end treated similarly. Since the frequencies are small no serious error would be involved.

Such a diagram does not usually of itself give any clear indication of the presence or otherwise of correlation and is open to the objection that it does not indicate the frequencies with which the pairs of associated values are observed.

This latter objection can be overcome by making a three-dimensional figure, using a third variable z to represent the frequency, but this conception is only of theoretical interest.

In order to condense the information conveyed by the scatter-diagram the following method is used.

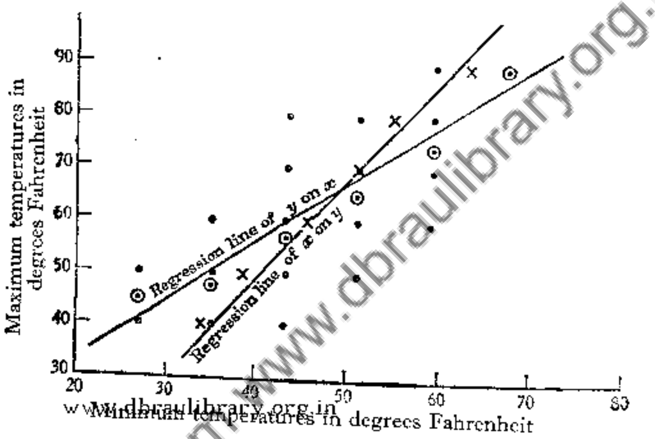


Fig. 2

Taking each observed value of x in turn, the mean value of y corresponding to it is plotted, the frequencies being allowed for in the usual way in calculating the means. Thus, corresponding to the assumed value of 35° minimum temperature, we find assumed maximum temperatures of 40°, 50° and 60° occurring with frequencies 30, 50 and 10 respectively, giving a mean temperature of

$$\tfrac{1}{90}[(30\times 40)+(50\times 50)+(10\times 60)]=48^\circ \text{ approx.}$$

Similarly, the mean maximum temperature corresponding to a minimum temperature of 43° is found to be 57°, and so on. These mean temperatures are indicated in the diagram by dots with rings round them.

When the data are extensive the simplification thus achieved by plotting the means is considerable and the means themselves will be found in general to lie on or near a smooth curve known as a *regression curve*.

In the same way we could take each observed value of y in turn and plot the mean value of x corresponding to it, thus obtaining a second regression curve.

In the simplest examples these curves can be assumed to be straight lines and the correlation is said to be linear.

In what follows, unless the contrary is stated, it will always be assumed that linear correlation is under discussion.

In Fig. 2 the data are so scanty that an attempt to fit a more elaborate curve would be unjustified and the *lines of regression* have been roughly sketched in. The mean values of x are indicated by crosses.

Having drawn the regression lines it is possible, as will be explained later, to deduce approximately the coefficient of correlation. The usual way of calculating this index is, however, by the analytical processes discussed in the next few paragraphs. At the same time it should be emphasized that the analysis is based on the assumption that the correlation is linear and does not give any indication of whether such an assumption is justified. By plotting the means we can throw considerable light on this very important point.

3. Analytical approach.

Suppose that we have N pairs of observed values $(x_1, y_1), (x_2, y_2), \ldots (x_r, y_r), \ldots (x_n, y_n)$ occurring with frequencies $f_1, f_2, \ldots f_n$ $(\sum_1^n f_r = N)$. Several of the x's may of course be equal while the y's differ (e.g. in the example dealt with above the assumed minimum temperature 43° is associated with assumed maximum temperatures 40°, 50°, 60°, 70° and 80°), and the same applies *mutatis mutandis* to the y's.

First consider the regression line which passes through or near to the mean values of y found for each observed value of x, and let its equation be

$$y = m_1 x + c_1,$$

where m_1 and c_1 have to be found.

Actually it is more convenient to revert to the original data rather than to deal with the various means.

The pair of values (x_t, y_t) denoted by P_t in Fig. 3 occurs with frequency f_t.

The ordinate through P_t cuts the line $y=m_1x+c_1$ at Q_t, the ordinate of which is $m_1x_t+c_1$.

Hence the distance $Q_tP_t=y_t-m_1x_t-c_1$.

If the line is to fit the data satisfactorily one obvious requirement is that the total of all distances such as Q_tP_t should be small, allowing for different signs cancelling and also for the frequencies involved, i.e. we should expect $\Sigma f_t(y_t-m_1x_t-c_1)$ to be small when the summation extends over all the observations.

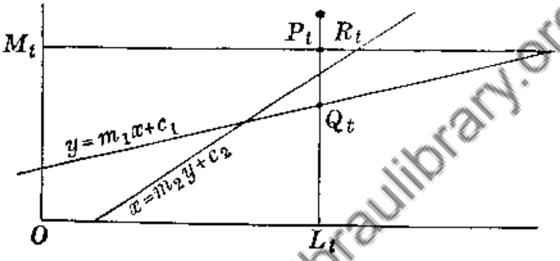


Fig. 3

If we make this sum zero we obtain

$$\Sigma y_tf_t=m_1\Sigma x_tf_t+c_1\Sigma f_t,$$

or, dividing by N, $\quad \frac{1}{N}\Sigma y_tf_t=\frac{m_1}{N}\Sigma x_tf_t+\frac{c_1}{N}\Sigma f_t,$

i.e.
$$\bar{y}=m_1\bar{x}+c_1,$$

where $\bar{y}$ is the mean of all the y's and $\bar{x}$ is the mean of all the x's.

Thus the regression line passes through the point $(\bar{x},\bar{y})$, which we take as a new origin, writing $x_t=\bar{x}+X_t$ and $y_t=\bar{y}+Y_t$.

The equation of the regression line is now $Y=m_1X$ and the distance
$$Q_tP_t=Y_t-m_1X_t.$$

The expression $\Sigma f_t(Y_t-m_1X_t)^2$, where the summation extends over all the observations, is essentially positive, but if the fit is good it should be small. We therefore choose m_1 so as to make it a minimum.

For a minimum (m_1 being the variable),

$$\frac{d}{dm_1}\Sigma f_t(Y_t-m_1X_t)^2=0,$$

i.e.
$$-2\Sigma f_tX_t(Y_t-m_1X_t)=0$$

or $$\Sigma f_t X_t Y_t = m_1 \Sigma f_t X_t^2$$

and $$m_1 = \frac{\Sigma f_t X_t Y_t}{\Sigma f_t X_t^2}. \qquad \text{......(1)}$$

If we divide the numerator and denominator by N, the number of pairs of observations, the denominator $\frac{1}{N}\Sigma f_t X_t^2$ is σ_x^2, where σ_x is the standard deviation of all the x's.

Denoting the numerator $\frac{1}{N}\Sigma f_t X_t Y_t$ by p, the slope of the regression line,

$$m_1 = \frac{p}{\sigma_x^2}. \qquad \text{......(2)}$$

The equation of the line is

$$Y = \frac{p}{\sigma_x^2} X,$$

or, referred to the original axes,

$$(y - \bar{y}) = \frac{p}{\sigma_x^2}(x - \bar{x}). \qquad \text{......(3)}$$

This is known as the line of regression of y on x.

Similarly, the line of regression passing through or near the mean values of x found for each value of y is known as the regression line of x on y.

Denote its equation by $x = m_2 y + c_2$.

In Fig. 3 the line $M_t P_t R_t$ parallel to the axis of x cuts this line at R_t.

The distance $R_t P_t = x_t - m_2 y_t - c_2$.

For a good fit we put

$$\Sigma f_t (x_t - m_2 y_t - c_2) = 0$$

and make $\Sigma f_t (x_t - m_2 y_t - c_2)^2$ a minimum.

The first of these equations readily reduces to

$$\bar{x} = m_2 \bar{y} + c_2,$$

so that the second regression line also passes through $(\bar{x}, \bar{y})$.

Taking axes through this point the second expression reduces to

$$\Sigma f_t (X_t - m_2 Y_t)^2.$$

Differentiating with respect to the variable m_2, and proceeding as before, we find that for the expression to be a minimum

$$m_2 = \frac{\Sigma f_t X_t Y_t}{\Sigma f_t Y_t^2}$$

$$= \frac{p}{\sigma_y^2}, \qquad \dots\dots(4)$$

where σ_y is the standard deviation of the y's.

Referred to the original axes the regression line of x on y is therefore

$$x - \bar{x} = \frac{p}{\sigma_y^2}(y - \bar{y}). \qquad \dots\dots(5)$$

The expressions $\frac{p}{\sigma_x^2}$ and $\frac{p}{\sigma_y^2}$ are known as *coefficients of regression* or *regression coefficients.*

Before proceeding to the coefficient of correlation let us consider exactly what the lines of regression enable us to do.

Take, for instance, the line of regression of y on x (Equation (3)). This gives, for any value of x, the "expected value" of y in the sense that "expected value" is used in the Theory of Probability. We infer, therefore, not that this value of y will in fact correspond to the chosen value of x in a given observation, but that if a large number of observations were made, always keeping x the same, the mean value of the various y's would approximate closely to the value derived from the equation, as the number of observations was increased.

If the value of x substituted in the equation is actually one of those included in the data the value of y found from the equation will not generally be equal to the mean of the y's in the data. For instance, we found on p. 54 that the mean value of y (max. temp.) corresponding to an assumed minimum temperature of 35° F. was about 48° F. If, however, we substituted 35 for x in the equation of the line of regression of y on x we should not expect to obtain 48° F. but a more reliable "expectation" based on the whole of the data, on the assumption that correlation was linear.

Therefore to answer such questions as "What maximum temperature would you expect to correspond to a minimum temperature of 35° F.?" we should use the equation of the line of regression of y on x. Similarly, the regression line of x on y would give the expected minimum temperature corresponding to a given maximum temperature.

4. Coefficient of correlation.

As was stated on p. 52, correlation is said to be

(i) positive if large values of x tend to correspond with large values of y, and vice versa;

(ii) negative if large values of x tend to correspond with small values of y, and vice versa.

Very often, however, there is no apparent tendency for x and y to vary together. Any given value of x seems to occur as often in conjunction with large as with small values of y, and similarly any given value of y is associated with large and small values of x. Correlation is then likely to be small, but we cannot be sure that it is negligible until we apply the tests derived from the Theory of Sampling (Chapter IV).

If we take our origin at the means of x and y and denote the values referred to this origin by X and Y as above:

(i) for positive correlation the values of XY will tend to be large and positive;

(ii) for negative correlation the values of XY will tend to be large and negative;

(iii) for a small degree of correlation the values of XY will be small and fairly evenly divided between positive and negative terms.

In other words, if f_r is the frequency with which the values X_r and Y_r are observed together and $\Sigma f_r = N$, the expression $p = \frac{1}{N}\Sigma f_r X_r Y_r$ seems a useful measure of the extent and sign of the correlation.

Unfortunately p reflects the scales used for x and y. By altering the scale of either variable we alter p, while the correlation is of course the same as before. Scale can best be allowed for by measuring the x's in terms of σ_x and the y's in terms of σ_y. This will be done extensively when we come to consider Sampling.

If we take $\frac{p}{\sigma_x \sigma_y}$ as our measure of correlation it will be seen that a change in the scale adopted for either x or y affects numerator and denominator alike and the expression can claim to be an absolute (as distinct from merely relative) measure of the correlation. It is of course symmetrical in x and y and is known as the *coefficient of correlation* (usually denoted by r).

Coefficient of correlation $r = \frac{p}{\sigma_x \sigma_y}$. ......(6)

In view of the above remarks about scale it seems more logical to write the equation of the lines of regression in the form:

$$\frac{y-\bar{y}}{\sigma_y} = \frac{p}{\sigma_x \sigma_y} \cdot \frac{x-\bar{x}}{\sigma_x} \qquad \text{......(7)}$$

and

$$\frac{x-\bar{x}}{\sigma_x} = \frac{p}{\sigma_x \sigma_y} \cdot \frac{y-\bar{y}}{\sigma_y}. \qquad \text{......(8)}$$

It will now be seen that the first factor on the right-hand side is r in both equations and the coefficients of regression can be written in the more usual forms:

$$m_1 \text{ (coefficient of regression of } y \text{ on } x) = r\frac{\sigma_y}{\sigma_x}, \qquad \text{......(9)}$$

$$m_2 \text{ (} \qquad ,, \qquad x \text{ on } y) = r\frac{\sigma_x}{\sigma_y}. \qquad \text{......(10)}$$

Finally we have $r = \sqrt{m_1 m_2}$. ......(11)

All these results should be memorized.

In practical work it is very desirable to set out the calculations in tabular form in such a way that $\bar{x}$, $\bar{y}$, σ_x, σ_y and p can all be derived in turn.

The student is already familiar with the method for calculating the first four of these values by selecting a convenient origin and scale and making subsequent adjustments. The same can be done in calculating p.

By definition, $p = \frac{1}{N}\Sigma f_t X_t Y_t$, where X_t and Y_t are measured from their respective means.

Suppose that we choose convenient origins so that the coordinates of the point representing the observation are $x_t = X_t + \bar{x}$ and $y_t = Y_t + \bar{y}$.

We first calculate $\frac{1}{N}\Sigma f_t x_t y_t$. This is generally called the *product moment* about the origin chosen, or, more correctly, about the axes chosen.

Now $\frac{1}{N}\Sigma f_t x_t y_t = \frac{1}{N}\Sigma f_t (X_t + \bar{x})(Y_t + \bar{y})$

$$= \frac{1}{N}[\Sigma f_t X_t Y_t + \bar{x}\Sigma f_t Y_t + \bar{y}\Sigma f_t X_t + \bar{x}\bar{y}\Sigma f_t],$$

since $\bar{x}$ and $\bar{y}$ are constants, viz. the distances of the mean from the origin chosen.

Since X_t and Y_t are measured from the mean, $\Sigma f_t X_t = \Sigma f_t Y_t = 0$, and we have

$$\frac{1}{N}\Sigma f_t x_t y_t = p + \bar{x}\bar{y}, \qquad \ldots\ldots(12)$$

$$p = \frac{1}{N}\Sigma f_t x_t y_t - \bar{x}\bar{y}. \qquad \ldots\ldots(13)$$

In other words, we find the product moment, using any convenient origin, and then derive p by deducting $\bar{x}\bar{y}$, where $\bar{x}$ is the distance of the mean of the x's from the origin (allowing for sign) and $\bar{y}$ is the distance of the mean of the y's from the origin.

The following examples should make this clearer, and it is essential that they should be closely studied and every stage of the working verified.

Example 1.

Calculate the coefficient of correlation for the following series of observations:

Table III

Year	Average yield of Consols during year	Average index number of wholesale commodity prices during year	Year	Average yield of Consols during year	Average index number of wholesale commodity prices during year
1810	4·5	171	1823	3·8	107
1811	4·7	164	1824	3·3	106
1812	5·1	147	1825	3·5	124
1813	4·9	138	1826	3·8	108
1814	4·5	137	1827	3·6	108
1815	5·0	131	1828	3·6	97
1816	4·8	109	1829	3·3	95
1817	4·1	141	1830	3·5	97
1818	3·9	160	1831	3·8	99
1819	4·2	135	1832	3·6	94
1820	4·4	124	1833	3·4	90
1821	4·0	113	1834	3·3	94
1822	3·8	106			

This is typical of the simpler problems where frequencies need not be specifically introduced as they are all unity, the table being simply a list. Take as origins:

average yield of Consols during year = 4·0 (variable x),
average index number during year = 120 (variable y).

Taking ·1 as the unit of x, the work can be arranged as follows:

x		y		x^2	y^2	xy	
+	−	+	−			+	−
5		51		25	2,601	255	
7		44		49	1,936	308	
11		27		121	729	297	
9		18		81	324	162	
5		17		25	289	85	
10		11		100	121	110	
8			11	64	121		88
1		21		1	441	21	
	1	40		1	1,600		40
2		15		4	225	30	
4		4		16	16	16	
—			7		49	—	
	2		14	4	196	28	
	2		13	4	169	26	
	7		14	49	196	98	
	5	4		25	16		20
	2		12	4	144	24	
	4		12	16	144	48	
	4		23	16	529	92	
	7		25	49	625	175	
	5		23	25	529	115	
	2		21	4	441	42	
	4		26	16	676	104	
	6		30	36	900	180	
	7		26	49	676	182	
62	58	252	257	784	13,693	2398	148

It will be noticed that positive terms are shown on the left and negative terms on the right of each column. This facilitates the additions and lessens the likelihood of arithmetical errors which may arise if this is not done.

Since the total frequency is 25 we have

$$\bar{x}=\frac{62-58}{25}=\cdot 16 \text{ units},$$

$$\bar{y}=\frac{252-257}{25}=-\cdot 20 \text{ units},$$

$$\sigma_x^2=\frac{784}{25}-(\cdot 16)^2=31\cdot 3344,$$

$$\sigma_x=5\cdot 6 \text{ units},$$

$$\sigma_y^2=\frac{13,693}{25}-(\cdot 20)^2=547\cdot 68,$$

$$\sigma_y=23\cdot 4 \text{ units},$$

$$p=\frac{2398-148}{25}-(\cdot 16)(-\cdot 20)=90\cdot 03 \text{ units}.$$

Coefficient of correlation $r=\dfrac{90\cdot 03}{(5\cdot 6)(23\cdot 4)}=\cdot 69$ approx.

All the work has been done in class units, but if the mean yield of Consols and the standard deviation are required in terms of the original scale and origin we have the following results.

Mean yield of Consols $=4\cdot 0+(\cdot 16)(\cdot 1)=4\cdot 016$.

Standard deviation $=5\cdot 6\times\cdot 1=\cdot 56$.

The mean index number is $120-\cdot 20=119\cdot 8$.

Example 2.

As a second example let us consider the data of Table II, which have been examined graphically earlier in this chapter.

The chief difficulty arises in determining the product moment denoted above by $\frac{1}{N}\Sigma f_t x_t y_t$.

To do this, a simple device, used extensively, is to write the product $x_t y_t$ close to each frequency and then to insert $f_t x_t y_t$, but not to attempt to write down $f_t x_t y_t$ in one step.

A little practical experience will convince the reader that the preliminary step of calculating $x_t y_t$ is well worth while.

In the following table $x_t y_t$ is shown in the top left-hand corner of each division and $f_t x_t y_t$ is shown in the bottom right-hand corner.

As before, x represents minimum temperatures and y maximum temperatures, but the origin has been taken at the centre of the 39–47 group (minimum temperatures) and the centre of the 55–65 group (maximum temperatures). The class-intervals have been taken as units, so that the frequencies, assumed concentrated at the mid-points of the intervals,

Table IV

Maximum temperatures in degrees Fahrenheit		Minimum temperatures in degrees Fahrenheit						Total frequencies	First moments of frequencies	Second moments of frequencies
		Below 31	31–39	39–47	47–55	55–63	63 and over			
	y \ x	−2	−1	0	+1	+2	+3			
Below 45	−2	+4 10 40	+2 30 60	0 5 0	—	—	—	45	− 90	180
45–55	−1	+2 10 20	+1 50 50	0 40 0	−1 10 −10	—	—	110	−110	110
55–65	0	—	0 10 0	0 50 0	0 40 0	0 5 0	—	105	0	0
65–75	+1	—	—	0 10 0	+1 30 30	+2 10 20	—	50	+ 50	50
75–85	+2	—	—	0 5 0	+2 15 30	+4 25 100	—	45	+ 90	180
85 and over	+3	—	—	—	—	+6 5 30	+9 5 45	10	+ 30	90
Total frequencies		20	90	110	95	45	5	365	+170 −200 = − 30	610
First moments of frequencies		−40	−90	0	+95	+90	+15	+200 −130 = +70		
Second moments of frequencies		80	90	0	95	180	45	490		
Product moments of frequencies		60	110	0	50	150	45	415		

occur for values of x from -2 to $+3$ and for y from -2 to $+3$, although the class-intervals are different.

The line below the data marked "Total frequencies" is self-explanatory. Since the "total frequency" 20 occurs for $x=-2$ its first moment about the origin of x is -40, the first entry in the next line.

Similarly, the second moment is $+80$, as shown immediately below.

The last line, "Product moments of frequencies", is obtained by adding vertically the products $f_t x_t y_t$ previously inserted.

The columns on the right headed "Total frequencies", etc. are obtained similarly, but now moments are about the origin of y, e.g. the total frequency 45 occurs for the value $y=2$, so that its first and second moments are 90 and 180. The "Product moments of frequencies" are not needed a second time, since the bottom line gives us $\Sigma f_t x_t y_t = 415$.

The other columns and rows are added as shown and we proceed as follows:

$$\bar{x} = \tfrac{70}{365} = \cdot 192 \text{ class units,}$$

$$\sigma_x^2 = \tfrac{490}{365} - (\cdot 192)^2 = 1\cdot 306, \text{ in terms of class units,}$$

$$\bar{y} = -\tfrac{30}{365} = -\cdot 082 \text{ class units,}$$

$$\sigma_y^2 = \tfrac{610}{365} - (-\cdot 082)^2 = 1\cdot 664, \text{ in terms of class units.}$$

$$\text{Product moment } p = \tfrac{415}{365} - (\cdot 192)(-\cdot 082)$$
$$= 1\cdot 153, \text{ in terms of class units.}$$

$$\text{Coefficient of correlation } r = \frac{p}{\sigma_x \sigma_y} = \frac{1\cdot 153}{\sqrt{1\cdot 306 \times 1\cdot 664}} = \cdot 782.$$

It will be noticed that all the work has been done in class units, which is the simplest plan if only r is required.

In terms of degrees Fahrenheit (as given in the original data)

Mean minimum temperature $= 43 + \cdot 192 \times 8 = 44\cdot 5°$ F. approx.,

since the origin is 43 and scale (class-interval) is 8.

Similarly,

Mean maximum temperature $= 60 + 10(-\cdot 082) = 59\cdot 2°$ F. approx.

To obtain σ_x, σ_y and p in terms of degrees we multiply by 8, 10 and 80 respectively; the value of r is clearly unaffected, thus illustrating the advantage of r rather than p as a measure of correlation.

No adjustment has been made for the error involved by assuming the frequencies to be concentrated at the mid-points of the intervals. Such an adjustment is not usually made in calculating a coefficient of correlation. It is not easy to adjust p and, unless the data are very extensive, to do so would be an unjustifiable refinement. The σ's can easily be corrected, but there is no point in doing this if the numerator is not dealt with.

This coefficient of correlation could be obtained approximately by inspection from the diagram in which the lines of regression were drawn (Fig. 2). These lines intersect at $(\bar{x}, \bar{y})$ and their slopes give the coefficients of regression.

The line of regression of y on x is

$$y-\bar{y}=r\frac{\sigma_y}{\sigma_x}(x-\bar{x}),$$

and hence the tangent of the angle which it makes with the x-axis is

$$r\frac{\sigma_y}{\sigma_x}=m_1.$$

Similarly, the tangent of the angle which the other regression line makes with the y-axis is

$$r\frac{\sigma_x}{\sigma_y}=m_2.$$

The product gives r^2 approximately; cf. equation (11).

5. Properties of r.

On p. 56 the coefficient of regression m_1 was found by making

$$\Sigma f_t(Y_t-m_1X_t)^2 \text{ a minimum.} \qquad \ldots\ldots(14)$$

Substituting for m_1, this becomes

$$\Sigma f_t\left(Y_t-r\frac{\sigma_y}{\sigma_x}X_t\right)^2=\Sigma f_t Y_t^2-2r\frac{\sigma_y}{\sigma_x}\Sigma f_t X_t Y_t+r^2\frac{\sigma_y^2}{\sigma_x^2}\Sigma f_t X_t^2. \qquad \ldots\ldots(15)$$

If there are N pairs of observations, the first and last terms are clearly

$$N\sigma_y^2 \quad \text{and} \quad r^2\frac{\sigma_y^2}{\sigma_x^2}N\sigma_x^2 \text{ respectively.}$$

Also

$$\Sigma f_t X_t Y_t = Np = Nr\sigma_x\sigma_y.$$

Hence the right hand side of (15) reduces to

$$N\{\sigma_y^2-2r^2\sigma_y^2+r^2\sigma_y^2\}=N(1-r^2)\sigma_y^2. \qquad \ldots\ldots(16)$$

But since f_t is always positive the expression (16) is essentially positive or zero.

Hence $1-r^2$ must be positive or zero and

$$-1\leqslant r\leqslant 1. \qquad \ldots\ldots(17)$$

In other words, r cannot be numerically greater than 1.

The expression $$Y_t - r\frac{\sigma_y}{\sigma_x}Xt,$$

previously denoted by $$Y_t - m_1 X_t,$$

gives the difference between the value Y_t associated with X_t and the ordinate of the point on the regression line $Y = m_1 X$ with the same abscissa.

Hence $\Sigma f_t (Y_t - m_1 X_t)^2$, which we have seen reduces to $N(1 - r^2)\sigma_y^2$, gives the sum of the squares of these differences or deviations.

If $r = \pm 1$ each of the deviations must be zero, since all the terms of the summation are positive. That is to say, when there is no "scattering", but all the observations lie on the regression lines, the correlation is perfect and $r = \pm 1$ according as correlation is positive or negative. (σ_y is not zero by hypothesis.)

If $r = 0$ the expression (16) reduces to $N\sigma_y^2$, i.e. the sum of the squares of the deviation from $\bar{y}$.

This can happen only if the regression line is parallel to the axis of x so that the average value of y for every value of x is $\bar{y}$.

When this happens there is no correlation and the points of the scatter-diagram do not tend to cluster round any straight line.

6. Standard deviation of the sum or difference of two variables.

Suppose that we have n values of a variable x with mean $\bar{x}$ and standard deviation σ_x and also m values of a variable y with mean $\bar{y}$ and standard deviation σ_y.

By definition $$n\bar{x} = \sum_1^n x; \quad n\sigma_x^2 = \sum_1^n (x - \bar{x})^2$$

and $$m\bar{y} = \sum_1^m y; \quad m\sigma_y^2 = \sum_1^m (y - \bar{y})^2.$$

Now suppose that a new variable z is formed, where $z = x + y$, and that every value of x is associated with each value of y, thus producing mn values of z.

We can find the mean and standard deviation of these values of z as follows:

$$\bar{z}\,(\text{the mean}) = \frac{1}{mn}\sum_1^{mn} z$$

$$= \frac{1}{mn}\sum_1^{mn} (x + y).$$

Although there are mn different values of $x+y$ on the right-hand side there are only n different values of x, each being repeated m times.

$$\therefore \quad \sum_1^{mn} x = m\sum_1^{n} x.$$

Similarly
$$\sum_1^{mn} y = n\sum_1^{m} y,$$

since there are only m different values of y.

Hence
$$\bar{z} = \frac{1}{mn}\left[m\sum_1^n x + n\sum_1^m y\right]$$
$$= \frac{1}{mn}[mn\bar{x} + nm\bar{y}]$$
$$= \bar{x} + \bar{y}. \qquad \ldots\ldots(18)$$

If σ_z is the standard deviation,

$$mn\sigma_z^2 = \sum_1^{mn}(z-\bar{z})^2 = \sum_1^{mn}(x-\bar{x}+y-\bar{y})^2$$
$$= \sum_1^{mn}(x-\bar{x})^2 + \sum_1^{mn}(y-\bar{y})^2 + 2\sum_1^{mn}(x-\bar{x})(y-\bar{y}). \qquad \ldots\ldots(19)$$

As before, the term $\sum_1^{mn}(x-\bar{x})^2$ reduces to $m\sum_1^{n}(x-\bar{x})^2$ and $\sum_1^{mn}(y-\bar{y})^2$ to $n\sum_1^{m}(y-\bar{y})^2$.

The third term is a "true" summation in that all the mn terms are different. They can, however, be divided into sections as follows.

A particular factor $(x_r-\bar{x})$ is associated with every term of the type $y_s-\bar{y}$ and the sum of these products is $(x_r-\bar{x})\sum_{s=1}^{m}(y_s-\bar{y})$.

But since $\bar{y}$ is the mean of the y's, $\sum_{s=1}^{m}(y_s-\bar{y})=0$.

In this way we can split up the third term of (19) into n groups each of m terms the sum of which is zero.

Hence
$$mn\sigma_z^2 = m\sum_1^n(x-\bar{x})^2 + n\sum_1^m(y-\bar{y})^2$$
$$= mn(\sigma_x^2+\sigma_y^2)$$

and
$$\sigma_z = \sqrt{\sigma_x^2+\sigma_y^2}. \qquad \ldots\ldots(20)$$

In the same way it can be shown that if $z=x-y$, then $\bar{z}=\bar{x}-\bar{y}$ and

$$\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2}.$$

More generally, if $z=x\pm y\pm w\pm \ldots$, where each value of each variable is associated with all the values of the other variables in turn,

$$\bar{z}=\bar{x}\pm\bar{y}\pm\bar{w}\pm\ldots$$

and

$$\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2+\sigma_w^2+\ldots} \qquad \ldots\ldots(21)$$

We sometimes have to deal, however, with a slightly different problem. Suppose that we have n values of x with mean $\bar{x}$ and standard deviation σ_x and n values of y with mean $\bar{y}$ and standard deviation σ_y.

If a new variable $z=x+y$ is formed by associating each value of x with *one and only one* value of y, correlation enters into the problem. As an example x might be the height of a father and y the height of his eldest son, so that there is what is sometimes called a "one-to-one correspondence".

$$\bar{z}=\frac{1}{n}\sum_1^n(x+y)=\bar{x}+\bar{y}, \qquad \ldots\ldots(22)$$

$$\begin{aligned} n\sigma_z^2 &= \sum_1^n(z-\bar{z})^2 \\ &= \sum_1^n(x-\bar{x}+y-\bar{y})^2 \\ &= \left[\sum_1^n(x-\bar{x})^2+\sum_1^n(y-\bar{y})^2+2\sum_1^n(x-\bar{x})(y-\bar{y})\right]. \end{aligned} \qquad \ldots\ldots(23)$$

The first two terms are clearly $n\sigma_x^2$ and $n\sigma_y^2$, but the third term is quite different from that dealt with before. We can no longer split it up into groups such as $(x_r-\bar{x})\Sigma(y_s-\bar{y})$, since each x is associated with one and only one y.

The term $\sum_1^n(x-\bar{x})(y-\bar{y})$ is, however, the expression previously denoted by np, i.e. $nr\sigma_x\sigma_y$, where r is the coefficient of correlation between x and y.

Note: Correlation cannot arise if the values of x and y are not paired off but each value of one is associated with all the values of the other.

Substituting in (23) we obtain:

$$n\sigma_z^2=n\sigma_x^2+n\sigma_y^2+2nr\sigma_x\sigma_y$$

and $$\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2+2r\sigma_x\sigma_y}. \qquad \ldots\ldots(24)$$

Similarly, if $z=x-y$,

$$\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2-2r\sigma_x\sigma_y},$$

and the result can be generalized for $z=x\pm y\pm w\pm\ldots$.

It should be noted that correlation between every pair of variables has to be allowed for.

Example 3.

In Example 5 of Chapter II let us suppose that 5000 of the lives are assured for £200 each and the other 5000 for £100 each.

The first group may give rise to claims for 0, £200, £400, £600, ... £1,000,000 according as 0, 1, 2, ... 5000 deaths occur.

The scale is clearly £200 and it was shown in Chapter II that for a binomial distribution such as this the mean is nqh and the standard deviation $h\sqrt{npq}$.

Hence the mean in this group of 5000 lives is

$$£200\times 5000\times \cdot 01=£10{,}000$$

and the standard deviation is

$$£200\sqrt{5000\times\cdot 01\times\cdot 99}.$$

The other group gives rise to claims for 0, £100, £200, ... £500,000, the scale being now £100.

The mean is therefore £5000 and the standard deviation is

$$£100\sqrt{5000\times\cdot 01\times\cdot 99}.$$

The total claim is of the form $x+y$, where x refers to claims in multiples of £200 and y to claims in multiples of £100.

The mean claim is $\bar{x}+\bar{y}$, i.e.

£10,000 + £5000 = £15,000, as we should expect.

The standard deviation

$$=\sqrt{\sigma_x^2+\sigma_y^2}$$
$$=\sqrt{(200^2+100^2)\,5000\times\cdot 01\times\cdot 99}$$
$$=£1573.$$

The probable error is therefore $\cdot 67\times £1573=£1054$ approx., and to reduce the chance of claims for more than £15,000 to one-quarter, the total premium should be

£(15,000 + 1054) or £1. 1*s*. 5*d*. per cent. approx.

No question of correlation arises here, since any life assured for £100 can be associated with every life assured for £200.

7. Non-linear regression.

When the means of the x's and y's of the scatter-diagram cannot reasonably be assumed to lie on straight lines the preceding analysis breaks down and r is misleading as a measure of the relationship between x and y.

It will be remembered that for linear regression we fitted a line $Y = m_1 X$ to the data, where $m_1 = \frac{\Sigma f_i X_t Y_t}{n\sigma_x^2}$, the value of Y thus obtained being the "best" value or expectation corresponding to the value X.

If we denote the "best" value corresponding to X_t by Y_t', while the actual observed values are Y_{t_1}, Y_{t_2}, Y_{t_3}, etc., then $Y_t' = m_1 X_t$ for all values of t.

$$\therefore \ \sigma_{y'} = m_1 \sigma_x \quad \text{and} \quad \frac{\sigma_{y'}}{\sigma_y} = m_1 \frac{\sigma_x}{\sigma_y} = \frac{\Sigma xy}{n\sigma_x\sigma_y},$$

the value previously denoted by r.

From this point of view r is the ratio of the standard deviation of the estimated values read off from the regression line to the standard deviation of the original observations Y_{t_1}, Y_{t_2}, etc.

Very often a curve is more suitable than a straight line for showing the relationship between x and y, and it will be seen later that the process of graduation is in effect the fitting of a curve showing correlation between age (x) and mortality (y), the graduated curve being a curvilinear regression line.

From such a curve the values of Y' corresponding to each x can be measured and $\sigma_{Y'}$ calculated. Actually this is the same as $\sigma_{y'}$, where y' is measured from the origin instead of from the mean.

$\sigma_{y'}/\sigma_y$ can be regarded as a measure of correlation and unless regression is linear it is known as the *index of correlation*.

In the special case where the regression curve passes through all the mean y's $\sigma_{y'}/\sigma_y$ is called the correlation ratio and is usually represented by η.

The reader will now be in a position to appreciate (and criticize) the following definitions of correlation:

(1) If two quantities vary in sympathy so that a movement in one tends to be accompanied by a movement in the other, they are said to be correlated.

(2) Two variables are said to be correlated when we do not find a fixed value of the one variable equally likely to be associated with different values of the other.

We close this chapter with an example which presents many points of interest.

Example 4.

The following data have been collected for the purpose of investigating the correlation between the duration of life of married and widowed women and the number of children. Calculate the correlation coefficient and mention any approximate measures which might be used to indicate the extent of the correlation, stating the objections to these measures.

Table showing ages at death in quinquennial age groups, of 1095 wives and widows, with particulars of the number of children				Table showing the distribution of the deaths according to number of children	
Central age at death	No. of deaths	Total no. of children	Average no. of children	No. of children	No. of deaths
20	29	36	1·2	0	24
25	87	151	1·7	1	130
30	99	261	2·6	2	122
35	109	478	4·4	3	134
40	90	450	5·0	4	111
45	87	437	5·0	5	106
50	64	370	5·8	6	85
55	54	331	6·1	7	91
60	69	430	6·2	8	81
65	73	447	6·1	9	77
70	83	547	6·6	10	58
75	77	590	7·7	11	23
80	78	547	7·0	12	24
85	59	398	6·7	13	15
90	26	212	8·2	14	6
95	7	50	7·1	15	2
100	4	35	8·8	16	2
—	—	—	—	17	2
—	—	—	—	18	2
Total	1095	5770	—	—	1095

Find an equation for the regression of the number of children on the length of life of the mother and plot this on a graph together with any information from the given data which will show whether the regression line is satisfactory. On consideration of your graph state whether the correlation coefficient may be regarded as the best measure in this particular case.

The following additional information is given:

Mean age at death	53·292
Standard deviation of age at death calculated on a unit of 5 years	4·091
Standard deviation of number of children	3·409

To investigate correlation we should expect the data to be given in the form of a double-entry table with (say) age at death along the top (x) and number of children down the left-hand side (y), the class-interval for x being taken as 5 years.

Actually we are not given the data to fill in the squares but we are given the total of each column, i.e. the number of deaths for a specified central age at death, and also the total of each line, i.e. the number of deaths for a specified number of children.

The column "Average number of children" is unnecessary, since it can be obtained by dividing the "Total number of children" at each central age by the number of deaths at that age. It will, however, be found useful later on.

We are given $\overset{y}{\Sigma} f$ for each value of x, i.e. the total frequency for each central age at death irrespective of number of children, and also $\overset{x}{\Sigma} f$ for each value of y, i.e. the total frequency for each number of children irrespective of age of mother at death. ($\overset{y}{\Sigma}$ is used to denote summation with regard to y and $\overset{x}{\Sigma}$ summation with regard to x.) Hence we could calculate $\bar{x}$, σ_x, $\bar{y}$ and σ_y in the usual way; three of these values are given, but they should be checked as an exercise.

The only difficulty is in finding the product moment Σfxy.

The method normally used breaks down because we do not know the individual values of f for every pair of associated values of x and y. We are given the total number of children for each central age at death but not how many mothers dying at that age left 0, 1, 2, 3, ... children.

The given data are in fact the values of Σfy for each central age at death, where y represents the number of children and f the frequency.

To deduce Σfxy is a simple matter, since x is the same for all the values in a given column.

We assume the frequencies concentrated at the mid-point of the intervals and take the class-interval as the unit. The origin of x (the age) is taken at 55. There is little to be gained by altering the origin of y. The work is as follows:

Table V

Age x (1)	No. of deaths $\overset{y}{\Sigma} f$ (2)	No. of children $\overset{y}{\Sigma} yf$ (3)	Product moments (1) × (3) $\overset{y}{\Sigma} fxy$ (4) −	 +
−7	29	36	252	
−6	87	151	906	
−5	99	261	1,305	
−4	109	478	1,912	
−3	90	450	1,350	
−2	87	437	874	
−1	64	370	370	
0	54	331		
1	69	430		430
2	73	447		894
3	83	547		1,641
4	77	590		2,360
5	78	547		2,735
6	59	398		2,388
7	26	212		1,484
8	7	50		400
9	4	35		315
Total	1095	5770	−6,969 + 12,647 = 5,678	

Clearly $\bar{y}$ (the mean number of children per death) $= \frac{5770}{1095} = 5{\cdot}269$.

There is thus no need to refer to the last two columns given in the data unless it is desired to check the given value of σ_y.

The product-moment about the chosen origins is $\frac{5678}{1095}$.

The values of the means are

$$\bar{x} = \tfrac{1}{5}\,(53{\cdot}292 - 55) \text{ class units, measured from } 55,$$

$$= -{\cdot}3416 \text{ class units,}$$

$$\bar{y} = 5{\cdot}269.$$

$$\therefore\ p \text{ (product moment about the means)} = \tfrac{5678}{1095} + ({\cdot}3416)(5{\cdot}269)$$

$$= 6{\cdot}985.$$

Coefficient of correlation $= \dfrac{p}{\sigma_x \sigma_y} = \dfrac{6{\cdot}985}{(4{\cdot}091)(3{\cdot}409)}$

$$= {\cdot}50 \text{ approx.}$$

Note that since p and σ_x are both expressed in class units there is no need to make any adjustment.

An approximate value of r can be obtained as follows.

For each value of x we are given not the frequencies with which each number of children was observed but the average number of children. We cannot therefore draw a scatter-diagram, but we can at once plot the mean value of y for each value of x and try to fit a straight line by inspection. As x increases from 20 to 100, i.e. 16 units of 5 years, the mean value of y increases from 1·2 to 8·8, i.e. by 7·6.

Hence m_1, the slope of the regression line, is roughly $\dfrac{7{\cdot}6}{16}$.

Since $m_1 = r\dfrac{\sigma_y}{\sigma_x}$, this gives

$$r = \frac{7{\cdot}6}{16} \cdot \frac{4{\cdot}091}{3{\cdot}409} = {\cdot}57.$$

The observations at ages 20 and 100 are, however, very scanty, and it seems preferable to base an estimate on the figures for ages 30 and 90, taking the mean value at age 90 as 8·0 instead of 8·2 to get a better run of the figures there.

On this basis $m_1 = \dfrac{5{\cdot}4}{12}$ and $r = \dfrac{5{\cdot}4}{12} \cdot \dfrac{4{\cdot}091}{3{\cdot}409} = {\cdot}54.$

The objection to both these estimates is of course that they are based on observations at two ages only and not on the general run of the means.

The equation of the regression line of y on x is

$$y - \bar{y} = r\frac{\sigma_y}{\sigma_x}(x - \bar{x}),$$

or

$$y - 5{\cdot}269 = {\cdot}5\frac{3{\cdot}409}{4{\cdot}091}(x + {\cdot}3416).$$

Reverting to the original scale of x and the original axes, this becomes

$$Y - 5{\cdot}269 = {\cdot}5\frac{3{\cdot}409}{4{\cdot}091}\left(\frac{X-55}{5} + {\cdot}3416\right),$$

or

$$Y = {\cdot}083X + {\cdot}84, \text{ approx.}$$

This line is drawn on the diagram on p. 76 together with the points representing the mean values of y for each value of x.

A glance at these means shows that they cannot in fact be represented satisfactorily by a straight line, i.e. that the regression is non-linear. The coefficient of correlation calculated is therefore misleading.

The curve shown seems to be a reasonably good approximation to the run of the values and may be taken as the curve of regression.

From general considerations we should expect that from age 20 to 45 the number of children would increase with the age at death, since the size of family must depend to a large extent on the duration of life during the child-bearing period. Once age 45 has been passed, however, we should not expect the number to increase except to the very small extent which reflects the superior vitality of healthy women who have larger families than the average.

The average number of children increases fairly regularly throughout. The explanation is almost certainly to be found in the fact that the given data relate in all probability to the deaths of married women and widows over a short period of time; if the average age at motherhood is taken as 30, this means that women dying at 90 had their children 60 years ago when large families were the rule rather than the exception. Similarly, those dying at 60 had their children on the average about 30 years ago.

Thus the regression curve reflects the variation of number of children not so much with age of mother at death as with the period when the children were born. The chief factor operating has thus been a steadily falling birth-rate.

This question illustrates the difficulties involved in interpreting the results of an investigation into correlation.

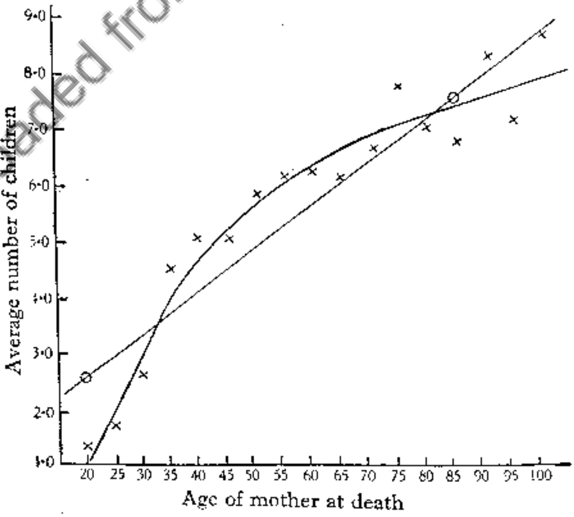


Diagram. Regression of number of children on length of life of mother.

BIBLIOGRAPHY

Frequency Curves and Correlation. Sir William P. Elderton. London, 1938.

An Introduction to the Theory of Statistics, chaps. 11, 13 and 16. G. Udny Yule and M. G. Kendall. London, 1948.

An Introduction to Medical Statistics. A. Bradford Hill. London, 1948.

EXAMPLES 3

1. A random sample of 170 cases has been taken from the new policies issued in 1936 by a certain life office, and the distribution of the sample with regard to age at entry and sum assured is found to be as follows:

Age group	Sum assured					Total no. of policies
	£50	£100	£200	£500	£1000	
15–24	18	20	6	2	—	46
25–34	21	26	6	5	1	59
35–44	10	9	3	6	1	29
45–54	7	8	5	4	—	24
55–64	8	3	1	—	—	12
Total no. of policies	64	66	21	17	2	170

Calculate the coefficient of correlation between age and sum assured (*a*) using only the data for ages at entry up to 44, and (*b*) using all the data. Comment on your results.

2. The following table shows the average Sum Assured under new assurances effected in 50 Insurance Offices in a particular year, and the expense ratio of these Offices for the same year.

Calculate the coefficient of correlation between the average new Sum Assured and the expense ratio.

Average sum assured £	Expense ratio	Average sum assured £	Expense ratio	Average sum assured £	Expense ratio
784	12·61	491	10·76	802	15·94
389	16·85	675	22·91	807	20·23
301	23·74	1002	14·58	1158	18·15
355	13·25	597	18·64	815	17·44
687	22·26	346	16·70	757	15·46
596	13·72	363	14·89	855	18·45
748	16·21	1097	15·33	718	14·32
660	23·75	646	13·81	793	21·48
900	15·90	498	16·98	553	16·21
629	24·28	621	13·98	668	15·42
699	12·28	626	15·96	631	19·04
474	21·71	922	14·17	710	14·73
932	17·75	404	11·96	195	17·30
941	13·65	289	13·61	867	13·82
689	16·05	606	17·77	656	20·90
797	15·80	675	13·94	1002	17·13
535	20·29	1011	16·49		

$$\text{The average Sum Assured} = \frac{\text{Total new Sum Assured (less re-assurances given off)}}{\text{Total number of new policies}}.$$

$$\text{Expense ratio} = \frac{\text{Total expenses for the year less 5 per cent of single premiums}}{\text{Total premium income for year (new and renewal) less single premiums}}.$$

Do you consider the coefficient, as calculated, a good measure of correlation (if any) between the class of business (as measured by size of policy) and the cost of conducting the business? Give reasons.

Assuming that any data you require are available, how would you calculate an improved coefficient?

3. Explain briefly the terms "line of regression" and "coefficient of correlation".

An investigation, based upon the data relative to 500 lives of the undernoted age distribution, has been made as to the degree of correlation between the age, x, at the date of the investigation and the total sum, y, for which each life is assured. As a result it has been found that the equations of the lines of regression of y on x and of x on y, taking the unit of y as £100, are respectively

$$\left.\begin{aligned} 5x-57y+\ 142&=0 \\ \text{and}\qquad 125x-57y-4652&=0 \end{aligned}\right\}.$$

Calculate (*a*) the mean value of y;
(*b*) the standard deviation of y;
(*c*) the coefficient of correlation between x and y.

Age	No. of lives	Age	No. of lives
31	29	41	25
32	28	42	26
33	28	43	24
34	28	44	24
35	27	45	24
36	26	46	23
37	26	47	23
38	26	48	22
39	26	49	21
40	25	50	19

4. You are given the following information in respect of the ages of husbands and wives:

Age group	No. of husbands	No. of wives	Difference between age of husband and age of wife*	Frequency of occurrence
15–	26	144	−15	2
20–	512	709	−10	17
25–	619	448	− 5	140
30–	212	134	0	699
35–	72	56	5	526
40–	39	32	10	131
45–	28	21	15	37
50–	23	13	20	15
55–	17	9	25	6
60–	13	5	30	3
65–	10	4	35	1
70–	5	2	Total	1577
75–	1	—		
Total	1577	1577		

* The difference has been calculated by deducting the central age of the age-group of the wife from the central age of the age-group of her husband.

Calculate the coefficient of correlation between the age of husband and the age of wife.

5. An office has investigated its mortality experience under Whole Life Policies With Profits. The rate of mortality of the non-medical business is q'_x and of the medical and non-medical business combined q_x. Given the exposed to risk in the non-medical class to be E'_x and in the combined classes E_x, find the coefficient of correlation between q'_x and q_x. Hence find the standard deviation of $q'_x - q_x$.

6. The table below gives the values of y observed for fixed values of x in eight separate investigations of similar data under similar conditions. Calculate the coefficient of correlation between y and x and plot the regression line of y on x.

Comment on your results in the light of the further information that y is $100{,}000q_x$, q_x having been arrived at by the normal methods of observations of exposures and deaths.

$x \backslash y$	1	2	3	4	5	6	7	8
30	238	217	247	237	248	246	244	239
40	385	414	412	401	349	413	350	400
50	782	746	818	710	755	800	746	728
60	1,983	1,993	1,933	2,013	2,053	1,953	1,853	2,012
70	5,302	5,352	5,127	5,227	5,126	5,027	5,304	5,306
80	12,972	12,785	12,660	13,410	12,846	12,410	12,850	13,409

CHAPTER IV

SAMPLING

1. We are all familiar with samples in everyday life and the purpose of sampling in the theory of statistics is very much what one would expect: to obtain information about a large body of data by examining a much smaller selection made in such a way as to be representative. The work falls into three main divisions, which may conveniently be dealt with separately:

(1) The construction of the sample.

(2) The analysis of the sample.

(3) Induction and inference from the results of this analysis.

The following terms will be used frequently in discussing sampling:

2. Definitions.

"Universe" or "Population". The large body of data from which the sample is assumed to have been drawn, is known as the *universe*, or in actuarial work as the *population*.

"Statistic" and "Parameter". A function such as a mean, standard deviation or coefficient of correlation calculated from a sample is known as a *statistic*; if it is based on the universe it is known as a *parameter*.

"Errors" and "Deviations". It will be found that we often have to consider the difference between the value of an index derived from a sample (a statistic) and the value derived from the universe (the parameter). This difference is referred to in statistics as the *error* and does not imply that any mistake has been made. The term *deviation*, which is sometimes used, is perhaps preferable.

3. When the sample is all the available data.

Quite often the selection of data has already been done by force of circumstance and we are given our sample "ready-made". For instance, actuaries wish to know how mortality changes from time to time, and in this country the samples with which they work consist of returns made by British Life Offices, the results of censuses and returns of births and deaths, and so on.

A point which usually causes difficulty to the beginner is that functions such as rates of mortality derived from the whole (or nearly the whole) of the available data should nevertheless be regarded as values derived from a sample and subject to sampling errors.

It should be borne in mind that the purpose of a mortality investigation is not to obtain a record of mortality actually experienced but to produce rates of mortality which would probably have been revealed if the data had been unlimited in extent.

For instance, the years 1924–29 had their own peculiarities, such as epidemics, unusually severe weather, freedom from any major war, and so on. The object of an investigation into the data for these years for producing the A 1924–29 Table was not to eliminate all these features but to form an estimate of the results in an "average" year. This "average" year, like the "average" man, does not exist. Although the whole of the available data were apparently used for the A 1924–29 Table and in similar investigations, sampling errors were bound to arise because of limitations of numbers and scope.

Any population will be subject, from year to year, to random fluctuations, which will appear as sampling errors in any investigation involving the population.

Similarly, in the construction of the English Life Tables the data should be regarded as a sample even if every life in England and Wales had been correctly observed and the facts relating to that life correctly incorporated. The force of this will be more fully appreciated when we come to consider graduation.

4. Random sampling.

When the sample has to be constructed the ideal would be to form a microcosm similar in all respects to the universe but, of course, on a very much smaller scale. As we have only limited information about the universe (otherwise we should not need a sample) this is impossible, but a good approximation can be obtained by selecting the constituent elements of the sample at random from the universe. We may say that a sample of n individuals is taken at random from a universe if all possible samples of n had an equal chance of being selected.

This sounds a simple matter, but in practice it is exceedingly difficult to avoid bias owing to personal idiosyncrasies or other factors even more difficult to identify or control. If it is possible to assign a number or other identifying symbol to each member of the universe a good way of forming a random sample is to draw tickets with the numbers on them from a drum in which they have been thoroughly mixed. This can rarely be done and bias nearly always creeps into the work, though its presence may not be detected. To appreciate the practical difficulties it is essential to read accounts of actual investigations, and the student is strongly recommended to study two very interesting papers on "Enquiry by Sample" contributed by Mr J. Hilton to the Royal Statistical Society and reproduced in "Reprints 1938".

5. Systematic methods of sampling.

The student should appreciate that a random sample may be constructed by a systematic process. For instance, a political agent who wished to form an estimate of the strength of the parties in a given district might instruct canvassers to call at every tenth house and provided there were no deviation from this strict rule the sample would probably be random since there is no reason to expect political views to be associated with number of the house.

So long as the sample is random *with respect to the character or characters to be measured* it is likely that this method will lead to a truly representative selection of individuals.

6. Stratified sampling.

Here the body of data is split into groups or "strata" by purposive means and one or more representatives from each group are selected at random.

Suppose, for example, that a manufacturer of electric-light bulbs wishes to know what current passes through them at a given standard voltage, what candle-power they develop, and how many hours "life" they have. To do this he might select a hundred bulbs by walking into the stock room and picking out a bulb here and there, thus hoping to obtain a random sample. Almost certainly bias would be much greater than he realized: for instance, he would probably tend to pick from the centre rather than the edges, from the top layer rather than any other and to pick clean bulbs rather than dirty ones. None of these points may be of importance, but the sample could not be described as random and results deduced from it might be misleading. For a purposive sample he might take every hundredth bulb turned out by each machine. On the face of it the sample thus produced would be representative in that personal bias seems to have been eliminated. The danger of this method is, however, that the "period" of the sampling may coincide with some "period" in the output. For instance, if a machine has five moulds for making the glass bulb itself, a sample formed by selecting every hundredth bulb would mean that only one mould was really being tested. Again if the count started each day when the first shift went on duty the first ninety-nine of each day's output would never be represented in the sample. If a guarantee of performance is given by the manufacturers it is important that all the bulbs, including those made at the beginning of the day, should be properly represented in the sample.

Alternatively the output of each machine might be taken as a stratum and specimens selected from each by some truly random method, as far as possible independent of any human operator. This would be "*stratified sampling*".

As an example of the pitfalls which would trap the unwary let us suppose that the works manager decided to take a bulb from each machine every time his telephone bell rang. As there is no apparent connection between his telephone calls and the quality of the output this might at first appear a satisfactory way of constructing a sample.

There are, however, many serious objections, such as the fact that telephone calls are not usually distributed over the working day in a random manner, while the selection of bulbs from all machines at a given time would tend to introduce correlation which would be difficult to assess or control.

A better method would be to decide the times at which drawings were to be made for each machine independently by drawing numbered tickets from a drum. In many factories, however, each article is numbered and a satisfactory and convenient way of constructing a sample would be as follows. Suppose the output of one machine for a given day had numbers H 43295 to H 43726 while the output of another had numbers K 83962 to K 84403. For the first machine discs with one of the numbers 0–9 on each could be drawn from a bag and replaced after the number had been recorded, thus producing a random series such as

4 3 9 6 2 5 4 3 8 7 ....

If these were marked off in threes we should obtain

439, 625, 438, ...

as the last three figures of the numbers of the bulbs to be drawn; i.e. we should select

H 43439, H 43625, H 43438, ...,

any number less than H 43295 or greater than H 43726 being ignored. Similarly, each of the other machines could be dealt with and the sample obtained in this way should be free from serious bias.

7. Simple sampling.

There is a particular form of random sampling, known as *simple sampling*, which is of special importance. The reader will be familiar with problems in probability involving the drawing of balls from an urn or cards from a pack when the object drawn is always replaced before the next drawing. The essential feature of such a problem is that the probability of drawing any one object is the same as that of drawing any other and this probability is the same when the last draw is made as it was at the beginning.

In simple sampling it is assumed that every individual in the universe is equally likely to be chosen in the sample and that the universe is so large that when the sample is extracted the remainder is to all intents and purposes the same as the original universe. Thus the chance of drawing a particular individual for inclusion in the sample at the *n*th draw is the same as it was when the sampling commenced. For simple sampling it is also assumed that the chance of drawing any individual is independent of that of drawing any other.

This should be borne in mind in considering mortality statistics, which are usually far from homogeneous and which are affected in such a way by wars, epidemics, etc. that it is not true to say that the chance of death of any one individual is independent of the chance of death of any other.

8. Analysis of the sample.

This is usually a straightforward matter and the methods previously described are used to calculate convenient indices to express the main characteristics. The algebraic functions, such as the mean, standard deviation and coefficient of correlation, are most commonly used, as they are good estimates of the parameters involved in normal theory.

9. Inference and deduction from the sample.

There are three main types of problem:

(*a*) when both the universe and the given sample are available for analysis;

(*b*) when only a sample is given;

(*c*) when two or more samples are given but the universe is unknown.

(*a*) is probably the least important, (*b*) is most commonly met with in actuarial work, and (*c*) is common in enquiries into social questions, education, housing, hygiene, and so on.

Under (*a*) the question to be considered is whether the given sample is likely to have been drawn from the given universe by random or other unbiased sampling. We might, for instance, be given records of the heights of a hundred adult males and also the mean height and standard deviation of the height of all adult males in England and Wales.

If we found the mean and standard deviation for the sample it is extremely unlikely that they would coincide exactly with the figures for all England and Wales (our "universe"). For a sample of ten we should be prepared for large discrepancies, for a sample of five hundred we should, as we say, expect "average" results. The point to be decided is what sort of discrepancies are likely to arise in a sample of a given size if it were selected in some unbiased way from the universe. If we found that the sample values differed from those for the universe by improbable amounts we should suspect that the sample was not taken from that universe (e.g. they might be a hundred Norwegians) or else that bias had crept in (e.g. by selecting the sample from a particular district or a particular trade).

(*b*) When only a sample is available we use it to estimate conditions in the universe by testing a hypothesis or set of hypotheses about the universe. A very common example is testing an estimate of a parameter (universe value) by comparing it with a statistic (sample value). We might, for instance, wish to test a theory that the average height of males over 21 years of age in England and Wales was 5 ft. 8·3 in. by measuring a representative sample of, say, two hundred and calculating the sample mean.

In mortality statistics the hypothesis specifies that the rates of mortality in the population are those given by the graduated table, and that any discrepancy between the rate derived from the data and the graduated rate is due merely to sampling errors. We can find approximately the probability that the observed discrepancy or one still greater would arise in this way, a large discrepancy corresponding to a small probability. If the probability is very small we suspect that the hypothesis about the population rates is fallacious; we suspect that the rates of mortality in the graduated table do not represent the facts as far as we can judge from the observations made. It may seem that (*a*) and (*b*) are almost indistinguishable, but there is a fundamental difference. In (*a*) the parameter is known and the sample is being tested. In (*b*) the parameter is not known but an estimate of it (often based on the sample) is being tested by means of the sample.

(*c*) When we are given two or more samples we try to ascertain whether the universes from which they were taken are likely to be the same (or similar) or whether they are different. For instance, the drug M and B 693 has been found to give good results in the treatment of pneumonia. As it is impossible to collect statistics of all cases, whether treated by the drug or not, a random sample might be taken of, say, two hundred cases so treated and another sample of two hundred as far as possible similar to the first in every respect (e.g. age distribution) and subjected to identical treatment except for the administration of the drug. As a result of the analysis of the samples we might be able to say that the differences between the two samples could not be explained as due to paucity of data and sampling errors, but were probably due to a real difference in the universes from which they were taken.

In (*a*), (*b*) and (*c*) we try to arrive at a probability that sampling would give rise to an error equal to or greater than the error observed. This usually causes difficulty to the student, who asks, "Why not try to estimate the probability of obtaining the observed error owing to paucity of data? Why include larger errors?" The reason is that the probability of obtaining a given error is meaningless if the variable is continuous (as it usually is).

It will be remembered that in applying Calculus to questions in Probability we had to deal with probabilities relating to small intervals. Thus on p. 308 of *Mathematics for Actuarial Students*, Part II, a typical example involves the probability that a man arrived in London between t and $t+dt$ from the beginning of the year. The probability that he arrived at exact time t is zero.

Similarly, in statistics there cannot be the probability of obtainin a given sampling error if the variable is continuous; what we evaluate is the probability of obtaining an error greater than that actually observed. As the student will no doubt expect, this function is represented by an integral. Before however we can deal with this integral we must consider sampling distributions.

10. Sampling distributions.

Suppose that from a universe of N we take every possible sample of n. There will be ${}^{N}C_{n}$ such samples because, although a particular value or event may be repeated several times in N, each of these values or events may be considered as a separate "individual" for the purposes of sampling. For each sample we can calculate some index such as σ, the standard deviation of the sample, and group the results into a frequency distribution, which we can if we wish represent graphically as a histogram. If the number of samples is very great the class-interval used in drawing the histogram can be reduced. By taking a very large number of samples, each of n

individuals, and calculating the standard deviation of each, we shall ultimately arrive at a smooth curve showing how the values of σ are distributed.

Similarly, we could find how the means of the samples were distributed and could arrive at, or at least approximate to, a smooth curve representing this distribution of the sample means—the "sampling distribution of the means". In the previous paragraph we discussed the sampling distribution of the standard deviation; if two variables were present we could imagine a sampling distribution of the coefficient of correlation, one value being calculated for each sample.

From the above it seems necessary not only to calculate a given statistic (say the mean) for the sample of n observations but to construct all the other possible samples of n individuals and calculate the statistic for each in order to form some idea of the sampling distribution.

11. Normal sampling distributions.

A considerable literature exists about sampling distributions. In most instances, however, it is assumed that the populations follow the normal curve. Actual experiments based on extensive data indicate that this assumption is a good approximation in most cases although it is unlikely that many distributions are really accurately represented by any such simple law.

For practical purposes we nearly always assume that sampling distributions are normal. (An important exception will be mentioned later when we come to the χ^2 test.) This assumption is usually approximately correct if n, the number of values involved in the statistic, is large but it should be made only in large sample theory.

12. Biased and unbiased estimates.

The expected value of a statistic can be expressed in terms of the population parameters. On intuitive grounds we use the observed values of statistics as estimates of population parameters. If the population parameter which it is desired to estimate is, in fact, equal to the expected value of the statistic employed the latter is said to be an *unbiased estimate* of the parameter. For example, the mean of a sample of n values is an unbiased estimate of the population mean.

On the other hand, the sample standard deviation, s, is not an unbiased estimate of the population standard deviation, σ. Such an estimate is called a *biased estimate.*

To appreciate why s is a biased estimate of σ we note that, in calculating s the deviation of each observed value is measured from the *sample mean* while, in calculating σ the corresponding deviations are measured from the *population mean*. The mean square deviation of n values is a minimum when measured from the mean of those values, so that s will, on the average, be less than the parameter σ and will be a biased estimate. For large samples, however, this bias is negligible and we shall assume in this book that all estimates used are unbiased.

13. Standard error.

Although, as we have seen in the previous paragraph, the expected value of a statistic in large sample theory may be assumed equal to the parameter a little consideration will show that the standard deviation of the sampling distribution, which is defined as the *standard error* of the statistic, will not approximate to the parameter σ, the standard deviation in the population. Thus the statistic "the mean of a sample of n individuals" will tend to be a much more stable quantity than the individuals themselves and hence the standard error of the mean should be much less than σ. As we shall see, the standard error of the mean is $\sigma/\sqrt{n}$.

It is important to notice that, apart from bias, the *accuracy* to be expected of an estimate is important. For an unbiased estimate this will be appropriately measured by the standard error of the statistic employed.

Expressions have been obtained for the standard errors of most of the well known statistics but the theoretical work involved is usually difficult and we shall content ourselves with dealing with two of the simplest: a class frequency and the mean. The results for other statistics are quoted in section 16 of this chapter.

14. Standard error of a class frequency.

Sometimes the universe is divided into different classes and we may be interested in the numbers in the various classes or class frequencies in a given sample. For instance, women may be divided into single, married and widowed (including divorced). Obviously the class frequencies shown in a sample will not be comparable with the class frequencies in the universe, and for this reason we usually deal with "proportionate class frequencies" found by dividing the class frequencies by the total number in the sample or the universe as the case may be. Thus, if out of a sample of 250 women 104 were married we should say that the class frequency for married women was 104 and the proportionate class frequency ·416.

Let us assume that for a particular class the proportionate frequency in the universe is q and that we draw a random sample of n.

Assuming that the laws of simple probability apply, the chance that all n are in the given class is q^n. Similarly, the chance that $n-1$ are in the given class and one is not is ${}^nC_1 q^{n-1}p$, and, generally, the chance that r are in the given class and the remaining $n-r$ are not is ${}^nC_r q^r p^{n-r}$, where $p=1-q$.

Hence, if we took a very large number S of random samples, each of n observations, we should expect the class frequencies to be distributed as follows:

Class frequency	No. of samples in which the frequency is observed
n	Sq^n
$n-1$	$S\,{}^nC_1 q^{n-1}p$
$n-2$	$S\,{}^nC_2 q^{n-2}p^2$
$\vdots$	$\vdots$
r	$S\,{}^nC_r q^r p^{n-r}$
$\vdots$	$\vdots$
2	$S\,{}^nC_{n-2} q^2 p^{n-2}$
1	$S\,{}^nC_{n-1} q p^{n-1}$
0	$S\,p^n$
Total	$S(p+q)^n=S$

Our sampling distribution is in fact the successive terms in the expansion of $S(q+p)^n$. We know therefore that the mean is nq and the standard error $\sqrt{npq}$ (see Chapter II, pp. 32, 33).

If instead we had considered the proportionate frequency we should have had a mean value q and standard error $\sqrt{\frac{pq}{n}}$ or $1/n$th of the values for the class frequency itself.

We therefore have the important results for a sample of n.

	Mean value	Standard error	
Class frequency	nq	$\sqrt{npq}$	......(1)
Proportionate class frequency	q	$\sqrt{\frac{pq}{n}}$	......(2)

Note. Although the mean value of the class frequency is not of course the class frequency in the universe, the general rule does apply for the proportionate class frequency, which is independent of the number concerned.

15. Normal approximation to sampling distribution of a proportionate class frequency.

Having obtained the mean and standard error of a proportionate class frequency as above, we usually assume that the sampling distribution roughly follows the normal curve. This seems at first inexplicable when we know that the actual distribution is the successive terms of $(q+p)^n$. The chief reasons for the assumption are as follows. By a suitable choice of origin and scale any normal distribution can be reduced to the standard form $y=e^{-\frac{1}{2}x^2}$, for which exhaustive sets of tables are available. The curve is in fact adequately "mapped". The distribution $(q+p)^n$ cannot, however, be reduced to a simple standard form and, except in special instances, the numerical work of expansion is prohibitive. The second reason is the fact that the normal curve is quite a good approximation to the binomial distribution provided that either q is roughly equal to p or n is very large. (The normal curve is of course $y=y_0 e^{-\frac{x^2 n}{2pq}}$, where the origin of x is taken at q, the mean.)

The following tables, given by H. L. Seal in his interesting paper "Tests of a Mortality Table Graduation" (*J.I.A.* Vol. LXXI), illustrate this point. The notation has been altered and the way in which the ordinates have been selected for comparison will be apparent only after the original paper has been read.

Comparison of terms of $(p+q)^n$ *with ordinates of the Normal Curve*

$q=·0025$ $n=4000$		$q=·005$ $n=2000$		$q=·01$ $n=1000$	
Binomial	Normal	Binomial	Normal	Binomial	Normal
·8747	·8742	·8746	·8741	·8743	·8741
·6356	·6348	·6352	·6344	·6343	·6336
·4281	·4286	·4275	·4280	·4264	·4269
·2651	·2678	·2645	·2672	·2633	·2660
·1500	·1542	·1495	·1537	·1486	·1527
·0776	·0816	·0773	·0812	·0766	·0805
·0371	·0396	·0369	·0393	·0365	·0388
·0169	·0176	·0168	·0174	·0165	·0171
·0076	·0071	·0075	·0070	·0074	·0069
·0035	·0026	·0034	·0026	·0033	·0025
·0016	·0009	·0015	·0009	·0015	·0008
·0007	·0003	·0007	·0003	·0007	·0003
·0003	·0001	·0003	·0001	·0003	·0001
·0001	·0000	·0001	·0000	·0001	·0000
·0000		·0000		·0000	

Comparison of terms of $(p+q)^n$ *with ordinates of the Normal Curve*

$q=·03$ $n=333$		$q=·05$ $n=200$		$q=·1$ $n=100$	
Binomial	Normal	Binomial	Normal	Binomial	Normal
·8730	·8724	·8716	·8711	·8681	·8676
·6308	·6299	·6272	·6265	·6178	·6171
·4216	·4219	·4168	·4173	·4042	·4047
·2584	·2609	·2536	·2562	·2410	·2434
·1444	·1483	·1405	·1443	·1301	·1336
·0735	·0773	·0708	·0744	·0636	·0668
·0345	·0368	·0328	·0350	·0284	·0303
·0154	·0160	·0144	·0150	·0119	·0124
·0067	·0063	·0062	·0058	·0049	·0046
·0030	·0023	·0027	·0021	·0020	·0015
·0013	·0007	·0012	·0007	·0008	·0005
·0006	·0002	·0005	·0002	·0003	·0001
·0002	·0001	·0002	·0001	·0001	·0000
·0001	·0000	·0001	·0000	·0000	
·0000		·0000			

It will be seen therefore that within fairly wide limits values derived from the normal curve are good approximations to the true binomial values. This is particularly so in the neighbourhood of the mean (the origin).

16. Standard error of the mean.

The second index with which we shall deal is the mean of the sample, and we shall first prove the almost self-evident fact that the mean of all the sample means is the mean of the universe. Suppose that the universe consists of N variates denoted by $u_1, u_2, u_3 \dots u_N$, the values being measured from some convenient origin.

The number of different samples of n each which can be constructed is ${}^NC_n$, which for convenience we shall denote by m. This applies even if several of the u's are equal, since each observation has a separate individuality although its values may coincide with that of some other observation.

Denote the values in the rth sample by

$$u_{r:1}, u_{r:2}, u_{r:3} \dots u_{r:s}, \dots u_{r:n}.$$

Then the sample mean

$$M_r = \frac{1}{n}\{u_{r:1} + u_{r:2} + \dots + u_{r:n}\}.$$

The mean of all the m sample means is therefore

$$\frac{1}{m}(M_1 + M_2 + \dots + M_r + \dots + M_m) = \frac{1}{mn}\{(u_{1:1} + u_{1:2} + \dots + u_{1:n}) + (u_{2:1} + \dots + u_{2:n}) + \dots + (u_{m:1} + u_{m:2} + \dots + u_{m:n})\}.$$

Clearly each u appears in ${}^{N-1}C_{n-1}$ of these brackets (once in each), thus making the total number of terms $N\ {}^{N-1}C_{n-1} = n\ {}^NC_n = nm$, as it should, there being m samples of n each.

Hence the expression for the mean of the sample means reduces to

$$\frac{1}{mn}\,{}^{N-1}C_{n-1}(u_1 + u_2 + \dots + u_N).$$

Since $m = {}^NC_n$, this becomes

$$\frac{1}{N}(u_1 + u_2 + u_3 + \dots + u_N)$$

or the mean of the universe.

To find the standard error of the sample means, take the universe mean as origin and denote the observed values measured from it by

$$v_1,\ v_2,\ \dots\ v_r,\ \dots\ v_N, \quad \text{so that} \quad \sum_1^N v_r = 0.$$

The mean of the rth sample measured from the new origin is given by

$$M'_r = \frac{1}{n}(v_{r:1} + v_{r:2} + \dots + v_{r:n}),$$

using an obvious extension of the previous notation.

The standard deviation of all the sample means, σ_M, is given by the relation

$$m\sigma_M^2 = \{M_1'^2 + M_2'^2 + M_3'^2 + \dots + M_m'^2\}$$

(since by the previous paragraph the mean of the M's is our new origin); i.e.

$$mn^2\sigma_M^2 = \{(v_{1:1} + v_{1:2} + \dots + v_{1:n})^2 + (v_{2:1} + v_{2:2} + \dots + v_{2:n})^2 + \dots \\ + (v_{m:1} + v_{m:2} + \dots + v_{m:n})^2\}. \quad \dots\dots(3)$$

If we expand the right-hand side and collect like terms it will reduce to the form

$$A(v_1^2 + v_2^2 + \dots + v_N^2) + 2B(v_1v_2 + v_1v_3 + \dots + v_rv_s + \dots),$$

since although for convenience we have used two suffixes to denote the sample considered and the observation in the sample, there are in fact only N different v's and each of them will occur in many samples.

Any given v will occur in ${}^{N-1}C_{n-1}$ samples and therefore in this number of brackets on the right-hand side of (3). Hence on collecting like terms the coefficient A of each v^2 will be ${}^{N-1}C_{n-1}$.

Similarly, any two given v's, v_r and v_s say, will occur in ${}^{N-2}C_{n-2}$ samples, and hence the coefficient $2B$ of each term such as v_rv_s will be $2\,{}^{N-2}C_{n-2}$.

We have therefore

$$mn^2\sigma_M^2 = {}^{N-1}C_{n-1}(v_1^2 + v_2^2 + \dots + v_N^2) + 2\,{}^{N-2}C_{n-2}(v_1v_2 + \dots). \quad \dots\dots(4)$$

Since the v's are measured from the mean:

$$v_1 + v_2 + \dots + v_N = 0.$$

$$\therefore\ 2(v_1v_2 + v_1v_3 + \dots + v_rv_s + \dots) = -(v_1^2 + v_2^2 + \dots + v_N^2).$$

Substituting in (4):

$$mn^2\sigma_M^2 = ({}^{N-1}C_{n-1} - {}^{N-2}C_{n-2})(v_1^2 + v_2^2 + \dots + v_N^2).$$

But $v_1^2 + v_2^2 + \dots + v_N^2 = N\sigma^2$, where σ is the standard deviation of the universe and $m = {}^NC_n$, so that we finally have

$$\sigma_M^2 = \frac{{}^{N-1}C_{n-1} - {}^{N-2}C_{n-2}}{{}^NC_n}\frac{N}{n^2}\sigma^2$$

$$= \frac{N-n}{N-1}\frac{\sigma^2}{n}. \qquad \dots\dots(5)$$

If N is large we can assume that $N-n$ is not very different from $N-1$ and write

$$\sigma_M = \frac{\sigma}{\sqrt{n}}. \qquad \dots\dots(6)$$

This gives us the standard error, i.e. the standard deviation of the sampling distribution of the mean in terms of σ the standard deviation of the universe.

Similarly, expressions have been obtained for the standard errors of most of the well-known statistics, but the proofs are generally difficult and beyond the scope of this book. They are set out for convenience in the following table and should be memorized.

Standard errors of well-known statistics based on samples of n

Index	Standard error	Remarks
Class frequency (nq)	$\sqrt{npq}$	q is the proportionate class frequency in the universe and $p = 1 - q$
Proportionate class frequency (q)	$\sqrt{\frac{pq}{n}}$	
Mean (M)	$\frac{\sigma}{\sqrt{n}}$	σ is the standard deviation of the universe
Standard deviation (σ)	$\frac{\sigma}{\sqrt{2n}}$ *	
Coefficient of correlation (r)	$\frac{1-r^2}{\sqrt{n}}$ *	r is the coefficient of correlation in the universe

The mean class frequency in nq and the means of the other indices are the corresponding values in the universe.

The formula $\frac{\sigma}{\sqrt{2n}}$ for the standard error of a standard deviation

* These formulae are approximate and should generally be used only if n is large.

applies exactly only if the standard deviation is itself derived from a normal distribution. It does not therefore strictly apply to the standard deviation of a binomial distribution, although in practice it is so applied. The general formula for the standard error of a standard deviation is

$$\sqrt{\frac{\mu_4-\mu_2^2}{4\mu_2 n}},$$

which however is largely of theoretical interest.

We are now in a position to deal with the types of problem mentioned on p. 86.

17. When information about the universe is available.

As has been stated, the problem is to find whether the sample is likely to have been obtained from the universe by random or other unbiased sampling. To do this we examine those statistics in which we are particularly interested; usually these include the mean and the standard deviation.

Suppose, for instance, that a headmaster wished to know how the standard of English teaching in his school compared with that of the general body of schools and decided to compare the results of (say) the School Certificate examination, treating his own school as a sample.

He might decide to take four "classes": under 25% marks, 25–49% marks, 50–74% marks and 75% marks and over. Suppose that he obtained the following results for his school, which sent in a hundred candidates for the English papers, for which the maximum possible marks were 250:

		School	Results for the whole country
Mean mark		165	170
Standard deviation of marks		22	20
Proportion with less than 25% marks		18 per cent	16 per cent
,,	25–49% marks	30 ,,	26 ,,
,,	50–74% marks	38 ,,	48 ,,
,,	75% and over	14 ,,	10 ,,

We know that the standard error of the mean is given by the formula $\sigma/\sqrt{n}$ which in this example gives $\frac{20}{\sqrt{100}}=2{\cdot}0$, as there are 100 candidates.

We now assume that if all possible samples of 100 were taken, the mean marks found from them would be normally distributed with a mean of 170 (the mean mark for the universe) and standard deviation 2·0.

The actual sample has a mean 165, i.e. 5 less than the universe mean. $\frac{5}{2}=2\cdot5$, so that the difference between the universe and sample means is two-and-a-half times the standard error.

Now we know from Chapter II, p. 36, that about 95·5 per cent of the area of a normal curve lies between the ordinates $\pm2\sigma$ and 99·7 per cent lies between the ordinates $\pm3\sigma$, so that the chance of a random value lying further from the mean than $2\cdot5\sigma$ is small (say $\frac{3}{100}$ approximately).

In other words, the results for that year indicate that the school is below the general level to an extent which is unlikely to be due to sampling errors.

Similarly, the standard error of the standard deviation is $\frac{\sigma}{\sqrt{2n}}$; in this example $\frac{20}{\sqrt{200}}=1\cdot4$ approx.

As the sampling distribution has a mean of 20 and a standard error of 1·4 we expect most of the sample values of σ to lie within $\pm3\times1\cdot4$ of the value 20. Actually the value 22 calculated for the given school lies about one-and-a-half times the standard error from the universe value and we cannot say that the school is abnormal in the "scatter" of its marks.

Similarly, we can compare the four proportionate class frequencies. Take, for example, the class 50–74 marks, for which the proportionate class frequencies are ·38 for the sample and ·48 for the universe.

The standard error of the sampling distribution of the proportionate class frequency is

$$\sqrt{\frac{pq}{n}}=\sqrt{\frac{(\cdot48)(\cdot52)}{100}}$$

$$\fallingdotseq\cdot05.$$

The absolute magnitude of the difference between the proportionate class frequencies for the sample and the universe is

$$\cdot48-\cdot38=\cdot10.$$

This difference is twice the standard error and is probably therefore significant of a real difference between the sample and the universe.

Similarly, the other classes could be dealt with and the final conclusion would be that as regards that year's results the school was below the average.

(In this last section we have tacitly assumed that the normal law holds for q. The inaccuracy cannot be great.)

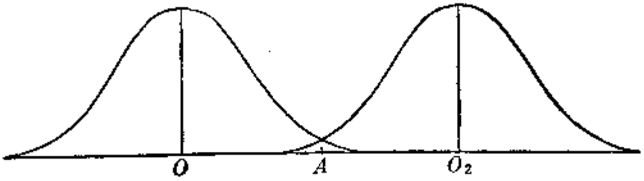
18. When only the sample is available for analysis.

In the second type of problem we cannot evaluate the standard errors accurately since we do not know the parameters. Accordingly we have to fall back upon the sample values as approximations and take, for example, $\frac{s'}{\sqrt{n}}$ for the standard error of the mean, where s' is the sample value of σ. It can be shown that it is more satisfactory to take $\frac{s'}{\sqrt{n-1}}$ for the standard error of the mean, using the form $\frac{\sigma}{\sqrt{n}}$ only when the parameter σ is known. The difference produced by replacing n by $n-1$ is only appreciable when n is small and for most purposes can be ignored in actuarial work. For a proof the reader is referred to *An Introduction to the Theory of Statistics* (see Bibliography).

Hitherto we have assumed the normal curve giving the sampling distribution to be known since we have had the mean and standard error given.

Now, however, the standard error is only approximate and we do not know the value which should be assumed for the mean of the sampling distribution.

Consider the following diagram:



A represents the value of the statistic considered. We calculate the standard error s' of the statistic, using the sample values of any necessary constants, and assume that this is a sufficiently good approximation to the true standard error.

We then measure distances O_1A, AO_2 equal to $2s'$ and draw normal curves with O_1 and O_2 as origins and with s' as standard deviation. The true sampling distribution might follow either of these curves or any curve of the same shape in between them, but the origin of the true curve is unlikely to lie outside the range O_1O_2.

To see why this is so consider the curve with mean O_1. If it were the sampling curve the error O_1A of the sample value would be twice the standard error.

Now although such an error may arise it is rather unlikely to be exceeded. Similarly, if the true curve were the one with mean O_2 we should have an error AO_2 equal to $2s'$.

It should not be assumed that $2s'$ is a sort of "foot-rule" to test whether an error is likely or unlikely, but for many purposes it is convenient to regard it as a likely upper limit. An error of as much as $3s'$, $4s'$ or even $150s'$ *may* occur, but while some authors use $3s'$ as an upper limit to the error likely to arise, such a criterion is severe. It will be remembered that if the sampling distribution is in fact normal the chance of an error arising as large as or larger than $3s'$ is only about ·0027 (see p. 36).

We may say therefore that, if A is the value of an index calculated from a sample and s' is the approximate standard error, the true value of the index is unlikely to lie outside the range

$$A-2s' \quad \text{to} \quad A+2s'$$

and very unlikely to lie outside the range

$$A-3s' \quad \text{to} \quad A+3s'.$$

Example 1

On p. 65 of Chapter III we found a coefficient of correlation ·782 for a sample of 365.

The approximate standard error is therefore $\dfrac{1-(\cdot 782)^2}{\sqrt{365}}=\cdot 020$. The true value of r found from unlimited observations may be as low as $\cdot 782-2\,(\cdot 020)$, i.e. ·742, or even $\cdot 782-3\,(\cdot 020)$, i.e. ·722, but is unlikely to be less. Thus very marked correlation does exist and the value of the parameter can be fixed within fairly narrow limits.

19. Test of a hypothesis.

Although the above method of approach is often used to make deductions from sample values, it is more logical to proceed on the lines previously indicated, i.e. to decide on a hypothesis about the universe and test its validity by means of the sample.

Thus in the previous example we might adopt the hypothesis that the true coefficient of correlation was only ·74. Taking the standard error as $\frac{1-(\cdot 74)^2}{\sqrt{365}} = \cdot 024$, we proceed as follows:

observed value − assumed value $= \cdot 782 - \cdot 74 = \cdot 042$.

The probability that a value chosen at random lies within a distance ·042 of the mean

$$= \frac{2}{\sqrt{2\pi}\sigma}\int_0^{\cdot 042} e^{-\frac{x^2}{2\sigma^2}}\,dx.$$

Putting $x = \sigma x' = \cdot 024x'$, this reduces to

$$\frac{2}{\sqrt{2\pi}}\int_0^{1\cdot 75} e^{-\frac{1}{2}x'^2}\,dx'.$$

Table I in the Appendix gives the value ·9205 for this integral.

The probability that a value differs from the mean by more than ·042 is therefore only $1 - \cdot 9205 = \cdot 0795$ and the chance that it is less than the mean by more than ·042 is half this, i.e. ·04 approx.

This is so small that our hypothesis seems improbable and the true coefficient of correlation is unlikely to be as low as ·74.

20. Probability levels.

For the normal curve in its simplest form

$$y = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

the probability that a value chosen at random differs from the mean by more than K, say, is given by the expression

$$1 - 2\int_0^K \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}\,dx. \qquad \text{......(7)}$$

Hitherto we have found the probabilities for given values of K, but for many purposes it is convenient to fix the probability and use equation (7) to find the corresponding value of K.

Tables have been prepared for various *probability levels* (as are called), the values most commonly used for the prol! draw being 50, 5, 1 and ·1 per cent. At the foot of Table I in the Apndard figures are given for twelve useful probability levels. her of

Probability levels are frequently met with in connection, but guarantees given by manufacturers as to the performance $O_1 O_2$ products. For instance, a firm selling wire cables might kn were a result of a large number of tests, that the breaking strain ould be ticular type of cable followed roughly the normal distributio mean M and standard deviation σ. A guarantee of performa to be usually essential and it may be decided to fix it so that not moran O_2 one rope in a thousand would be returned as unsatisfactory.

Suppose that the appropriate guaranteed strain is $M-K$ o test hypothesis the probability that a cable chosen at random has a s it is breaking strain is $\frac{1}{1000}$. Since the distribution is roughly nmuch this is also the probability that a cable has a breaking strain gse $3s'$ than $M+K\sigma$. on is

Hence the chance that a random value differs from the on is by $K\sigma$ or more is $\frac{2}{1000}$ or ·002. rger

The prepared tables show for the normal curve with unit stan deviation what value of K corresponds to the probability level ated Having obtained the value of K in this way the manufacturer could guarantee a breaking strain of $M-K\sigma$ with the knowledge that only about one rope in a thousand would be rejected as below standard. Incidentally only about one rope in a thousand would stand a higher breaking strain than $M+K\sigma$, but this would not be of practical importance.

It is important to remember that prepared tables do not distinguish between values greater or less than the mean. They give probabilities that a random value will lie within a given distance of the mean (either above or below it) and care should be taken to allow for this. For instance, in the above example the probability of $\frac{1}{1000}$ taken as the standard had to be doubled so as to include the probability of breaking strains in excess of the mean.

21. Further remarks on correlation.

It will be noticed that Example 1 deals with correlation. It is in this connection that tests for what is known as "significance" are particularly important. If any two series of numbers are written down at random, or if two quite unconnected sets of pairs are taken (e.g. number of the house and salary of the owner), the usual process will give us a coefficient of correlation which may be positive or negative, but is very unlikely to be exactly zero. A significance test would almost certainly show, however, that the whole of this apparent correlation may be due to errors of sampling and that the universe value of r is probably zero.

Before attempting to draw any conclusions it is essential therefore to test all coefficients of correlation by comparing them with the standard error, $1/\sqrt{n}$ which is appropriate in this case since our hypothesis is that correlation in the universe is zero. Care must be taken in the interpretation of correlation coefficients even when they are found to be significant.

Suppose that an unskilled investigator wished to find if births and deaths were correlated and obtained his data by taking the births and deaths in a number of towns in a given year. The births and deaths might quite well show a marked degree of correlation which, while not spurious in itself, might cause some very misleading results to be drawn. In actual fact the number of births is the result of two factors, the birth-rate and the population of the town (or more correctly the female population of child-bearing ages), and similarly the number of deaths reflects the product of the death-rate and the population of the town. The result of the investigation is very probably due therefore to the correlation of the "weights", viz. the female population of child-bearing age and the total population of the town, and not to any interdependence of the fundamental birth-rates and death-rates. When the functions dealt with are composite and reflect the combined operation of a fundamental variable and sets of "weights" such as the populations above, any correlation shown may arise simply from the "weights" and should be further analysed.

In statistics we sometimes meet time series which may be defined as records of the values of variables taken at regular intervals of time. Correlation between two series is often spurious in that it reflects only the result of changing conditions on two quite unrelated variables. In an address given to the Royal Statistical Society (*J.R.S.S.* Vol. LXXXIX, p. 1) G. Udny Yule gives an interesting example of what he calls "nonsense correlation" between two series. The first showed the proportion of Church of England marriages to all marriages in the years 1866 to 1911, while the second was formed by the standardized mortality of 1000 persons in England and Wales for the same years. The coefficient of correlation was +·95, with a standard error of only ·014. This is, however, due to the fact that over the period considered both the variables decreased fairly steadily, so that both might be said to be "correlated" with *time*. It is probable that some of the correlation in Example 1 of Chapter III is spurious, since we are dealing with two time series over the years 1810 to 1834.

22. Comparison of two samples.

For the satisfactory comparison of two samples it is usually desirable to compare means, standard deviations, and, in some cases, class frequencies. Generally, tests for significance have to be restricted to the functions whose standard errors can be found. One obvious way of comparing the two sample values of a given index (say, the mean) is as follows:

If M_1 is the value for one sample, and σ_1 its standard error, then the "true" value of the index is unlikely to be outside the range $M_1-2\sigma_1$ to $M_1+2\sigma_1$, and still less likely to lie outside $M_1-3\sigma_1$ to $M_1+3\sigma_1$. Similarly, if M_2 and σ_2 refer to the second sample, the "true" value for the universe from which it was drawn should be between $M_2-2\sigma_2$ and $M_2+2\sigma_2$ or, at any rate, between $M_2-3\sigma_2$ and $M_2+3\sigma_2$. If the ranges $M_1-2\sigma_1$ to $M_1+2\sigma_1$ and $M_2-2\sigma_2$ to $M_2+2\sigma_2$ do not overlap the two samples are very unlikely to be drawn from the same universe, because the parameter cannot lie within a distance $2\sigma_1$ of M_1 and at the same time lie within a distance $2\sigma_2$ of M_2. This test is, however, much too severe for general use, for the following reason. We know from Chapter II, p. 36, that in a normal distribution the chance of a value selected at random lying further than 2σ from the mean is $1-{\cdot}9545$, i.e. ·0455, which is small; but if the ranges $M_1-2\sigma_1$ to $M_1+2\sigma_1$ and $M_2-2\sigma_2$ to $M_2+2\sigma_2$ in the above test did not overlap, the true or universe values of the index would be the same only if M_1 and also M_2 lay more than twice the standard error from the parameter. The probability of this is $({\cdot}0455)^2$, or about 1 in 480.

To arrive at a more satisfactory test let us consider the difference M_1-M_2 between the two sample indices.

In Chapter III it was shown that if z is the difference between two variables x and y, $\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2}$ (formula (20)), if no correlation exists between x and y, or

$$\sigma_z=\sqrt{\sigma_x^2+\sigma_y^2-2r\sigma_x\sigma_y} \quad \text{(formula (24))},$$

if correlation does exist.

Now a standard error is merely another name for a standard deviation and it follows therefore that in our previous notation the standard error of M_1-M_2 is $\sqrt{\sigma_1^2+\sigma_2^2}$, provided that we are satisfied that correlation does not exist between the two samples.

A direct comparison of M_1-M_2 with $\sqrt{\sigma_1^2+\sigma_2^2}$ will then indicate whether there is any strong evidence that the means of the universes from which the samples are drawn are unequal.

Example 2.

The heights of 100 men are recorded and it is found that the mean height is 5 ft. 9 in. and the standard deviation 2·24 in. A second sample of 100 men gives a mean height 5 ft. 7 in. and standard deviation 2·23 in.

To form an opinion as to whether the samples are drawn from the same or different universes we compare both the means and the standard deviations.

Using the formula $\frac{\sigma}{\sqrt{n}}$, the standard error of the first mean is

$$\frac{2{\cdot}24}{\sqrt{100}} \text{ in.} = {\cdot}224 \text{ in.}$$

Similarly, the standard error of the second mean is ·223 in. Assuming therefore that there is no correlation between the samples the standard error of the difference in the means is $\sqrt{({\cdot}224)^2+({\cdot}223)^2}={\cdot}315$ approx. The actual difference is 2 in. or about 6 times the standard error and this cannot therefore be accounted for by sampling errors.

To test the standard deviations we use the formula $\frac{\sigma}{\sqrt{2n}}$ for the standard error of a standard deviation. The standard error of the first sample standard deviation $= \frac{2\cdot24}{\sqrt{200}}$ in. $= \cdot158$ in., and the same value applies approximately for the second sample. Hence the standard error of the difference of the standard deviations is approximately

$$\sqrt{(\cdot158)^2+(\cdot158)^2} = \cdot22 \text{ in. roughly.}$$

The actual difference is only ·01 in., so that there is certainly no evidence here of significant difference.

The test of the standard deviations is only of theoretical interest as the difference in the means has already given convincing evidence that the samples have not been drawn from the same universe. If this test had been inconclusive we should have tested the hypothesis that the samples came from universes with similar standard deviations. This is an instance when the argument discussed on p. 104 could have been employed as follows:

Using three times the standard error as the maximum error likely to arise we find that the mean height in the universe from which the first sample is drawn is unlikely to lie outside the range 5 ft. 9 in. ± ·672 in., i.e. it is unlikely to be greater than 5 ft. 9·67 in. or less than 5 ft. 8·33 in. Similarly, the mean height in the universe from which the second sample is drawn is unlikely to lie outside the range 5 ft. 7 in. ± ·669 in. The upper limit of this range, i.e. 5 ft. 7·67 in., does not fall within the first range and hence it is very unlikely that the two universes are the same.

23. Amalgamation of samples.

It will be remembered that, in evaluating the formulae for the standard errors $\sqrt{npq}, \frac{\sigma}{\sqrt{n}}, \frac{1-r^2}{\sqrt{n}}$, and so on, the values of the indices q, σ and r should be taken from the universe.

Unfortunately this is impossible because the only available data are those included in the sample or samples; we are therefore faced with the difficulty of making the best estimate we can. In large sample work it is usually sufficient to take the sample value as an approximation to the corresponding parameter but if the number in the sample is small it is necessary to reduce bias; thus, if s is the standard deviation in a sample of n, we should use $s/\sqrt{n-1}$ and not $s/\sqrt{n}$ as an estimate of the standard error of the mean.

If we are considering more than one sample it is important to specify clearly the hypothesis which is being tested since this affects the way in which any parameters are estimated. Suppose, for instance, that two samples of n_1 and n_2 observations respectively with means m_1 and m_2 and standard deviations s_1 and s_2 are being considered. We might decide to test whether the means differ significantly and to do this we set up the hypothesis that the samples have been drawn from universes with the same mean (we do not specify that they should have the same standard deviation as we are not testing the difference between s_1 and s_2).

The best estimate we can make of the common mean of the two populations is formed by amalgamating them and taking the mean $(n_1m_1+n_2m_2)/(n_1+n_2)$ of the combined data. The estimates of the standard errors of the means of the two populations are then

$$s_1/\sqrt{n_1-1} \quad \text{and} \quad s_2/\sqrt{n_2-1}$$

or simply

$$s_1/\sqrt{n_1} \quad \text{and} \quad s_2/\sqrt{n_2}$$

if the n's are large.

If we assume no correlation we then take $\sqrt{s_1^2/n_1+s_2^2/n_2}$ as the estimate of the standard error of the difference in the means by which to test the observed difference $|m_1-m_2|$.

If, on the other hand, we were testing the same hypothesis assuming that the samples were drawn from populations with the same standard deviation (σ, say) we should amalgamate the samples in order to form an estimate of this parameter σ.

Example 3.

The following data have been obtained relating to schoolchildren:

	Total	No. with medium hair colour
Edinburgh	9,743	4,008
Glasgow	39,764	17,529

We proceed to test the hypothesis that the samples are drawn from populations in which the proportion with medium coloured hair is the same. To form an estimate of this common proportion we amalgamate the two samples and assume the common parameter q to be equal to:

$$\frac{4008+17,529}{9743+39,764}=\cdot 4350.$$

The standard error of this value of q is

$$\sqrt{\frac{\cdot 4350 \times \cdot 5650}{9743}} \quad \text{for Edinburgh (sample size 9743).}$$

and

$$\sqrt{\frac{\cdot 4350 \times \cdot 5650}{39{,}764}} \quad \text{for Glasgow (sample size 39,764).}$$

Hence, if our hypothesis were correct and we took samples of 9743 and 39,764 respectively from the two populations, the expected value of the difference between the proportions of children with medium coloured hair would be zero and the standard error would be:

$$\sqrt{\cdot 4350 \times \cdot 5650 \left(\frac{1}{9743} + \frac{1}{39{,}764}\right)} = \cdot 0056.$$

As the observed difference is

$$\left| \frac{4008}{9743} - \frac{17{,}529}{39{,}764} \right| = \cdot 0294.$$

As this is more than five times the standard error we conclude that our hypothesis is untenable.

Before drawing any conclusions, however, we should need to be satisfied that the description "medium hair colour" is uniform between the two cities. Owing to the difficulties of obtaining a satisfactory definition it is likely that different standards were adopted.

24. Loss of degrees of freedom.

It should not be assumed that the values of parameters are always estimated from the sample. Sometimes the hypothesis itself specifies such a value. For instance, in the middle of p. 103 we considered the hypothesis that two variables were really uncorrelated (a very common hypothesis) and therefore assumed a standard error of $1/\sqrt{n}$ which is obtained by putting $r=0$ in the formula quoted on p. 96.

When the values of parameters have to be estimated from the sample the student may feel that this tends to be unduly favourable to the hypothesis since the more the test depends on sample values the more likely is it to indicate a good "fit" with the sample. This point is unimportant if the number of values fitted is large but it is of general interest and must be allowed for if only a few values are being fitted. We say that "*degrees of freedom*" are lost when the values of parameters are estimated from the sample. This phrase will be discussed further in Chapter V where examples will be found of the adjustments made in applying a certain test for this loss of degrees of freedom.

Example 4.

The following example illustrates how correlation may sometimes be dealt with:

	Total live births in 1937	Male live births in 1937	Proportion of male births to total births
Town *A*	956	502	·525
Towns *A* and *B* combined	1406	697	·496

Is there any significant difference in these results?

If the proportionate class frequencies are compared directly, correlation is bound to arise, since the proportion in town *A* influences that for the combined towns, and the standard deviation of the difference is of the more general form $\sqrt{\sigma_1^2+\sigma_2^2-2r\sigma_1\sigma_2}$. If the true value of r is unknown the problem may sometimes be solved by giving r its maximum and minimum values $+1$ and -1 respectively; a more refined method of approach is however discussed in the next section.

The simplest method of dealing with this problem is to eliminate the obvious correlation by subtracting the data for town *A* from the combined data. We thus obtain for town *B* only:

Total live births in 1937	...	...	450
Male live births in 1937	...	...	195
Proportion of male births to total births			·433

Standard error of proportion of male births in town $A=\sqrt{\dfrac{\cdot525\times\cdot475}{956}}$,

„ „ „ „ $B=\sqrt{\dfrac{\cdot433\times\cdot567}{450}}$.

Standard error of difference of these proportions

$$=\sqrt{\frac{\cdot525\times\cdot475}{956}+\frac{\cdot433\times\cdot567}{450}}=\cdot028.$$

The actual difference is $\cdot525-\cdot433=\cdot092$.

As this is between three and four times the standard error it is significant.

If it is desired to compare the proportion for town *A* with the proportion for the combined towns, the correlation can be allowed for as follows, bearing in mind that a given proportion for town *A* can be associated with one and only one proportion for the combined towns.

Let q_A, q_B and q_{A+B} represent the proportions of male births for A, B and the combined towns respectively, and let n_A and n_B denote the total number of live births in A and B respectively. Let σ_A, σ_B and σ_{A+B} denote the standard errors of q_A, q_B and q_{A+B}, and let r denote the coefficient of correlation between q_A and q_{A+B}.

Then the standard error of $q_A - q_{A+B}$ is

$$\sqrt{\sigma_A^2+\sigma_{A+B}^2-2r\sigma_A\sigma_{A+B}}. \qquad \ldots\ldots(8)$$

Now
$$q_{A+B}=\frac{n_Aq_A+n_Bq_B}{n_A+n_B}.$$

Hence, if capital letters denote deviations from the mean,

$$Q_{A+B}=\frac{n_AQ_A+n_BQ_B}{n_A+n_B},$$

since in considering standard errors we assume samples of the same size to be taken, so that n_A and n_B are constants.

$$\therefore\ Q_{A+B}Q_A=\frac{n_A(Q_A)^2+n_BQ_BQ_A}{n_A+n_B}. \qquad \ldots\ldots(9)$$

If we sum for all the N imaginary samples included in the sampling distribution we have

$$\Sigma Q_{A+B}Q_A=Nr\sigma_{A+B}\sigma_A; \quad \text{(from definition of } r\text{)}$$

while from equation (9) we have

$$\Sigma Q_{A+B}Q_A=\frac{n_A}{n_A+n_B}\Sigma(Q_A)^2+\frac{n_B}{n_A+n_B}\Sigma(Q_AQ_B).$$

The second term on the right vanishes because by hypothesis the samples from A and B are independent,

and
$$\frac{\Sigma Q_AQ_B}{\sigma_A\sigma_B},$$

the coefficient of correlation between q_A and q_B, must be zero.

Also
$$\Sigma(Q_A)^2=N\sigma_A^2,$$

so that we have, finally,

$$Nr\sigma_{A+B}\sigma_A=\frac{n_A}{n_A+n_B}N\sigma_A^2,$$

$$r=\frac{n_A}{n_A+n_B}\cdot\frac{\sigma_A}{\sigma_{A+B}}.$$

But $$\sigma_A = \sqrt{\frac{\cdot 496 \times \cdot 504}{956}}$$

and similarly $$\sigma_{A+B} = \sqrt{\frac{\cdot 496 \times \cdot 504}{1406}}$$

(taking the observed value of q_{A+B} as the true mean value of the q's for both towns).

Hence the expression (8) becomes:

$$\sqrt{\cdot 496 \times \cdot 504 \left\{\frac{1}{956} + \frac{1}{1406} - 2\,\frac{956}{1406}\cdot\frac{1}{956}\right\}}$$
$$= \sqrt{\cdot 496 \times \cdot 504 \times \frac{450}{956 \times 1406}}$$
$$= \cdot 0091 \text{ approx.}$$

The observed difference $q_A - q_{A+B}$ is $\cdot 525 - \cdot 496 = \cdot 029$.

As this is more than three times the standard error it is very probably significant. If the calculations had been carried out to a greater degree of accuracy it would in fact be found that the ratio of the difference of the q's to the standard error of the difference is the same whether we compare q_A with q_B or q_{A+B} with either q_A or q_B.

BIBLIOGRAPHY

Methods of Statistical Analysis. C. H. GOULDEN. New York, 1939.

An Introduction to the Theory of Statistics, chaps. 20 and 21. G. UDNY YULE and M. G. KENDALL. London, 1948.

Statistical Methods for Research Workers. R. A. FISHER. London, 1938.

An Introduction to Medical Statistics. A. BRADFORD HILL. London, 1948.

EXAMPLES 4

1. A factory turns out an article by mass production methods. From past experience, which is sufficiently extensive to be reliable, it appears that 10 articles on the average are rejected out of every batch of 100. Find the standard deviation of the number of rejects in a batch. What is the approximate probability of 7 or more being rejected?

It is reported that several batches have recently been turned out containing 20 to 30 rejects. What inference would you draw?

2. An examination is held in several different centres and the following data have been extracted from the results:

	Number of candidates	Mean percentage of marks	Standard deviation of percentage of marks
Centre *A*	127	44·8	8·3
All other centres combined	2346	47·3	6·5

Do you consider these data significant of any difference between the candidates at Centre *A* and at other centres and, if so, why?

3. An assurance company, a large proportion of whose business consists of 15- and 20-year term endowment assurances for sums assured of £50 and £100, has in the past kept hand-written valuation cards on which net premiums are inserted to three places of decimals. These cards are filed according to year of birth for whole life assurances, and according to year of maturity for endowment assurances, and are arranged within each group according to date of entry.

Machine-punched cards are now to be adopted and it is desired to save space on the card by tabulating the net premiums to the nearest integer. You are asked what the error is likely to be in the total of a number of net premiums and how you would proceed to construct samples for the purpose of testing whether any bias in the figures is likely to cause a divergence from your theoretical results.

Investigate the problem and draft a short reply in non-technical language.

4. An office has on its books 4000 policies of all classes subject to monthly premiums, the average sum assured being £300 per policy and deviations from the average being negligible in number and amount. The monthly premium for each policy is obtained by adding $2\frac{1}{2}$ per cent to the annual premium rate per £100 sum assured (this rate being calculated to the nearest penny), dividing the amount so obtained by twelve, fractions of a penny counting as one penny, and multiplying by the number of £100's assured.

Estimate the amount of loading in excess of $2\frac{1}{2}$ per cent secured, indicating the statistical error involved.

5. Under a group life assurance scheme the employees of a large industrial combine are on each 1st January assured against death during the ensuing year for sums of £100, £200, £300, £400 or £500 according to their salaries on the 1st January. The following schedule shows the number of employees assured on 1st January 1937—subdivided according to the nearest quinquennial age at that date, and, in the last column, the rate of mortality used in calculating the premiums:

Age	No. of employees assured for					Total no. of employees	Total sum assured	q_x
	£100	£200	£300	£400	£500			
25	2000	1000	300	100	—	3,400	530,000	·003
30	1800	1100	500	300	100	3,800	720,000	·004
35	1600	1100	700	400	200	4,000	850,000	·005
40	1300	1000	800	500	300	3,900	920,000	·006
45	1000	900	700	600	400	3,600	930,000	·008
50	700	700	600	500	400	2,900	790,000	·012
55	300	400	400	400	500	2,000	640,000	·020
Total	8700	6200	4000	2800	1900	23,600	5,380,000	—

The death claims arising during 1937 amounted to £51,000.

To what extent do you consider that the difference between the mortality experienced during 1937 and that used in calculating the premiums was significant?

(You may ignore any correction for the grouping according to nearest quinquennial age.)

6. (*a*) An office investigating its mortality experience among lives aged x exactly observes that, among E_x lives, θ_x deaths during the year of age give a rate of mortality

$$q_x = \theta_x / E_x.$$

Assuming that the true rate of mortality q'_x is known, what deviation $|q_x - q'_x|$ would you consider significant? Give reasons.

State the test for significance of the difference between the observed deaths θ_x and those expected by the true mortality θ'_x $[=E_x \times q'_x]$ and give a convenient practical approximation that you would make if q'_x were in the neighbourhood of ·01.

(*b*) The office, for the sake of convenience, would prefer to enumerate the number of policies on lives aged x and the number of such policies becoming claims, but it is suggested that the normal tests for significance would be invalidated by the existence of lives on which two or more policies are in force.

On what grounds is this criticism based? Illustrate your answer by assuming:

(1) That there are two policies in force on each of the E_x lives.

(2) That one-half of the E_x lives have one policy each and the other half two policies. For this purpose you may assume that the rate of mortality is the same among lives having one policy and lives having two policies.

7. A specified universe consists of N measurements and its standard deviation is σ_N. A group of n measurements is selected at random and its standard deviation is obtained as σ_n. Show that the mean value of $(\sigma_n)^2$, when many such samples are taken, approximates to

$$\frac{n-1}{n}\frac{N}{N-1}(\sigma_N)^2.$$

CHAPTER V

GRADUATION. GENERAL CONSIDERATIONS AND TESTS

1. Before we proceed to the description of individual methods of graduation there are several matters which can conveniently be dealt with as they arise in the application of several, or even all, methods.

2. As has been said previously, any body of data examined by the actuary should be regarded as a sample (even though all the available data have been included in the investigation) and hence any results deduced will be subject to errors of sampling.

By speaking of parameter values of, say, q_x or sickness rates z_x, we mean the ideal values which would have been obtained had there been unlimited data available (and the machinery for handling them) and had the years considered been themselves free from any accidental peculiarities. It is difficult to decide which of the peculiarities can properly be regarded as accidental and on this point opinions may differ. For instance, a severe influenza epidemic may render a given year abnormal, but the effect of epidemics in general should be allowed for in the universe values. Only the intensity of the attack may be regarded as abnormal. Again, while the years 1914–18 were abnormal it does not necessarily follow that, in considering our picture of the universe rates of mortality, sickness, fertility, etc., such upheavals as occurred in those years should be ignored.

3. The statistical rationale of graduation.

It may be said that the art of graduation is to arrive at an estimate of the true or universe values from the values derived from a particular investigation. For the purpose of most of the tests discussed in this chapter it will be assumed that the data examined are a random sample. In actual practice other considerations arise, for, in order to formulate estimates for the future, the actuary records and analyses data which must of necessity relate to the past. The graduated rates may not therefore indicate what results would have been obtained but for the limitations of sampling. The best example of this is given by the $a(m)$, $a(f)$ tables, which were derived from an investigation covering the years 1900–20. The graduated table finally adopted did not reflect the actual experience of those years.

There are other factors which sometimes apply, such as a desire to err on the side of safety either over the whole table or over a particular range; for the purpose of this chapter, however, it will be assumed that the graduated series of values represents an estimate of true or universe values. We shall develop tests to be applied on the assumption that the data used in the investigation represent a random sample, i.e. that no bias has been introduced accidentally or deliberately.

4. Properties of a well-graduated series.

In the language of the previous chapter we shall set up as a hypothesis that the graduated rates are the "true" rates and that the observed rates differ from them only owing to sampling errors.

There is a saying "natura non agit per saltum" expressing the fundamental fact that natural forces operate gradually and that their effects become apparent continuously and not in sudden jerks. In its application to mortality and sickness data it implies that any rates which may reflect the operation of purely natural causes should not exhibit any discontinuities, breaks, or sudden and unexpected changes. In other words, we expect any set of true values to follow a smooth curve, or, as we usually say, the graduated series must possess a high degree of smoothness.

We know the sampling distribution of functions such as q_x; this is only a proportionate class frequency, and if we assume that its distribution is roughly normal we can form some idea of the probabilities of errors of various sizes arising from the operation of random sampling.

Taking the graduated values as true values, we can calculate the discrepancies between the true values and the sample values (i.e. those revealed by the data) and examine whether they are reasonable or not.

We thus have two main sets of tests: (i) tests for smoothness and (ii) tests of fidelity to data.

5. Tests for smoothness.

For practical purposes any table which is to be extensively used should have a very high degree of smoothness: otherwise the more complicated functions based on it, such as policy values, will show alarming and even embarrassing irregularities.

It is usually found desirable therefore to examine the first three orders of differences of the graduated values. Generally speaking the third order of differences will be very small and it has been suggested that, in comparing two different graduations for smoothness, the sum of the third differences should be found for each table, the summation including all the values. On this basis it is usual to accept, as the better graduation, that which gives rise to the smaller total. On the other hand, it should be remembered that smoothness in the successive orders of differences is more important than smallness. It is quite easy in fact to write down any number of ideally smooth series derived from mathematical formulae for which the successive orders of differences, although smooth, show little or no tendency to diminish.

It should be remembered, moreover, that although natural causes are unlikely to produce irregularities, other factors may be operating to do so. Irregularities which they produce are often inherent in the data and no attempt should then be made to eradicate or reduce them. For instance, many pension schemes provide for retirement between ages 60 and 65 or before age 60 for reasons of ill health. The retirement rates for such a scheme would probably show a steady increase up to age 60 but a sudden jump when that age was reached. There might well be two peaks, one at age 60 (the first age for normal retirement) and the other at age 65, with a trough in between. These discontinuities, or sudden changes of curvature, are due to the operation of the rules, and unless there is good evidence that special circumstances, unlikely to recur, have exaggerated them no attempt should be made to reduce them so as to produce a more regular curve.

6. "Errors" and "mistakes".

As hitherto, we shall use the word "error" to indicate the discrepancy between a parameter and the corresponding value derived from a random sample. The error is solely due to the smallness of the sample. Unfortunately, other factors such as human fallibility have to be reckoned with, and we shall use the word "mistakes" to refer to inaccuracies or discrepancies between universe and sample values due to causes other than random sampling.

It may be said at once that there is no satisfactory way of dealing with mistakes except their complete eradication before graduation commences. Some of the tests described later may draw attention to mistakes and all methods of graduation do in fact reduce their disturbing effect. If mistakes can be traced the whole graduation should be done again after they have been corrected. Unfortunately this is often impracticable.

7. Bias.

As has been previously stated, bias does not introduce errors in the statistical sense and the foregoing theory does not apply. It may arise through some personal factor, such as misstatement of age, and sometimes the statistical processes may themselves introduce bias.

If bias cannot be eliminated, the method of graduation should be chosen so as to reduce the effect to a minimum. The method used for the English Life Tables in Chapter X is an excellent example.

If bias is at all extensive the ordinary tests fail, except that the examination of the deviations and accumulated deviation is still very valuable (see para. 11).

8. Special objects of the investigation.

In deciding upon a method of graduation, and in testing the results, the object of the investigation should be borne in mind. For instance, if a mortality table is required for use in ordinary life assurance it is important that the mortality should not be underestimated, while if it is to be used for calculating annuity purchase money the converse holds. For valuation purposes the gradient should not be underestimated over the important range of ages 40 to 70, since net premium reserves generally vary with the gradient of the mortality curve rather than with the lightness or heaviness of the rates.

9. Choice of the function to be operated upon.

The rough data derived from the investigation are usually available in the form of the exposed to risk at each age or group of ages and the corresponding decrements (deaths, retirements, marriages, etc.).

We may either

(i) attempt to graduate the exposed to risk and the decrements separately, finally deriving the rates of decrement by division;

or (ii) obtain ungraduated rates of decrement direct from the rough data and then graduate them.

The first method has certain advantages in special problems, notably those connected with the English Life Tables. Above all it enables the operator to keep in view the weight of the data at each age while he is performing the graduation.

If the second method is used, one rate of decrement is very much like another, irrespective of the volume of the data on which it is based; and although in testing the graduation the weight of the observations is allowed for this is not done during the actual process of graduation.

On the other hand the following considerations should be borne in mind:

(*a*) in many experiences the exposed to risk is essentially a discontinuous function; and

(*b*) a slight distortion of the exposed to risk may coincide with a slight distortion of the decrement in the opposite direction and the combined effect may be quite appreciable. Moreover, if the rate of decrement is increasing slowly, a slight distortion of the decrements may produce a graduated rate decreasing with an increase in age over a range where such a feature is unlikely to represent the facts.

As regards (*b*) it is not uncommon to graduate the exposed to risk and then to adjust the decrements so as to produce the same crude rates as before. The adjusted values are then graduated and serious distortion of the resulting rates is unlikely. The extra work involved is usually well worth while.

On the whole the method (ii) above, viz. the calculation of crude rates of decrement which are then graduated is almost invariably adopted. By this means the work is reduced to about one-half of that involved in method (i). Moreover, a quotient such as a rate of decrement is much more likely to progress smoothly from age to age and to follow approximately a mathematical curve than a function such as the exposed to risk.

In considering mortality tables the rate of mortality q_x is usually chosen for graduation, but the form of the curve with its gentle gradient at the younger ages and its very rapidly increasing gradient at higher ages makes it unsuitable for some purposes.

For this reason $\log q_x$ and $\log(q_x + \cdot 1)$ have sometimes been used, since these functions tend to be much flatter and can therefore be represented more easily in the graphic form.

Again, the reader will be familiar with tables in which μ_x is of the form $A + Bc^x$. Because of the practical advantages of such a table many attempts have been made to find satisfactory graduations of μ_x or colog p_x (a closely allied function) using Makeham's Law or some modification of it, such as $A + Bx + Dc^x$.

10. Comparison of graduation methods.

In graduation, more than in any other branch of actuarial science, it may be said that "the proof of the pudding is in the eating". No method, however unorthodox or crude it may appear, should be condemned in its application to a particular set of data provided that the results produced are satisfactory. It may be felt that the method is unlikely to be equally successful if used for other tables; that is, however, no criticism of its adoption for the particular table considered.

It will be seen later that each of the well-known methods has peculiarities which make it likely to be valuable in special circumstances. For example, the graphic method is especially useful if the data are scanty. The experimenter should not, however, be deterred from trying any process which appears likely to suit the particular problem in hand. The important point to remember is that the graduated table should satisfy the two essentials of smoothness and adherence to data.

11. Deviation and accumulated deviation.

In discussing tests for adherence to data it will be assumed for clearness that we are dealing with rates of mortality. Most of the results, however, apply to any tables of rates which can be regarded as proportionate frequencies (e.g. rates of withdrawal or marriage); they do not apply to sickness rates, which allow for duration of incapacity and are not therefore related to frequencies of occurrence.

In order to allow for the weight of the data at each age it is usual to compare actual frequencies and not proportionate frequencies. Thus the ungraduated rate of morality q_x' is not usually compared with the graduated rate q_x: both are weighted by the exposed to risk E_x and the actual deaths θ_x are then compared with the value of $E_x q_x$. $E_x q_x$ is defined as "the expected deaths at age x". In terms of statistics the E_x exposed in the sample are imagined as divided into two classes, those dying before age $x+1$ and those surviving to that age. The observed class frequency in the "deaths" category is θ_x, while the true value which would have resulted if the universe rate of mortality had applied is $E_x q_x$.

The expression "actual deaths minus expected deaths" is usually referred to as the "deviation" and will often be denoted by $\theta_x - E_x q_x$ or, more briefly, by $A-E$.

Table VI on pp. 122, 123 is typical of a graduation based on scanty data and shows how most of the functions needed in testing it are calculated. The functions "accumulated deviation" and "approximate standard error" will be discussed later.

The columns for Δq_x and $\Delta^2 q_x$ are inspected for smoothness, but on this occasion $\Delta^3 q_x$ has not been found, since only three significant figures of q_x are available and $\Delta^2 q_x$ is clearly smooth.

In column (8) the deviation is shown on the left if it is negative and on the right if it is positive. Allowing for sign the total of the column is ·3, which checks the numerical work, since the total actual deaths are 373 (column (3)) while the total expected deaths are 372·7 (column (7)). The sum of the actual deaths should always be nearly equal to the sum of the expected deaths, i.e. the total deviation allowing for sign should be approximately zero.

Table VI

Age x	Exposed to risk E_x	Actual deaths θ_x	$10^4 \times q_x$ graduated	Δ (4)	Δ^2 (4)	Expected deaths $E_x \times q_x$	Deviation: actual – expected (3) – (7)		Accumulated deviation Σ (8)		Approximate standard error $\sqrt{(7)}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)		(9)		(10)
							–	+	–	+	
47	166	2	144	1		2·4	·4		·4		1·5
48	187	2	145	2	1	2·7	·7		1·1		1·6
49	218	4	147	1	1	3·2		·8	·3		1·8
50	243	6	148	2	1	3·6		2·4		2·1	1·9
51	276	2	150	2	0	4·1	2·1		—		2·0
52	302	4	152	2	0	4·6	·6		·6		2·1
53	347	7	154	3	1	5·3		1·7		1·1	2·3
54	390	3	157	4	1	6·1	3·1		2·0		2·5
55	430	9	161	5	1	6·9		2·1		·1	2·6

56	494	9	166	6	1	8·1		·9		1·0	2·8
57	558	8	172	8	2	9·6	1·6		·6		3·1
58	628	11	180	10	2	11·3	·3		·9		3·4
59	701	14	190	12	2	13·3		·7	·2		3·6
60	813	18	202	15	3	16·4		1·6		1·4	4·0
61	917	18	217	18	3	19·9	1·9		·5		4·5
62	1040	24	235	21	3	24·4	·4		·9		4·9
63	1182	30	256	25	4	30·3	·3		1·2		5·5
64	1299	43	281	28	3	36·5		6·5		5·3	6·0
65	1432	41	309	31	3	44·2	3·2			2·1	6·6
66	1596	54	340	34	3	54·3	·3			1·8	7·4
67	1752	64	374			65·5	1·5			·3	8·1
Total		373	—	—	—	372·7	−16·4	+16·7	−8·7	+15·2	78·2

If the curve representing the graduated values closely followed the curve representing the observed values we should expect the two curves to cross and re-cross at frequent intervals. In other words we should expect the deviation to change sign frequently. In the table above, the deviation changes sign ten times in twenty-one values, and this must be considered very satisfactory.

The mere crossing and re-crossing of the curves is not enough, however, since large positive deviations at one point may be followed by only small negative deviations. To investigate this point the column headed "accumulated deviation" is formed by adding the previous column from the top downwards, i.e. the second entry is the sum of the first two in column (8), the third is the sum of the first three in column (8), and finally the last is the sum of the whole of column (8): this again checks the numerical work.

Clearly any item in column (9) represents the difference between the total actual deaths up to that age and the total corresponding expected deaths. The figures, therefore, should never be large and the same applies to their total, while the sign should change fairly frequently. In the table shown the total of column (9), allowing for sign, is 6·5 and there are nine changes of sign; this indicates satisfactory agreement with the data.

Most of the tests of a graduation relate to the size of the deviations. Before we leave the question of changes of sign, however, the following section may be of interest.

12. Changes of sign in the deviation and accumulated deviation.

If the graduated rates are regarded as parameters of the universe from which the given data were obtained by random sampling, any given deviation is equally likely to be positive or negative although strictly this assumes that the graduated value is a median rather than a mean.

If the signs are random the number of changes of sign in the column of deviations should be roughly equal to the number of non-changes; the same applies to the signs in the column of accumulated deviations.

This forms the basis of a test suggested by H. W. Haycocks in the discussion which followed the reading of H. L. Seal's paper, "Tests of a Mortality Table Graduation".

In that paper two more refined tests for changes of sign in the deviations were discussed. The first of these was devised by Makeham (*J.I.A.* Vol. XXVIII), and although the test is rather inelastic, the paper in which it appeared forms a useful introduction to a more modern paper by W. L. Stevens, "Distribution of groups in a sequence of alternatives", *Ann. Eugen., Lond.*, Vol. IX, p. 10. Seal showed how Stevens's technique could be applied to changes of sign in a series of deviations, but Stevens's original paper must be studied by anyone who wishes to make use of the method.

13. Standard error of a deviation.

We know that if a particular class had a proportionate frequency q in the universe the class frequency in random samples of n will have a binomial frequency distribution with mean nq and standard error $\sqrt{npq}$. If q is the rate of mortality q_x and the number in the sample is E_x this means that the expected deaths would form a binomial frequency distribution with mean $E_x q_x$ and standard error $\sqrt{E_x p_x q_x}$, where q_x is the true rate of mortality, which we may take as the graduated rate. In practical work p_x is usually so near to unity that the standard error is taken as $\sqrt{E_x q_x} = \sqrt{\text{expected deaths}}$. This function is tabulated in the last column of Table VI and gives a means of testing the size of the deviations.

Since the expected deaths $E_x q_x$ are the mean of the sampling distribution the deviation is merely a sampling error, and if we apply the results derived from the normal curve we can say that the deviation should not exceed twice the standard error, or at any rate, three times the standard error. Accordingly, we compare each deviation with the value in the last column, regardless of sign, and estimate roughly the probability of its arising through sampling errors. In Table VI only four deviations exceed the standard error—an excellent result which is probably fortuitous.

Although we have so far compared each individual deviation with its approximate standard error more information is often obtained by considering several values together. For instance, if we find a succession of, say, three positive deviations, we can compare their sum with the sum of the standard errors; the same applies still more forcibly if we find a succession of four or five deviations of like sign.

Thus the last three deviations in Table VI total 5·0, while the corresponding standard errors total approximately 22·1, about four times as much.

Similarly, we may group the data in order to obtain a group rate of mortality q',

$$\text{where } q' = \frac{\text{Total deaths for group}}{\text{Total "exposed" for group}} = \frac{\Sigma\theta_x}{\Sigma E_x},$$

and compare the group deviation with $\sqrt{(\Sigma E_x)\,q'p'}$ or, more simply, with $\sqrt{(\Sigma E_x)\,q'}$. This test is not strictly accurate, since the data at each age form separate samples with different sampling distributions, although as a practical device it has its uses.

It should be emphasized that, although deviations in excess of two or three times the standard error indicate distortion of the data, in graduation the converse does not hold. Deviations can be, and often are, only a small fraction of the standard error in good graduations, and should not then be regarded as evidence of undergraduation. Data are said to be undergraduated if the graduated curve has adhered too slavishly to the ungraduated values.

14. Application of the probable error.

It will be remembered that in the normal curve the probable error is approximately two-thirds of the standard deviation. Hence, if a sampling distribution is approximately normal, we may take the probable error as roughly $\frac{2}{3}$ (standard error) and from the definition of probable error we should expect roughly half the observed deviations to fall short of this value and the rest to exceed it. Thus at age 61 in Table VI we may take the probable error as about 3·0, and if we took a great many samples each of 917 lives we should expect the deviation (irrespective of sign) to exceed 3 in about half of the samples.

Actually, although we have only one sample at each age, the argument leads us to expect about half the deviations to fall short of the respective probable errors. It must be borne in mind that the deviations are not part of one sampling distribution; each is a single representative of its own sampling distribution. In Table VI, fifteen deviations are less than $\frac{2}{3}\sqrt{\text{expected deaths}}$, while only six exceed that value. This means that, on the whole, the deviations are less than we should expect and suggests that the data may be undergraduated. This test alone would not cause the graduation to be condemned.

15. Application of the mean deviation.

Although comparison of deviations with the probable errors as above may draw attention to the possibility of undergraduation, a more satisfactory test is needed. For this, the mean deviation may be used. In a normal distribution the mean deviation is roughly four-fifths of the standard deviation. Hence, if we make the usual assumption that in suitable conditions a binomial distribution is approximately normal we can take the mean deviation $(A-E)$ in a great many samples of E_x lives aged x as

$$\cdot 8\sqrt{E_x p_x q_x} \quad \text{or, approximately,} \quad \cdot 8\sqrt{E}.$$

In other words the deviation regardless of sign should on the average be $\cdot 8\sqrt{E}$. We have only one sample of lives aged x and cannot expect the observed deviation to approximate to the average value. If, however, we sum the deviations regardless of sign for all the available ages the total should approximate to $\cdot 8\Sigma\sqrt{E}$, because, although some deviations will be greater than their mean value and some less, these differences will tend to cancel out when a great many values are amalgamated.

In Table VI the total of the deviations regardless of sign is $16\cdot 4 + 16\cdot 7 = 33\cdot 1$. Four-fifths of the total of the last column is $62\cdot 5$, so that the total deviations are about half the expected value, thus tending to confirm our previous impression that the data have been undergraduated.

It should be emphasized however that no one test is conclusive and that in any event it is unlikely that a graduation would be discarded merely because it seemed to adhere too closely to the data, provided the smoothness were satisfactory. In nearly all such instances it will be found that the differences of the graduated rates progress irregularly and for this reason an attempt would be made to find a better graduation.

16. Limitations of the foregoing tests.

The tests for adherence to data, so far discussed, may be summarized as follows:

(1) The smallness of the individual values and of the totals in the columns showing the deviations and the accumulated deviations.

(2) The number of changes of sign of the deviations and the accumulated deviations.

(3) Comparison of each deviation with the standard error both age by age and by suitable age-groups.

(4) Comparison of each deviation with its approximate probable error.

(5) Comparison of the total of the deviations, irrespective of sign, with four-fifths of the total of the standard errors.

Although for most practical purposes these tests are sufficient they are subject to the following limitations:

(1) The smallness of the deviations and accumulated deviations is necessary for a good graduation; each of them may, however, be very small indeed without giving any indication that the adherence to data is too good. The test can give evidence of distortion of the rough data but it gives no evidence of undergraduation.

(2) It is difficult to say how many changes of sign are to be expected or how many are to be regarded as satisfactory. The problem often arises of how far the number of changes can differ from the number of non-changes before we may regard the graduation as suspect.

(3) The deviations should bear reasonable ratios to the standard errors, although here again we have no evidence of undergraduation however small the ratios may be.

(4) In theory this test reveals undergraduation as well as overgraduation (i.e. distortion), but it is rather insensitive.

(5) Evidence of undergraduation is obtained rather more reliably than by (4), but the mere addition of deviations irrespective

of sign is not really satisfactory. What is required is some means of finding a combined probability that the observed deviations would arise from random sampling.

Two such tests are discussed in Seal's paper and one of them, known as the χ^2 test, is of very general application outside actuarial work. The theoretical work involved in the χ^2 test is rather difficult and beyond the scope of this book, but it is hoped that the following introductory notes will enable the student to read one of the standard textbooks on the subject more profitably after obtaining a preliminary grasp of the fundamental principles involved.

17. The χ^2 test.

To combine deviations by straightforward addition is clearly unsatisfactory since positive and negative items will tend to cancel out. Test (5) discussed above avoids this difficulty by ignoring signs. Another obvious way is to square the deviations before adding. The student will by now be familiar with the idea of measuring discrepancies such as deviations in terms of their standard errors, as was done for instance in considering regression lines. It would seem logical therefore to divide each deviation by its standard error before squaring and adding, thus arriving at a function

$$\Sigma\frac{(\theta_x-E_xq_x)^2}{E_xp_xq_x}$$

or more generally

$$\Sigma\left(\frac{\text{actual value}-\text{expected value}}{\text{standard error}}\right)^2.$$

This function is known as χ^2.

Clearly χ^2 will be small if the graduation adheres closely to the data and large if the deviations are large. By means of prepared tables it is possible to find the probability that the χ^2 actually found, or one even greater, is likely to arise from simple sampling. If this probability is small, i.e. if a value of χ^2 as great as or greater than the one observed is unlikely to occur, the graduation has departed too far from the data. If the probability is large we know that a greater value of χ^2 was very likely to occur, so that the small value actually obtained was probably due to causes other than sampling errors. The graduation has then adhered too closely to the rough data.

The prepared tables are very extensive, since one is required for each *degree of freedom*. The exact meaning of this term is explained in statistical textbooks, but the following remarks, though not strictly accurate in detail, may help the reader to grasp the underlying ideas. Suppose that we have any body of data split into *cells*. For instance, the A 1924–29 data could be split into eight cells according to the following classifications: (i) with or without profits, (ii) medical or non-medical, (iii) whole-life or endowment assurance, or again the combined data could be regarded as split into cells, one for each age, or one for each quinquennial group of ages. The number in each cell is known as the *cell frequency* and χ^2 is defined as

$$\Sigma \frac{(\text{actual cell frequency} - \text{expected cell frequency})^2}{(\text{standard error of cell frequency})^2},$$

the summation extending over all the cells.

The cell frequencies are rarely independent of each other: in particular the method of graduation involved often ensures for a mortality table that the total number of actual deaths and the total number of expected deaths shall be equal. Any relationship between cell frequencies is known as a *constraint* or, if the relationship is of the first degree, as a *linear constraint*.

If there are n cells and k linear constraints the function $f=n-k$ is called the *number of degrees of freedom*.

Suppose that we took a great many samples of the same size with the same values of n and k, and that we calculated χ^2 for each. The resulting values could be grouped into a frequency distribution, namely the sampling distribution of χ^2. It should be noted that in applying the χ^2 test the whole of the data are regarded as one sample, whereas previously we have regarded the data at each age as a sample.

It can be shown that the sampling distribution of χ^2 follows the curve

$$y=y_0 x^{\frac{1}{2}f-1} e^{-\frac{1}{2}x}, \qquad \ldots\ldots(1)$$

where f is the number of degrees of freedom.

The range of x is 0 to ∞, so that the total area is

$$y_0 \int_0^\infty x^{\frac{1}{2}f-1} e^{-\frac{1}{2}x}\, dx.$$

Substituting $\frac{1}{2}x=t$, this becomes

$$y_0 2^{\frac{1}{2}f}\int_0^\infty t^{\frac{1}{2}f-1}e^{-t}\,dt=y_0 2^{\frac{1}{2}f}\Gamma(\tfrac{1}{2}f).$$

(*Mathematics for Actuarial Students*, Part I, p. 148.)

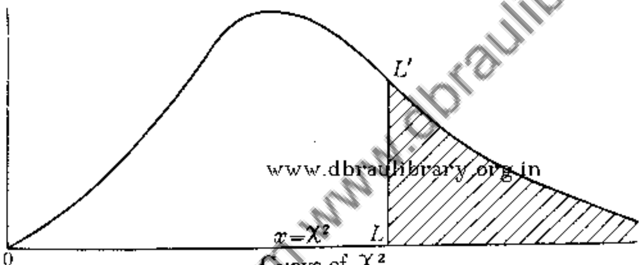
To make the total area unity we put

$$y_0=\frac{1}{2^{\frac{1}{2}f}\Gamma(\frac{1}{2}f)},$$

and write the equation to the curve of χ^2:

$$y=\frac{1}{2^{\frac{1}{2}f}\Gamma(\frac{1}{2}f)}x^{\frac{1}{2}f-1}e^{-\frac{1}{2}x},$$

where x represents χ^2.



Curve of χ^2

The above figure shows roughly the form of the frequency curve if $f>2$. It rises to a maximum where x (i.e. χ^2) $=f-2$ and falls away more gradually to the right as x increases. If f is large the curve is roughly symmetrical and it can be shown that $\sqrt{2\chi^2}$ follows approximately a normal curve with mean $\sqrt{2f-1}$ and unit standard deviation. Consequently for large values of f the tables prepared for the normal curve can be used if a new variable x' is taken, so that

$$\sqrt{2\chi^2}=\sqrt{2x}=x'.$$

The values given in standard tables for degrees of freedom 40 and over have usually been obtained from the tables for the normal curve by means of this approximate method.

If $OL=x_0$ the shaded area to the right of the ordinate LL' is equal to

$$\frac{1}{2^{\frac{1}{2}f}\Gamma(\frac{1}{2}f)}\int_{x_0}^\infty x^{\frac{1}{2}f-1}e^{-\frac{1}{2}x}\,dx=P,\text{ say.} \qquad \ldots\ldots(2)$$

This expression for P, although somewhat involved, can be evaluated and extensive sets of tables have been prepared. They are more complicated than those based on the normal curve, since each value of f has to be treated separately.

There are two main sets of tables:

(1) Tables giving the value of P for certain values of x_0. In *Tables for Statisticians and Biometricians*, Part I, these tables are given for each value of f from 2 to 29 and for higher values of f at somewhat greater intervals.

(2) Tables giving the values of x_0 for certain useful values of P. Table II in the Appendix is typical. It is reproduced from H. L. Seal's paper, "Tests of a Mortality Table Graduation", by kind permission of the author.

Since the total area is unity, P is clearly the probability that a value of x (i.e. a value of χ^2) obtained from a random sample will be greater than x_0. Having calculated χ^2 for a sample, we can use tables in the first form to obtain the probability that a random sample would give rise to a value of χ^2 as great as or greater than the one obtained.

Tables in the second form start with probability levels (see p. 101) and enable the values of χ^2 corresponding to them to be obtained.

Thus from Table II in the Appendix we find that for a sample with 40 degrees of freedom the value of χ^2 corresponding to the probability level ·01 is 62·88. If in testing a graduation we grouped the data so as to give 40 degrees of freedom and obtained a value of χ^2 of, say, 65 we should infer that the data had been distorted because the chance of so high a value as 65 arising from sampling errors is less than ·01. If on the other hand χ^2 lay between, say, 28 and 50 we should be satisfied with the graduation as regards fidelity to data since the probability P would be more reasonable, i.e. nearer $\frac{1}{2}$.

18. Alternative hypotheses as to the graduated rates.

In applying the χ^2 test it is important to distinguish between a hypothesis being tested by the data of a sample but based on other observations, and a hypothesis based to some extent on the sample by which it is being tested.

For instance, we might collect data showing the degrees of immunization against colds produced by a certain treatment for men of different age-groups in England. If the resulting figures, after any necessary graduation, were applied to a sample of Scotsmen we could compare the actual and expected results and calculate χ^2. If the figures were applied instead to the data on which they were based we should naturally expect a better fit and a smaller value of χ^2. It is for this reason that, in such cases, a deduction is made from the number of cells in finding the degrees of freedom.

A similar point arises in testing a mortality table graduation. We might assume that the graduated rates were the universe rates and would remain the same if we tested thousands of samples of the same size as the given data. We can imagine a value of χ^2 to be found from each sample and the results tabulated.

A second hypothesis would be that a separate curve was fitted to each of the imaginary samples before the actual and expected results were compared and the values of χ^2 calculated. On this assumption a much better agreement would be obtained and the values of χ^2 should be much smaller, corresponding to what we should expect with fewer degrees of freedom.

In actual practice we have only one sample and one set of graduated rates. Which hypothesis are we to adopt? Are we to assume that the expected deaths are fixed unalterable values to be compared with the actual deaths in a great many samples of similar constitution, or are we to regard them as applying only to the data on which they are based?

The practice in the past has been to adopt the first hypothesis, but modern statisticians, led by Prof. R. A. Fisher, favour the second and make an appropriate deduction in finding the number of degrees of freedom.

The point will be referred to later.

19. Example 1.

The data of Table VI may be set out as on p. 134, after grouping where necessary to ensure an expected class frequency of about 10 as the minimum.

The method of graduation did not impose any constraints, although the approximate equalising of the totals of actual and expected deaths suggests that one degree of freedom should be deducted so that the

Age (1)	Exposed to risk (2)	Assumed $q_x \times 10^4$ (3)	Expected frequency (4)	Actual frequency (5)	(Actual minus expected)^2 (6)	(Standard error)^2 $E_x p_x q_x$ (7)	$\frac{(6)}{(7)}$ (8)
47–50	814	146	11·9	14	4·41	11·7	·38
51–53	925	152	14·1	13	1·21	13·8	·09
54–55	820	159	13·0	12	1·00	12·8	·08
56–57	1,052	169	17·8	17	·64	17·5	·04
58	628	180	11·3	11	·09	11·1	·01
59	701	190	13·3	14	·49	13·1	·04
60	813	202	16·4	18	2·56	16·1	·16
61	917	217	19·9	18	3·61	19·5	·18
62	1,040	235	24·4	24	·16	23·9	·01
63	1,182	256	30·3	30	·09	29·5	·00
64	1,299	281	36·5	43	42·25	35·5	1·19
65	1,432	309	44·2	41	10·24	42·9	·24
66	1,596	340	54·3	54	·09	52·4	·00
67	1,752	374	65·5	64	2·25	63·1	·03
Total	14,971	—	372·8	373	—	—	$\chi^2 = 2·45$

number of degrees of freedom is the same as the number of cells, i.e. 14. Reference to tables shows that the probability of obtaining a value of χ^2 greater than 2·45 is ·99. Hence the low value obtained for χ^2 is less than we should expect from sampling errors, thus confirming our previous conclusion that the table was undergraduated.

Example 2.

The following example is taken from Seal's paper in *J.I.A.* Vol. LXXI *supra* and is based on data for ages 46½ and 51½ included in the A 1924–29 mortality investigation.

Class of policy	Age 46½				
	E	θ	Eq	$\theta - Eq$	$\frac{(\theta - Eq)^2}{Epq}$
Whole life					
With profits	32,768¼	221	189·01	31·99	5·446
Without profits	9,307¾	49	53·69	− 4·69	·412
Endowment assurance					
With profits	175,736	997	1013·68	−16·68	·276
Without profits	30,964¼	168	178·61	−10·61	·634
Total	248,776¼	1435	1434·99	—	6·768

Class of policy	Age 51½				
	E	θ	Eq	$\theta-Eq$	$\frac{(\theta-Eq)^2}{Epq}$
Whole life					
With profits	41,444¼	404	347·86	56·14	9·137
Without profits	9,814¾	87	82·38	4·62	·261
Endowment assurance					
With profits	137,610¼	1124	1155·03	−31·03	·841
Without profits	27,607½	202	231·73	−29·73	3·847
Total	216,476¾	1817	1817·00	—	14·086

Working on the hypothesis that the differences in the mortality in the various classes may be due merely to sampling errors, the rates of mortality are first calculated as follows:

$$q_{46\frac{1}{2}}=\frac{\text{Total deaths}}{\text{Total exposed}}=\frac{1435}{248,776\frac{1}{4}}=\cdot 0057682,$$

$$q_{51\frac{1}{2}}=\frac{1817}{216,476\frac{3}{4}}=\cdot 0083935.$$

The expected frequencies in each cell (Eq) are then calculated and χ^2 is found in the usual way, the results being 6·768 and 14·086 respectively. Owing to the way q was found the total expected deaths equal the total actual deaths, so that a linear constraint is introduced and the number of degrees of freedom is not four (the number of cells) but three. Entering the table for three degrees of freedom we find that the probability of obtaining a value of χ^2 as great as 6·768 or greater is about ·05, while for $\chi^2 \geqslant 14\cdot086$ the probability is about ·003.

For age 46½ the probability ·05 is small but not really significant. The value of ·003 for age 51½ is, however, so small that the value of χ^2 obtained cannot be explained as being due to sampling errors. It seems therefore that the four classes of policy differ to a significant extent. Why this result was not shown at age 46½ is difficult to explain. This question is discussed in the original paper, to which reference should be made.

20. For convenience we have considered rates of mortality only. The tests described apply to other functions of the same type, e.g. proportionate frequencies or probabilities, and can easily be adapted for any functions which can be expressed approximately in probability form.

In considering other functions, such as sickness rates, we can proceed in the normal way as far as obtaining the deviations and accumulated deviations and examining the results for changes of sign. Statistical tests for measuring the size of the deviations are, however, beyond the scope of this book.

21. Limitation of statistical tests for adherence to data.

The tests described for estimating the goodness or otherwise of the graduation of a proportionate frequency should not be applied too automatically nor should the results be interpreted too rigidly.

The use of 2σ or 3σ as a significance test is based on the assumption that the sampling distribution is normal. This does not rest upon rigid theory when applied to a proportionate class frequency for which the deviations usually follow a skew curve representing the binomial expansion $(p+q)^n$. The error involved should not be great if nq is large.

The use of $\cdot 8\sigma$ as a measure of the mean deviation is open to the same objection. Moreover, although with this approximate measure we can compare the total deviations irrespective of sign, we have no satisfactory criterion by which to interpret the difference. For instance, if the total irrespective of sign is 54·6, while $\cdot 8\Sigma\sqrt{npq}$ is 46·8, we cannot readily decide whether the difference of 7·8 is too great to be considered as attributable to sampling errors. Although an Italian actuary discovered the sampling distribution of the mean deviation in 1937 it is too complicated to be considered here.

22. Comparison of two or more graduations.

A method sometimes used for comparing the results of two or more graduations is to calculate for each graduation the sum of the squares of the deviations, a small sum indicating good adherence to data.

This test is not very satisfactory. A comparison of the values of χ^2, which takes into account the standard error at each age, is to be preferred, and this method will probably be used more extensively in future.

It should be remembered, however, that the graduation with the smallest χ^2 is not necessarily the best from a practical point of view. A simple graduation, easy to apply and having practical advantages (*e.g.* a Makeham graduation if joint-life functions are required), will usually be the best provided the value of χ^2 produced is satisfactory and the smoothness adequate.

BIBLIOGRAPHY

An Introduction to the Theory of Statistics. G. UDNY YULE and M. G. KENDALL. London, 1948.

"Tests of a Mortality Table Graduation." H. L. SEAL. *J.I.A.* Vol. LXXI.

Frequency Curves and Correlation. Sir WILLIAM P. ELDERTON. London, 1938.

"On the interpretation of χ^2 from contingency tables and the calculation of *P*." R. A. FISHER. *J. Roy. Stat. Soc.* Vol. LXXXV (1922), p. 87.

Tables for Statisticians and Biometricians, Part I. Ed. by K. PEARSON. London, 1914.

"Distribution of groups in a sequence of alternatives." W. L. STEVENS. *Ann. Eugen., Lond.*, Vol. IX (1939), p. 10.

EXAMPLES 5

1. The rates of mortality observed among a body of lives have been graduated in three ways. The following table shows in quinary age-groups the actual numbers of deaths experienced and the expected numbers according to each graduation:

Age-group	Actual deaths	Expected deaths by		
		Graduation A	Graduation B	Graduation C
20–24	15	10	12	14
25–29	28	21	22	26
30–34	39	41	42	40
35–39	54	64	62	58
40–44	84	85	81	81
45–49	93	107	102	99
50–54	97	90	93	93
55–59	80	75	78	78
60–64	55	52	57	54
Total	545	545	549	543

Examine the graduations with regard to their fidelity to the experience (ignoring smoothness) and comment on their relative merits.

2. Apply the χ^2 test to the above data, assuming that the agreement between the actual deaths and those expected according to Graduation A is not fortuitous but that otherwise no constraints have been imposed.

3. A small Life Office has examined its mortality experience over a recent period of time. The total actual deaths number 471, whilst the total expected according to the A 1924–29 Table was 450.

In respect of the Without Profit business the actual deaths were 31 and the expected deaths 50.

It is accordingly suggested that the Without Profit business must attract a much better class of life than the With Profit business.

Criticize this suggestion and state carefully the assumptions underlying any tests which you might consider it necessary to make.

4. What is the justification for graduation of mortality statistics?

Apply the usual tests to the graduated rates of mortality given in the following table, which represents a section of a mortality experience. What special features do the graduated rates show?

Age x	Exposed to risk E_x	Deaths θ_x	$10^5 \times \theta_x / E_x$ $= 10^5 \times q_x$	Graduated $10^5 \times q_x$
50	2000	10	500	616
51	2000	14	700	658
52	1900	6	316	704
53	1800	17	944	753
54	1700	9	529	805
55	1400	7	500	861
56	1300	5	385	921
57	1300	12	923	985
58	1300	18	1385	1055
59	1250	24	1920	1138
60	1050	8	762	1241
61	950	15	1579	1371
62	950	12	1263	1534
63	950	20	2105	1736
64	900	15	1667	1980
65	850	18	2118	2267
66	800	27	3375	2597
67	800	26	3250	2965
68	750	34	4533	3361
69	750	24	3200	3774
70	700	35	5000	4196
71	700	25	3571	4623
72	650	37	5692	5065
73	650	25	3846	5538
74	600	41	6833	6058
75	550	29	5273	6628

5. Criticize the following graduation from the point of view of fidelity to the data:

Age-group	Exposed to risk	Actual deaths	Expected deaths
40–44	15,518	65	73·9
45–49	19,428	144	134·6
50–54	21,594	219	223·9
55–59	21,890	378	346·3
60–64	19,174	465	468·1
65–69	15,775	557	600·2
70–74	11,414	685	675·5
75–79	6,993	644	637·4
80–84	3,276	471	458·7
85–89	1,096	217	240·6
90–94	201	67	61·4

6. Criticize the following section of a graduation from the point of view of smoothness and test the adherence to the data:

Age	Exposed to risk	Actual deaths	Expected deaths	Graduated rate of mortality
71	8292	401	423·7	·0511
72	8156	432	461·0	·0565
73	7905	518	497·3	·0629
74	7588	532	528·8	·0697
75	7214	592	556·1	·0771
76	6713	567	566·1	·0843
77	6205	548	571·5	·0921
78	5704	565	568·3	·0996
79	5196	562	563·7	·1085
80	4664	549	548·4	·1176
81	4142	558	532·2	·1285
82	3576	479	497·5	·1391
83	3070	481	467·0	·1521
84	2575	399	425·4	·1652
85	2146	404	388·4	·1810
86	1706	342	332·8	·1951

7. The following table shows the results of another attempt to graduate the data shown in Example 6. Calculate, for both sets of graduated values, any functions which enable (*a*) the smoothness and (*b*) the adherence to data of the two graduations to be compared.

Age	Graduated rate of mortality	Expected deaths	Age	Graduated rate of mortality	Expected deaths
71	·050	414·6	79	·108	561·2
72	·056	456·7	80	·118	550·3
73	·063	498·0	81	·129	534·3
74	·070	531·1	82	·140	500·6
75	·077	555·5	83	·152	466·6
76	·084	563·9	84	·166	427·5
77	·091	564·7	85	·183	392·7
78	·099	564·7	86	·200	341·2

8. The following table shows part of the Manchester Unity sickness experience 1893–97 for Occupation Groups A, H, J.

Age	No. of years of life exposed to risk of sickness	No. of sickness claims. First two years of sickness	Graduated proportion of members sick. First two years of sickness	No. of weeks of sickness claim. First three months of sickness	Graduated weeks of sickness per member per annum. First three months of sickness
	(1)	(2)	(3)	(4)	(5)
30	70,903	15,051	·213	45,742	·649
31	69,389·5	14,866	·213	46,185	·656
32	67,639·5	14,507	·214	45,034	·663
33	66,397·5	14,212	·214	44,661	·670
34	64,043·5	13,629	·213	42,848	·677
35	61,632·5	13,108	·213	42,432	·685
36	59,473	12,701	·214	41,436	·696
37	57,380·5	12,406	·215	40,518	·709
38	54,741·5	11,777	·217	39,411	·726
39	52,911	11,642	·220	39,767	·744
40	51,478	11,434	·222	39,323	·764
41	49,835·5	11,378	·226	39,613	·784
42	48,199	10,929	·227	38,737	·803
43	46,818·5	10,737	·230	38,413	·821
44	45,418	10,563	·232	38,105	·839
45	43,483	10,168	·235	37,115	·858
46	41,654·5	9,828	·237	36,675	·879
47	40,328·5	9,749	·241	36,521	·904
48	39,107	9,481	·244	35,735	·932
49	37,723·5	9,441	·249	36,304	·962
50	36,510	9,194	·252	37,251	·994
51	35,237·5	9,082	·257	35,873	1·027
52	33,876·5	8,827	·260	36,023	1·061
53	32,727·5	8,673	·266	36,218	1·096
54	31,190·5	8,362	·271	34,181	1·134
55	29,664·5	8,327	·278	35,306	1·175
56	27,969·5	7,949	·284	34,385	1·218
57	26,461·5	7,726	·292	33,306	1·262
58	24,377	7,314	·300	32,169	1·308
59	22,872	6,985	·308	30,178	1·356
60	21,318	6,756	·315	30,175	1·409

The proportion of members sick, first two years of sickness, is obtained by dividing the figures in column (2) by the corresponding figures in column (1). Column (3) shows the results of graduating the proportions thus obtained.

The number of weeks of sickness per member per annum, first three months of sickness, is obtained by dividing the figures in column (4) by the corresponding figures in column (1). Column (5) shows the results of graduating the results.

Criticize the two graduations and point out why tests applicable to one graduation are not applicable to the other.

CHAPTER VI

THE GRAPHIC METHOD

1. The student of physics will be familiar with plotting on squared paper the results of experiments and subsequently drawing smooth curves through or near to the points in an effort to arrive at some underlying law connecting the variables.

The graphic method of graduation as used by actuaries is merely a more refined version of this process. It can conveniently be considered in three main stages.

1. The rough data are grouped.
2. The grouped data are represented graphically in some form or another and a smooth curve is drawn reproducing the general trend of the data, but not adhering too closely to local fluctuations.
3. Values are read off from the smooth curve and subsequently adjusted and readjusted until they satisfy the two requirements of smoothness and adherence to data. This last process is usually referred to as hand-polishing.

There are two main ways in which the method can be used: it can be applied to the exposed to risk and decrements separately or the crude rates of decrement can be found from the data and subsequently graduated.

2. Separate graduation of exposed to risk and decrements.

As broad groupings may be necessary an obviously satisfactory method of representing the data graphically is by a histogram. It should be remembered that the rectangles of the histogram must have bases proportionate to the ranges of the separate groups, which will not always be equal. It is rather difficult to decide on the preliminary grouping, but as a general guide it may be said that sufficiently wide or "coarse" groups should be adopted to ensure a fairly regular outline for the histogram. The drawing of the smooth curve is a very difficult matter and unless the grouping is satisfactory it may become almost impossible.

In drawing the smooth curve care should be taken that in each of the vertical strips of the original diagram the area of the rectangle of the histogram is roughly equal to the area bounded by the curve and the ordinates forming the sides of the rectangle. The areas should not be exactly equal, as the process would then be merely a graphical interpolation and not a graduation at all. We know in fact from the previous chapters that the graduated values should differ from the ungraduated by amounts equal to reasonable sampling errors. The difficulty, therefore, of ensuring that the smooth curve departs sufficiently but not too much from the data as represented by the histogram is one of the chief drawbacks of this method.

In reading off the graduated values from the smooth curve thus drawn it is desirable to tabulate them either for every age or, if this is impossible, for small groups of ages which do not correspond with the groupings adopted for the histogram. This is necessary because in testing for adherence to data the histogram groupings must not be used. The way in which the curve was drawn ensures reasonable agreement within those groups and we need to know whether the agreement is satisfactory over all ranges.

The tests for smoothness of the graduated rates call for no special comment, and the student should remember to make an inspection of the deviations and accumulated deviation, preferably for individual ages.

As mentioned in the previous chapter it is desirable to recalculate the decrements after the exposed to risk figures have been graduated so as to retain the same rates of decrement as before. The adjusted decrements are then graduated in the same way.

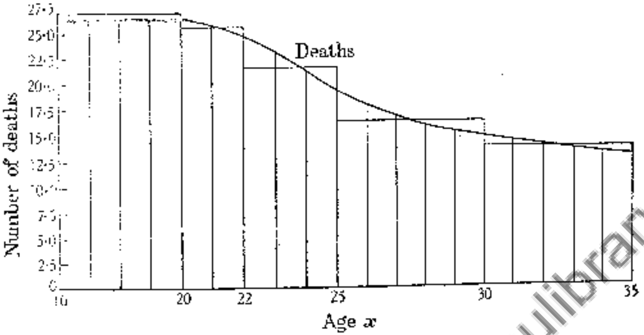
The following example will serve to illustrate the method:

Example 1.

Graduate by a graphic process the deaths and marriages in the following table:

Recorded age x (exact)	Decrements between x and next recorded age shown in the first column	
	Deaths	Marriages
16	106	979
20	50	1281
22	64	2069
25	81	2033
30	68	756
35	—	—

It will be assumed that the exposed to risk figures have already been graduated and the decrements re-calculated as mentioned above. As the data are so scanty no further grouping seems possible or desirable. It should be noted that the ranges of the groups are unequal.



In drawing the histogram the heights of the rectangles are obtained as follows:

No. of deaths between ages 16 and 20 = 106 = area of rectangle.
Base of rectangle (range of group) = 4.
∴ Height of rectangle $= \frac{106}{4} = 26·5$.
Similarly, height of second rectangle $= \frac{50}{2} = 25$, and so on.

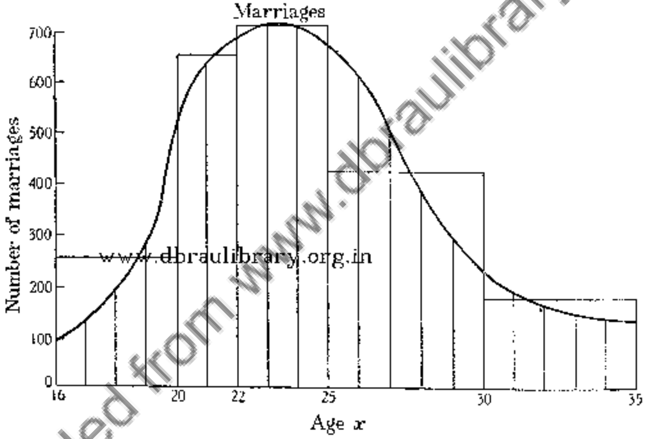
The following values of the deaths at individual ages can then be derived from the curve (i.e. the areas of unit strips). It is impossible to apply a satisfactory test for adherence to data in view of the way in which the figures are given.

Age x	16	17	18	19	20	21	22
Deaths between age x and $x+1$	26·7	26·5	26·3	26·0	25·5	24·7	23·5
Age x	23	24	25	26	27	28	29
Deaths between age x and $x+1$	21·8	19·8	18·2	16·9	15·9	15·3	14·8
Age x	30	31	32	33	34		
Deaths between age x and $x+1$	14·4	14·0	13·6	13·3	13·0		

A rough comparison in the given groups is as follows:

Ages	Actual	Graduated
16–20	106	105·5
20–22	50	50·2
22–25	64	65·1
25–30	81	81·1
30–35	68	68·3
Total	369	370·2

The marriages can be dealt with similarly. In this case it will be noticed that the histogram has quite a different shape, thus:



The values derived from the curve are as follows:

Age x	16	17	18	19	20	21	22
Marriages between age x and $x+1$	140	165	250	430	595	665	690
Age x	23	24	25	26	27	28	29
Marriages between age x and $x+1$	700	675	600	500	390	305	240
Age x	30	31	32	33	34		
Marriages between age x and $x+1$	195	165	140	130	125		

It is almost impossible to test for smoothness and the comparison in the original groups is not a satisfactory test of adherence to data. The figures are:

Ages	Actual	Graduated
16–20	979	985
20–22	1281	1260
22–25	2069	2065
25–30	2033	2035
30–35	756	755
Total	7118	7100

Because of the practical difficulties of replacing a histogram by a smooth curve it is often easier to deal with an ogive curve and represent the data by points.

Thus the data in the above example might be written in the form:

Age x	Decrements occurring below age x	
	Deaths	Marriages
16	—	—
20	106	979
22	156	2260
25	220	4329
30	301	6362
35	369	7118

In this form the data can be represented by points and graduated by the method explained in the next few paragraphs. The objection to the use of an ogive in this way is that the numbers tend to mount up very rapidly; this renders it difficult to find a suitable scale. The graduated values of the decrements are found by differencing.

3. The Carlisle Table.

Before we leave the histogram method of graduating by the graphic process it is appropriate to mention some of the principal features of the Carlisle Table. This was the first standard table constructed on sound lines.

Previously the Northampton Table had been constructed by a Dr Price, who investigated the registers of the four parishes of Northampton. He concentrated on the parish of All Saints, where the records were most complete, and ignoring the exposed to risk based a table on the recorded deaths during the years 1735 to 1780. In effect, as he thought that the population had been sufficiently stable for a long time, he took the average deaths as the d_x column of a life table. The mortality, especially at the younger ages, was overstated, but the table was used for many years for calculating rates of annuity granted by the National Debt Office.

Joshua Milne based his table on statistics for the town of Carlisle. His data related to the parishes of St Mary and St Cuthbert and consisted of an enumeration of the population in January 1780 and December 1787 and particulars of the deaths in the years 1779 to 1787 inclusive.

The exposed to risk and deaths were graduated by the method described in the previous paragraph and the resulting table was a great improvement on any previously available. The sexes were not dealt with separately and the large female population included caused the mortality to be unusually light. As mortality has steadily improved since Milne's time this light mortality of the Carlisle Table prevented it from becoming out of date so soon as it would otherwise have done.

The extensive sets of joint-life tables published by Milne are sometimes used for unusual rates of interest, although some adjustment of the ages is necessary.

4. Rates of decrement graduated graphically.

When the rates of decrement are calculated from the rough data they can be represented graphically by points and the graduating curve merely has to pass near to these points. Such a curve is consequently much easier to draw than a curve which seeks to keep areas substantially unaltered. Moreover, a quotient such as a rate of decrement tends to progress more smoothly than a function such as the exposed to risk. In the graduation of rates of mortality, other standard tables may be available giving an indication of the general trend of q_x. This is particularly useful at the ends of the table, where the data are bound to be so scanty that the rates brought out are unreliable, and the curve has to be sketched in the light of previous knowledge of other tables.

5. Preliminary grouping.

This is usually a difficult matter and calls for considerable experience and skill. As will be seen in Example 2 different groupings may result in widely differing results. The following remarks are, however, of general application. Some authorities, such as G. F. Hardy, maintain that grouping with a fixed class-interval (e.g. quinquennial groups) gives the best results. The general opinion seems to be however that it is preferable to select groups in such a way that the group rates progress smoothly. This almost invariably involves the use of groups of unequal range; these can be found approximately as follows.

(i) Plot the rough rates of decrement, without grouping, as a series of points and sketch in lightly a smooth curve representing the general run of the values. This may prove difficult, but refinement is unnecessary as a moderate degree of smoothness is quite sufficient.

(ii) Now choose groupings in such a way that points above this guide curve are balanced by points below it; i.e. so that a point some distance above may be offset by grouping it with the next two, or even three, points lying slightly below the curve. The aim should be to ensure that the group rate will lie close to the general run of the values. Generally speaking, if the curvature seems to be changing rapidly, groups should be of short range; if the curve rises to a peak and falls again care should be taken to ensure that this peak is not cut off by the method of grouping adopted. When the ranges have been decided upon the guide curve should be erased.

6. Plotting the data and sketching the curve.

For each group the total deaths (or decrements) are divided by the total exposed to risk and the resulting rate is plotted as corresponding to the middle age of the range. As we have seen in Chapter I, para. 12, this is not strictly accurate, and formula (18) of that chapter should be used to find the deaths and exposed to risk corresponding to the central age of each group.

On the other hand, such a refinement is not justifiable unless the fourth and higher differences are negligible, and the slight systematic distortion involved when it is omitted is automatically corrected in the process of hand-polishing.

The points representing the group rates should progress more regularly than those representing the original data and it should not be difficult to draw a smooth curve passing close to them. The importance of this step should not be underestimated, because, although it may be said that the final graduation takes place in the hand-polishing, this latter process is at the best of times a laborious business and may well become very lengthy unless a good curve has produced a set of values which are fairly satisfactory before any further adjustment is carried out.

If it is found that an appreciable section of the table is unsatisfactory by, say, overstating the mortality, it will probably be quicker in the long run to re-draw that part of the curve and deduce fresh values rather than to attempt to adjust the discrepancies by hand-polishing.

Inexperienced operators usually tend to adhere to the data too closely in sketching their curves and, as a result, the rates are undergraduated.

The graduation discussed in Chapter V, Table VI, etc. was obtained by the graphic process and it will be remembered that there was evidence of undergraduation.

7. Hand-polishing.

The rates of decrement should be read from the curve for individual ages, and the first two or three orders of differences should be tabulated. Scrutiny of these reveals not only where the smoothness is unsatisfactory but also where points of inflexion occur. If the first differences are positive, a positive second difference shows that the gradient is increasing, while a negative second difference shows that the curve is becoming flatter. Consequently, a change in the sign of the second difference indicates a point of inflexion and this still applies if all the first differences are negative. The reader should satisfy himself as to the truth of this by drawing a few curves. Although points of inflexion are not unknown in mortality curves, particularly between the ages of 15 and 35, they always require investigation and they can often be eliminated by a bolder drawing of the curve; this will also improve the graduation generally. If data are grouped, divided differences can be used for testing for points of inflexion.

For testing adherence to data the processes previously discussed should be applied and, though grouping is usually necessary because of scanty data, care should be taken to avoid the groupings adopted in drawing the curve.

In the light of both these sets of tests, i.e. for smoothness and for adherence to data, the rates are adjusted and readjusted until a satisfactory table is obtained. If, for instance, there is a run of positive deviations at one section greater than could be accounted for by sampling errors the curve would appear to be too low and might with advantage be re-drawn. In practice the operator tends to concentrate in the first place on a good progression of second differences and devotes too little attention to the matter of adherence to data.

8. Advantages of the method.

Although not generally used for standard tables and therefore seldom discussed in actuarial literature the graphic method is more widely used than any other, with the possible exception of the method described in the next chapter.

It is extremely adaptable and can be used for almost any function. Above all, it can give good results when the data are so scanty that other methods would be out of the question. This is its supreme advantage.

It is commonly used in connection with pension funds and friendly societies for functions such as rates of withdrawal or retirement. In these cases each society is a law unto itself and it is usually impossible to use a standard table without considerable adjustment.

The graphic method allows great scope for individual judgment, based very often on wide experience, and in this connection it should be pointed out that the ends of the table, which always cause difficulty because of scanty data, can usually be dealt with satisfactorily by sketching those portions of the curve in the light of knowledge gained from other tables of a similar type.

There is no reason why the graduation should not be adequate, since the hand-polishing is assumed to continue until the criteria for both smoothness and adherence to data have been satisfied.

9. Disadvantages of the method.

In actual practice the method is much more difficult to apply than it would seem to be and demands considerable skill and patience from the operator.

It is unsuitable for standard tables based on extensive data, since a very high degree of smoothness is difficult to achieve. It is usually impossible to obtain sufficient places of decimals in the graduated rates because of the difficulty of reading more than three figures from a graph.

In this connection it should be pointed out that, owing to difficulties of scale, it is usually necessary in graduating rates of mortality to draw the curve in two parts, namely, one curve up to age 65 or 70 and another from age 60 or 65 to the end of life. When this is done it is desirable to make the curves overlap so as to ensure continuity.

In Example 2 it will be noticed that the two curves used did not overlap but that the comparatively smooth progression of the ungraduated rates ensured a smooth junction. Nevertheless it is unwise to rely on such an uncertain result.

By leaving scope for individual judgment the method also leaves scope for individual bias and prejudice, and it must be confessed that by means of the graphic method equally eminent and experienced workers might obtain widely differing results from the same data.

Although a graphic process can be used to graduate the ultimate portion of a mortality table it is not very satisfactory for dealing with select data. These are usually so scanty that it is impossible to form an idea of the trend until they have been grouped in quinquennial or decennial groups of ages.

If the group rates are plotted on the same sheet as the curve representing the graduated ultimate rates, it is often possible to form an idea of how the select rates run into the ultimate rates, although the select rates themselves may be largely a matter of speculation.

10. Application of the χ^2 test.

It was pointed out in Chapter VI, para. 18, that two hypotheses were possible in applying the χ^2 test. If in testing a graphic graduation we adopt the first, we need make no deduction in finding the number of degrees of freedom, which in this event would be the number of groups.

The second hypothesis involves the fact that the curve was to some extent "forced" to fit the rough data; the necessary deduction that should be made on this account from the total number of groups is a difficult problem to decide and there may well be different opinions on the subject. A full discussion of this difficulty is outside the scope of this book.

Example 2. Female Government Annuitants Table.

The classic paper on the graphic method was given by Dr T. B. Sprague in 1886 (*J.I.A.* Vol. XXVI, and reproduced in *J.S.S.* Special Number on Graduation) and related to the graduation of the mortality of Female Government Annuitants four years and upwards after purchase.

The following results are taken from this paper.

The rough data are shown in Table VII.

As a first step the data up to age 51 were grouped in three ways, as shown in Tables VIII, IX, and X.

Comparison of the rates of mortality in these three tables illustrates very clearly the need for experience and sound judgment in deciding on the grouping to be adopted.

The grouping of Table VIII does not appear to be satisfactory since the rates produced are far from smooth. The rates in Tables IX and X show that either of the groupings adopted would be suitable. It will be seen, however, that the rates in these two tables suggest widely differing types of curve. From Table IX it might be inferred that the rates of mortality from ages 19 to 49 inclusive are approximately constant; in fact by amalgamating the data for the whole of this range this constant rate is found to be ·0116, which does not differ to a significant extent from the rates brought out for each of the five groups in the table.

An examination of the rates in Table X would seem to indicate that q_x exceeds ·0159 at age 19, falls to a minimum at an age in the group 30–35 and thereafter increases steadily.

In order to decide which of the sets of groupings was preferable Sprague investigated the data for each of the first four years' duration. As the features of each duration were very similar it will be sufficient to show the data for durations 0–3 combined (Table XI).

Table VII. *Data for Female Government Annuitants, 4 years and over after purchase*

Age	Exposed to risk	Deaths	Rate of mortality	Age	Exposed to risk	Deaths	Rate of mortality
19	9	—	—	61	6038	116	·019
20	14	—	—	62	6422	157	·024
21	18	—	—	63	6762	182	·027
22	23	—	—	64	7247	209	·029
23	31	—	—	65	7599	258	·034
24	37	—	—	66	7863	254	·032
25	54	2	·037	67	8061	317	·039
26	66	2	·030	68	8197	353	·043
27	85	1	·012	69	8307	343	·041
28	99	1	·010	70	8372	396	·047
29	130	3	·023	71	8292	401	·048
30	163	2	·012	72	8156	432	·053
31	196	—	—	73	7905	518	·066
32	223	2	·009	74	7588	532	·070
33	248	3	·012	75	7214	592	·082
34	280	3	·011	76	6713	567	·084
35	319	5	·016	77	6205	548	·088
36	360	5	·014	78	5704	565	·099
37	416	3	·007	79	5196	562	·108
38	473	6	·013	80	4664	549	·118
39	571	2	·004	81	4142	558	·135
40	653	13	·020	82	3576	479	·134
41	733	9	·012	83	3070	481	·157
42	833	12	·014	84	2575	399	·155
43	941	9	·010	85	2146	404	·188
44	1097	12	·011	86	1706	342	·200
45	1252	17	·014	87	1332	299	·224
46	1416	21	·015	88	1018	210	·206
47	1578	13	·008	89	788	208	·264
48	1761	14	·008	90	569	160	·281
49	1958	26	·013	91	391	115	·294
50	2160	22	·010	92	265	72	·272
51	2387	37	·016	93	189	63	·333
52	2669	43	·016	94	121	39	·322
53	2909	51	·018	95	78	31	·397
54	3330	59	·018	96	47	12	·255
55	3682	57	·015	97	34	18	·529
56	4104	83	·020	98	16	9	·563
57	4473	86	·019	99	7	4	·571
58	4885	90	·018	100	3	—	—
59	5281	123	·023	101	3	2	·667
60	5644	138	·024	102	1	—	—

Table VIII

Ages	Exposed to risk	Deaths	Rate of mortality
19–26	252	4	·0159
27–29	314	5	·0159
30–33	830	7	·0084
34, 35	599	8	·0133
36, 37	776	8	·0103
38	473	6	·0127
39, 40	1224	15	·0123
41, 42	1566	21	·0134
43, 44	2038	21	·0103
45–47	4246	51	·0120
48, 49	3719	40	·0108
50, 51	4547	59	·0130
19–51	20584	245	—

Table IX

Ages	Exposed to risk	Deaths	Rate of mortality
19–33	1396	16	·0115
34–37	1375	16	·0117
38–40	1697	21	·0124
41–44	3604	42	·0117
45–49	7965	91	·0114
50, 51	4547	59	·0130

Table X

Ages	Exposed to risk	Deaths	Rate of mortality
19–29	566	9	·0159
30–35	1429	15	·0105
36–38	1249	14	·0112
39–44	4828	57	·0118
45–51	12512	150	·0120

Table XI

Ages	Exposed to risk	Deaths	Rate of mortality
15–20	102·5	2	·0195
21–27	501·6	6	·0120
28–32	675·3	7	·0104
33–37	1356·0	10	·0074
38–42	2548·3	19	·0075
43–46	3179·9	35	·0110
47–49	3196·0	29	·0091
Total	11,559·6	108	—

The trend of these rates was not materially affected by other groupings tried and Sprague accordingly decided that the mortality did decrease up to say age 30. On this assumption the best of the three methods of grouping the ultimate data was the third, which he therefore adopted.

For ages 52 to 70 the rates progressed more regularly and the choice of groups was much easier. The following groupings were used:

Table XII

Ages	Exposed to risk	Deaths	Rate of mortality
52–55	12,590	210	·0167
56–58	13,462	259	·0192
59–61	16,963	377	·0222
62	6,422	157	·0245
63	6,762	182	·0269
64	7,247	209	·0288
65, 66	15,462	512	·0333
67	8,061	317	·0393
68	8,197	353	·0431
69, 70	16,679	739	·0443
Total	111,845	3315	—

Finally, with certain exceptions, the data over age 70 were used for individual ages; the data were amalgamated for the following ages: 81 and 82; 83 and 84; 88 and 89; 92 and 93; 95–97; and 98–102.

Because of the difficulty of finding a suitable scale for the whole table

Sprague used two curves. He used one curve for ages up to 70 and a second curve for ages 71 and upwards. Although, as stated on p. 152, it is usually desirable to construct two such curves so that there is a slight overlap, Sprague succeeded in effecting a smooth junction without this precaution.

The reader should plot the points representing the data, grouped as above, and should graduate them graphically, including the final hand-polishing.

In order to enable the student to criticize his own work Sprague's graduated rates are set out in Table XIII.

It is a useful exercise to analyse these results, testing the smoothness by differencing and also testing adherence to data by calculating the deviations, accumulated deviations, square root of the expected deaths and finally χ^2 by the methods described in the previous chapter.

Table XIII. *Sprague's graduated values of q_x*

Female Government Annuitants. 4 years and upwards after purchase

Age	q_x	Age	q_x	Age	q_x	Age	q_x
19	·0180	40	·0122	61	·0232	82	·1400
20	·0178	41	·0121	62	·0249	83	·1520
21	·0176	42	·0119	63	·0270	84	·1660
22	·0173	43	·0116	64	·0294	85	·1830
23	·0169	44	·0114	65	·0321	86	·2000
24	·0162	45	·0113	66	·0350	87	·2170
25	·0151	46	·0113	67	·0380	88	·2340
26	·0143	47	·0114	68	·0410	89	·2500
27	·0135	48	·0115	69	·0440	90	·2650
28	·0128	49	·0121	70	·0470	91	·2830
29	·0121	50	·0130	71	·0500	92	·3000
30	·0116	51	·0140	72	·0560	93	·3200
31	·0110	52	·0155	73	·0630	94	·3390
32	·0105	53	·0163	74	·0700	95	·3600
33	·0103	54	·0170	75	·0770	96	·3820
34	·0103	55	·0177	76	·0840	97	·4100
35	·0106	56	·0184	77	·0910	98	·4390
36	·0109	57	·0191	78	·0990	99	·4660
37	·0112	58	·0199	79	·1080	100	·5000
38	·0118	59	·0208	80	·1180	101	·5450
39	·0120	60	·0219	81	·1290	102	·6050

Example 3.

Table XIV. *Graduation of marriage rates*

Age	Exposed to risk of marriage	No. of marriages	Rate of marriage
20	1100	0	—
21	1100	6	·0055
22	1100	2	·0018
23	1100	13	·0118
24	1000	17	·0170
25	1000	18	·0180
26	900	36	·0400
27	800	34	·0425
28	700	49	·0700
29	700	49	·0700
30	600	48	·0800
31	600	45	·0750
32	500	40	·0800
33	500	44	·0880
34	400	35	·0875
35	400	30	·0750
36	400	25	·0625
37	300	21	·0700
38	300	15	·0500
39	300	21	·0700
40	300	15	·0500
Total	14100	563	—

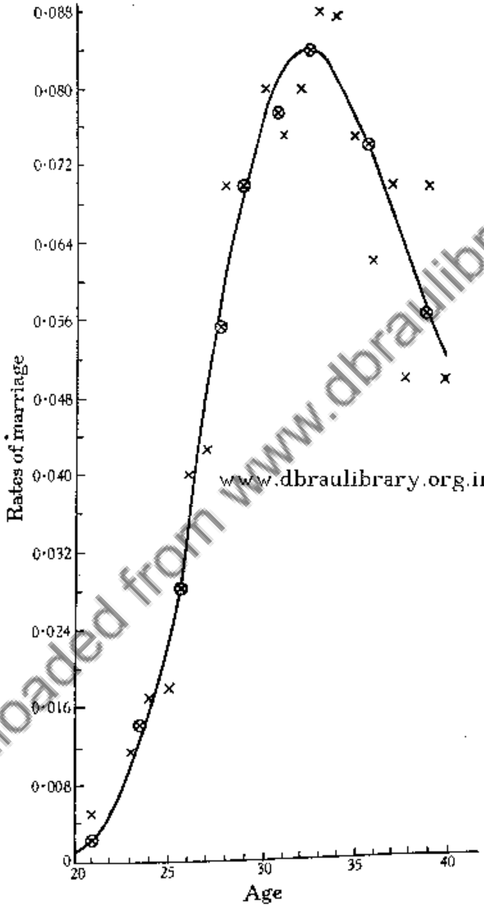
There appears to be no advantage in graduating the exposed to risk and marriages separately and we shall accordingly operate on the rate of marriage.

Suitable groupings can be arrived at only after repeated trials, but a rough sketch of the curve suggests that the groupings shown in the following table should give good results.

It should be borne in mind that these may not necessarily be the best groupings; there may be others which will give equally satisfactory results.

Age range	Central age of group	Exposed to risk	No. of marriages	Rate of marriage
20–22	21	3,300	8	·0024
23–24	$23\frac{1}{2}$	2,100	30	·0143
25–26	$25\frac{1}{2}$	1,900	54	·0284
27–28	$27\frac{1}{2}$	1,500	83	·0553
29	29	700	49	·0700
30–31	$30\frac{1}{2}$	1,200	93	·0775
32–33	$32\frac{1}{2}$	1,000	84	·0840
34–37	$35\frac{1}{2}$	1,500	111	·0740
38–40	39	900	51	·0567
Total	—	14,100	563	—

Sometimes, in a particular group, it may be desirable to use a weighted mean age instead of the central age if the numbers exposed to risk are very unevenly distributed. In this example such refinement was unnecessary and the rates shown in the last column were plotted as corresponding to the central ages shown in the second column and a curve was drawn passing near these points. The position of the maximum ordinate causes difficulty in this particular graduation.



The graduated values read from the curve are shown in the following table, only three decimal places being recorded (the fourth would be quite unreliable). The first three orders of differences are shown so that the smoothness can be examined. It should be noted that the two points of inflexion which from the general shape of the curve seem to be actual features of the experience to be retained in the final table.

Table XV

Age	Rate $(mq)_x$	$10^3 \times \Delta\,(mq)_x$	$10^3 \times \Delta^2\,(mq)_x$	$10^3 \times \Delta^3\,(mq)_x$
20	·001			
		+1		
21	·002		+2	
		+3		0
22	·005		+2	
		+5		−1
23	·010		+1	
		+6		+1
24	·016		+2	
		+8		+1
25	·024		+3	
		+11		+1
26	·035		+4	
		+15		−7
27	·050		−3	
		+12		0
28	·062		−3	
		+9		+1
29	·071		−2	
		+7		−1
30	·078		−3	
		+4		+1
31	·082		−2	
		+2		−1
32	·084		−3	
		−1		+1
33	·083		−2	
		−3		+1
34	·080		−1	
		−4		0
35	·076		−1	
		−5		+1
36	·071		0	
		−5		0
37	·066		0	
		−5		+1
38	·061		+1	
		−4		−1
39	·057		0	
		−4		
40	·053			

The following table is also needed for testing adherence to data:

Table XVI

Age	Graduated rate	Exposed to risk	Expected marriages	Actual marriages	Deviations (5) − (4)	Accumulated deviations	$\sqrt{(4)}$ approx.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
20	·001	1,100	1·1	0	− 1·1	− 1·1	1
21	·002	1,100	2·2	6	+ 3·8	+ 2·7	1
22	·005	1,100	5·5	2	− 3·5	− ·8	2
23	·010	1,100	11·0	13	+ 2·0	+ 1·2	3
24	·016	1,000	16·0	17	+ 1·0	+ 2·2	4
25	·024	1,000	24·0	18	− 6·0	− 3·8	5
26	·035	900	31·5	36	+ 4·5	+ ·7	6
27	·050	800	40·0	34	− 6·0	− 5·3	6
28	·062	700	43·4	49	+ 5·6	+ ·3	7
29	·071	700	49·9	49	− ·9	− ·6	7
30	·078	600	46·8	48	+ 1·2	+ ·6	7
31	·082	600	49·2	45	− 4·2	− 3·6	7
32	·084	500	42·0	40	− 2·0	− 5·6	6
33	·083	500	41·5	44	+ 2·5	− 3·1	6
34	·080	400	32·0	35	+ 3·0	− ·1	6
35	·076	400	30·4	30	− ·4	− ·5	5
36	·071	400	28·4	25	− 3·4	− 3·9	5
37	·066	300	19·8	21	+ 1·2	− 2·7	4
38	·061	300	18·3	15	− 3·3	− 6·0	4
39	·057	300	17·1	21	+ 3·9	− 2·1	4
40	·053	300	15·9	15	− ·9	− 3·0	4
Total	—	14,100	566·0	563	+28·7 −31·7	+ 7·7 −42·2	100

$\sqrt{\text{Expected marriages}}$ has been taken as an approximation to the standard error of the deviation, but the more accurate expression $\sqrt{E_x p_x q_x}$ would not materially affect the results. It is seen that the deviations change sign 16 times as compared with only 4 continuations of sign; this is not unsatisfactory, although the accumulated deviations are negative above age 30.

The individual deviations never exceed twice the standard error except at age 21, where the number of marriages is so small as to make the test unreliable. We do not get any runs of deviations with the same sign and of substantial amount which would need to be tested as a group, while the net deviation of −3 is satisfactory.

The total of the deviations irrespective of sign is $28{\cdot}7+31{\cdot}7=60{\cdot}4$, while $\cdot 8\,\Sigma\sqrt{\text{expected marriages}}$ is 80. These figures are sufficiently close to give no indication of under- or over-graduation.

Looking at the table of differences we see that $\Delta^3(mq)_{25}$ is ·007, which is large as compared with the other third differences. In altering this it should be borne in mind that the accumulated deviations above age 30 are all negative, suggesting that the curve is rather too high around ages 25–30.

If we alter $(mq)_{27}$ from ·050 to ·048 and $(mq)_{28}$ from ·062 to ·061 the run of the differences is improved and the largest third difference, namely, $\Delta^3(mq)_{26}$, is now only $-{\cdot}003$. What is perhaps more important is the surprisingly large effect which these slight modifications have on the accumulated deviations.

The revised figures for ages 27 and 28 are as follows:

Age	Graduated rate	Exposed to risk	Expected marriages	Actual marriages	Deviations (5)−(4)	Accumulated deviation
(1)	(2)	(3)	(4)	(5)	(6)	(7)
27	·048	800	38·4	34	−4·4	−3·7
28	·061	700	42·7	49	+6·3	+2·6

Thus the accumulated deviations from age 28 onwards are all 2·3 greater than they were before and several changes of sign are thus introduced. The net deviation is now only ·7, while the accumulated deviations add up to $+18{\cdot}2-21{\cdot}2$, a result much nearer to zero than before. The total deviations irrespective of sign become 59·5.

The graduation may now be considered satisfactory.

BIBLIOGRAPHY

"Paper on the Graphic Method of Graduation." T. B. Sprague. *J.I.A.* Vol. XXVI, or *J.S.S.* Special Number.

"Tests of a Mortality Table Graduation." H. L. Seal. *J.I.A.* Vol. LXXI.

EXAMPLES 6

1. A Life Office which does not grant surrender values until three complete years' premiums have been paid has obtained the following data from its experience of withdrawals (lapses and surrenders combined) under Whole Life Policies.

Graduate the rates of withdrawal graphically, and write down the graduated values. Comment on any special points arising.

Curtate duration n	Exposed to risk of withdrawal E_x	Withdrawals W_x	Rate of withdrawal $w_x = W_x/E_x$
0	11,000	209	·019
1	10,000	650	·065
2	9,000	1215	·135
3	7,000	679	·097
4	5,500	352	·064
5	4,000	196	·049
6	3,000	102	·034
7	2,500	92	·037
8	2,000	43	·022
9	1,600	56	·035
10	1,200	36	·030
11	1,000	21	·021
⋮	⋮	⋮	⋮

2. Given the following particulars, find interpolated values for durations 0 to 26 weeks by the graphic method:

No. of weeks sickness since accident	Under 2 weeks	2 but under 3 weeks	3 but under 4 weeks	4 but under 5 weeks	5 but under 13 weeks	13 but under 26 weeks
No. of cases	470	1970	1510	120	203	159

3. Employ the graphic method of graduation to obtain from the data below the adjusted rates of mortality for ages 47–67. Apply all the tests you know to the results of your graduation.

Age	Exposed to risk	Deaths	$10^5 \times q_x$	Age	Exposed to risk	Deaths	$10^5 \times q_x$
47	166	2	1205	58	628	11	1752
48	187	2	1070	59	701	14	1997
49	218	4	1835	60	813	18	2214
50	243	6	2469	61	917	18	1963
51	276	2	725	62	1040	24	2308
52	302	4	1325	63	1182	30	2538
53	347	7	2017	64	1299	43	3310
54	390	3	769	65	1432	41	2863
55	430	9	2095	66	1596	54	3383
56	494	9	1822	67	1752	64	3653
57	558	8	1434				

4. A series of values of q_x which were obtained from deaths, θ_x, and exposed to risk, E_x, is to be graduated by a graphic method. It has been suggested that if the two sets of points given by $q_x \pm \dfrac{2\sqrt{\theta_x}}{E_x}$ are plotted on the graph, then any smooth curve lying entirely between these two sets of points will give a satisfactory graduation. Comment on the suggestion.

5. Graduate the following experience by the graphic process. What limitations as regards (*a*) the graduation, (*b*) the tests of the graduation, are imposed by the way in which the data are given?

Age-group	Exposed to risk of death	Actual deaths	Age-group	Exposed to risk of death	Actual deaths
30–34	15	—	70–74	4,150	195
35–39	50	—	75–79	3,568	247
40–44	189	2	80–84	2,516	286
45–49	475	8	85–89	1,284	198
50–54	1,020	9	90–94	365	111
55–59	2,032	34	95–99	56	23
60–64	3,300	54	100 and over	3	2
65–69	4,203	113			
			Total	23,226	1282

6. Find, by the graphic process, graduated values of q_x from the following data:

(i) by separate graduation of the exposed to risk and deaths,

(ii) by graduating the group values of q_x.

Age-group	Exposed to risk of death	Actual deaths	Age-group	Exposed to risk of death	Actual deaths
30–34	9	—	65–69	829	40
35–39	22	—	70–74	864	51
40–44	24	—	75–79	796	85
45–49	54	—	80–84	488	69
50–54	194	9	85–89	217	54
55–59	395	10	90–94	49	16
60–64	678	27	95 and over	3	2
			Total	4622	363

7. From the following data obtain by the graphic method graduated values of q_x for ages 45 to 60 inclusive, and apply the usual tests to your graduation:

Age x	E_x	θ_x	q_x	Age x	E_x	θ_x	q_x
21 25	20	1	·0500	46	100	2	·0200
26	10	—	—	47	110	3	·0273
27	20	—	—	48	110	1	·0091
28	20	—	—	49	100	3	·0300
29	30	1	·0333	50	100	1	·0100
30	40	—	—	51	100	1	·0100
31	40	2	·0500	52	100	—	—
32	40	—	—	53	100	1	·0100
33	50	—	—	54	100	1	·0100
34	50	2	·0400	55	100	2	·0200
35	40	—	—	56	100	3	·0300
36	60	—	—	57	100	2	·0200
37	60	—	—	58	110	3	·0273
38	70	—	—	59	110	—	—
39	70	1	·0143	60	110	3	·0273
40	80	—	—	61	110	3	·0273
41	80	1	·0125	62	110	2	·0182
42	90	2	·0222	63	110	2	·0182
43	90	2	·0222	64	100	4	·0400
44	90	—	—	65	100	3	·0300
45	100	1	·0100	66–70	500	20	·0400

CHAPTER VII

GRADUATION BY REFERENCE TO A STANDARD TABLE

1. In this chapter we shall deal only with rates of mortality, but if a suitable standard table can be found for other rates of decrement (e.g. withdrawals or marriages) the method can be applied quite satisfactorily. It is rarely, however, that such a standard table exists.

All mortality tables commence with a portion where dy/dx, the slope of the curve, is small and finish with a portion where it is steep. A method which may be satisfactory for one part of the curve may be quite unsuitable for the other; for instance we have mentioned in the previous chapter that, in applying the graphic method, two curves drawn to different scales usually have to be employed.

In order to produce a flatter curve than that given by $y=q_x$, a function such as $\log(q_x+\cdot 1)$ is sometimes calculated from the data. This function is then graduated instead of q_x.

The most satisfactory method for restricted data is, however, to use a standard table as a "base curve". There are many ways in which this can be done. One of the simplest is to calculate the ratios of the q's derived from the data to the corresponding q's of the standard table and to graduate these ratios. Clearly if a suitable table is taken for the base curve the ratio should not vary very greatly from unity. We are here in fact graduating q'_x/q_x, where q'_x is the rate derived from the data and q_x is obtained from the standard table. The ratio may quite well be graduated graphically on one diagram.

It is interesting to note that the first recorded instance of the use of a standard table was in a graduation by Griffith Davis. In this graduation the ratio q'_x/q_x was adjusted graphically (*J.I.A.* Vol. XI).

2. Lidstone's graphic method.

In a paper in *J.I.A.* Vol. XXX, G. J. Lidstone greatly improved on this method by dealing not with the ratio q'_x/q_x but with $\log(p_x/p'_x)$, where as before unaccented symbols refer to the standard table. This function produces values which are not only smaller than q'_x/q_x, but which usually progress more smoothly.

The following quotation from Lidstone's paper is important: "Considering first the mortality table to be used in calculating the expected deaths, it is evident that smoothness of graduation is most essential, more so, in fact, than close agreement with the observed rates of mortality, since any irregularities in the standard table would be reproduced and possibly exaggerated in our final results; for this reason a table following a mathematical law—as, for example, Makeham's—will generally be the most suitable to employ."

If $\mu'_x = \mu_x + c$ (where c is a constant) all the values of p'_x will bear a constant ratio to the corresponding values of p_x and the function $\log p_x - \log p'_x$ will be a straight line parallel to the axis of x. The special case when both μ'_x and μ_x follow Makeham's law was also investigated, but most of this work is now of less interest than formerly because it is rarely that a modern experience can be graduated successfully by that law.

3. Formulae methods.

In order to obtain values which progress smoothly enough for a graphic graduation to be successful it is usually necessary to deal with functions such as p_x and q_x rather than with the exposed to risk and deaths—although these are to be preferred on general grounds because they give effect to the weight of the data at successive ages or age-groups.

For this reason it is usual to assume some algebraic relationship between, say, q'_x and q_x and to determine the constants in the relationship by reference to the exposed to risk and deaths.

The following are a few of the formulae which might be used:

1. $q'_x = aq_x + b$. (If $b = 0$ this becomes $q'_x/q_x =$ constant.)

2. $\mu'_x = a\mu_x + b$.

3. $q'_x = q_x(ax + b)$.

4. $\mu'_x = \mu_{x+n} + K$.

5. $q'_x = aq_x^{(1)} + bq_x^{(2)}$, where $q_{(x)}^{(1)}$ refers to one standard table and $q_x^{(2)}$ to a second. Usually the mortality rates to be graduated are intermediate between those of the two standard tables.

In these equations a, b, K and n are constants.

The following points arise in the application of the formulae.

4. Formulae 1 and 2.

If we assume for every age that $q'_x = aq_x + b$, it follows that

$$E_x q'_x = aE_x q_x + bE_x,$$

$$\left.\begin{aligned} \Sigma E_x q'_x &= a\Sigma E_x q_x + b\Sigma E_x \\ \text{and} \qquad \Sigma^2 E_x q'_x &= a\Sigma^2 E_x q_x + b\Sigma^2 E_x \end{aligned}\right\}. \qquad \ldots\ldots(1)$$

$E_x q'_x$ is merely the actual deaths observed at age x, while $E_x q_x$ is the expected deaths according to the standard table. When these expected deaths have been calculated it is a simple matter to derive the equations (1), which can then be solved for a and b. The example later in this section will give the details of the calculations involved.

When the data are grouped and E_x is not available for individual ages the expected deaths are usually calculated by using q_x for the central age of the group, e.g. for a group of five ages 30–34 q_{32} would be used, while for a group of four ages 30–33 $q_{31\frac{1}{2}}$ would be used. The slight error thus introduced is not a serious matter when the data are scanty and sampling errors are therefore considerable.

Similar remarks apply to Formula 2.

5. Formula 3.

There are several ways in which the constants a and b can be found. The following method is probably the best.

Put $q'_x/q_x = y$, so that the equation can be written:

$$y = ax + b. \qquad \ldots\ldots(2)$$

We have to determine the constants so as to secure the best fit, having regard to the weight of the data at each age or in each age-group.

This is precisely the problem which arises in fitting a line of regression to data when frequencies have to be allowed for; the same method can therefore be employed.

First tabulate y for each value of x or for the central age x of each group and take as the corresponding frequency some convenient number roughly proportionate to the exposed to risk on which the value of y is based. (The actual values of the exposed to risk would make the work very laborious without achieving any material improvement in the fit.)

Treating y and x as correlated variables we can then find the line of regression, $y=ax+b$, and hence find q'_x for each age from the known value q_x at that age.

6. Formula 4.

Unless Makeham's law applies approximately it is impossible to find n except by trial and error. The method in which the formula can be applied and K determined will be apparent when we consider Makeham's law in Chapter IX. For the moment it is sufficient to note that the formula $\mu'_x=\mu_{x+n}+K$ has practical advantages.

It will be remembered that

$$\frac{d}{dx}\log D_x=-(\mu_x+\delta),$$

so that an addition of K to μ_x has the same effect on the values of $d\log D_x/dx$ as the same addition to the force of interest δ.

It follows that the values of D_x, N_x and a_x are the same if we add K to the force of interest as if we add K to the force of mortality.

If, therefore, we find it possible to assume that

$$\mu'_x=\mu_{x+n}+K,$$

many functions can be obtained from tables based on the standard mortality by adding n years to the age and by adding K to the force of interest. Annuity values can be derived in this way and the single and annual net premiums can then be deduced from Premium Conversion Tables using the true rate of interest and not the rate produced as a result of adding K to δ.

In actual practice refinements are out of place in using a table based on scanty data and it is usual to add K to the rate of interest instead of to δ and to round off the result to the nearest rate of interest for which functions are already tabulated.

7. Formula 5.

If the decremental rates of the special experience seem intermediate between those according to two standard tables this method can give good results.

Since $$q'_x = aq_x^{(1)} + bq_x^{(2)},$$

and $$\left.\begin{aligned} \Sigma E_x q'_x &= a\Sigma E_x q_x^{(1)} + b\Sigma E_x q_x^{(2)} \\ \Sigma^2 E_x q'_x &= a\Sigma^2 E_x q_x^{(1)} + b\Sigma^2 E_x q_x^{(2)} \end{aligned}\right\}. \qquad \ldots\ldots(3)$$

The summations extend over all ages, and once the expected deaths according to the two standard tables have been calculated the constants a and b can easily be found from equations (3).

It should be pointed out that when the available constants are connected by linear equations such as (3), linear constraints are said to be imposed. In applying the χ^2 test, therefore, it should be remembered that two degrees of freedom are lost.

As an example we shall deal with a graduation by a slight modification of Formula 1. The method illustrates most of the points involved.

8. Example.

Graduate the following experience by means of the formula

$$q'_x = aq_{x-n} + b,$$

where $n = 10$ and the values of q_x are taken from the H^{M} table (Makeham Graduation). θ_x denotes the number of deaths between age x and $x+1$ and E_x is the corresponding exposed to risk.

Age x	E_x	θ_x	Age x	E_x	θ_x	Age x	E_x	θ_x
30	1000	5	40	1600	7	50	1600	12
31	1100	4	41	1700	12	51	1500	13
32	1200	5	42	1700	5	52	1500	14
33	1300	6	43	1700	14	53	1400	12
34	1400	6	44	1800	12	54	1300	11
35	1500	8	45	1800	12	55	1200	13
36	1500	6	46	1800	14	56	1100	8
37	1500	8	47	1800	14	57	1000	11
38	1600	10	48	1700	10	58	900	9
39	1600	7	49	1600	13	59	800	9
						60	800	8

The necessary calculations are shown in the following table, column (4) of which shows the expected deaths according to the H^{M} table. Column (3) is obtained by summing column (2) successively from the top downwards: i.e. the *r*th entry in column (3) is the sum of the first *r* entries of column (2).

Table XVII

Age (1)	E_x (2)	ΣE_x (3)	$E_x q_{x-10}$ (4)	Σ (4) (5)	θ_x (6)	$\Sigma\theta_x$ (7)
30	1,000	1,000	5·7	5·7	5	5
31	1,100	2,100	6·7	12·4	4	9
32	1,200	3,300	7·7	20·1	5	14
33	1,300	4,600	8·7	28·8	6	20
34	1,400	6,000	9·7	38·5	6	26
35	1,500	7,500	10·6	49·1	8	34
36	1,500	9,000	10·8	59·9	6	40
37	1,500	10,500	11·0	70·9	8	48
38	1,600	12,100	11·9	82·8	10	58
39	1,600	13,700	12·1	94·9	7	65
40	1,600	15,300	12·3	107·2	7	72
41	1,700	17,000	13·3	120·5	12	84
42	1,700	18,700	13·7	134·2	5	89
43	1,700	20,400	14·0	148·2	14	103
44	1,800	22,200	15·1	163·3	12	115
45	1,800	24,000	15·5	178·8	12	127
46	1,800	25,800	15·9	194·7	14	141
47	1,800	27,600	16·4	211·1	14	155
48	1,700	29,300	15·9	227·0	10	165
49	1,600	30,900	15·5	242·5	13	178
50	1,600	32,500	16·0	258·5	12	190
51	1,500	34,000	15·6	274·1	13	203
52	1,500	35,500	16·2	290·3	14	217
53	1,400	36,900	15·7	306·0	12	229
54	1,300	38,200	15·2	321·2	11	240
55	1,200	39,400	14·7	335·9	13	253
56	1,100	40,500	14·1	350·0	8	261
57	1,000	41,500	13·5	363·5	11	272
58	900	42,400	12·7	376·2	9	281
59	800	43,200	11·9	388·1	9	290
60	800	44,000	12·6	400·7	8	298
Total	44,000	729,100	400·7	5855·1	298	4282

Similarly, column (5) is derived from column (4) and column (7) from column (6). The sums of columns (3), (5) and (7) then give the values of $\Sigma^2 E_x$, $\Sigma^2 E_x q_{x-10}$ and $\Sigma^2 \theta_x$ required.

Note that the sum of column (2) automatically checks the last entry in column (3) and similarly for the other columns. Columns (3), (5) and (7) could have been obtained by summing from the bottom upwards. The values of *a* and *b* would be the same whichever method were adopted.

The equations corresponding to those numbered (1) on p. 168 are:

$$\left.\begin{aligned} 298 &= 400{\cdot}7a + 44{,}000b \\ \text{and} \qquad 4282 &= 5855{\cdot}1a + 729{,}100b \end{aligned}\right\}.$$

These give $$a = {\cdot}8363 \quad \text{and} \quad b = -{\cdot}00084.$$

The graduated rates of mortality can be derived from the equation

$$q_x' = {\cdot}8363 q_{x-10} - {\cdot}00084.$$

The actual deaths and those expected according to the graduated table (not the standard table) should be compared; owing, however, to the small number of deaths involved any elaborate tests would be out of the question. There is no need to test for smoothness, since the standard table was graduated by a mathematical formula.

9. Advantages of the method.

The method of graduation by reference to a standard table is particularly valuable when the data are scanty, so that most other methods are out of the question. In such cases even a graphic graduation would be largely guesswork.

If the standard table is smooth (as it certainly should be) the results are satisfactory as far as smoothness is concerned and it is possible to concentrate on tests for adherence to data.

Knowledge of other tables based on similar experience is automatically brought into use in the process of graduation.

The method can be adapted to select tables, but with scanty data the select rates themselves are suspect, as the sampling errors are so great.

The ends of the table cause little difficulty, but the reliability of the results may of course be doubtful.

10. Disadvantages of the method.

It is not always possible to find a suitable standard table, so that even if the constants in the graduation formula are chosen properly the adherence of the results to the rough data is not satisfactory.

BIBLIOGRAPHY

"Graduation by reference to a Standard Table." G. J. LIDSTONE. *J.I.A.* Vol. xxx: "Reprints", 1935.

EXAMPLES 7

1. Show how a comparatively small mortality experience may be graduated by reference to a standard table:

(*a*) by a graphic method;
(*b*) by a formula so that the first and second summations of the actual deaths and the expected deaths are equal.

Give any conditions necessary for, or restrictions to be imposed upon, the use of the methods stated.

2. The values of q_x given in the following schedule are to be graduated by the formula

$$q_x = a + bq'_x,$$

where q'_x, values of which are given in the last column of the schedule, is the rate of mortality according to a standard table.

Obtain values for the constants a and b.

Age x	Exposed to risk	Deaths	Ungraduated rate of mortality q_x	Standard rate of mortality q'_x
80	250	35	·140	·134
81	200	25	·125	·144
82	150	22	·147	·154
83	120	21	·175	·165
84	100	18	·180	·176
85	70	15	·214	·188
86	50	9	·180	·200
87	30	10	·333	·213
88	20	5	·250	·227
89	10	5	·500	·242

3. Graduate the rates of mortality of Example 3 at the end of Chapter VI by reference to a suitable standard table, adjusting the ages if necessary.

4. Obtain graduated rates of mortality from the following experience with reference to any standard table which you consider to be suitable.

Age-group	Exposed to risk	Deaths	Age-group	Exposed to risk	Deaths
10–20	77	—	51–55	5891	61
21–25	1472	5	56–60	5415	68
26–30	5449	8	61–65	3907	77
31–35	8087	16	66–70	1687	41
36–40	7739	19	71–75	451	18
41–45	7111	24	76–80	71	6
46–50	6237	37	81–85	1	1

5. Obtain graduated rates of mortality from the following experience by the formula

$$aq_x^{(1)} + bq_x^{(2)}$$

of para. 7, using as the standard tables the A 1924–29 and the A 1924–29 Light tables.

Age-group	Exposed to risk	Deaths	Age-group	Exposed to risk	Deaths
10–20	33	—	61–65	43,244	1084
21–25	617	2	66–70	34,013	1349
26–30	5,312	17	71–75	22,237	1326
31–35	13,811	31	76–80	12,516	1240
36–40	21,293	63	81–85	5,243	716
41–45	28,403	147	86–90	1,944	327
46–50	37,014	242	91–95	415	70
51–55	42,789	442	96–100	78	16
56–60	45,484	731	101–	15	2

6. The following data are derived from the Continuous Investigation into the Mortality of Assured Lives over the 5-year period 1934–38:

Whole-Life with profit policies. Medical section.
Duration 3 and over

Age-group	Exposed to risk	Deaths	Age-group	Exposed to risk	Deaths
10–20	110	—	61–65	47,151	1161
21–25	2,089	7	66–70	35,700	1390
26–30	10,761	25	71–75	22,688	1344
31–35	21,898	47	76–80	12,587	1246
36–40	29,033	82	81–85	5,244	717
41–45	35,514	171	86–90	1,944	327
46–50	43,251	279	91–95	415	70
51–55	48,681	503	96–100	78	16
56–60	50,899	799	101–	15	2

Taking the A 1924–29 table as a standard, use a graphic method to graduate the values of

$$\frac{q'_x\,(\text{observed})}{q_x\,(\text{A }1924\text{–}29)}.$$

7. Using the data of the previous question, obtain graduated values of q_x, using the $a(m)$ ult. table instead of the A 1924–29.

Comment on the results, bearing in mind (*a*) the relative smoothness, and (*b*) the suitability as regards similarity of the experiences of the two standard tables used.

CHAPTER VIII

GRADUATION BY A SUMMATION FORMULA

1. It is convenient in this chapter to regard any ungraduated value u'_x as consisting of two parts, the true or universe value u_x and a superimposed error ϵ_x.

Thus $$u'_x = u_x + \epsilon_x.$$

ϵ_x may be positive or negative and although we shall for the most part consider errors of sampling, the ϵ's will in practice contain inaccuracies as well, which will be dealt with by the formulae in exactly the same way as the sampling errors.

An ideal graduation would of course eliminate all the ϵ's, but it will be seen that the most that can be attained by the use of a summation formula is a reduction of the ϵ's and a smooth progression of the graduated values. Although we shall concentrate much of our attention on reduction of error and smoothness, it should be borne in mind that all the usual tests of adherence to data should be applied to the results of any graduation. This is a point the importance of which is sometimes overlooked because actuarial literature on the subject of summation formulae tends to deal with methods rather than results.

2. Running averages. Reduction of irregularities.

In analysing a series of observations which show irregularities in the form of ripples or undulations statisticians often tabulate moving or "running" averages as a means of showing the general trend of the observations. By taking an average of, say, five consecutive values, the ripples are greatly reduced. This can best be illustrated by a consideration of the series shown in Table XVIII.

The first column is a series of values written down at random. The next column gives moving averages, each of which is the average of the corresponding value in the first column and the values on either side. Thus the average of the first three items in the first column is 4; this is written on the second line. Similarly, the average of the second, third and fourth is $3{\cdot}6\dot{6}$, and so on. These are examples of moving averages as used by statisticians; it is usual, however, to take the average of more than three terms, twelve being a common number in connection with monthly observations.

Table XVIII

4			
1	4		
7	$3{\cdot}6\dot{6}$	$4{\cdot}8\dot{8}$	
3	7	$6{\cdot}1\dot{1}$	$6{\cdot}33$
11	$7{\cdot}6\dot{6}$	8	$7{\cdot}26$
9	$9{\cdot}3\dot{3}$	$7{\cdot}6\dot{6}$	$7{\cdot}70$
8	6	$7{\cdot}4\dot{4}$	$7{\cdot}52$
1	7	$7{\cdot}4\dot{4}$	$8{\cdot}11$
12	$9{\cdot}3\dot{3}$	$9{\cdot}4\dot{4}$	
15	12		
9			

We need not confine our method to the figures in the first column, and if we perform the same operation on those in the second column we obtain the values shown in the third column. Similarly, the last column can be derived from the third. The irregularities of the first column have been greatly reduced by the averaging. It will be seen later that this would have been more marked if the averages had not always related to three values, e.g. if in deriving the third column from the second the average of four consecutive terms had been taken, and the average of two consecutive terms in arriving at the fourth column from the third.

3. Distortion of smooth values.

In Table XIX the values shown in the first column have been dealt with in a similar manner. The values in the second column have been found by averaging each set of four consecutive values of u_x, and here it will be noticed that they are not in alignment with the values of u_x. This is because the average of an even number of values cannot be regarded as corresponding to either of the two central values but to a hypothetical value lying between them. Consequently the values in the second column correspond to values between those in the first column. The third column is derived from the second by averaging in fives and the values are in alignment with those in the second column. Finally, the values in the last column are derived from those in the third by averaging six at a time and are therefore out of alignment with those values, i.e. they are back in alignment with the original values of u_x.

Table XIX

u_x	Average of 4 terms	Average of 5 terms	Average of 6 terms
1097			
1084			
	1067·5		
1061			
	1039·5		
1028			
	1001·5	991·5	
985			
	953·5	943·5	
932			
	895·5	885·5	
869			838·1$\dot{6}$
	827·5	817·5	
796			765·1$\dot{6}$
	749·5	739·5	
713			682·1$\dot{6}$
	661·5	651·5	
620			
	563·5	553·5	
517			
	455·5	445·5	
404			
	337·5		
281			
	209·5		
148			
5			

These original values were actually obtained from the expression

$$u_x = 1100 + 2x - 5x^2$$

by giving x the successive values 1, 2, 3, ... 15, and were therefore ideally smooth. The process of averaging has distorted these smooth values quite appreciably, the difference in each case being 30·8$\dot{3}$. The reason for this will be seen later, but the two examples have shown that by averaging successive values we tend (i) to reduce irregularities and fluctuations and, (ii) to distort values already smooth. We shall deal first with the second of these two features.

4. [n] or "summation n".

In *Mathematics for Actuarial Students*, Part II, pp. 114 *et seq.*, the operator [n] or "summation n" is defined thus:

$$[n]\,u_0 = u_{-\frac{n-1}{2}} + u_{-\frac{n-3}{2}} + \ldots + u_{\frac{n-3}{2}} + u_{\frac{n-1}{2}},$$

whether n is odd or even.

If n is odd the central term is u_0, and $[n]\,u_0$ is simply the sum of n consecutive terms, the central one of which is u_0.

Thus $$[5]\,u_0 = u_{-2}+u_{-1}+u_0+u_1+u_2.$$

If n is even the two middle terms are $u_{-\frac{1}{2}}$ and $u_{\frac{1}{2}}$, and u_0 itself does not appear. Nevertheless $[n]\,u_0$ still represents the sum of n consecutive u's with an equal number lying on each side of u_0, e.g.

$$[6]\,u_0 = u_{-\frac{5}{2}}+u_{-\frac{3}{2}}+u_{-\frac{1}{2}}+u_{\frac{1}{2}}+u_{\frac{3}{2}}+u_{\frac{5}{2}}.$$

In Table XIX the second column was obtained by the operation $\frac{[4]}{4}$, the next was obtained by the operation $\frac{[5]}{5}$ and the last by the operation $\frac{[6]}{6}$.

Gauss's formula may be written

$$u_r = u_0 + r\Delta u_0 + r_{(2)}\Delta^2 u_{-1} + (r+1)_{(3)}\Delta^3 u_{-1} + (r+1)_{(4)}\Delta^4 u_{-2} + \dots.$$

Also $\Delta^{-1}\,(r+s)_{(K)} = (r+s)_{(K+1)}$, r being the argument.

Hence

$$[n]\,u_0 = \Big[\Delta^{-1}u_r\Big]_{-\frac{n-1}{2}}^{\frac{n+1}{2}} = \Big[ru_0 + r_{(2)}\Delta u_0 + r_{(3)}\Delta^2 u_{-1} + (r+1)_{(4)}\Delta^3 u_{-1} + (r+1)_{(5)}\Delta^4 u_{-2}\Big]_{-\frac{n-1}{2}}^{\frac{n+1}{2}}$$

$$= nu_0 + \frac{n(n^2-1)}{2^2\,3\,!}\Delta^2 u_{-1} + \frac{n(n^2-1)(n^2-9)}{2^4\,5\,!}\Delta^4 u_{-2},$$

ignoring sixth and higher differences.

It is usual to write b for the operator $\Delta^2 E^{-1}$ and the above result is then written in the form

$$\frac{[n]}{n}u_x = \left\{1 + \frac{n^2-1}{24}b + \frac{(n^2-1)(n^2-9)}{1920}b^2\right\}u_x, \quad \dots\dots(1)$$

on replacing u_0 by the more general symbol u_x. This result is important and should be memorized.

We see therefore that by taking the average of successive values we introduce distortions of $\frac{n^2-1}{24}bu_x$ (the second difference error) and $\frac{(n^2-1)(n^2-9)}{1920}b^2u_x$ (the fourth difference error).

5. Second and fourth difference errors.

Most of the well-known formulae for graduation by summation involve three operators, often denoted by [l], [m] and [n].

Since finite difference operators obey the ordinary laws of algebra within certain well-defined limits:

$$\frac{[l]\,[m]\,[n]}{lmn} \equiv \left\{1 + \frac{l^2-1}{24}b + \frac{(l^2-1)(l^2-9)}{1920}b^2\right\}$$

$$\left\{1 + \frac{m^2-1}{24}b + \frac{(m^2-1)(m^2-9)}{1920}b^2\right\}\left\{1 + \frac{n^2-1}{24}b + \frac{(n^2-1)(n^2-9)}{1920}b^2\right\},$$

ignoring sixth and higher differences,

$$\equiv 1 + \frac{l^2+m^2+n^2-3}{24}b + \lambda b^2,$$

where λ is an algebraic function of l, m, n which can readily be evaluated numerically in any given case.

For the moment we shall ignore the fourth difference error and deal only with the second term $\frac{l^2+m^2+n^2-3}{24}bu_x$.

In the example on p. 178, l, m and n were 4, 5 and 6.

Hence $$\frac{l^2+m^2+n^2-3}{24} = \frac{74}{24} = \frac{37}{12}.$$

Also bu_x (i.e. $\Delta^2 u_{x-1}$) $= -10$, third and higher orders of differences vanishing.

Hence for all values of x the second difference error is $-\frac{370}{12}$. This reduces to $-30{\cdot}83$, the actual distortion produced.

6. Choice of operand.

With one important exception it may be said that formulae used in practice introduce no second difference error. Even in the exception which will be discussed later the second difference error is small.

To eliminate the error we operate not on u_x itself but on an expression known as the *operand*; this operation will counterbalance the second difference error introduced by the successive summations.

For instance, for three summations $[l]$, $[m]$ and $[n]$ the operand should reduce to

$$\left\{1-\frac{l^2+m^2+n^2-3}{24}b\right\}u_x,$$

ignoring fourth and higher differences.

The result of the operations denoted by $[l]\,[m]\,[n]/lmn$ on this function will be

$$\frac{[l]\,[m]\,[n]}{lmn}\left\{1-\frac{l^2+m^2+n^2-3}{24}b\right\}u_x$$
$$=\left\{1+\frac{l^2+m^2+n^2-3}{24}b\right\}\left\{1-\frac{l^2+m^2+n^2-3}{24}b\right\}u_x$$
$$=u_x,$$

ignoring fourth and higher differences.

It will be evident that this gives an unlimited choice of functions which satisfy the criterion, and even if we restrict ourselves to those which are practical and easily handled the number available will be considerable.

The term *operator* is loosely applied to the combination of the $[n]$ operators and the dividing factors (e.g. $[l]\,[m]\,[n]/lmn$), although the distinction between operator and operand is somewhat arbitrary. The order in which the operations are effected is immaterial.

Consider the operator $\dfrac{[5]^3}{125}$.

Putting $l=m=n=5$ we see that the second difference error in the formula is $3b$.

Hence the operand must reduce to $1-3b$.

Now $$(1-3b)\,u_x=u_x-3\,(u_{x-1}-2u_x+u_{x+1})$$
$$=-3u_{x-1}+7u_x-3u_{x+1},$$

which may be written $\{10\,[1]-3\,[3]\}\,u_x$.

Thus the formula

$$\frac{[5]^3}{125}\{10\,[1]-3\,[3]\}\,u_x \qquad \text{......(2)}$$

will not introduce any second difference error.

This is known as King's form of Woolhouse's formula and is of historic interest.

Alternatively, it will be found that the operand

$$-u_{x-2}+u_{x-1}+u_x+u_{x+1}-u_{x+2}$$

also reduces to $\{1-3bu_x\}$, ignoring fourth and higher differences.

Writing this operand in the form $2[3]-[5]$, we have the formula

$$\frac{[5]^3}{125}\{2[3]-[5]\}u_x \qquad \ldots\ldots(3)$$

(Higham's formula).

These formulae have the same operator and neither introduces any second difference error; the second however is greatly superior to the first in the way in which it enables superimposed errors to be dealt with.

We have seen that the operator $[n]$ does not involve first or third differences: hence the operand must also exclude them and must therefore be of the form

$$u_x+c_1(u_{x-1}+u_{x+1})+c_2(u_{x-2}+u_{x+2})+\ldots;$$

i.e. terms equidistant from the central value u_x must have the same coefficient.

From Gauss's formula

$$u_{x+r}+u_{x-r}=\left\{2+r^2b+\frac{r^2(r^2-1)}{12}b^2\right\}u_x, \qquad \ldots\ldots(4)$$

another result which should be memorized.

By means of this formula it is a simple matter to find operands which, when combined with a given operator, will produce no second difference error.

The following are examples of the use of both formulae.

Example 1.

Find a suitable operand to combine with $\frac{[4][5][6]}{120}$.

Now $$\frac{[4][5][6]}{120}\equiv 1+\frac{4^2+5^2+6^2-3}{24}b$$

$$\equiv 1+3\tfrac{1}{12}b.$$

Hence the operand should reduce to $\{1-3\tfrac{1}{12}b\}u_x$, which is rather an awkward expression. If we are prepared to ignore the small distortion of $\tfrac{1}{12}bu_x$ we obtain a much simpler formula.

Suppose that we decide to have an operand involving u_{x-2} to u_{x+2}, i.e. of the form

$$c_0u_x+c_1(u_{x-1}+u_{x+1})+c_2(u_{x-2}+u_{x+2}).$$

Using formula (4) this reduces to

$$(c_0+2c_1+2c_2)\,u_x+(c_1+4c_2)\,bu_x+\text{terms in } b^2u_x, \text{ etc.}$$

If this is to reduce to $(1-3b)\,u_x$, we have

$$\left.\begin{aligned} c_0+2c_1+2c_2 &= 1 \\ c_1+4c_2 &= -3 \end{aligned}\right\}.$$

Solutions are $\quad c_0=c_1=1 \quad$ and $\quad c_2=-1.$

We have therefore the formula

$$\frac{[4]\,[5]\,[6]}{120}\{-u_{x-2}+u_{x-1}+u_x+u_{x+1}-u_{x+2}\},$$

or

$$\frac{[4]\,[5]\,[6]}{120}\{2\,[3]-[5]\}\,u_x \qquad \text{......(5)}$$

(Hardy's Friendly Society formula).

Although there is a slight second difference error here, this formula was used very successfully for the purpose for which it was devised and is the best-known formula involving such an error.

Alternatively, we could have written the unknown operand in the form

$$\{K_1\,[1]+K_3\,[3]+K_5\,[5]+\dots\}\,u_x, \text{ the } K\text{'s being constants.}$$

If we decide to restrict its range to five terms (i.e. to ignore $[7]\,u_x$, etc.) we can use formula (1) to reduce this to the form

$$\{K_1+K_3\,3\,(1+\tfrac{1}{3}b)+K_5\,5\,(1+b)\}\,u_x=\{(K_1+3K_3+5K_5)+(K_3+5K_5)\,b\}\,u_x.$$

If this is to reduce to $\{1-3b\}\,u_x$, we have

$$\left.\begin{aligned} K_1+3K_3+5K_5 &= 1 \\ K_3+5K_5 &= -3 \end{aligned}\right\},$$

giving as possible solutions $K_1=0$, $K_3=2$, $K_5=-1$, i.e. the operand $\{2\,[3]-[5]\}\,u_x$ as before.

Example 2.

Find a suitable operand, involving terms u_{x-3} to u_{x+3}, for use with the operator $\dfrac{[5]\,[13]}{65}$.

This is perhaps the best-known example of a two-term operator.

As has been said previously, to achieve smooth results, operators involving three summations are usual. This two-term operator was designed for a special purpose.

$$\frac{[5]\,[13]}{65}\equiv 1+\frac{5^2+13^2-2}{24}\,b\equiv 1+8b.$$

The operand

$$c_0 u_x + c_1(u_{x-1}+u_{x+1}) + c_2(u_{x-2}+u_{x+2}) + c_3(u_{x-3}+u_{x+3})$$

reduces to

$$(c_0+2c_1+2c_2+2c_3)\,u_x + (c_1+4c_2+9c_3)\,bu_x,$$

to third differences.

If this is equivalent to $(1-8b)\,u_x$, we have

and
$$\left.\begin{aligned} c_0+2c_1+2c_2+2c_3 &= 1 \\ c_1+4c_2+9c_3 &= -8 \end{aligned}\right\}.$$

Convenient integral solutions are $c_0=c_1=1$, $c_2=0$, $c_3=-1$, giving the formula

$$\frac{[5][13]}{65}\{-u_{x-3}+u_{x-1}+u_x+u_{x+1}-u_{x+3}\}$$

or
$$\frac{[5][13]}{65}\{[3]+[5]-[7]\}\,u_x \qquad \ldots\ldots(6)$$

(Hardy's "wave-cutting" formula).

7. Calculation of fourth difference error.

Hitherto we have concentrated on the elimination of the second difference error. It is, however, important to know what fourth difference is introduced.

As an example we shall consider Spencer's 21-term formula, probably the most famous and generally satisfactory of all summation formulae:

$$\frac{[5]^2[7]}{350}\{[1]+[3]+[5]-[7]\}\,u'_x. \qquad \ldots\ldots(7)$$

The operand can also be written thus:

$$\{-u'_{x-3}+u'_{x-1}+2u'_x+u'_{x+1}-u'_{x+3}\},$$

and formula (4) of this chapter can be used to expand this in terms of differences. We shall, however, use the formula

$$[n] \equiv n\left\{1+\frac{n^2-1}{24}\,b+\frac{(n^2-1)(n^2-9)}{1920}\,b^2\right\},$$

ignoring b^3, etc.; this formula would in any case have to be used for the operator.

Substituting in (7) we have

$$u_x = \tfrac{1}{2}\{1+b+\tfrac{1}{5}b^2\}^2\{1+2b+b^2\}\{1+3(1+\tfrac{1}{8}b)+5(1+b+\tfrac{1}{5}b^2)$$
$$-7(1+2b+b^2)\}u'_x$$
$$=\tfrac{1}{2}\{1+2b+1\tfrac{2}{5}b^2+\ldots\}\{1+2b+b^2\}\{2-8b-6b^2\ldots\}u'_x$$
$$=\{1-\tfrac{63}{5}b^2\}u'_x. \qquad \ldots\ldots(8)$$

Thus there is no second difference error, but a fourth difference error of

$$-\tfrac{63}{5}\Delta^4 u'_{x-2}.$$

8. The range.

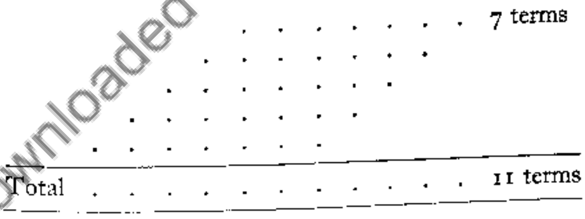
The meaning of the "range" of a formula is almost self-evident. Briefly, though not strictly accurately, it may be said to be the number of ungraduated u's involved in the calculation of a single graduated value. The exception to this definition arises if it is found that some of the coefficients are zero when the formula is fully expanded (see para. 10). For instance, the range of the expression

$$-u'_{x-3}+u'_{x-1}+u'_x+u'_{x+1}-u'_{x+3}$$

is seven, although only five terms are apparently involved.

The range can be found easily as follows. Find the range of the operand by inspection and add $l-1$, $m-1$, $n-1$, ... etc. for the operations $[l]\,[m]\,[n]\ldots$.

The following diagram illustrates the effect of the operator [5] on a 7-term operand, each term being represented by a dot.



The first line represents the operand and each subsequent line represents the effect of increasing the argument by 1. The total therefore represents the effect of the operator [5] and it will be seen that the original 7 terms have been increased to 11. The process is, of course, quite general and can be applied to several operators in succession.

Thus, in Spencer's formula above, the range of the operand is 7.

Hence the range of the formula $= 7+4+4+6 = 21$ terms.

Similarly, the range of Hardy's "wave-cutting" formula is 23 terms.

It is sometimes said that in using a summation formula one assumes that the underlying true function is a polynomial of the third degree. This is not correct. Nearly all these formulae involve no second difference error, but they will involve a fourth difference error unless the true values of the function being graduated have negligible fourth differences over the range of values covered by a single application of the formula. Thus the use of Spencer's 21-term formula does not imply that the function follows a third-degree curve over its entire range, but merely that any given set of 21 consecutive values can be represented with sufficient accuracy by a polynomial of the third degree. A different polynomial will usually be implied for each different set of 21 terms.

It will be seen therefore that, other things being equal, the shorter the range of a formula the better, because

(1) it is easier to apply;

(2) the assumption that fourth and higher differences are negligible over the range is more likely to be accurate; and

(3) a smaller number of terms at the ends remain to be filled in by other methods.

This last point is important and will be dealt with more fully in a later section. For the time being it is sufficient to point out that if a formula has a range of n, the first graduated value produced corresponds to the $\frac{n+1}{2}$th ungraduated value, leaving $\frac{n-1}{2}$ values at the beginning and similarly $\frac{n-1}{2}$ values at the end to be filled in by other methods.

Thus in Table XVIII the operator $\frac{[3]^3}{27}$ with a range of 7 left 3 terms blank at the beginning and end of the graduated values. Similarly in Table XIX, where the range of the formula was 13, only 3 graduated values could be obtained out of 15.

9. Effect of a summation formula on superimposed errors.

Hitherto we have considered only the way in which a summation formula affects the underlying true values and we have seen that it is a simple matter to ensure that, apart from the fourth difference error, it will reproduce them without distortion.

The whole purpose of the graduation is to eliminate the superimposed errors as far as possible. The tests now to be discussed deal with this aspect of the problem.

A complete analysis of a formula includes an investigation of the following features:

(1) The range.
(2) The second difference error.
(3) The fourth difference error.
(4) The error-reducing power of the formula.
(5) Its smoothing power.
(6) Its "wave-cutting" properties.

Of these, (1), (2) and (3), which deal with the underlying true values, have already been described, and (3) is probably the least important. To investigate all the points detailed it is necessary to expand the formula.

10. Expansion of a formula.

Any summation formula can be written in the simple forms

$$K_0 u'_x + K_1(u'_{x+1} + u'_{x-1}) + K_2(u'_{x+2} + u'_{x-2}) + \dots \\ + K_r(u'_{x+r} + u'_{x-r}) \quad \text{(range odd)}$$

or

$$K_{\frac{1}{2}}(u'_{x+\frac{1}{2}} + u'_{x-\frac{1}{2}}) + K_{\frac{3}{2}}(u'_{x+\frac{3}{2}} + u'_{x-\frac{3}{2}}) + \dots \\ + K_{r+\frac{1}{2}}(u'_{x+r+\frac{1}{2}} + u'_{x-r-\frac{1}{2}}) \quad \text{(range even).}$$

Incidentally it is interesting to note that, although every summation formula can be so expressed, there are an infinite number of formulae of the expanded type which cannot be derived from summation formulae. They may nevertheless be excellent for graduation purposes and anyone interested in the subject is referred to Dr Sheppard's paper in *J.I.A.* Vol. XLVIII.

Summation formulae owe their importance to the ease with which they can be applied, but with modern mechanical aids this is not now as important as formerly.

The actual process of expansion can best be demonstrated by examples. We shall first consider Spencer's 15-term formula (not to be confused with his 21-term formula given previously)

$$\frac{[5]\,[4]\,[4]}{320}\{-3\,[5]+6\,[3]+1\}\,u'_x. \qquad \ldots\ldots(9)$$

The operand can be expressed as

$$-3u'_{x+2}+3u'_{x+1}+4u'_x+3u'_{x-1}-3u'_{x-2}.$$

This is the most convenient form for our purpose. The method of detached coefficients is almost always used and care should be taken to insert zero coefficients for missing terms. In the first method of expansion demonstrated it is also assumed that there are zero coefficients at each end of the operand. The work is best set out in tabular form as follows:

Coefficients of operand (1)	Operator [5] applied to previous column (2)	Operator [4] applied to previous column (3)	Operator [4] applied to previous column (4)
.	.	.	
.	.	.	
.	.	0	
.	.	0	−3
0	0	−3	−6
0	0	−3	−5
0	−3	1	3
0	0	8	21
−3	4	15	46
3	7	22	67
4	4	22	74
3	7	15	67
−3	4	8	46
0	0	1	21
0	−3	−3	3
0	0	−3	−5
0	0	0	−6
.	.	0	−3
.	.		
.	.		

Hence the expanded formula is

$$\tfrac{1}{320}\{74u'_x+67(u'_{x+1}+u'_{x-1})+46(u'_{x+2}+u'_{x-2}) \\ +21(u'_{x+3}+u'_{x-3})+3(u'_{x+4}+u'_{x-4})-5(u'_{x+5}+u'_{x-5}) \\ -6(u'_{x+6}+u'_{x-6})-3(u'_{x+7}+u'_{x-7})\},$$

where u'_x as before represents the ungraduated value ($=u_x+\epsilon_x$ in the previous notation).

It will be seen that in effect the formula has been applied to 1 with the important proviso that an unlimited number of noughts could be assumed at either end. In applying the formula to observed data it is not of course possible to make this assumption, so that we finish with far fewer terms than we started. The assumption of zero coefficients at each end in the above table has the effect of apparently increasing the number of terms.

In actual practice it would not be necessary to obtain the final column below the entry 74 or at any rate the second 67, since the coefficients then repeat in reverse order. The previous columns could have been abbreviated accordingly.

11. Alternative method.

To illustrate another method of expansion we shall consider Woolhouse's formula

$$\frac{[5]^3}{125}(-3u'_{-1}+7u'_0-3u'_1). \qquad \text{......(10)}$$

This formula is now chiefly of historic interest.

First we develop the operator as follows, writing coefficients only:

[5] gives	1, 1, 1, 1, 1,
$[5]^2$,,	1, 2, 3, 4, 5, 4, 3, 2, 1,
$[5]^3$,,	1, 3, 6, 10, 15, 18, 19, 18, 15, 10, 6, 3, 1.

Each line is derived from the previous one by summing, very much as in the previous example.

We now have to incorporate the operand, the coefficients of which are -3, 7, -3, as follows:

Operator gives	1	3	6	10	15	18	19	18	15	10...
−3	−3	−9	−18	−30	−45	−54	−57	−54	−45	−30...
7		7	21	42	70	105	126	133	126	105...
−3			−3	−9	−18	−30	−45	−54	−57	−54...
Total	−3	−2	0	3	7	21	24	25	24	21...

The terms repeat in the reverse order after the coefficient 25 and need not be written down.

The expanded formula is therefore

$$\tfrac{1}{125}\{-3u'_{x-7}-2u'_{x-6}+3u'_{x-4}+7u'_{x-3}+21u'_{x-2}+24u'_{x-1}$$
$$+25u'_x+24u'_{x+1}+\dots\}.$$

12. Analysis of an expanded formula.

It has been stated previously that formulae of the type

$$u_x = K_{-2}u'_{x-2}+K_{-1}u'_{x-1}+K_0u'_x+K_1u'_{x+1}+K_2u'_{x+2}$$

can be constructed so as to produce a satisfactory graduation though they cannot be arrived at by summations. This section applies to these formulae as well as to summation formulae.

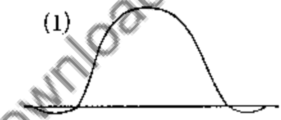
A great deal can be learnt merely from the inspection of the expanded formula without the calculation of any indices of smoothing power etc.

The range, for instance, is found by considering the first and last terms, so that in the examples of paras. 10 and 11 the range is 15 in each case.

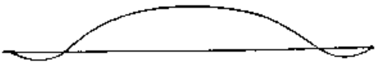
13. The coefficient curve.

An important conception in connection with an expanded formula is the *curve of coefficients*, or *coefficient curve* as it is sometimes called. This is obtained by plotting graphically the coefficients $K_{-r}\dots K_0\dots K_r$ and joining them by a smooth curve of which the following are typical.

(1)



(2)



The general characteristics are that the curve is symmetrical, rises to a peak in the middle, cuts the axis towards each end and thereafter lies below it.

The sum of the coefficients must be unity, so that they will tend to be mainly positive proper fractions with a few negative ones.

Remembering also that $u'_{x-r}+u'_{x+r}$ gives a second difference term of $r^2bu'_x$ (i.e. $r^2\Delta^2u'_{x-1}$), it follows that for zero second difference error

$$\Sigma r^2K_r=0.$$

Hence some of the K's must be negative, and in order to counteract the predominating positive K's, they must occur for the higher values of r; i.e. the negative coefficients will occur at the ends of the coefficient curve where they are weighted with the largest values of r. This is not a rigid demonstration; it is a discussion of the general form of the coefficient curve. Unusual formulae may prove exceptions.

From an examination of the run of the coefficients (the coefficient curve is not actually drawn in practice) it is possible to form an idea of how the formula will smooth the superimposed errors and also of its wave-cutting power. For clearness we shall deal with numerical examples; the reasoning is, however, quite general.

Consider first Woolhouse's formula:

$$\epsilon_x = \tfrac{1}{125}\{-3\epsilon'_{x-7} - 2\epsilon'_{x-6} + 3\epsilon'_{x-4} + 7\epsilon'_{x-3} + 21\epsilon'_{x-2} \\ + 24\epsilon'_{x-1} + 25\epsilon'_x + 24\epsilon'_{x+1} \cdots\},$$

where ϵ'_x is the error superimposed on the true value u_x, thus giving the observed value u'_x; ϵ_x is the graduated error thrown up by the use of the formula, ignoring any distortion of the u's such as fourth difference errors.

Consider a particular observed value

$$u'_{17} = u_{17} + \epsilon'_{17}.$$

This error ϵ'_{17} will first appear on the extreme right of the formula giving the graduated error ϵ_{10}. Subsequently it will appear in ϵ_{11}, ϵ_{12}, ..., rising to maximum importance in ϵ_{16}, ϵ_{17} and ϵ_{18} and finally disappearing from the formula after ϵ_{24} has been calculated.

Similar remarks apply to the other errors, and it will be seen that any graduated error ϵ_x differs from the previous graduated error ϵ_{x-1} for the following reasons:

(1) ϵ'_{x-8} has disappeared and ϵ'_{x+7} has appeared for the first time.

(2) ϵ'_{x-7} to ϵ'_{x+6} inclusive now appear with different coefficients, having "moved up" one; thus every coefficient K_r has been changed to K_{r+1}.

(1) is only a special case of (2) if we imagine zero coefficients at each end.

It follows therefore that if the coefficient curve is smooth, i.e. if the successive coefficients K change only gradually, the graduated errors ϵ_x will themselves change only gradually however irregular the ungraduated errors ϵ'_x may be. Since the underlying "true" values are supposed to be smooth it follows that a formula which when expanded has a smooth run of coefficients will produce smoothly progressing graduated values.

The coefficients $-3, -2, 0, 3, 7, 21, 24, 25, 24, \ldots$ which we have been considering do not progress smoothly except at the centre, and it is not surprising therefore that Woolhouse's formula is unsatisfactory from the point of view of smoothness.

14. Wave-cutting.

We can also learn something from the shape of the coefficient curve quite apart from its regular or irregular progression.

In the diagrams on p. 190, the first curve rises steeply to a narrow peak, while the second rises very gradually to a broad flat top. The use of a formula represented by the first curve will mean that any particular ungraduated error will have a marked influence on the graduated values close to it but very little effect on the others. A formula typified by the second curve will spread the effect over a wide field. Any given ϵ'_x will have only moderate influence on ϵ_x and values near it, and appreciable effect on more distant values.

Since the ϵ''s are random errors they will tend to change sign frequently, although sometimes a run of several consecutive errors of the same sign may arise. An ideal graduation would eliminate them completely; a summation formula gives them full weight although it spreads them over a larger range of values. A formula with a coefficient curve of the first type will tend to localize the effect of the errors and a wave in the ungraduated errors will be repeated although to a less extent in the graduated values.

If, however, the coefficient curve is of the second type any graduated value depends to a larger extent on the more distant values and far less on the near ones. Consequently, a wave in the ungraduated errors will begin to have an appreciable effect on the graduated values much earlier than with the other type of formula and will continue to have an appreciable effect much later, while its maximum importance, corresponding to the peak of the coefficient curve, will be greatly reduced. Such a formula is therefore said to be a good wave-cutter.

Woolhouse's formula has coefficients:

$$\tfrac{1}{125}\{-3, -2, 0, 3, 7, 21, 24, 25, 24, \ldots\},$$

and because of the marked peak is a very poor wave-cutter.

Hardy developed a special formula, given on p. 184, known as his "wave-cutting" formula, to distinguish it from his Friendly Society formula (p. 183, formula (5)).

The coefficients of the expanded "wave-cutting" formula are:

$$\tfrac{1}{65}\{-1, -2, -2, -1, 1, 4, 6, 7, 7, 6, 5, 5^*, 5, 6, \ldots\}.$$

The centre coefficient is marked with an asterisk and it will be noticed that the coefficient curve actually has a trough instead of the usual peak and the central eleven coefficients are all 5, 6 or 7. This formula is efficient therefore in dealing with long waves.*

The action of a summation formula in this respect is similar to that of a roller on uneven ground; the local irregularities are almost entirely removed by flattening out ridges and filling up troughs with earth taken from those ridges, while more extensive mounds and hollows are reduced but not eliminated.

15. Wave-cutting index.

We now come to the calculation of certain well-known indices or "coefficients", as they are usually called. The use of the word "coefficient" in this connection seems misleading and it is proposed therefore to use the word "index" in this book.

The wave-cutting index is defined as the sum of the *five* central coefficients. This is somewhat arbitrary and breaks down if there is an even number of terms; in this case the sum of the four middle coefficients and the next one at either end is taken.

The rationale is clear from the preceding section, for if the sum is large the coefficient curve is sharply peaked and the effect of the ungraduated errors is localized (unsatisfactory wave-cutter), while if the sum is small the curve is flat topped and the effect of the ungraduated errors is widely spread (good wave-cutter).

The wave-cutting index of Woolhouse's formula is $\frac{115}{125} = \cdot 92$.

* Vaughan has pointed out that this formula, when applied to short waves, may actually "reverse" them because of the high shoulders of the trough referred to in this paragraph.

The wave-cutting index of Hardy's "wave-cutting" formula is $\frac{27}{65} = \cdot 42$.

As a general rule it may be said that a combination of two operators, one of short and the other of long range, will deal effectively with waves.

16. Error-reducing power.

Clearly one of the most important functions of a summation formula is to reduce the superimposed errors ϵ'. The expanded formula however does not itself give the required information.

Consider, for instance,

$$\epsilon_x = K_0\epsilon'_x + K_1(\epsilon'_{x-1} + \epsilon'_{x+1}) + \ldots + K_r(\epsilon'_{x-r} + \epsilon'_{x+r}).$$

We are not interested in a particular set of ϵ's and the resulting ϵ, but rather in the results obtained on the average if the formula were applied very many times in similar circumstances. In considering a large number of values of ϵ'_x, ϵ'_{x+1}, etc. we are faced with the difficulty that some will be positive and some negative. The only satisfactory way of overcoming this difficulty is to deal with the root-mean-square deviation or standard deviation of each superimposed error.

We imagine a large number of observations made under similar conditions, thus producing a whole series of values of ϵ' at each age for which we can then calculate the standard deviations, taking each age separately.

Denote the standard deviation of the various values of ϵ'_x by S'_x, the standard deviation of the various values of ϵ'_{x+r} by S'_{x+r}, and so on.

By the application of the formula we can calculate the graduated errors ϵ_x and their standard deviation S_x for each age.

It remains to find a relationship between S_x and the various S''s.

We have shown in Chapter III that if

$$z = x + y + t + \ldots,$$
$$\sigma_z^2 = \sigma_x^2 + \sigma_y^2 + \sigma_t^2 + \ldots$$

(if the variables x, y, t, etc. are independent, i.e. not correlated).

By a simple extension it follows that if

$$z = K_0x + K_1y + K_2t + \ldots,$$
$$\sigma_z^2 = K_0^2\sigma_x^2 + K_1^2\sigma_y^2 + K_2^2\sigma_t^2 + \ldots.$$

To show this we merely have to take K_0x as a new variable X, K_1y as a new variable Y, and so on, and as this is equivalent to altering the scale it follows that the standard deviation of X is $K_0\sigma_x$, and so on. The result then follows, and we deduce as a special case that

$$S_x^2 = K_0^2 S_x'^2 + K_1^2(S'^2_{x-1} + S'^2_{x+1}) + K_2^2(S'^2_{x-2} + S'^2_{x+2}) + \dots \qquad \dots\dots(11)$$

If each S' is based on a very large number of ϵ's the values involved in a single application of formula (11) should not differ greatly, and if we assume them all equal to S' we find that

$$S^2 = S'^2\{K_0^2 + 2K_1^2 + 2K_2^2 + \dots\}. \qquad \dots\dots(12)$$

Hence the formula will on the average reduce the superimposed errors in the ratio $\dfrac{S}{S'} = \sqrt{K_0^2 + 2K_1^2 + 2K_2^2 + \dots}$. $\quad\dots\dots(13)$

The right-hand side is known as the error-reducing coefficient or index.

In the more general type of formula (not derived by summation), in which coefficients equidistant from the centre are not equal, the same argument applies, but the error-reducing index is

$$\{K_{-r}^2 + K_{-r+1}^2 + \dots + K_{-1}^2 + K_0^2 + K_1^2 + \dots + K_r^2\}^{\frac{1}{2}},$$

i.e. the root-mean-square of the coefficients. (K_{-r} is not necessarily equal to K_r.)

The smaller the error-reducing index the more powerful the formula.

For Woolhouse's formula:

Error-reducing index $= \frac{1}{125}\{2(3^2 + 2^2 + 3^2 + 7^2 + 21^2 + 24^2) + 25^2\}^{\frac{1}{2}}$, as the central coefficient does not occur twice

$$= \cdot 423. \quad \text{(unsatisfactory.)}$$

For Hardy's "wave-cutting" formula:

Error-reducing index

$$= \tfrac{1}{65}\sqrt{6 \times 1^2 + 4 \times 2^2 + 2 \times 4^2 + 4 \times 6^2 + 4 \times 7^2 + 3 \times 5^2}$$

$$= \cdot 333.$$

(Quite satisfactory, considering the range of 23 terms.)

17. Smoothing power.

This may at first seem synonymous with error-reducing power, since errors which are brought closer to zero will tend to progress more smoothly. The error-reducing index depends only on the size of the coefficients, not on their order, and the formula operates by grouping together errors of opposite sign, which tend to cancel out. When we come to consider the smoothness of the graduated results, the order of the coefficients and the coefficient curve become of great importance. As we have seen, the ungraduated errors "move up one" each time the formula is applied to give successive graduated values; i.e. each ungraduated error is multiplied in turn by each of the K's. Provided that these progress smoothly, so will the graduated values, and this effect is independent of any reduction in the errors.

The obvious way of testing smoothness is to consider the various orders of differences. It has become conventional to take the third order of differences in calculating a smoothing index. This choice is arbitrary, but, as will be seen later, it has one important practical advantage in that most well-known formulae have an operator consisting of three summations.

As before, we are concerned not with a particular set of errors but with the result of a great many applications. We therefore consider not the ϵ's but their standard deviations S', which for convenience we shall again assume to be equal.

The general algebraical discussion is rather involved and we can best illustrate the argument by a numerical example, using Woolhouse's formula

$$\epsilon_x = \tfrac{1}{125}\{-3\epsilon'_{x-7} - 2\epsilon'_{x-6} + 3\epsilon'_{x-4} + 7\epsilon'_{x-3} + 21\epsilon'_{x-2} + 24\epsilon'_{x-1} + 25\epsilon'_x + \ldots\}.$$

To find $\Delta^3\epsilon$, the third difference of the right-hand side, we remember that $\Delta^3 \equiv (E-1)^3 \equiv E^3 - 3E^2 + 3E - 1$. Writing coefficients only we arrange the work as shown on p. 197:

The denominator 125 is introduced only in the last line, and although the whole formula has been developed for the purpose of illustration, it is not necessary to go beyond half-way, as coefficients then are repeated in reverse order but with changed signs.

1		−3	−2	0	3	7	21	24	25	24
−3			9	6	0	−9	−21	−63	−72	−75
3				−9	−6	0	9	21	63	72
−1					3	2	0	−3	−7	−21
Δ^3	$\frac{1}{125}\{$	−3	7	−3	0	0	9	−21	9	0
		21	7	3	0	−2	−3			
		−72	−63	−21	−9	0	6	9		
		75	72	63	21	9	0	−6	−9	
		−24	−25	−24	−21	−7	−3	0	2	3
Δ^3		0	−9	21	−9	0	0	3	−7	3}

If we assume a standard deviation of the errors at each age equal to S', the standard deviation of the third difference error will be

$$\frac{S'}{125}\sqrt{3^2+7^2+3^2+0^2+\ldots+3^2+7^2+3^2},$$

or, since each coefficient is repeated (there being no central term),

$$=\frac{S'}{125}\sqrt{2(3^2+7^2+3^2+\ldots+9^2+0^2)},$$

the summation stopping at the middle.

The argument is perfectly general; we first of all expand the formula in the usual way and then find the coefficients of $\Delta^3\epsilon_x$ expressed in terms of the ungraduated errors.

The standard deviation of $\Delta^3\epsilon_x$ is then

$$S'\sqrt{\text{sum of the squares of the coefficients}}.$$

Now $$\Delta^3\epsilon'_x=\epsilon'_{x+3}-3\epsilon'_{x+2}+3\epsilon'_{x+1}-\epsilon'_x.$$

Hence, on the same assumption as before, the standard deviation of the ungraduated error is

$$S'\sqrt{1^2+3^2+3^2+1^2}=S'\sqrt{20}.$$

The graduation has therefore reduced the standard deviation of the third difference of the errors in the ratio

$$\sqrt{\frac{\text{sum of the squares of the coefficients}}{20}}, \quad \ldots\ldots(14)$$

where the coefficients referred to are those in the expansion of $\Delta^3\epsilon$ in terms of the ungraduated ϵ''s.

In the above example the *smoothing index* is

$$\frac{1}{125}\sqrt{\frac{2(3^2+7^2+3^2+9^2+21^2+9^2)}{20}}=\frac{\sqrt{67}}{125}=\cdot 0655 \quad \text{or} \quad \frac{1}{15} \text{ approx.}$$

Having regard to modern formulae this cannot be considered satisfactory; it should, however, be remembered that the range is only 15 terms and that the formula was one of the earliest ever constructed.

Although the assumption made in this and the preceding section that the standard deviations of the errors are all equal is not likely to be realized in practice, the error-reducing index and the smoothing index will still provide useful relative measures for comparing two or more formulae.

18. Second and fourth difference errors.

When the formula has already been expanded, the second and fourth difference errors can conveniently be found by the following method.

From most of the well-known central difference formulae it follows that

$$u'_{x-r}+u'_{x+r}=2u'_x+r^2\Delta^2 u'_{x-1}+\frac{r^2(r^2-1)}{12}\Delta^4 u'_{x-2}+\dots$$

$$=\left\{2+r^2b+\frac{r^2(r^2-1)}{12}b^2+\dots\right\}u'_x. \qquad \dots\dots(15)$$

Hence

$$K_0u'_x+K_1(u'_{x-1}+u'_{x+1})+K_2(u'_{x-2}+u'_{x+2})+\dots+K_r(u'_{x-r}+u'_{x+r})$$

$$=\left\{K_0+2\sum_1^r K_r+b\Sigma r^2K_r+b^2\Sigma\frac{r^2(r^2-1)}{12}K_r\right\}u'_x.$$

It should be remembered that we are now concerned with the underlying true values u'_x and not with the superimposed errors ϵ', although the coefficients are the same.

Owing to the construction of the formula

$$K_0+2(K_1+K_2+\dots+K_r)=1.$$

The second difference error is

$$\Sigma r^2K_rbu'_x. \qquad \dots\dots(16)$$

The fourth difference error is

$$\Sigma \frac{r^2(r^2-1)}{12} K_r b^2 u'_x, \qquad \text{......(17)}$$

where the summations extend from the central term of the expanded formula to either end and not over the entire range, so that each coefficient K occurs once only.

If there is no second difference error we have $\Sigma r^2 K_r = 0$ and the expression for the fourth difference error reduces to

$$\Sigma \frac{r^4}{12} K_r b^2 u'_x. \qquad \text{......(18)}$$

For numerical work it will often be found that the more general expression (17) is preferable, as $r^2(r^2-1)$ is always divisible by 12. An example of the use of this method will be given later.

19. Alternative method of finding the smoothing index.

If only the second and fourth difference errors and the smoothing index are required it is unnecessary to expand the formula.

The two errors can be found by the method described in para. 5, 6 and 7, while the smoothing index can be found by the following very elegant method which also throws considerable light on the construction of summation formulae. The method is due to G. J. Lidstone, who described it in two papers to be found in *J.I.A.* Vols. XLI, pp. 348 *et seq.* and XLII, pp. 106 *et seq.* Both these should be read as classic examples of actuarial literature and because of the masterly analysis of summation formulae which they contain.

$$\begin{aligned}[n] &\equiv E^{-\frac{n-1}{2}} + E^{-\frac{n-3}{2}} + \dots + E^{\frac{n-3}{2}} + E^{\frac{n-1}{2}} \\ &\equiv E^{-\frac{n-1}{2}}\{1 + E + E^2 + \dots + E^{n-1}\} \\ &\equiv \frac{E^n - 1}{\Delta E^{\frac{n-1}{2}}}. \qquad \text{......(19)}\end{aligned}$$

Hence a typical formula involving three summations in the operator, say,

$$\frac{[l][m][n]}{lmn}\{\text{operand}\},$$

may be written

$$\frac{(E^l-1)(E^m-1)(E^n-1)}{lmn\,\Delta^3 E^{\frac{l+m+n-3}{2}}}\{\text{operand}\}.$$

To find the smoothing index we find the third difference; i.e. we operate on the formula with Δ^3, giving

$$\frac{(E^l-1)(E^m-1)(E^n-1)}{lmn\,E^{\frac{l+m+n-3}{2}}}\{\text{operand}\}. \qquad \ldots\ldots(20)$$

As the operator in the denominator affects merely the suffix and not the coefficient of each term it can be ignored. We are left with

$$\frac{1}{lmn}\{E^{l+m+n}-E^{m+n}-E^{n+l}-E^{l+m}+E^l+E^m+E^n-1\}\{\text{operand}\}. \qquad \ldots\ldots(21)$$

It should be pointed out in passing that the practice of taking the *third* order of differences means that an operator involving three summations can be dealt with very easily. If only two summations are involved (e.g. Hardy's "wave-cutting" formula) it is easier to find the coefficients of Δ^2 by the method and then difference the result so as to give the coefficients of Δ^3.

From the above expression (21) it is a simple matter to evaluate the necessary coefficients of $\Delta^3\epsilon$ in terms of the ϵ''s. It should be borne in mind that in the second half of the expansion the coefficients are repeated in reverse order but with changed signs. If the range of the formula is R the number of terms in $\Delta^3\epsilon$ is $R+3$ (usually an even number), so that only the first $\frac{R+3}{2}$ need be evaluated if R is odd, or the first $\frac{R+4}{2}$ if R is even.

The following example illustrates the method:

Example 3.

Find the second and fourth difference errors and the smoothing index of Spencer's 21-term formula

$$\frac{[5]^2[7]}{350}\{[1]+[3]+[5]-[7]\}u_x.$$

All the required information can be obtained without expanding the formula.

It was shown on p. 185 that this formula has no second difference error but a fourth difference error of $-\frac{63}{5}b^2u_x$.

To find the smoothing index we proceed as follows:

$$\Delta^3\frac{[5]^2[7]}{350}\{[1]+[3]+[5]-[7]\}u'_x$$

$$=\Delta^3\frac{(E^5-1)^2(E^7-1)}{350\Delta^3E^7}\{-u'_{x-3}+u'_{x-1}+2u'_x+u'_{x+1}-u'_{x+3}\}.$$

Writing only coefficients of the operand so that E^7 in the denominator can be ignored we obtain:

$$\tfrac{1}{350}(E^{17}-2E^{12}-E^{10}+E^7+2E^5-1)(-1, 0, 1, 2, 1, 0, -1).$$

The expansion can be carried out as follows, evaluating only $\frac{R+3}{2}$, i.e. 12 terms:

E^{17}	-1	0	1	2	1	0	-1						
$-2E^{12}$						2	0	-2	-4	-2	0	2	…
$-E^{10}$								1	0	-1	-2	-1	…
E^7											-1	0	…
……	……												
……	……												
Total	-1	0	1	2	1	2	-1	-1	-4	-3	-3	1	…

The term in E^5 and the constant do not affect the first 12 terms. The sum of the squares of the coefficients

$$=\frac{2}{350^2}\{6\times 1^2+2\times 2^2+2\times 3^2+4^2\}=\frac{96}{350^2}.$$

Hence the smoothing index $=\frac{1}{350}\sqrt{\frac{96}{20}}=\cdot 00626$ or $\frac{1}{160}$ approx.

Example 4.

Analyse fully Spencer's 21-term formula.

For a full analysis it is necessary to expand the formula. This can be done as follows, remembering that only the first eleven coefficients are needed:

$[5]$ gives 1, 1, 1, 1, 1,
$[5]^2$ gives 1, 2, 3, 4, 5, 4, 3, 2, 1,
$[5]^2[7]$ gives 1, 3, 6, 10, 15, 19, 22, 23, 22, 19, 15 ….

Since the coefficients of the operand are −1, 0, 1, 2, 1, 0, −1, we have

−1	−1	−3	−6	−10	−15	−19	−22	−23	−22	−19	−15	...
1			1	3	6	10	15	19	22	23	22	...
2				2	6	12	20	30	38	44	46	...
1					1	3	6	10	15	19	22	...
−1							−1	−3	−6	−10	−15	...
Total	−1	−3	−5	−5	−2	6	18	33	47	57	60	...

As a check on the accuracy of the work we have:

$$\text{sum of the last line} = 2(-1-3-5-\ldots+47+57)+60 = 350,$$

so that the sum of the coefficients is unity, as it should be.

Note: although the denominator of 350 in the formula is left out of account in much of the numerical work it should never be overlooked.

Range. The range is 21 terms.

Coefficient curve. The coefficients progress quite smoothly and the peak of the curve is fairly broad. We should therefore expect good smoothing power and fair wave-cutting properties, in spite of the fact that the range of the operators is 5, 5 and 7, while for good wave-cutting the ranges should differ widely.

The second and fourth difference errors and the error-reducing index can conveniently be calculated at the same time as follows, where K_r has its usual meaning as a coefficient in the expanded formula.

r	$350K_r$	$r^2 \times 350K_r$	$\frac{r^2(r^2-1)}{12} \times 350K_r$	$(350K_r)^2$
(1)	(2)	(3)	(4)	(5)
10	−1	−100	−825	1
9	−3	−243	−1620	9
8	−5	−320	−1680	25
7	−5	−245	−980	25
6	−2	−72	−210	4
5	6	150	300	36
4	18	288	360	324
3	33	297	198	1089
2	47	188	47	2209
1	57	57	—	3249
0	60	—	—	3600
Total	—	−980 + 980 = 0	−5315 + 905 = −4410	3600 + 6971

Column (4) was derived from (3) by multiplying by $\frac{r^2-1}{12}$, as this reduces the numbers involved. Since $\Sigma r^2 K_r = 0$, we could however have calculated $\Sigma r^4 K_r$ to arrive at the fourth difference error (see formula 18).

Second difference error. This is equal to $\Sigma r^2 K_r = 0$.

Fourth difference error. This is given by $\Sigma \frac{r^2(r^2-1)}{12} K_r$ or by $\Sigma \frac{r^4}{12} K_r$ if $\Sigma r^2 K_r$ is zero.

From column (4) the fourth difference error

$$= -\frac{4410}{350} b^2 u_x = -\tfrac{63}{5} b^2 u_x, \text{ as before.}$$

It should be remembered, that in finding the second and fourth difference errors by this method, the summations extend over only half the whole range, i.e. from the centre to either end. This does not apply to any other index.

Error-reducing index. $\sqrt{\Sigma K_r^2}$ over the *whole* range $= \sqrt{\frac{2 \times 6971 + 3600}{(350)^2}}$.

Note: the total of the last column is shown in two parts, since the square of the central coefficient ($r = 0$) does not need to be doubled.

$$\therefore \text{ Error-reducing index} = \tfrac{1}{350}\sqrt{17542} = \cdot 378.$$

Wave-cutting index. The sum of the five central coefficients is $\frac{268}{350}$, or about ·766, indicating only very moderate wave-cutting power. (Cf. ·42 for Hardy's "wave-cutting" formula.)

Smoothing index. Since $\Delta^3 \equiv E^3 - 3E^2 + 3E - 1$, we find the coefficients of $\Delta^3 \epsilon$ as follows, writing only the first twelve values and ignoring the denominator of 350.

E^3	−1	−3	−5	−5	−2	6	18	33	47	57	60	57 ...
$-3E^2$		3	9	15	15	6	−18	−54	−99	−141	−171	−180 ...
$3E$			−3	−9	−15	−15	−6	18	54	99	141	171 ...
-1				1	3	5	5	2	−6	−18	−33	−47 ...
Total	−1	0	1	2	1	2	−1	−1	−4	−3	−3	1 ...

Hence the sum of the squares of the coefficients of $\Delta^3 \epsilon$

$$= \frac{2}{350^2}\{6 \times 1^2 + 2 \times 2^2 + 2 \times 3^2 + 4^2\} = \frac{96}{350^2}.$$

The smoothing index is therefore $\frac{1}{350}\sqrt{\frac{96}{20}}$ as before

$$= \cdot 00626 \quad \text{or} \quad \tfrac{1}{160} \text{ approx.}$$

20. Choice of the operators $[l]$, $[m]$, $[n]$.

The smoothing index is

$$\sqrt{\text{sum of squares of coefficients} \div 20}.$$

Hence for the formula

$$\frac{[l]\,[m]\,[n]}{lmn}\,\{\text{operand}\}$$

the smoothing index will have the product lmn in the denominator. For a given operand and a given range the sum $l+m+n$ is fixed, since

$$\text{range} = \text{range of operand} + l + m + n - 3 \quad \text{(see para. 8).}$$

We can increase the product lmn by making the factors more nearly equal, and it might at first seem that the most efficient formula would be obtained by making

$$l = m = n.$$

Although this tends to improve the error-reducing power it tends to impair the smoothing power and the same applies if two only of the values l, m, n are made equal.

To prove this we consider expression (21), which can be used in finding the smoothing index. The smaller the coefficients found by expanding this expression the better the smoothing power, especially when it is remembered that the coefficients have to be squared in finding the index.

Now the operator

$$E^{l+m+n} - E^{m+n} - E^{n+l} - E^{l+m} + E^{l} + E^{m} + E^{n} - 1$$

will have unit coefficients if l, m and n are unequal. If, however, two of them (say m and n) are equal it reduces to

$$E^{l+2m} - E^{2m} - 2E^{l+m} + E^{l} + 2E^{m} - 1,$$

which tends to produce greater coefficients when combined with the operand.

If $l=m=n$ it becomes $E^{3l} - 3E^{2l} + 3E^{l} - 1$, so that the coefficients in the expanded formula will be relatively large because of the coefficients 3 which are subsequently squared. It follows therefore that l, m and n should be nearly equal. [4] [5] [6] as used in Hardy's Friendly Society formula is therefore a good operator.

Vaughan has pointed out that the result is further improved if

$l+m=n$, since then the terms $-E^{l+m}$ and E^n cancel, leaving only six terms in the product.

He has devised some very interesting formulae on these lines and the reader is strongly advised to consult two papers which he read to the Institute (reproduced in *J.I.A.* Vols. LXIV, p. 428 and LXVI, p. 463).

One objection to formulae in which $l+m=n$ is that either one or all of the summations must be even; the effect of this is discussed in the next section.

A further point clearly brought out by Mr Vaughan's work is that a good operator and a good operand do not necessarily combine to produce a good formula; it is in fact impossible to predict how they will blend. There is no really satisfactory method of judging the efficacy of a formula except by analysing it fully as explained in previous sections. As regards its application to a particular experience the best test is the examination of the results that it produces.

21. Even summations.

As previously explained the result of a summation of an even number of terms is to produce a result lying midway between the two central ones. A second even summation brings the resulting values back into alignment, and hence a formula involving, say, the operators [4] [5] [6] produces values for integral arguments. If, for instance, we operate on crude values of q_x for integral ages the graduated values will also relate to integral ages. Occasionally, however, it may be an advantage to have one or three even summations; for example we can then make $l+m=n$, as mentioned above, and if we are operating on values for "ages last birthday" the formula will automatically produce values for exact integral ages, provided that it can be assumed that on the average

$$\text{exact age} = \text{age last birthday} + \tfrac{1}{2}.$$

One important disadvantage of even summations in general is that the coefficients tend to be more complicated and it is therefore difficult to eliminate a second difference error.

It will be remembered that the second difference error in $[n]u_x$ is $\frac{n(n^2-1)}{24}\,bu_x$. Now if n is odd, $n-1$ and $n+1$ are both even and one is a multiple of 4. Moreover, one of the consecutive numbers $n-1$, n, $n+1$ must be divisible by 3, so that the expression $\frac{n(n^2-1)}{24}$ is an integer. If n is even the numerator may only be divisible by 6, so that awkward fractions arise. In order to produce a convenient working formula it may be necessary to introduce a small second difference error. For instance, using the operators [4] [5] [6], Hardy produced the formula

$$\frac{[4]\,[5]\,[6]}{120}\{2\,[3]-[5]\}\,u'_x,$$

which has a second difference error of $\frac{1}{12}\Delta^2 u'_{x-1}$ or $\frac{1}{12}\delta u'_x$.

Actually a small positive second difference error may be an advantage, because the fourth difference error is always negative and the two tend to offset each other. Spencer in fact devised a formula for which the second and fourth difference errors would cancel in this way if the function operated on were of the form $A+Hx+BC^x$.

It is impossible to ensure that this will occur in general; much depends on the particular function under investigation.

If we consider the coefficient curve we can see easily why the fourth difference error is always negative. The negative values occur at each end where the values of r (in our previous notation) are greatest (see para. 13).

This ensures that $\Sigma r^2 K_r$ shall approximate to zero, thus producing little or no second difference error, but making $\Sigma \frac{r^2(r^2-1)}{12} K_r$ negative, since weighting with r^4 gives the negative values of K_r great importance.

22. Maximum smoothing power and maximum error-reducing power are mutually exclusive.

It is obvious that any formula which reduces the errors effectively will automatically produce smooth results. If a formula is devised so as to have the maximum error-reducing power for its range it will always be possible to construct a formula of the same range with greater smoothing power. The first concentrates on eliminating the errors; the second on smoothing them.

Consider, for instance, the formula

$$\frac{[5]^3}{125}\{xu_0+y(u_{-1}+u_1)+z(u_{-2}+u_2)\},$$

where x, y and z are to be determined so as to produce no second difference error and the minimum smoothing index.

The coefficients of $\Delta^3\epsilon$ are given by $\dfrac{(E^5-1)^3}{125}\{z+y+x+y+z\}$.

(Incidentally, the weakness of operators of equal range has been mentioned previously and the operator $[5]^3$ is unlikely to give good results.)

Writing down only the first ten coefficients and ignoring the denominators we obtain the following:

E^{15} gives $\qquad z+y+x+y+z,$

$-3E^{10}$ gives $\qquad -3z-3y-3x-3y-3z.$

The operators E^5 and the constant do not affect the first ten terms and the sum of the squares of the coefficients of $\Delta^3\epsilon$ is

$$\frac{2}{125^2}\{20z^2+20y^2+10x^2\},$$

giving a smoothing index of $\frac{1}{125}\sqrt{x^2+2y^2+2z^2}$.

Hence we have to make $x^2+2y^2+2z^2$ a minimum.

Differentiating with respect to x, we obtain

$$x+2y\frac{dy}{dx}+2z\frac{dz}{dx}=0. \qquad \text{......(22)}$$

Also, since the formula must reduce to the form $(1+Kb^2)$ if it has no second difference error, we must have

$$x+2y+2z=1 \qquad \text{......(23)}$$

and

$$y+4z=-3. \qquad \text{......(24)}$$

This last equation arises from the fact that the second difference term in the operand is $\qquad b(\Sigma r^2K_r),$

and this has to neutralize the second difference error of $3b$ in the operator.

Differentiating the last two equations with respect to x, we have

$$\left.\begin{aligned}1+2\frac{dy}{dx}+2\frac{dz}{dx}&=0\\ \text{and}\qquad \frac{dy}{dx}+4\frac{dz}{dx}&=0\end{aligned}\right\}.$$

Eliminating $\frac{dy}{dx}$ and $\frac{dz}{dx}$ from these two equations and equation (22) we obtain

$$3x-4y+z=0. \qquad \text{......(25)}$$

Equations (23), (24) and (25) can now be solved giving:

$$x=\tfrac{47}{35}, \quad y=\tfrac{27}{35}, \quad z=-\tfrac{33}{35}.$$

The formula is therefore

$$\frac{[5]^3}{4375}\{47u'_0+27(u'_{-1}+u'_1)-33(u'_{-2}+u'_2)\}.$$

In actual practice x would probably be taken as $\frac{3}{2}$, y as $\frac{3}{4}$ and z as -1, the theoretical result being too cumbersome.

For maximum error-reducing power the formula would have been expanded, and instead of $x^2+2y^2+2z^2$ the sum of the squares of the coefficients would have been made a minimum.

The resulting formula would have been different, thus illustrating the fact that maximum error-reducing power is incompatible with maximum smoothing power.

23. Recent practical developments.

The practical aspect of these formulae as distinct from the theoretical aspect had been very largely ignored until G. J. Lidstone and D. C. Fraser contributed two interesting notes to *J.I.A.* Vol. LXVII giving some neat labour-saving devices.

The reader is referred to the original articles for details, but in the numerical example which follows use has been made of an artifice suggested by Fraser. If the formula does not involve any second difference error then the values obtained by graduating

$$u_x-(a+bx+cx^2),$$

where a, b and c are constants, will be equal to the values obtained by graduating u_x and afterwards subtracting $(a+bx+cx^2)$. This means that the numerical values of u_x can be reduced appreciably by a suitable choice of the function $a+bx+cx^2$; after the remainders have been graduated we merely have to add back the values previously deducted to produce the required graduation.

If the formula involves a second difference error $a+bx$ can be deducted, but a term cx^2 will itself contribute a second difference error.

24. Graduation of the A 1924–29 Table, Ultimate Rates.

The rates for durations 3 and over, "All Classes Combined", in the A 1924–29 experience were given by the data for half-ages $20\frac{1}{2}$, $21\frac{1}{2}$, etc., and values for integral ages might have been obtained by the use of a summation formula involving one or three even summations. Actually, in order to produce results quickly, it was decided to use Spencer's 21-term formula, as this is one of the best formulae for general purposes. Values for integral ages were deduced by visual interpolation for most of the table, but where second and higher differences were appreciable a simple finite difference formula was used at higher ages.

The ends of the table were completed by third difference extrapolation. It is obvious that any summation formula of range R will leave $\frac{R-1}{2}$ terms at each end to be filled in by other methods. Sometimes Makeham's or Gompertz's law is assumed—particularly at high ages—but the precise method to be adopted depends on the run of the data, and for the A 1924–29 rates a finite difference method was found to be satisfactory. In this connection it should be pointed out that the report states that the graduation was "made to show a more distinct increase age by age in the rates of mortality above age 85 than that shown by the statistics".

As a numerical example we shall calculate the graduated rates (duration 3 and over) for ages $30\frac{1}{2}$ to $64\frac{1}{2}$, using the rough values of q_x for ages $20\frac{1}{2}$ to $74\frac{1}{2}$ for the combined data for the years 1927–9.

The ungraduated values of $10^5 \times q_x$ are shown in the first column. Spencer's 21-term formula is

$$\frac{[5]^2[7]}{350}\{[1]+[3]+[5]-[7]\}u'_x$$

$$=\frac{[5]^2[7]}{350}\{2u'_x+(u'_{x-1}+u'_{x+1})-(u'_{x-3}+u'_{x+3})\}.$$

Example 5. *Continuous Experience* 1927–9. *Durations* 3 *and over. All classes combined*

Age x	$10^5 \times q_x$	u_x $=(2)-f_x^*$	$2u_x+u_{x-1}$ $+u_{x+1}$	$u_{x-3}+u_{x+3}$	(4) − (5)	[7] (6)	[5] (7)	$\frac{(8)}{50}$	[5] (9)	$\frac{(10)}{7}$	(11) + f_x^*
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
20½	280	170									
21½	215	95	491								
22½	261	131	451								
23½	234	94	399	225	174						
24½	230	80	348	139	209						
25½	254	94	323	175	148						
26½	225	55	248	116	132	824					
27½	224	44	187	129	58	734					
28½	234	44	154	132	22	575	2,932	59			
29½	222	22	137	56	81	446	2,414	48			
30½	259	49	158	74	84	353	2,033	41	214	31	241
31½	258	38	126	76	50	306	1,741	35	184	26	246
32½	231	1	70	51	19	353	1,556	31	166	24	254
33½	270	30	93	54	39	283	1,452	29	157	22	262
34½	282	32	123	112	11	261	1,491	30	163	23	273
35½	289	29	95	26	69	249	1,595	32	187	27	287
36½	275	5	113	102	11	345	2,043	41	232	33	303
37½	354	74	178	116	62	457	2,745	55	298	43	323
38½	315	25	196	158	38	731	3,696	74	388	55	345
39½	372	72	253	138	115	963	4,785	96	494	71	371
40½	394	84	369	218	151	1,200	6,082	122	609	87	397
41½	449	129	475	190	285	1,434	7,332	147	729	104	424
42½	463	133	539	238	301	1,754	8,518	170	855	122	452
43½	484	144	586	338	248	1,981	9,712	194	992	142	482
44½	515	165	640	344	296	2,149	11,119	222	1,148	164	514
45½	526	166	751	393	358	2,394	12,928	259	1,335	191	551
46½	624	254	889	547	342	2,841	15,173	303	1,559	223	593

47½	595	215	944	625	319	3,563	17,860	357	1,820	260	640
48½	650	260	1,138	608	530	4,226	20,904	418	2,110	301	691
49½	803	403	1,526	778	748	4,836	24,170	483	2,428	347	747
50½	870	460	1,765	795	970	5,438	27,451	549	2,778	397	807
51½	862	442	1,868	909	959	6,107	31,030	621	3,171	453	873
52½	954	524	2,070	1,102	968	6,844	35,373	707	3,618	517	947
53½	1020	580	2,333	1,389	944	7,805	40,542	811	4,136	591	1031
54½	1099	649	2,577	1,589	988	9,179	46,504	930	4,751	679	1129
55½	1159	699	2,976	1,709	1,267	10,607	53,353	1067	5,478	782	1242
56½	1399	929	3,704	1,995	1,709	12,069	61,785	1236	6,329	904	1374
57½	1627	1147	4,408	2,064	2,344	13,693	71,688	1434	7,313	1045	1525
58½	1675	1185	4,932	2,545	2,387	16,237	83,125	1662	8,435	1205	1695
59½	1915	1415	5,430	3,000	2,430	19,082	95,706	1914	9,681	1383	1883
60½	1925	1415	6,091	3,523	2,568	22,044	109,447	2189	11,039	1577	2087
61½	2366	1846	7,178	3,646	3,532	24,650	124,107	2482	12,499	1786	2306
62½	2601	2071	8,364	4,252	4,112	27,434	139,595	2792	14,071	2010	2540
63½	2916	2376	9,284	4,613	4,671	30,897	156,109	3122	15,788	2255	2795
64½	3011	2461	10,135	5,185	4,950	34,570	174,288	3486	17,677	2525	3075
65½	3397	2837	11,333	6,162	5,171	38,558	195,306	3906			
66½	3768	3198	12,572	6,679	5,893	42,829	218,570	4371			
67½	3919	3339	13,967	7,726	6,241	48,452					
68½	4681	4091	15,824	8,304	7,520	54,161					
69½	4903	4303	17,962	9,579	8,383						
70½	5875	5265	20,300	10,006	10,294						
71½	6087	5467	22,580	11,921	10,659						
72½	7011	6381	24,896								
73½	7307	6667	27,545								
74½	8480	7830									

• It was difficult to find a satisfactory second-degree function to deduct from $10^5 \times q_x$ so as to reduce the numbers involved in the graduation. Eventually it was decided to deduct the linear function $110 + 10(x - 20\frac{1}{2}) = f_x$. To save space this function is not shown.

The same values of the function are added to the graduated values given in column (11), giving column (12), from which the rates for integral ages could be interpolated.

25. The A 1924–29 Table, Select rates.

It is convenient to deal here with the select portion of the A 1924–29 table although the rates were not graduated by a summation formula.

The select rates for durations 0, 1 and 2 were expressed as percentages of the ultimate rates for the same attained age.

Thus $q_{[47]+2}$ was expressed as a percentage of q_{49} ult. After an investigation of the actual percentages derived from the crude select rates it was found possible at duration 0 to assume a percentage of 61 at age 45, increasing by ·4 for each year of age under 45 to a maximum of 68·2 for ages 27 and under and decreasing by ·3 for each year of age over 45.

For duration 1 it was assumed for all ages that

$$q_{[x]+1}=\tfrac{1}{2}\left(q_{[x+1]}+q_{x+1}\right). \qquad \ldots\ldots(26)$$

For duration 2 it was assumed for all ages that

$$q_{[x]+2}=\cdot6q_{x+2}+\cdot4q_{[x+1]+1}. \qquad \ldots\ldots(27)$$

It will be noticed that, in each of these equations, the q's relate to the same attained age.

26. The χ^2 test applied to a graduation by a summation formula.

It is impossible by any analytical method to find what constraints are imposed by a summation formula, but in Seal's paper referred to earlier an experiment was carried out by the author with Spencer's 21-term formula. As a result he decided to assume that about five degrees of freedom were lost. The original paper should be referred to for details of the method. The assumption made later that Kenchington's formula results in the loss of six degrees of freedom is rather controversial and should be accepted with some reserve.

27. Advantages of the summation method of graduation.

Once a suitable formula has been chosen or constructed the process is purely mechanical and does not require a highly skilled operator as does the graphic method.

The method is suitable for standard tables based on large experiences and can be relied upon to give adequate smoothness, provided that the unadjusted rates themselves progress fairly smoothly. An advantage of the method is that results are produced quickly.

Perhaps its greatest merit, possessed by no other method, arises in connection with functions such as sickness rates.

Suppose that

$$f(x)=a_1\phi_1(x)+a_2\phi_2(x)+a_3\phi_3(x)+\dots+a_r\phi_r(x),$$

where the a's are constants and x is variable.

If the functions f, ϕ_1, ϕ_2, ... ϕ_r are graduated separately by the same summation formula it will be seen that the same equation will connect the graduated rates.

Sickness rates analysed according to period of attack (z^3, $z^{3/3}$, etc. in the usual notation) are from their very nature additive before graduation. For instance

$$z^3+z^{3/3}=z^6 \quad \text{and} \quad z^6+z^{6/6}=z^{12}.$$

If they are graduated by summation the resulting rates will still be additive because each graduated rate is a linear function of ungraduated rates (see para. 10).

Thus z^{12} (graduated) $= z^6$ (graduated) $+ z^{6/6}$ (graduated).

To illustrate this we shall consider the formula

$$u_x=K_{-r}u'_{-r}+K_{-r+1}u'_{-r+1}+\dots+K_{-1}u'_{-1}+K_0u'_0+K_1u'_1+\dots+K_{r-1}u'_{r-1}+K_ru'_r.$$

In a summation formula $K_{-t}=K_t$, but the argument still applies to the more general formula for which this relationship does not hold.

If we apply the same formula to graduate z^6 and $z^{6/6}$ we deduce that $z^6+z^{6/6}$ (graduated)

$$=K_{-r}(z'^6_{-r}+z'^{6/6}_{-r})+K_{-r+1}(z'^6_{-r+1}+z'^{6/6}_{-r+1})+\dots+K_{r-1}(z'^6_{r-1}+z'^{6/6}_{r-1})+K_r(z'^6_r+z'^{6/6}_r),$$

where all the functions on the right-hand side are ungraduated.

But

$$z'^6_{-r}+z'^{6/6}_{-r}=z'^{12}_{-r},$$
$$z'^6_{-r+1}+z'^{6/6}_{-r+1}=z'^{12}_{-r+1}.$$
$$\dots\dots\dots\dots\dots\dots$$

Hence the right-hand side reduces to

$$K_{-r}z'^{12}_{-r}+K_{-r+1}z'^{12}_{-r+1}+\dots+K_{r-1}z'^{12}_{r-1}+K_rz'^{12}_r=z^{12} \text{ (graduated).}$$

28. Disadvantages of the method.

(1) Although skill is needed in choosing a suitable formula, once the choice has been made there is no scope for individual judgment. As a result it is impossible to retain any special feature of the experience (such as a discontinuity in withdrawal rates) which the operator might feel to be an essential feature. It is certain to be greatly modified in the graduation and might even disappear altogether.

(2) Unless the unadjusted rates progress fairly smoothly the results will be unsatisfactory. This means in effect that the experience must be large.

(3) The ends of the table always have to be completed by some other method.

(4) The method cannot be used satisfactorily for select rates, and since it assumes that the function operated on has negligible fourth differences over the range of the formula, its use is in practice restricted to ratios such as q_x, μ_x, etc. It is therefore impossible to take into account the weight of the exposed to risk at each age.

29. Illustrative example.

We shall conclude this chapter with an example which illustrates many of the points discussed.

Example 6.

The following table is a representative extract of ten values from a complete table in which x ranges from 20 to 100. Column (2) gives the values of a certain function of x calculated by a mathematical formula. Column (3) gives the results of an experimental approximation thereto and differs from column (2) only in small superimposed errors. Columns (4) and (5) are the results of graduating column (3) by Woolhouse's formula and Spencer's 21-term formula.

(*a*) Test roughly the agreement between the theoretical smoothing index of each formula and the smoothing power as disclosed by the figures in the table. Give possible reasons for any anomaly.

(*b*) Suggest very briefly any reason for the fact (not confined to the ten values shown) that the graduation by Woolhouse's formula, which is theoretically less powerful than Spencer's, produces results nearer to the true values.

x	True value of function	Experimental value	Graduation of column (3) by	
			Woolhouse	Spencer
(1)	(2)	(3)	(4)	(5)
⋮	⋮	⋮	⋮	⋮
50	691	663	681	675
51	696	719	688	681
52	706	702	698	692
53	726	728	723	714
54	762	749	760	751
55	821	844	819	811
56	911	919	908	902
57	1041	1028	1040	1031
58	1221	1220	1218	1211
59	1462	1473	1457	1451
⋮	⋮	⋮	⋮	⋮

To deal with (*a*) we must first find the ungraduated errors ϵ', i.e. the differences between the true values given in column (2) and the observed values, and also the graduated errors, taken as the difference between the graduated and the true values. The results including the third differences are are shewn on p. 216.

ϵ_1 and ϵ_2 represent graduated errors produced by the use of Woolhouse's and Spencer's formulae respectively.

The smoothing index deals not with the third differences of the actual errors $\Delta^3\epsilon$ but with the standard deviation of the values of $\Delta^3\epsilon$ deduced from a great many applications of the formulae. As a rough test we may however compare the actual values of $\Delta^3\epsilon$ and $\Delta^3\epsilon'$ taken positively.

$\Sigma\Delta^3\epsilon'$ (disregarding signs) $= 456$.

$\Sigma\,|\,\Delta^3\epsilon_1|$ for Woolhouse graduation $= 34$.

$\Sigma\,|\,\Delta^3\epsilon_2|$ for Spencer graduation $\quad = 11$.

A comparison item by item would be useless.

Thus $\Sigma\,|\,\Delta^3\epsilon\,|$ has been reduced in the ratio $\frac{1}{13}$ approx. by Woolhouse's formula and in the ratio $\frac{1}{41}$ approx. by Spencer's formula.

Bearing in mind the limitations imposed on the test by the fact that only seven values are available, it may be said that the result for the first graduation may be said to be consistent with the smoothing index of $\frac{1}{15}$ calculated on p. 198.

On p. 201, however, we showed that the smoothing index for Spencer's formula is about $\frac{1}{160}$, so that even allowing for the roughness of the test the result of $\frac{1}{41}$ produced requires some explanation. An examination of the difference table shows that the numbers involved are very small and

				Woolhouse graduation				Spencer graduation			
Un-graduated errors $=\epsilon'$ (3)−(2)	$\Delta\epsilon'$	$\Delta^2\epsilon'$	$\Delta^3\epsilon'$	Graduated error $=\epsilon_1$ (4)−(2)	$\Delta\epsilon_1$	$\Delta^2\epsilon_1$	$\Delta^3\epsilon_1$	Graduated error $=\epsilon_2$ (5)−(2)	$\Delta\epsilon_2$	$\Delta^2\epsilon_2$	$\Delta^3\epsilon_2$
(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
⋮											
−28				−10				−16			
	51				2				1		
23		−78		−8		−2		−15		—	
	−27		111		—		7		1		1
−4		33		−8		5		−14		1	
	6		−54		5		−9		2		−2
2		−21		−3		−4		−12		−1	
	−15		72		1		3		1		1
−13		51		−2		−1		−11		—	
	36		−102		—		—		1		—
23		−51		−2		−1		−10		—	
	−15		45		−1		4		1		−2
8		−6		−3		3		−9		−2	
	−21		39		2		−7		−1		3
−13		33		−1		−4		−10		1	
	12		−33		−2		4		—		−2
−1		—		−3		—		−10		−1	
	12				−2				−1		
11				−5				−11			

indicate a lack of significant figures rather than deficiency in smoothing power. Had another significant figure been retained the result would almost certainly have been improved. Even if a series of values is ideally smooth the third differences will exhibit irregularities if the values are curtailed or rounded off, so that they no longer give the exact figures.

(*b*) It is debatable whether the word "powerful" applied to a summation formula refers to error-reducing power or smoothing power, properties which cannot both exist to the greatest degree in any one formula. In (*a*) we considered smoothing power, but error-reducing power remains to be examined. We have shown on pp. 195 and 203 that the error-reducing index of Woolhouse's formula is ·423, while that of Spencer's 21-term formula is ·378.

As a rough test we may consider the total of the errors regardless of sign, although the error-reducing index relates to standard deviations and not to the results of one experiment.

From the previous table we have

Total of ungraduated errors, irrespective of sign = 126.
Total of graduated errors, irrespective of sign (Woolhouse graduation) = 45.
Total of graduated errors, irrespective of sign (Spencer graduation) = 118.

x	True value u_x	Δu_x	$\Delta^2 u_x$	$\Delta^3 u_x$	$\Delta^4 u_x$
50	691				
		5			
51	696		5		
		10		5	
52	706		10		1
		20		6	
53	726		16		1
		36		7	
54	762		23		1
		59		8	
55	821		31		1
		90		9	
56	911		40		1
		130		10	
57	1041		50		1
		180		11	
58	1221		61		
		241			
59	1462				

As we know that Spencer's formula should reduce the unadjusted errors ϵ' to about one-third of their former value we are led to suspect that the total of 118 is due in part to distortion of the true values.

The table on p. 217 shows the differences of the true values.

From our previous work we know that neither formula introduces a second difference error. Woolhouse's has, however, a fourth difference error of $-\frac{27}{5}b^2u_x$, while Spencer's has a fourth difference error of $-\frac{63}{5}b^2u_x$.

Hence the graduated errors we have been considering are not the true errors which we previously denoted by ϵ, but are of the form $-5{\cdot}4+\epsilon$ (Woolhouse) and $-12{\cdot}6+\epsilon$ (Spencer). Fortunately the fourth difference errors are constant so that in the first part of the solution the columns headed $\Delta\epsilon$, $\Delta^2\epsilon$ and $\Delta^3\epsilon$ in the table are correct, although the values from which $\Delta\epsilon$ was obtained are misleading. The smoothness of the results is unaffected.

To find the graduated errors ϵ we must eliminate the fourth difference errors. The following values will be produced.

x	Woolhouse	Spencer
50	−5	−4
51	−3	−3
52	−3	−2
53	2	0
54	3	1
55	3	2
56	2	3
57	4	2
58	2	2
59	0	1
Total	27	20

It would be illogical to introduce decimals, as the values were all recorded as integers and the fourth difference errors were taken as -5 and -12.

If allowance is made for the small number of values tested the totals, irrespective of sign, viz. 27 and 20, are not inconsistent with the smoothing powers of the two formulae.

The relative power is more consistent with the error-reducing indices than the absolute power of the formulae.

BIBLIOGRAPHY

"Graduation by Summation." G. J. LIDSTONE. *J.I.A.* Vols. XLI and XLII.

"Summation formulae of graduation with a special type of operator, etc." H. VAUGHAN. *J.I.A.* Vols. LXIV and LXVI.

"The Arithmetic of Graduation by Summation." G. J. LIDSTONE. *J.I.A.* Vol. LXVII.

"A Note on Graduation Formulae." D. C. FRASER. *J.I.A.* Vol. LXVII.

"Tests of a Mortality Table Graduation." H. L. SEAL. *J.I.A.* Vol. LXXI.

"The Smoothing of Time Series." F. R. MACAULAY. National Bureau of Economic Research, N.Y.

EXAMPLES 8

1. Analyse and criticize fully the following summation formula, evaluating any criteria which would enable you to compare (*a*) its smoothing power, (*b*) its wave-cutting properties, with those of another formula

$$u_x = \frac{[2][3][5]}{360}\{-2u_0 + 15(u_1 + u_{-1}) - 8(u_2 + u_{-2})\}.$$

Illustrate its use by graduating the central terms of the following series:

112, 103, 124, 118, 106, 127, 115, 115,
127, 124, 112, 130, 127, 118, 133, 130.

2. Analyse and criticize the following summation formula:

$$\frac{[5][13]}{65}\{u_0 + (u_1 + u_{-1}) - (u_3 + u_{-3})\}.$$

3. Calculate the values of *l*, *m* and *n* in the summation formula

$$\frac{[5][7]}{35}\{l[3] + m[n]\}.$$

The formula contains 15 terms and does not involve any second difference error.

Show how to apply the formula by graduating the central terms of the series

0, 4, 4, 4, 5, 4, 2, 4, 3, 5, 6, 4, 7, 5, 3, 5, 7, 6, 9, 8, 5.

4. State briefly the theoretical basis of formulae for summation graduation and mention the circumstances in which the formulae give satisfactory results.

Find the missing part of the following 17-term formula, calculate the smoothing coefficient and discuss the merits and demerits of the formula as an instrument for graduation:

$$\frac{[5]\,[7]}{35}\{\tfrac{3}{5}[3]+?\}.$$

5. A summation formula of graduation has been wrongly written in the last two terms of the operand as

$$\frac{[5]\,[11]}{110}\{-[7]+2\,[5]-1\}\,u_0.$$

Correct the mistakes, analyse and criticize the corrected formula and compare it with any standard formula of the same type of which you are aware.

6. Explain the following statement, which refers to Woolhouse's summation formula

$$\frac{[5]^3}{125}\{-3u_{-1}+7u_0-3u_1\}.$$

"The errors in the ungraduated values are reduced by the graduation to about the values they would have in an ungraduated experience of five times the magnitude. The smoothness of the graduated curve would, however, be much greater than that of an ungraduated curve based on the larger experience."

7. In graduating mortality rates by summation formulae, what advantages are gained by graduating separately the exposed to risk and deaths?

Mention the principal objections to graduation by means of summation formulae. It has been suggested that graduated values of $q_{[x]+t}$ can be obtained from the graduated values of $q_{[x+t]}$ and q_{x+t} by use of the formula

$$q_{[x]+t}=a_t q_{[x+t]}+(1-a_t)\,q_{x+t}.$$

What is the rationale of the suggestion and how would you calculate the values of a_t if it were desired to secure equality between the expected deaths obtained by use of the graduated select q's and the actual deaths, in each year of Assurance within the select period?

8. (*a*) A mortality experience has been graduated by the following process:

(i) The exposed to risk were replaced by a smooth curve of similar shape calculated by a mathematical formula.

(ii) Adjusted deaths were obtained by multiplying the substituted exposed to risk by the ungraduated rates of mortality.

(iii) The adjusted deaths were graduated by a summation formula.

(iv) Graduated rates of mortality were obtained by dividing the graduated deaths by the substituted exposed to risk.

Discuss the advantages and disadvantages of the method.

(*b*) In the graduation of the above experience the unadjusted and the adjusted deaths over the range of ages from 45 to 70 were approximately constant at 200 deaths in each year of age. The summation formula used was

$$\frac{[5]\,[13]}{65}\{[3]+[5]-[7]\}\,u_x.$$

What are the approximate probable errors, expressed as percentages in the ungraduated and graduated rates of mortality at age 57?

9. The table below shows index-numbers of the price of a staple commodity over the years 1872 to 1911 (1891 = 1000). For the purpose of comparison with the trend, over the period, of the price of another commodity, the price of which exhibits similar features, it is suggested that the series be graduated by a summation formula. Do you approve?

Devise a formula which you would consider suitable in the circumstances and calculate its smoothing index.

Year	Index no.	Year	Index no.	Year	Index no.	Year	Index no.
1872	1356	1882	1235	1892	988	1902	907
1873	1452	1883	1255	1893	973	1903	882
1874	1493	1884	1266	1894	998	1904	880
1875	1519	1885	1288	1895	1024	1905	930
1876	1469	1886	1310	1896	1047	1906	950
1877	1404	1887	1276	1897	1069	1907	970
1878	1371	1888	1173	1898	1011	1908	990
1879	1352	1889	1121	1899	994	1909	961
1880	1311	1890	1070	1900	935	1910	915
1881	1250	1891	1000	1901	946	1911	895

10. What is the effect of applying a summation formula to a series which is already smooth? Illustrate your answer by considering the application of the formula

$$\frac{[4]\,[5]\,[6]}{120}\{2\,[3]-[5]\}$$

to a series whose *n*th term is of the form

$$a+bn+cn^2+dn^3+en^4.$$

CHAPTER IX

GRADUATION BY MATHEMATICAL FORMULAE. MAKEHAM AND ALLIED CURVES

1. Preliminary considerations.

Hitherto we have started from the data and derived a more or less smooth series of values from them. In this chapter we commence with a smooth curve and adjust the constants in the equation to the curve so as to secure the best adherence to data.

Before attempting this line of approach in practice it is necessary to bear in mind the source from which the crude data were derived, any heterogeneous features and other peculiarities.

Heterogeneity is usually the greatest stumbling-block, because any rate of decrement derived from data open to this criticism is unlikely to follow a single mathematical curve over its whole range. Further, any graduated table, however derived from the data, is liable to be suspect unless used in conjunction with a population similar in constitution to that on which the table is based.

As with the graphic method we can operate either on the exposed to risk and decrements separately or on the crude rates of decrement. The advantages and disadvantages of each method were discussed in Chapter VI. In dealing with mathematical formulae the graduation of exposed to risk and decrements separately has the added disadvantage that the rates finally obtained will be represented by complicated expressions difficult to handle in theoretical work, e.g. in dealing with joint-life functions.

2. Makeham and Gompertz curves.

The first important contribution towards finding a "law of mortality" was made by Benjamin Gompertz, who found that μ_x could be represented approximately by the formula Bc^x, i.e. by the successive terms of a geometric progression. He then proceeded to find the best values of the available constants B and c.

A development of Gompertz's law was subsequently made by Makeham, who adopted the now well-known formula

$$\mu_x = A + Bc^x.$$

With its three constants A, B and c this formula was found for most tables to give a satisfactory agreement with the facts, and more standard tables have been produced by its use than by any other method. A mass of literature, mathematical and otherwise, has grown up around the formula, and the student will be familiar with the ways in which joint-life functions and problems involving complicated multiple-life statuses can be dealt with if the table used has been graduated by Makeham's formula. So convenient is it that its adoption is often justified at the expense of a certain amount of distortion of the facts.

Of recent years it has been increasingly difficult to obtain satisfactory graduation by this simple formula. Allied formulae, such as

$$\mu_x = A + Hx + Bc^x,$$

have been tried without any very great success.

3. Preliminary tests.

Usually the rough data are given in quinquennial or decennial groups. Even when this is not so it is desirable to amalgamate the entries into one of these groups.

By doing this we reduce irregularities very considerably, particularly at the ends of the table, and it is much easier to form an opinion as to whether a Makeham graduation is likely to be successful by examining group rates than by examining rates for individual ages.

It will be assumed therefore that the exposed to risk and decrements are available in groups of five years.

If the exposed to risk is given in the initial form E_x, it must be adjusted to the central form E_x^c by the deduction of half the corresponding decrements.

Note. The c in E_x^c is not in any way connected with the c in Makeham's formula.

In Chapter I it was shown that Hardy's formula

$$u_0 = \frac{1}{n}\{w_0 - \tfrac{1}{24}\Delta^2 w_{-1}\}$$

could be applied to the central exposed to risk to give P_x for the central point of age of a group and to the deaths to give $P_x\mu_x$ for the same age. It cannot, however, be used for the exposed to risk in the "initial" form, since this is essentially a discontinuous function. This accounts for the change to the "central" exposed E_x^c.

The application of Hardy's formula then gives a succession of values of P_x and $P_x\mu_x$ and hence, by division, a set of values of μ_x.

From these values it is possible to form an opinion whether a Gompertz or a Makeham graduation is likely to prove successful.

If the ratio of successive terms is roughly constant a Gompertz curve is indicated, with μ_x of the form Bc^x. To test if a Makeham curve is appropriate we form the first differences of successive terms and then find the ratios of the differences. The first step eliminates the constant A and should produce values roughly in geometric progression.

Sometimes it is more convenient to deal with colog p_x instead of μ_x, and since these functions are of the same form the same process can be applied to both.

In the preliminary tests for the graduation of the $a(m)$ and $a(f)$ Tables, based on British Offices' annuity experience over the years 1900–20, the function colog p_x was used, with the following results:

Males

Age last birthday (1)	colog p_x (2)	Δ colog p_x (3)	$\frac{\text{colog } p_{x+5}}{\text{colog } p_x}$ (4)	$\frac{\Delta \text{ colog } p_{x+5}}{\Delta \text{ colog } p_x}$ (5)
50	·00436	·00176	1·40	2·01
55	·00612	·00354	1·58	1·26
60	·00966	·00446	1·46	1·32
65	·01412	·00588	1·42	1·96
70	·02000	·01152	1·58	1·79
75	·03152	·02056	1·65	1·28
80	·05208	·02623	1·50	1·34
85	·07831	·03520	1·45	1·62
90	·11351	·05719	1·51	1·68
95	·17070	·09630	1·56	—
100	·26700	—	—	—

The report comments on this table as follows: "It will be seen that column (4) would not be much distorted if an average value of 1·51 were assumed; this value is too high up to 65, then too low for the important ages 70 and 75 and a little too high afterwards. The average of column (5) is 1·58 (or 1·51 if we exclude the first and last entries), so that the evidence of the table leads us to say that the mortality follows the Gompertz law sufficiently nearly to justify an attempt at graduation and that c^5 is about 1·51, or log c about ·036...."

The table for females was as follows:

Females

Age last birthday (1)	colog p_x (2)	Δ colog p_x (3)	$\dfrac{\text{colog } p_{x+5}}{\text{colog } p_x}$ (4)	$\dfrac{\Delta \text{ colog } p_{x+5}}{\Delta \text{ colog } p_x}$ (5)
50	·00349	·00087	1·25	1·52
55	·00436	·00132	1·30	2·00
60	·00568	·00265	1·47	1·68
65	·00833	·00445	1·53	1·83
70	·01278	·00813	1·64	1·82
75	·02091	·01483	1·71	1·87
80	·03574	·02775	1·78	1·46
85	·06349	·04054	1·64	1·25
90	·10403	·05087	1·49	—
95	·15490	—	—	—

The report continues: "This table shows that the later values in column (4) are considerably greater than the earlier values. It follows, therefore, that there is little hope of a graduation on Gompertz's hypothesis. The next column, which supplies a test for Makeham's formula, is better and the values opposite ages 55 to 75 could be averaged at 1·84 without leading to distortion."

As will be seen later, it was found impossible to produce a satisfactory graduation for either sex, using a single curve for the whole range. The above tables are, however, interesting, since they show the sort of results that preliminary tests are likely to give in practice.

4. The Makeham constant c.

Tests on the above lines give an indication of the constant c, but the following method is better for finding a more exact value.

Suppose that by the use of Hardy's formula we have obtained crude values of μ_x for ages $27\frac{1}{2}$, $32\frac{1}{2}$, $37\frac{1}{2}$, ..., $62\frac{1}{2}$, $67\frac{1}{2}$, the values at younger and higher ages being unreliable.

It is desirable to base the value of c on all these observations. To do so we form three composite functions:

$$S_1=\mu_{27\frac{1}{2}}+3\mu_{32\frac{1}{2}}+5\mu_{37\frac{1}{2}}+6\mu_{42\frac{1}{2}}+5\mu_{47\frac{1}{2}}+3\mu_{52\frac{1}{2}}+\mu_{57\frac{1}{2}},$$
$$S_2=\mu_{32\frac{1}{2}}+3\mu_{37\frac{1}{2}}+5\mu_{42\frac{1}{2}}+6\mu_{47\frac{1}{2}}+5\mu_{52\frac{1}{2}}+3\mu_{57\frac{1}{2}}+\mu_{62\frac{1}{2}},$$
$$S_3=\mu_{37\frac{1}{2}}+3\mu_{42\frac{1}{2}}+5\mu_{47\frac{1}{2}}+6\mu_{52\frac{1}{2}}+5\mu_{57\frac{1}{2}}+3\mu_{62\frac{1}{2}}+\mu_{67\frac{1}{2}}.$$

The coefficients 1, 3, 5, 6, 5, 3, 1 are quite arbitrary and might equally well have been taken as 1, 2, 3, 4, 3, 2, 1. The reason for their introduction will be made clear at a later stage.

If $\mu_x=A+Bc^x$,

$$S_3-S_2=Bc^{32\frac{1}{2}}(c^5-1)(1+3c^5+5c^{10}+6c^{15}+5c^{20}+3c^{25}+c^{30})$$

and $$S_2-S_1=Bc^{27\frac{1}{2}}(c^5-1)(1+3c^5+5c^{10}+6c^{15}+5c^{20}+3c^{25}+c^{30}).$$

Hence $$\frac{S_3-S_2}{S_2-S_1}=c^5. \qquad \text{......(1)}$$

One reason for the introduction of the coefficients rising to a maximum in the centre and diminishing towards each end is that less weight is thus given to the values $\mu_{27\frac{1}{2}}$ and $\mu_{67\frac{1}{2}}$ which will generally be based on fewer data. Most weight is therefore attached to the values in the middle of the available range.

Another important point is that, without some such coefficients being introduced, S_3-S_2 would reduce to $\mu_{67\frac{1}{2}}-\mu_{32\frac{1}{2}}$ and S_2-S_1 to $\mu_{62\frac{1}{2}}-\mu_{27\frac{1}{2}}$, the remaining values disappearing. The estimate of c would not be based on all the available values of μ but on four values only.

Having obtained an estimate of c, $\log_{10} c$ is found and is usually rounded off to two or three significant figures.

5. The Makeham constants A and B.

Wherever possible it is desirable to allow for the size of the exposed to risk at each age or group of ages, and although this is impossible in finding a trial value of c it can be and should be done in finding the remaining constants A and B.

By the application of Hardy's formula we have a set of values of

$P_x\mu_x$ and P_x from which the values of μ_x used in the preceding section were derived.

If Makeham's law applies these are connected by the equation

$$P_x\mu_x = P_x(A+Bc^x). \qquad \ldots\ldots(2)$$

Once c has been found as above the only unknown quantities are A and B.

Thus, using any two ages, we could form a pair of simultaneous equations such as (2) and solve for A and B. This, however, would not make use of all the data so we use instead the equations

$$\Sigma P_x\mu_x = A\Sigma P_x + B\Sigma c^x P_x \qquad \ldots\ldots(3)$$

and

$$\Sigma\Sigma P_x\mu_x = A\Sigma\Sigma P_x + B\Sigma\Sigma c^x P_x. \qquad \ldots\ldots(4)$$

The first of these equations is formed by summing all the available values and the second by taking the second summations.

The two simultaneous equations can be solved for A and B.

The graduation can now be completed, and although there is no need to test for smoothness the usual tests for adherence to data must be applied. These may indicate unsatisfactory features, such as a large discrepancy between the third summations of the actual and expected deaths. In view of the way in which A and B were found the first and second summations should show no discrepancies.

As a result of the tests a new trial value of c may be chosen and the constants A and B re-calculated, thus giving a new graduation. Finally, the best value of c may be found by interpolation from the results of the two trials. For instance, if the discrepancies in the third summations of the actual and expected deaths are of opposite signs, a value of c might be adopted so as to make the total discrepancy approximately zero.

At the same time it should be borne in mind that the advantages of a Makeham graduation are so great that the statistical tests of adherence to data should not be applied too rigorously, and much greater discrepancies than would normally be allowed may well be counterbalanced by the practical convenience of the formula.

6. The Gompertz law.

As the Gompertz law is the Makeham law in the special case where $A=0$ any method used for the latter can be applied to the former.

The following simple method is, however, preferable.

Since $$P_x\mu_x = Bc^x P_x,$$

we have $$\log P_x\mu_x = x\log c + \log B + \log P_x. \qquad \text{......(5)}$$

Summing for all available values we have

$$\left.\begin{aligned} \Sigma\log P_x\mu_x &= \log c\Sigma x + \Sigma\log B + \Sigma\log P_x \\ \Sigma\Sigma\log P_x\mu_x &= \log c\Sigma\Sigma x + \Sigma\Sigma\log B + \Sigma\Sigma\log P_x \end{aligned}\right\}. \qquad \begin{aligned}&\text{......(6)}\\&\text{......(7)}\end{aligned}$$

Thus $\log c$ and B can be found in one process.

$\Sigma\log B$ is simply $n\log B$, where n is the number of observations summed; and

$$\Sigma\Sigma\log B = \log B(1+2+3+\ldots n) = \frac{n(n+1)}{2}\log B.$$

Another method is described in the report on the graduation of the $a(m)$ and $a(f)$ Tables. This method does not make any allowance for the weight of the exposed to risk at each age and may for this reason occasionally give poor results.

7. Example 1.

Assuming that preliminary tests have indicated that a Makeham graduation is likely to be successful, graduate the following data in that way:

Table XX

Age-groups	Exposed to risk	Deaths
40–44	15,518	65
45–49	19,428	144
50–54	21,594	219
55–59	21,890	378
60–64	19,174	465
65–69	15,775	557
70–74	11,414	685
75–79	6,993	644
80–84	3,276	471
85–89	1,096	217
90–94	201	67

In the absence of information to the contrary it must be assumed that the exposed to risk are in the "initial" form and must first be expressed in the "central" form by deducting half the deaths.

Hardy's formula $u_0 = \frac{1}{n}\{w_0 - \frac{1}{24}\Delta^2 w_{-1}\}$ then gives P_x and $P_x \mu_x$ for the central point of age of each group.

The work can be arranged as follows:

Table XXI

Age-group (1)	Exposed to risk E_x^c (2)	Δ (2) (3)	Δ^2 (2) (4)	$5P_x$ (5)	Central point of age x (6)
40–44	15,485·5				
		3,870			
45–49	19,356		−1,742	19,430	$47\frac{1}{2}$
		2,128			
50–54	21,484·5		−1,911	21,564	$52\frac{1}{2}$
		217			
55–59	21,701		−2,976	21,825	$57\frac{1}{2}$
		−2,759			
60–64	18,941·5		− 686	18,970	$62\frac{1}{2}$
		−3,445			
65–69	15,496·5		− 980	15,537	$67\frac{1}{2}$
		−4,425			
70–74	11,071·5		24	11,070	$72\frac{1}{2}$
		−4,401			
75–79	6,671		770	6,639	$77\frac{1}{2}$
		−3,631			
80–84	3,040·5		1,578	2,975	$82\frac{1}{2}$
		−2,053			
85–89	987·5		1,233	936	$87\frac{1}{2}$
		− 820			
90–94	167·5				

If the differences are set out as above $\Delta^2 w_{-1}$ appears on the same line as w_0. Column (5) is obtained by subtracting $\frac{1}{24}$th of the entries in column (4) from the corresponding entries in column (2). Although decimals were retained in column (2) to show more clearly how the figures were derived, all other entries have been rounded off to the nearest integer.

Table XXII is obtained in a similar manner to Table XXI. In that Table column (6) is derived by dividing the entries in column (5) by the corresponding entries in column (5) of Table XXI.

Table XXII

Age-group (1)	Deaths θ_x (2)	Δ (2) (3)	Δ^2 (2) (4)	$5P_x\mu_x$ (5)	μ_x (6)	Central point of age x (7)
40–44	65					
		79				
45–49	144		−4	144	·0074	$47\frac{1}{2}$
		75				
50–54	219		84	215	·0100	$52\frac{1}{2}$
		159				
55–59	378		−72	381	·0175	$57\frac{1}{2}$
		87				
60–64	465		5	465	·0244	$62\frac{1}{2}$
		92				
65–69	557		36	555	·0357	$67\frac{1}{2}$
		128				
70–74	685		−169	692	·0625	$72\frac{1}{2}$
		−41				
75–79	644		−132	650	·0979	$77\frac{1}{2}$
		−173				
80–84	471		−81	474	·1593	$82\frac{1}{2}$
		−254				
85–89	217		104	213	·2276	$87\frac{1}{2}$
		−150				
90–94	67					

As we now have nine ungraduated values of μ_x from which to find a trial value of c, we proceed as follows:

$$S_1 = \mu_{47\frac{1}{2}} + 3\mu_{52\frac{1}{2}} + 5\mu_{57\frac{1}{2}} + 6\mu_{62\frac{1}{2}} + 5\mu_{67\frac{1}{2}} + 3\mu_{72\frac{1}{2}} + \mu_{77\frac{1}{2}} = \cdot 7352,$$

$$S_2 = \mu_{52\frac{1}{2}} + 3\mu_{57\frac{1}{2}} + 5\mu_{62\frac{1}{2}} + 6\mu_{67\frac{1}{2}} + 5\mu_{72\frac{1}{2}} + 3\mu_{77\frac{1}{2}} + \mu_{82\frac{1}{2}} = 1\cdot 1642,$$

$$S_3 = \mu_{57\frac{1}{2}} + 3\mu_{62\frac{1}{2}} + 5\mu_{67\frac{1}{2}} + 6\mu_{72\frac{1}{2}} + 5\mu_{77\frac{1}{2}} + 3\mu_{82\frac{1}{2}} + \mu_{87\frac{1}{2}} = 1\cdot 8392.$$

Hence $$c^5 = \frac{S_3 - S_2}{S_2 - S_1} = \frac{\cdot 6750}{\cdot 4290} = 1\cdot 573,$$

giving $$\log_{10} c = \cdot 03936.$$

For convenience we take

$$\log_{10} c = \cdot 04.$$

Then, by using five-figure logarithms, $P_{47\frac{1}{2}}c^{47\frac{1}{2}}$, $P_{52\frac{1}{2}}c^{52\frac{1}{2}}$, etc. can readily be found as follows:

x (1)	$\log c^x P_x$ (2)	$\frac{c^x P_x}{10^2}$ (3)	$\frac{\Sigma c^x P_x}{10^2}$ (4)	P_x (5)	ΣP_x (6)	$P_x \mu_x$ (7)	$\Sigma P_x \mu_x$ (8)
$47\frac{1}{2}$	5·48951	3,087	3,087	3,886	3,886	29	29
$52\frac{1}{2}$	5·73478	5,430	8,517	4,313	8,199	43	72
$57\frac{1}{2}$	5·93999	8,710	17,227	4,365	12,564	76	148
$62\frac{1}{2}$	6·07910	11,998	29,225	3,794	16,358	93	241
$67\frac{1}{2}$	6·19233	15,572	44,797	3,107	19,465	111	352
$72\frac{1}{2}$	6·24516	17,585	62,382	2,214	21,679	138	490
$77\frac{1}{2}$	6·22319	16,719	79,101	1,328	23,007	130	620
$82\frac{1}{2}$	6·07452	11,872	90,973	595	23,602	95	715
$87\frac{1}{2}$	5·77184	5,914	96,887	187	23,789	43	758
—	—	96,887	432,196	23,789	152,549	758	3425

The figures P_x and $P_x \mu_x$ are one-fifth of the values in column (5) of Tables XXI and XXII. Although it is not necessary to divide by 5 in this way in finding A and B, it has been done to draw the attention of the reader to the fact that the factor $1/n$ in Hardy's formula had previously been ignored.

The equations for A and B are

$$\left.\begin{aligned} 758 &= 23{,}789A + 96{,}887 \times 10^2 B \\ 3425 &= 152{,}549A + 432{,}196 \times 10^2 B \end{aligned}\right\}. \qquad \ldots\ldots(8)$$

The solutions of the equations are $A = \cdot 000{,}910$, $B = \cdot 000{,}076$.

So that: $\mu_x = \cdot 000{,}910 + \cdot 000{,}076 c^x$

where $\log_{10} c = \cdot 04.$

The following graduated values are then easily calculated:

Central age x	$47\frac{1}{2}$	$52\frac{1}{2}$	$57\frac{1}{2}$	$62\frac{1}{2}$	$67\frac{1}{2}$	$72\frac{1}{2}$	$77\frac{1}{2}$	$82\frac{1}{2}$	$87\frac{1}{2}$
μ_x	·00695	·01048	·01607	·02494	·03900	·06128	·09659	·15255	·24124
Expected deaths $5P_x\mu_x$	135	226	351	473	606	678	641	454	226

The total of the Expected deaths, 3790, is one more than the total of the corresponding ungraduated values. The student will find it an instructive exercise to test this graduation thoroughly by the methods explained in Chapter V.

8. Curves allied to the Makeham curve.

When the general trend of mortality altered and it was no longer possible to graduate successfully by Makeham's law, attempts were made to modify the formula. The most important modification is

$$\mu_x = A + Hx + Bc^x. \qquad \ldots\ldots(9)$$

To fit this curve to the data the work is the same as before as far as the calculation of the rough values of μ_x.

The first differences of these values are of the form $\alpha + \beta c^x$, where α and β are constants, and by working with first differences instead of μ's the constant c can be found in the usual way.

A, H and B can then be found from the equations

$$\left.\begin{aligned} \Sigma P_x \mu_x &= A\Sigma P_x + H\Sigma x P_x + B\Sigma c^x P_x \\ \Sigma\Sigma P_x \mu_x &= A\Sigma\Sigma P_x + H\Sigma\Sigma x P_x + B\Sigma\Sigma c^x P_x \\ \Sigma\Sigma\Sigma P_x \mu_x &= A\Sigma\Sigma\Sigma P_x + H\Sigma\Sigma\Sigma x P_x + B\Sigma\Sigma\Sigma c^x P_x \end{aligned}\right\} \ldots\ldots(10)$$

Another assumption which has been experimented with is

$$\mu_x = ma^x + nb^x. \qquad \ldots\ldots(11)$$

To find the constants a and b we need four values of μ_x at equal intervals. We might for instance group the data decennially instead of quinquennially; convenient age-groups are 25–35, 35–45, 45–55 and 55–65.

The application of Hardy's formula will then enable crude values of μ_x for ages 30, 40, 50 and 60 to be found:

$$\left.\begin{aligned} \mu_{30} &= ma^{30} + nb^{30} = L + M \\ \mu_{40} &= ma^{40} + nb^{40} = L\kappa + M\lambda \\ \mu_{50} &= ma^{50} + nb^{50} = L\kappa^2 + M\lambda^2 \\ \mu_{60} &= ma^{60} + nb^{60} = L\kappa^3 + M\lambda^3 \end{aligned}\right\} \quad \begin{aligned} &(\text{where } \kappa = a^{10}, \\ &\quad \lambda = b^{10}). \end{aligned} \qquad \begin{aligned} &\ldots\ldots(12) \\ &\ldots\ldots(13) \\ &\ldots\ldots(14) \\ &\ldots\ldots(15) \end{aligned}$$

From these we obtain

$$\left.\begin{aligned} &\mu_{30}\mu_{50} - \mu_{40}^2 &&= LM(\kappa^2 + \lambda^2 - 2\kappa\lambda) &&= LM(\kappa - \lambda)^2 \\ &\mu_{40}\mu_{60} - \mu_{50}^2 &&= LM(\kappa\lambda^3 + \kappa^3\lambda - 2\kappa^2\lambda^2) &&= LM(\kappa - \lambda)^2 \kappa\lambda \\ &\text{and} \\ &\mu_{30}\mu_{60} - \mu_{40}\mu_{50} &&= LM(\kappa^3 + \lambda^3 - \kappa^2\lambda - \kappa\lambda^2) &&= LM(\kappa - \lambda)^2(\kappa + \lambda) \end{aligned}\right\}, \qquad \ldots\ldots(16)$$

$$\left.\begin{aligned} \text{giving} \qquad \kappa\lambda &= \frac{\mu_{40}\mu_{60} - \mu_{50}^2}{\mu_{30}\mu_{50} - \mu_{40}^2} \\ \text{and} \qquad \kappa + \lambda &= \frac{\mu_{30}\mu_{60} - \mu_{40}\mu_{50}}{\mu_{30}\mu_{50} - \mu_{40}^2} \end{aligned}\right\}. \qquad \ldots\ldots(17)$$

Since the μ's are known κ and λ can be found and thence a and b.

The remaining constants m and n are best calculated from the equations

$$\left.\begin{aligned}\Sigma P_x\mu_x&=m\Sigma a^xP_x+n\Sigma b^xP_x\\ \Sigma\Sigma P_x\mu_x&=m\Sigma\Sigma a^xP_x+n\Sigma\Sigma b^xP_x\end{aligned}\right\}. \qquad ......(18)$$

There is an infinity of curves of this general type which can be fitted to mortality rates.

The general procedure is usually the same. Crude values of μ_x are calculated after suitable groupings of the exposed to risk and deaths, and constants having the age as index (such as c in Makeham's formula) are calculated from the values of μ, suitable arbitrary weights being introduced where possible to lessen the importance of the μ's based on more scanty data at the ends of the experience. Considerable ingenuity is often required and no general rules can be given.

Once these constants having the age as index are known, or at least suitable trial values adopted, the remaining constants can be found by equating the first, second and even third summations of the actual and expected deaths, as in equations (10) above. It is usually unwise to introduce summations beyond the third, since, apart from the large numbers involved, too much importance is thereby attached to the rates at high ages where the rates are least reliable. As an alternative, summations over half the range might be used instead of a single summation over the whole range.

9. Application to m_x and q_x.

Sometimes it is desired to fit a curve to m_x, the central rate of mortality, instead of to μ_x. The only modification of the above is in the initial stage. From grouped data we can always find the central term by the formula

$$u_0=\frac{1}{n}\left\{w_0-\frac{n^2-1}{24n^2}\Delta^2w_{-1}\right\},$$

where n is the class interval.

Hence from the grouped values of E^c_x and θ_x we obtain not P_x and $P_x\mu_x$ at the central point of age but E^c_x and θ_x for the central age, e.g. age 32 for the quinquennial group 30–35. Crude values of m_x are thus obtained and the rest of the work is as before.

Similarly, if grouped values of E_x and not E_x^c are used, the values of E_x for the central age obtained by the use of the formula enable q_x to be calculated. It will be remembered that the formula

$$u_0 = \frac{1}{n}\left\{w_0 - \frac{n^2-1}{24n^2}\Delta^2 w_{-1}\right\}$$

can be applied to any function, continuous or otherwise, provided that fourth and higher differences are negligible.

10. Perks's formulae.

In a paper in *J.I.A.* Vol. LXIII, W. Perks has given some interesting formulae which have produced good results with modern data and which represent the most promising attempt of recent years to fit a single curve to the whole range of a table. The paper should be read carefully by anyone interested in modern developments.

The principal formulae discussed are

$$\mu_x(\text{or } q_x) = \frac{A+Bc^x}{1+Dc^x} \quad \text{(both functions are used)}$$

and

$$\mu_x = \frac{A+Bc^x}{Kc^{-x}+1+Dc^x}.$$

Perks himself assumed an arbitrary value of c, but it is possible to calculate one from the data on the lines of the previous examples. As an exercise the reader is advised to evolve a suitable method.

The remaining constants can be found as usual by the method of successive summations, although the denominator presents difficulty. To overcome this the equations can be written in the form

$$\left.\begin{aligned} \mu_x + Dc^x\mu_x &= A+Bc^x \\ \text{and} \qquad Kc^{-x}\mu_x + \mu_x + Dc^x\mu_x &= A+Bc^x \end{aligned}\right\}.$$

First, second and, if necessary, third summations can then be formed, and if c is known the remaining constants can readily be found.

A word of warning is advisable in this connection. The third summation is open to objection because of the weight given to the observations at ages remote from the mean; this applies with still more force to the fourth summation.

To make up the required number of equations therefore it may be necessary to take the second and third summations over each half of the range. This is to some extent fitting the curve in sections rather than as a whole, but the tests for adherence to data will reveal any weakness in the results.

11. Example 2.

The following is a good example of how a variable denominator can be dealt with.

Graduate the withdrawal rate in the following schedule by the formula

$$w_n = \frac{a}{b+n}.$$

Duration n	Exposed to risk of withdrawal E_n	Withdrawals W_n	Rate of withdrawal $w_n = \frac{W_n}{E_n}$
0	1600	240	·150
1	1800	162	·090
2	1800	117	·065
3	1600	80	·050
4	1200	54	·045
5	800	28	·035
6	300	9	·030
7	100	4	·040

The exposed to risk varies from 1800 at durations 1 and 2 to 100 at duration 7, so that any attempt to find a and b from the rates given in the last column would be unlikely to give good results unless proper allowance were made for the weight of the data at each duration.

It is easier to deal with the values of E_n and W_n rather than the rates w_n and we first write the equation

$$\frac{W_n}{E_n} = \frac{a}{b+n}$$

in the form

$$(b+n)\,W_n = aE_n.$$

By summing twice, we obtain the equations for a and b:

$$\left.\begin{aligned} b\Sigma W_n + \Sigma n W_n &= a\Sigma E_n \\ b\Sigma\Sigma W_n + \Sigma\Sigma n W_n &= a\Sigma\Sigma E_n \end{aligned}\right\}.$$

The necessary calculations are as follows:

n (1)	E_n (2)	ΣE_n (3)	W_n (4)	ΣW_n (5)	nW_n (6)	ΣnW_n (7)	Graduated rate (8)
0	1600	1,600	240	240	—	—	·1465
1	1800	3,400	162	402	162	162	·0911
2	1800	5,200	117	519	234	396	·0661
3	1600	6,800	80	599	240	636	·0518
4	1200	8,000	54	653	216	852	·0427
5	800	8,800	28	681	140	992	·0362
6	300	9,100	9	690	54	1046	·0315
7	100	9,200	4	694	28	1074	·0278
Total	9200	52,100	694	4478	1074	5158	—

Hence and

$$\left.\begin{aligned} 694b + 1074 &= 9200a \\ 4478b + 5158 &= 52,100a \end{aligned}\right\},$$

giving

$$\left.\begin{aligned} a &= \cdot 2407 \\ b &= 1\cdot 6428 \end{aligned}\right\}.$$

The graduated rates can then be calculated and the usual tests for adherence to data carried out. The graduated rates are shown in the last column above.

12. Example of the χ^2 test.

The application of the χ^2 test to a graduation by curve-fitting presents no special features, except that each constant in the equation found from the given data results in the loss of one degree of freedom.

For instance, in the example in the previous paragraph two degrees of freedom were lost as a and b were found from the given data.

We should therefore proceed in amalgamating the data for durations 6 and 7.

Since the number of cells is 7 and two constraints have been imposed there are five degrees of freedom.

The value of χ^2 at the foot of the last column is ·570 and when there are five degrees of freedom the probability of obtaining a value equal to or larger than this is about ·99. The fit therefore seems too good to be true, due no doubt to the fact that the question is artificial and was not based on actual data which, by virtue of the case, would have included sampling errors such as are met with in practice. No actuary, however, is likely to reject a graduation by a formula method because the fit seems too good.

n	E_n	Graduated rate q_n	Expected withdrawals	Actual withdrawals	(4) − (3)	$(5)^2$	$E_n p_n q_n$	(6)/(7)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0	1600	·1465	234·4	240	5·6	31·4	200·1	·157
1	1800	·0911	164·0	162	−2·0	4·0	149·0	·027
2	1800	·0661	119·0	117	−2·0	4·0	111·1	·036
3	1600	·0518	82·9	80	−2·9	8·4	78·6	·107
4	1200	·0427	51·2	54	2·8	7·8	49·0	·156
5	800	·0362	29·0	28	−1·0	1·0	28·0	·036
6, 7	400	·0306*	12·2	13	·8	·6	11·8	·051
Total	9200	—	692·7	694	1·3	—	—	·570

13. The N.H.I. Table.

For the purposes of the National Health Insurance Act 1911 a table of mortality (males and females separately) was required for the calculation of Reserve Values. The table was based on the recorded deaths in England and Wales for the three years 1908–10 and an estimated population as at 30th June 1909.

The results of the Census as at 31st March 1911 were not available, but an estimate in decennial age-groups was provided. The required figures for 30th June 1909 were interpolated from these figures and the corresponding figures for the 1901 Census on the assumption that the numbers increased in arithmetical progression. Thus the 1909 figure was taken as

$$\text{population } 1901 + \cdot 825\,(\text{population } 1911 - \text{population } 1901).$$

The figures operated upon were not the usual grouped populations and deaths but the population at ages x and over and the deaths occurring at ages x and over as used in plotting an ogive curve. The processes used are very clearly set out in the part of the Report

* Allowing for the weight of the exposed to risk at durations 6 and 7 the graduated rate on amalgamation was taken as $\frac{3}{4}\cdot 0315 + \frac{1}{4}\cdot 0278$.

which is reproduced in *J.I.A.* Vol. XLVII, pp. 548–59. The reader is strongly advised to study this extract, which explains both the theoretical and practical aspects very lucidly.

14. The $O^{[M]}$, $O^{M(5)}$ and $O^{[NM]}$ Tables.

The $O^{[M]}$ table was based on the experience of whole-life with profit policies over the years 1863–93 and has a select period of ten years. The ultimate part of the table is therefore sometimes called the $O^{M(10)}$ table.

The $O^{M(5)}$ and O^{M} tables are both interesting in that the full select period of ten years used in the $O^{[M]}$ table was abandoned. For the O^{M} table, sometimes known as the O^{M} aggregate table, the data for all durations were amalgamated, while for the $O^{M(5)}$ the experience for the first five years after entry was excluded. Thus the O^{M}, $O^{M(5)}$ and $O^{M(10)}$ tables differ from each other because of the exclusion of data relating to the early durations.

The $O^{[NM]}$ table was based on the experience of whole-life non-profit policies over the years 1863–93. In this experience it was found that a select period of five years was appropriate.

The $O^{M(10)}$, $O^{M(5)}$ and the ultimate portion of the $O^{[NM]}$ tables were all graduated by Makeham's formula with $\log_{10} c$ taken as ·039.

For the select portion of the $O^{[NM]}$ table it was found possible to assume that

$$\mu_{[x]+t} = A_t + B_t c^x,$$

where A_t and B_t are independent of x.

For details the reader is referred to *J.I.A.* Vol. XXXVIII, pp. 501 et seq., reproduced in Reprints 1935.

15. Re-graduation of the O^{M} Table by Makeham's formula.

The O^{M} table was not graduated by Makeham's formula. For special purpose Mr G. J. Lidstone re-graduated it by that formula, although he realized that a certain amount of distortion would arise. The method used is interesting in that it was not necessary to refer to the original data; further, it had the merit of giving speedy results. Its use should, however, be restricted to the type of problem for which it was derived and the graduation of rough data should be carried out by the methods previously described. Mr Lidstone put

$$\operatorname{colog}_{10} p_x = \alpha + \beta c^x$$

and took the values of $\text{colog}_{10} p_x$ from the O$^{\text{M}}$ table that he was re-graduating.

He then used the following three equations for finding α, β and c:

$$\sum_{x=25}^{64} \text{colog}_{10} p_x = 40\alpha + \beta c^{25}\frac{c^{40}-1}{c-1},$$

$$\sum\sum_{x=25}^{64} \text{colog}_{10} p_x = \frac{40\times 41}{2}\alpha + \beta c^{25}\frac{40c^{40}-\dfrac{c^{40}-1}{c-1}}{c-1},$$

and

$$\sum\sum\sum_{x=25}^{64} \text{colog}_{10} p_x = \frac{40\times 41\times 42}{2\times 3}\alpha + \beta c^{25}\frac{\dfrac{40\times 41}{2}c^{40}-\dfrac{40c^{40}-\dfrac{c^{40}-1}{c-1}}{c-1}}{c-1},$$

where the successive summations of the tabular values were taken for ages 25 to 64 inclusive. This was the range of ages for which it was desired to fit the Makeham curve; higher ages were relatively unimportant.

16. The use of two curves. Blending.

Often, particularly in recent years, it has been found impossible to fit a single curve to the whole of the data, although one curve may have been satisfactory at the younger ages and a second curve at the higher ages. Clearly such a graduation is not as satisfactory as when a single curve is used; it has, however, the great advantage that the rates progress smoothly. The chief difficulty is in passing from one curve to the other and this brings us to the question of *blending* and *blending functions*.

17. Blending functions.

Suppose that a curve has been fitted to the data at the younger ages, giving graduated values

$$u_0^a,\ u_1^a,\ \dots\ u_r^a,\ u_{r+1}^a,\ \dots\ u_{s-1}^a,$$

and a second curve at the higher ages, giving graduated values

$$u_{r+1}^b,\ u_{r+2}^b,\ \dots\ u_{s-1}^b,\ u_s^b,\ \dots\ u_n^b.$$

In other words the two curves are assumed to overlap and there are two graduated values for

$$u_{r+1},\ u_{r+2},\ \dots\ u_{s-1}.$$

The problem is to combine or fuse these two values in such a way that the final values pass smoothly from the first curve, which gives values up to and including u^a_r, to the second curve, which gives values from u^b_s onwards.

Assume that a typical blended function u'_{r+t} is given by the equation

$$u'_{r+t}=\kappa_{r+t}\,u^a_{r+t}+\lambda_{r+t}\,u^b_{r+t}, \qquad \ldots\ldots(19)$$

where κ_{r+t} and λ_{r+t} are not constants but functions of t.

If u'_{r+t} is to be a blend of varying proportions of the two u's we make

$$\kappa_{r+t}+\lambda_{r+t}=1. \qquad \ldots\ldots(20)$$

Also, in order to make the blended function merge with the main graduations at each end ($t=0$ and $t=s-r$), we have

$$\left.\begin{array}{ll}\kappa_r=1, & \lambda_r=0\\ \kappa_s=0, & \lambda_s=1\end{array}\right\}. \qquad \ldots\ldots(21)$$

Between these extremes κ_{r+t} should clearly diminish steadily and λ_{r+t} increase; and it is usual, though not essential, to make

$$\kappa_{r+t}=\lambda_{s-t}. \qquad \ldots\ldots(22)$$

In this event the values of λ are merely the values of κ in reverse order.

κ_{r+t} (and therefore λ_{r+t}) should be a continuous function. Further, since it is unity if t is zero or negative and commences to diminish as t becomes positive, it is essential in order to make a smooth transition that $\dfrac{d\kappa_{r+t}}{dt}$ should be zero when $t=0$, and that $\dfrac{d^2\kappa_{r+t}}{dt^2}$, and preferably $\dfrac{d^3\kappa_{r+t}}{dt^3}$, should be either zero or small for the same value, so that the blending curve has little curvature at either end.

To produce a smooth transition from one curve to the other the range of values over which blending is carried out must be fairly large. Much will depend on the differences between the pairs of overlapping values and also on the differences in gradient and curvature of the two main curves at the ends of the blending range.

18. The curve of sines.

A natural blending function is the sine of an angle because of its smoothness and the zero gradient when the angle is a multiple of π.

In order to make κ_{r+t} unity when $t=0$ and zero when $t=s-r$, we put

$$\left.\begin{aligned}\kappa_{r+t}&=\frac{1}{2}\left\{1+\sin\left(\frac{\pi}{2}+\frac{t\pi}{s-r}\right)\right\}\\&=\frac{1}{2}\left\{1+\cos\frac{t\pi}{s-r}\right\}\\ \text{Then}\qquad\lambda_{r+t}&=\frac{1}{2}\left\{1-\cos\frac{t\pi}{s-r}\right\}\end{aligned}\right\}.\qquad\ldots\ldots(23)$$

$$\begin{aligned}\text{Hence}\qquad\lambda_{s-t}&=\frac{1}{2}\left\{1-\cos\frac{(s-r-t)\pi}{s-r}\right\}\\&=\frac{1}{2}\left\{1+\cos\frac{t\pi}{s-r}\right\}\\&=\kappa_{r+t}.\end{aligned}$$

The values of λ are therefore the same as the values of κ, but in reverse order.

We can therefore concentrate our attention on κ_{r+t}.

$$\frac{d}{dt}\kappa_{r+t}=-\frac{1}{2}\cdot\frac{\pi}{s-r}\sin\frac{t\pi}{s-r}$$

and vanishes for the values $t=0$ and $t=s-r$.

$$\frac{d^2}{dt^2}\kappa_{r+t}=-\frac{1}{2}\left(\frac{\pi}{s-r}\right)^2\cos\frac{t\pi}{s-r}.$$

When $t=0$, this becomes $\frac{-\pi^2}{2(s-r)^2}$ or $\frac{-5}{(s-r)^2}$ approximately; when $t=s-r$, it has the same value but the opposite sign.

Provided that $s-r$ is fairly large (say 10 or over) the curvature of the blending function is therefore small at both ends of the range.

19. The curve of squares.

The curve $y=\kappa x^2$ is a parabola. If κ is negative the position of the curve is as in Fig. 4, while if κ is positive the position of the curve is as in Fig. 5.

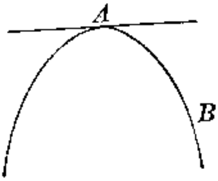


Fig. 4.

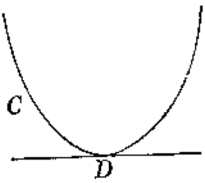
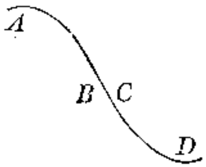


Fig. 5

If we take the section AB of the first curve and run it into the section CD of the second we obtain a curve of the form



which should be satisfactory for blending.

To do this we take

$$\left.\begin{aligned}\kappa_{r+t}&=1-2\left(\frac{t}{s-r}\right)^2\\ \lambda_{r+t}&=2\left(\frac{t}{s-r}\right)^2\end{aligned}\right\}\ \begin{gathered}\text{for the portion } AB\\ \text{when } t\leqslant\tfrac{1}{2}(s-r)\end{gathered}\qquad\ldots\ldots(24)$$

and

$$\left.\begin{aligned}\kappa_{r+t}&=2\left(1-\frac{t}{s-r}\right)^2\\ \lambda_{r+t}&=1-2\left(1-\frac{t}{s-r}\right)^2\end{aligned}\right\}\ \begin{gathered}\text{for the portion } CD\\ \text{when } t\geqslant\tfrac{1}{2}(s-r).\end{gathered}\qquad\ldots\ldots(25)$$

It will be noticed that, as before, $\lambda_{s-t}=\kappa_{r+t}$, so that the values of λ and κ are the same but in reverse order.

κ_{r+t} decreases from unity when $t=0$ to zero when $t=s-r$.

$\dfrac{d\kappa_{r+t}}{dt}=-\dfrac{4t}{(s-r)^2}$ if $t\leqslant\frac{1}{2}(s-r)$, and therefore vanishes when $t=0$; and

$=-\dfrac{4}{s-r}\left(1-\dfrac{t}{s-r}\right)$ if $t\geqslant\frac{1}{2}(s-r)$, and therefore vanishes when $t=s-r$.

$\dfrac{d^2\kappa_{r+t}}{dt^2}=-\dfrac{4}{(s-r)^2}$ when $t\leqslant\frac{1}{2}(s-r)$ and $\dfrac{4}{(s-r)^2}$ when $t\geqslant\frac{1}{2}(s-r)$.

Here again a reasonable value of $s-r$ will make the curvature quite small at both ends of the range.

When $t=\frac{1}{2}(s-r)$, i.e. at the point B or C in the third diagram,

$$\kappa_{r+t} \text{ is continuous (value } \tfrac{1}{2});$$

$$\frac{d}{dt}\kappa_{r+t} \text{ is continuous } \left(\text{value} -\frac{2}{s-r}\right).$$

For the curve AB, $\dfrac{d^2}{dt^2}\kappa_{r+t} = -\dfrac{4}{(s-r)^2}$,

whereas for the curve CD,

$$\frac{d^2}{dt^2}\kappa_{r+t} = \frac{4}{(s-r)^2}.$$

From the theoretical point of view, therefore, the curve of sines, which is a natural blending function, is to be preferred to the more artificial curve of squares.

In practice the use of either method is likely to give similar results.

20. Polynomial blending functions.

The polynomial $\kappa_{r+t} = a + bt + ct^2 + dt^3$ can be used as a blending function if the values of a, b, c and d are found as follows.

Since $\kappa_{r+t} = 1$, when $t = 0$ and 0 when $t = s - r$;

$$a = 1 \quad \text{and} \quad a + b(s-r) + c(s-r)^2 + d(s-r)^3 = 0.$$

Also, since $\dfrac{d}{dt}\kappa_{r+t} = 0$ at both ends of the range,

$$b + 2ct + 3dt^2 = 0, \quad \text{when} \quad t = 0 \quad \text{or} \quad s - r.$$

Solving, we find that

$$\left.\begin{aligned}\kappa_{r+t} &= 1 - 3\left(\frac{t}{s-r}\right)^2 + 2\left(\frac{t}{s-r}\right)^3\\ \lambda_{r+t} &= 3\left(\frac{t}{s-r}\right)^2 - 2\left(\frac{t}{s-r}\right)^3\end{aligned}\right\}. \qquad \ldots\ldots(26)$$

Finally,

$$\frac{d^2}{dt^2}\kappa_{r+t} = 2c + 6dt$$

$$= -\frac{6}{(s-r)^2}, \quad \text{when} \quad t = 0$$

and

$$= \frac{6}{(s-r)^2}, \quad \text{when} \quad t = s - r.$$

Thus the curve of squares and, still more, the curve of sines are superior as regards curvature at the ends of the range of blending. Nevertheless as a natural, instead of a hybrid, blending function the polynomial possesses certain advantages over the curve of squares.

21. Modified blending functions.

In using a blending function κ_{r+t} for values of t from 0 to $s-r$ we are in effect assuming that κ_{r+t} is unity if t is negative and zero if t is greater than $s-r$. Hence we must take care to ensure zero gradient and small curvature when the value begins to change.

In Tables XXIII to XXVI the values of the functions for $t=-1$ and $t=13$ are inserted to illustrate this point.

In order to ease the junction with the main curves some writers prefer to effect the blending from

$$t=\tfrac{1}{2} \quad \text{to} \quad t=s-r-\tfrac{1}{2},$$

i.e. over $s-r-1$ values instead of $s-r$.

Thus
$$\kappa_{r+\frac{1}{2}}=1, \quad \kappa_{s-\frac{1}{2}}=0.$$
$$\frac{d}{dt}\kappa_{r+t}=0 \quad \text{if} \quad t=\tfrac{1}{2} \quad \text{or} \quad s-r-\tfrac{1}{2}.$$

The following values of κ_{r+t} result:

curve of sines: $$\frac{1}{2}\left\{1+\cos\frac{(t-\frac{1}{2})\pi}{s-r-1}\right\};$$

curve of squares:

$$1-2\frac{(t-\frac{1}{2})^2}{(s-r-1)^2} \quad \text{if} \quad t \leqslant \tfrac{1}{2}(s-r)$$

or $$2\left\{1-\frac{t-\frac{1}{2}}{s-r-1}\right\}^2 \quad \text{if} \quad t \geqslant \tfrac{1}{2}(s-r);$$

polynomial: $$1-3\left(\frac{t-\frac{1}{2}}{s-r-1}\right)^2+2\left(\frac{t-\frac{1}{2}}{s-r-1}\right)^3.$$

Table XXVI shows the values for the curve of squares when $s-r=12$. The values $\Delta\kappa_{r+t}$ and $\Delta^2\kappa_{r+t}$ will enable the reader to judge to what extent the junction with the main curves is eased by shortening the blending range.

Table XXIII. *Blending function—curve of sines*

$(s-r=12)$

$$\kappa_{r+t}=\frac{1}{2}\left[1+\cos\frac{t\pi}{12}\right]$$

t	κ_{r+t}	$\Delta\kappa_{r+t}$	$\Delta^2\kappa_{r+t}$
(-1)	(1·000)		(·000)
		(·000)	
0	1·000		(−·017)
		−·017	
1	·983		−·033
		−·050	
2	·933		−·029
		−·079	
3	·854		−·025
		−·104	
4	·750		−·017
		−·121	
5	·629		−·008
		−·129	
6	·500		−·000
		−·129	
7	·371		+·008
		−·121	
8	·250		+·017
		−·104	
9	·146		+·025
		−·079	
10	·067		+·029
		−·050	
11	·017		+·033
		−·017	
12	·000		(+·017)
		(·000)	
(13)	(·000)		(·000)

Table XXIV. *Blending function—curve of squares*

$(s-r=12)$

$$\kappa_{r+t}=1-2\left(\frac{t}{12}\right)^2=1-\frac{t^2}{72},$$

when $t\leqslant\frac{1}{2}(s-r)$;

$$\kappa_{r+t}=2\left(1-\frac{t}{12}\right)^2=\frac{(12-t)^2}{72},$$

when $t\geqslant\frac{1}{2}(s-r)$

t	κ_{r+t}	$\Delta\kappa_{r+t}$	$\Delta^2\kappa_{r+t}$
(-1)	(1·000)		(·000)
		(·000)	
0	1·000		(−·014)
		−·014	
1	·986		−·028
		−·042	
2	·944		−·027
		−·069	
3	·875		−·028
		−·097	
4	·778		−·028
		−·125	
5	·653		−·028
		−·153	
6	·500		·000
		−·153	
7	·347		+·028
		−·125	
8	·222		+·028
		−·097	
9	·125		+·028
		−·069	
10	·056		+·027
		−·042	
11	·014		+·028
		−·014	
12	·000		(+·014)
		(·000)	
(13)	(·000)		(·000)

For explanation of figures in brackets see p. 244.

Table XXV. *Blending function—polynomial*

$$(s-r)=12$$

$$\kappa_{r+t}=1-3\left(\frac{t}{12}\right)^2+2\left(\frac{t}{12}\right)^3$$

$$=1-\frac{t^2}{48}+\frac{t^3}{864}$$

t	κ_{r+t}	$\Delta\kappa_{r+t}$	$\Delta^2\kappa_{r+t}$
(−1)	(1·000)		(·000)
		(·000)	
0	1·000		(−·020)
		−·020	
1	·980		−·034
		−·054	
2	·926		−·028
		−·082	
3	·844		−·021
		−·103	
4	·741		−·014
		−·117	
5	·624		−·007
		−·124	
6	·500		·000
		−·124	
7	·376		+·007
		−·117	
8	·259		+·014
		−·103	
9	·156		+·021
		−·082	
10	·074		+·028
		−·054	
11	·020		+·034
		−·020	
12	·000		(+·020)
		(·000)	
(13)	(·000)		(·000)

Table XXVI. *Modified blending function—curve of squares*

$$(s-r=12)$$

$$\kappa_{r+t}=1-2\left(\frac{t-\frac{1}{2}}{11}\right)^2,\ \text{when}\ t\leqslant\tfrac{1}{2}(s-r)$$

$$\kappa_{r+t}=2\left(1-\frac{t-\frac{1}{2}}{11}\right)^2,\ \text{when}\ t\geqslant\tfrac{1}{2}(s-r)$$

t	κ_{r+t}	$\Delta\kappa_{r+t}$	$\Delta^2\kappa_{r+t}$
(−1)	(1·000)		(·000)
		(·000)	
(0)	(1·000)		(−·004)
		(−·004)	
1	·996		(−·029)
		−·033	
2	·963		−·033
		−·066	
3	·897		−·033
		−·099	
4	·798		−·033
		−·132	
5	·666		−·034
		−·166	
6	·500		·000
		−·166	
7	·334		+·034
		−·132	
8	·202		+·033
		−·099	
9	·103		+·033
		−·066	
10	·037		+·033
		−·033	
11	·004		(+·029)
		(−·004)	
(12)	(·000)		(+·004)
		(·000)	
(13)	(·000)		(·000)

For explanation of figures in brackets see p. 244.

22. Limitations of blending—alternative methods

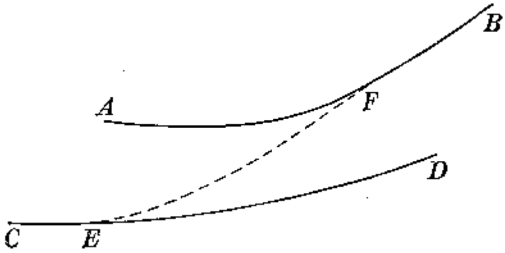


Fig. 6.

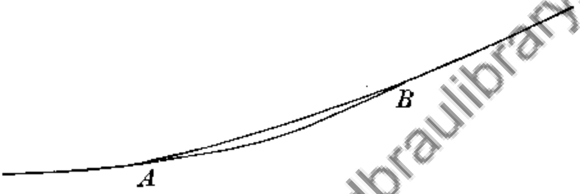


Fig. 7.

Blending can be relied upon to give good results if the two main curves do not intersect as in Fig. 6 or if they intersect twice as in Fig. 7. The curve represented by the dotted line *EF* should have the same tangent as *CD* at *E* and the same tangent as *AB* at *F*.

In the second instance the curve passes through *A* and *B*.

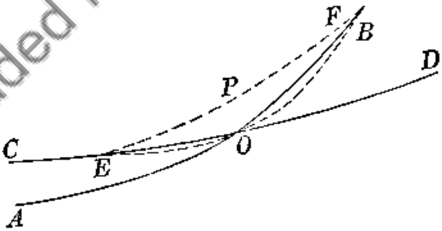


Fig. 8.

When the two main curves intersect only once, as at the point *O* in Fig. 8, a blending process is unlikely to give the best results. A good blend is indicated by a curve such as *EPF*, but any blending process normally employed makes the curve pass through the point of intersection *O* and drags it out of its natural course. (This does not apply of course if there are two points of intersection.)

Generally speaking, therefore, when the two main curves intersect only once near the point where blending is to be effected it is preferable to pass from one curve to the other by a process of osculatory interpolation. This process is described in *Mathematics for Actuarial Students*, Part II, Chapter VII, and will be met with again in Chapter X of this book. Briefly it may be said to be a process of interpolation which ensures a smooth join at each end of the interval in which the values are being inserted.

23. Offices Annuitants Experience, 1900–20. The a (m) and a (f) Tables.

The rates of mortality adopted as a basis were obtained by a process of extrapolation from the rates for the 1863–93 experience and those for 1900–20. They were therefore fairly smooth without any graduation, but apart from the desire for more places of decimals it was decided for reasons of practical convenience to attempt to fit a Makeham or Gompertz curve to the rates.

The problem was more akin to the re-graduation of an existing table than to the graduation of rough data. In any event the exposed to risk and deaths for the years 1900–20 did not relate to the extrapolated rates which formed the basis of the tables.

The constants were therefore found from the rough rates (colog p_x) and not by reference to observed data.

More than one attempt was made unsuccessfully and for details the reader is referred to the official report, *The Mortality of Annuitants*, 1900–1920.

The function operated on was colog p_x, which was assumed to be of the form $A + Bc^x$. A and B therefore have not the same meaning as the constants in the formula for μ_x.

The methods of graduation finally adopted were as follows:

Females. A Makeham curve was fitted to the data at ages 55, 60, 65, 70, 75 and 80.

This curve gave good values also for ages 50 and 85, but at higher ages it overstated the mortality as had the other curves experimented with.

A second Makeham curve was therefore found which would reproduce the values for ages 80 and 85 given by the other curve and also give a value for age 100 approximating to the ungraduated rate at that age. The constants were found direct from the three equations

$$\left.\begin{aligned} \operatorname{colog} p_{80} &= A + Bc^{80} \\ \operatorname{colog} p_{85} &= A + Bc^{85} \\ \operatorname{colog} p_{100} &= A + Bc^{100} \end{aligned}\right\}.$$

The two curves consequently overlapped over the range 80–85 and intersected at each end. A blending method was therefore adopted. Ages last birthday were used in the investigation and it was found that on the average the exact age exceeded the age last birthday by $4\frac{1}{2}$ months or ·375 year. The points of intersection of the graduating curves were therefore given by exact ages 80·375 and 85·375.

This accounts for the factor $\left(\dfrac{85{\cdot}375 - x}{5}\right)^2$ in the following table, which shews how the blending was effected.

Table XXVII

Exact age x	Main graduation	Old age graduation	Difference (3) − (2)	Adjusting factor $\left(\dfrac{85{\cdot}375 - x}{5}\right)^2$	(4) × (5)	Blended value (3) − (6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
81	·038534	·038883	·000349	·7656	·000267	·038616
82	·043159	·043906	·000747	·4556	·000340	·043566
83	·048383	·049286	·000903	·2256	·000204	·049082
84	·054285	·055057	·000772	·0756	·000058	·054999
85	·060952	·061234	·000282	·0056	·000002	·061232

Males. As the preliminary tests had suggested a Gompertz graduation it was assumed that

$$\operatorname{colog} p_x = Bc^x.$$

Hence $$\log(\operatorname{colog} p_x) = \log B + x \log c,$$

and $$\sum_{-n}^{n} \log(\operatorname{colog} p_x) = (2n+1)\log B,$$

for an odd number of terms, or

$$= 2n \log B, \qquad \text{......(27)}$$

for an even number of terms.

Also $$\sum_{-n}^{n} x \log(\operatorname{colog} p_x) = \log c \sum_{-n}^{n} x^2. \qquad \ldots\ldots(28)$$

Taking the origin at age 75 and using 5 years as unit the work was as follows:

Table XXVIII

Age	log (colog p_x)	n	n log (colog p_x)	
			+	−
50	− 2·360	−5	11·800	
55	− 2·213	−4	8·852	
60	− 2·015	−3	6·045	
65	− 1·850	−2	3·700	
70	− 1·699	−1	1·699	
75	− 1·501	0		
80	− 1·283	1		1·283
85	− 1·106	2		2·212
90	− 0·945	3		2·835
95	− 0·768	4		3·072
100	− 0·573	5		2·865
	−16·313		32·096 − 12·267 = 19·829	

The equations for B and c were therefore

$$11 \log B = -16{\cdot}313 \quad \text{or} \quad \log B = -1{\cdot}48300;$$

and since $\sum_{-5}^{5} n^2 = 2(1^2+2^2+3^2+4^2+5^2) = 110,$

equation (28) became

$$110 \log C = 19{\cdot}829 \quad \text{or} \quad \log C = {\cdot}180264,$$

where $C = c^5$.

Hence $$\log c = {\cdot}0360528.$$

This graduation was unsatisfactory, and an attempt was made to fit a Makeham curve by using the equations

$$\sum_{-n}^{n} \operatorname{colog} p_x = (2n+1)A + B \sum_{-n}^{n} c^x \qquad \ldots\ldots(29)$$

and $$\sum_{-n}^{n} x \operatorname{colog} p_x = B \sum_{-n}^{n} xc^x. \qquad \ldots\ldots(30)$$

Taking 5 years as unit and $C=c^5=1{\cdot}58$ the work was arranged as follows:

Table XXIX

Age	colog p_x	n	n colog p_x		C^n	nC^n	
			−	+		−	+
50	·00436	−5	·02180		·102	·510	
55	·00612	−4	·02448		·160	·640	
60	·00966	−3	·02898		·254	·762	
65	·01412	−2	·02824		·401	·802	
70	·02000	−1	·02000		·633	·633	
75	·03152	0	—		1·000	—	
80	·05208	1		·05208	1·580		1·580
85	·07831	2		·15662	2·496		4·992
90	·11351	3		·34053	3·944		11·832
95	·17070	4		·68280	6·232		24·928
100	·26700	5		1·33500	9·847		49·235
	·76738		−·12350 + 2·56703 = 2·44353		26·649	−3·347 + 92·567 = 89·220	

The equations for the constants were therefore

$$89{\cdot}220B=2{\cdot}44353 \quad \text{and} \quad 11A+26{\cdot}649B={\cdot}76738.$$

This graduation was also unsuccessful. The above tables have been reproduced to show how practical work should be set out and to emphasize that to find the constants the rates and not the exposed to risk and deaths were used.

Eventually one Gompertz curve was fitted at ages 50, 55, 60, 65 and 70 and a second Gompertz curve at ages 80, 85, 90, 95 and 100.

Actually the first curve was continued up to age 81 and the second carried back to age 70, so that the position was as shown in Fig. 6.

The rates (or rather values of log (colog p_x)) were made to relate to exact ages on the same assumption as before and were then blended by the curve of squares as shewn in Table XXX.

The central entry in column (2) is $\sqrt{\cdot 5}$ and the other entries are $\frac{1}{10}$, $\frac{3}{10}$, $\frac{5}{10}$, $\frac{7}{10}$ and $\frac{9}{10}$ of this value. Column (7) represents the additions to be made to the values of Graduation I for ages 71–75 and the deductions to be made from the values of Graduation II for ages 76–80.

Table XXX

Age	Factor	(Factor)^2 h	log (colog p_x)		Difference	Difference × h	Interpolated log (colog p_x)
			Graduation I	Graduation II			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
71	·0707107	·005	$\bar{2}$·3307	$\bar{2}$·38432	·05362	·00027	$\bar{2}$·33097
72	·2121321	·045	·3644	·41944	·05504	·00248	·36688
73	·3535535	·125	·3981	·45456	·05646	·00706	·40516
74	·4949749	·245	·4318	·48968	·05788	·01418	·44598
75	·6363963	·405	·4655	·52480	·05930	·02402	·48952
...	·7071069	·500					
76	·6363963	·405	·4992	·55992	·06072	·02459	·53533
77	·4949749	·245	·5329	·59504	·06214	·01523	·57981
78	·3535535	·125	·5666	·63016	·06356	·00795	·62221
79	·2121321	·045	·6003	·66528	·06498	·00292	·66236
80	·0707107	·005	·6340	·70040	·06640	·00033	·70007

The rates at ages under 50 were not taken from Graduation I, which gave very low rates, but were arranged so that they ran reasonably as compared with the female rates.

24. Government Life Annuitants Table, 1900–20.

Because of the practical advantages where joint-life functions are concerned an attempt was made to fit a single curve to the whole of the data. This, however, was unsuccessful. The rates at young ages were very low, particularly for females, and increased slowly up to age 70 and from age 90 onwards.

Between ages 70 and 90 the rise was fairly rapid.

As with a Makeham curve the accumulated deviations were too large to be ignored, the double Gompertz curve, for which

$$\operatorname{colog} p_x = Ma^x + Nb^x,$$

was tried, but although a fairly good fit was obtained for the female experience between ages 50 and 90 the constant b was less than unity. This meant that the second term had little effect over age 70 but made the rates below age 46 or 47 increase with a decrease in age. The male experience was still less adaptable.

Finally, the following method was used.

Males. A Makcham curve was fitted at ages 44 to 70 and an ordinary Gompertz curve at ages 64 to 89. These curves intersected once only at about age 65 and had to be blended over the range 61 to 69 inclusive.

The curve of sines was used with $\kappa = \frac{1}{2}\left(1 + \cos\frac{n\pi}{10}\right)$, but as was to be expected with the curves intersecting only once the results were not entirely satisfactory. It is understood they were subsequently hand-polished to produce a better transition from one curve to the other.

At ages over 90 the agreement with the crude rates was unsatisfactory and an attempt to improve it by re-calculating the constants of the curve merely produced distortion elsewhere. Finally, the limiting age was taken as 103 and the rates for ages over 90 were inserted so as to give a reasonable agreement with the ungraduated rates.

Females. A Makeham curve was fitted at ages 40 to 67 and a Gompertz curve at ages 65 to 86. The curves intersected once between ages 66 and 67 and the overlapping values of colog p_x at ages 63 to 71 were blended by the curve of sines as before. The same difficulty arose at high ages and the rates at ages over 90 were inserted in the light of the ungraduated rates, the limiting age being taken as 105.

To produce a table extending to young ages, English Life Table No. 8 was used (males and females separately) to find values of q_x for ages 5 and 25. The values for ages 45 and 46 were taken from the main graduation, thus ensuring a fairly smooth join.

It was assumed that q_x at ages under 46 was a polynomial of the third degree, and the rates from age 6 to age 44 were inserted by assuming constant third differences.

q_5 (from English Life No. 8) $= u_0$,

q_{25} ,, ,, $= (1 + 20\Delta + 190\Delta^2 + 1140\Delta^3)\,u_0$,

q_{45} (from main graduation) $= (1 + 40\Delta + 780\Delta^2 + 9880\Delta^3)\,u_0$,

q_{46} ,, ,, $= (1 + 41\Delta + 820\Delta^2 + 10660\Delta^3)\,u_0$,

whence Δu_0, $\Delta^2 u_0$ and $\Delta^3 u_0$ were found.

The rates for ages 5 to 10 were assumed to be constant, although the third-difference curve produced a slight dip at this section.

25. Curve fitting. General remarks.

As used by the statistician the term "curve-fitting" is usually applied to the process whereby observed data (as distinct from ratios such as rates of mortality derived from the data) have a mathematical curve fitted to them. There may be a variety of reasons for such a step, but we need not consider them in this book.

A very common example, viz. the fitting of a normal curve, has already been considered. It is only necessary to find the mean of the distribution (which is taken as origin) and the standard deviation.

Another typical example of actuarial curve-fitting is afforded by the National Health Insurance table mentioned on p. 237. Here a mathematical curve was fitted to the exposed to risk in ogive form. The actual deaths were then adjusted so as to leave the group rate of mortality the same as before and finally a curve was fitted to the deaths so adjusted.

The most important curves for fitting to statistical data were developed by Karl Pearson and bear his name.

They are solutions of the differential equation

$$\frac{1}{f(x)}\frac{df(x)}{dx}=\frac{x+a}{b+cx+dx^2}.$$

From the available data constants such as the mean, standard deviation, measures of skewness, etc. have first to be calculated. The theory underlying the reasons for these calculations is too involved for discussion here. There are many excellent books dealing specially with the subject and the reader is particularly recommended to read *Frequency Curves and Correlation* by Sir William Elderton, which deals very fully with the practical difficulties likely to arise.

It will be sufficient here to mention briefly two general methods of approach often referred to in actuarial literature.

26. Method of least squares.

Suppose that for a given value of the variable (e.g., the age) the difference between the observed value and the value according to the curve fitted to the data is ϵ. Suppose further that we can obtain a measure of the standard deviation σ of all the ϵ's if a large number of values were available instead of one observed value.

We then have an observed deviation of ϵ with a standard error σ, and if the normal curve applied we could say that the probability of an observed error ϵ was $\kappa e^{-\frac{\epsilon^2}{2\sigma^2}}$.

This would apply at every value of the variable, so that the combined probability of a whole set of independent errors arising would be $\kappa e^{-\Sigma\frac{\epsilon^2}{2\sigma^2}}$ and would be a maximum if $\Sigma\frac{\epsilon^2}{2\sigma^2}$ were a minimum.

Consequently the most acceptable curve fitted to the data, i.e. the one which produces the most likely set of errors or discrepancies, would be the one which made $\Sigma\frac{\epsilon^2}{2\sigma^2}$ a minimum.

This is the basis of the Method of Least Squares.

F. M. Redington has suggested (*J.S.S.* Vol. IV, No. 4) that instead of dividing each ϵ by σ we could instead divide by the square root of the "expected" value of the function, i.e. the value given by the curve.

That is because σ^2 is usually of the form npq and in actuarial work is very nearly nq, the "expected" value. As Redington points out, however, some of the assumptions made may be wide of the mark.

In fitting a mathematical curve the errors ϵ produced are not likely to be random and to follow the normal law.

Consequently theoretical niceties are out of place; it is quite usual to assume that all the σ's are approximately equal and to make $\Sigma\epsilon^2$ a minimum.

Theoretically this is simple.

$\Sigma\epsilon^2$ is written down in terms of the observed values and the values according to the curve.

For example, in fitting the curve $\frac{A+Bc^x}{Kc^{-x}+1+Dc^x}$ we should write

$$\Sigma\epsilon^2=\Sigma\left\{u_x-\frac{A+Bc^x}{Kc^{-x}+1+Dc^x}\right\}^2,$$

where u_x is the observed value.

To make this a minimum we equate to zero the partial differential coefficients with respect to A, B, K, D and c, thus producing five equations for the five unknowns. The practical side of the whole subject is, however, much more complicated than is apparent from this brief outline.

27. Method of moments.

If f_x represents the frequency with which a value x of the variable is observed it is a relatively simple matter to calculate the successive moments.

$$\mu_1 = \frac{\Sigma x f_x}{\Sigma f_x}, \quad \mu_2 = \frac{\Sigma x^2 f_x}{\Sigma f_x}, \quad \mu_3 = \frac{\Sigma x^3 f_x}{\Sigma f_x},$$

$$\cdots\cdots\cdots\cdots\cdots\cdots$$

where the x's are measured from the mean.

If ϕ_x is the expected frequency with which the value x occurs in the curve to be fitted, the successive moments can be calculated in terms of the constants in the equation of the curve. The constants can then be found by equating the successive moments.

Suppose, for instance, that it is desired to fit the curve $A+Bc^x$ to a set of values of μ_x.

The equations would be

$$\Sigma \mu_x = \Sigma A + \Sigma B c^x, \qquad \ldots\ldots(31)$$

$$\Sigma x \mu_x = \Sigma A x + \Sigma B x c^x, \qquad \ldots\ldots(32)$$

$$\Sigma x^2 \mu_x = \Sigma A x^2 + \Sigma B x^2 c^x. \qquad \ldots\ldots(33)$$

A, B and c can then be found from these equations, which are greatly simplified in numerical work if the origin is taken at the centre.

In actuarial work the method of moments is seldom employed in this form; instead we equate successive summations.

As will be seen from the following, the two methods are equivalent if the argument x is a linear function such as the age.

Suppose that we have a series of values

$$u_1, u_2, u_3, \ldots u_n, \quad \text{where} \quad \sum_1^n u_r = N.$$

By summing col. (1) of Table XXXI continuously from the end we obtain the terms

$$(u_1+u_2+\ldots+u_n), \quad (u_2+u_3+\ldots+u_n), \quad \ldots, \quad (u_{n-1}+u_n), \quad u_n$$

shown in column (2).

Table XXXI

Function (1)	First summation (2)	Second summation (3)	Third summation (4)
u_1	$u_1+u_2+\dots+u_n$	$u_1+2u_2+\dots$	$u_1+\frac{2\cdot 3}{2}u_2+\frac{3\cdot 4}{2}u_3+\dots+\frac{n(n+1)}{2}u_n$
u_2	$u_2+u_3+\dots+u_n$	$u_2+2u_3+\dots+(n-1)u_n$	$u_2+\frac{2\cdot 3}{2}u_3+\dots+\frac{n(n-1)}{2}u_n$
u_3	$u_3+\dots+u_n$	....	....
...	....	....	....
u_r	$u_r+\dots+u_n$	$u_r+2u_{r+1}+\dots+(n-r+1)u_n$	$u_r+\frac{2\cdot 3}{2}u_{r+1}+\dots+\frac{(n-r+1)(n-r+2)}{2}u_n$
$\vdots$			
u_n	u_n	u_n	u_n
$u_1+u_2+u_3+\dots+u_n$	$u_1+2u_2+\dots+nu_n$	$u_1+\frac{2\cdot 3}{2}u_2+\dots+\frac{r(r+1)}{2}u_r+\dots+\frac{n(n+1)}{2}u_n$	$u_1+\frac{2\cdot 3\cdot 4}{6}u_2+\frac{3\cdot 4\cdot 5}{6}u_3+\dots+\frac{n(n+1)(n+2)}{6}u_n$

The total of these is $nu_n+(n-1)u_{n-1}+(n-2)u_{n-2}+\dots+u_1$ and $=Nm_1$, where m_1 is the first moment about the origin.

Summing again continuously from the end we obtain

$$(u_1+2u_2+3u_3+\dots+nu_n),$$
$$(u_2+2u_3+\dots+(n-1)u_n)\dots(u_{n-1}+2u_n),\quad u_n,$$

shown in column (3).

The total of these is

$$\frac{n(n+1)}{2}u_n+\frac{(n-1)n}{2}u_{n-1}+\dots+\frac{2\cdot 3}{2}u_2+u_1$$
$$=\frac{1}{2}\left\{\sum_1^n r^2u_r+\sum_1^n ru_r\right\}=\frac{N}{2}(m_2+m_1),$$

where m_2 is the second moment about the origin.

Similarly, it is possible to express the rth summation in terms of the first r moments of the distribution.

Since

$$\sum_1^n \frac{r(r+1)\dots(r+t)}{t!} = \sum_1^n (r+t)_{(t)}$$
$$= \Big[(r+t)_{(t+1)}\Big]_1^{n+1} = \frac{n(n+1)\dots(n+t+1)}{(t+1)!}$$

the way in which each summation can be written down from the previous one is obvious.

For convenience we have assumed the summations to be made from the end, but it is clearly immaterial whether we do this or, as is more usual, sum from the beginning.

By equating the successive summations of the observed values and the expected values according to the curve, we are ensuring that the successive moments about the mean (or any convenient origin) are also equal.

The most common example is that frequently met with earlier in this chapter, where the successive summations of actual and expected deaths are made equal. It is sometimes referred to as "the method of moments applied by way of successive summations of the actual and expected deaths".

The method of moments and the method of least squares can be shown to give identical results in many cases and on most problems likely to arise in practice the methods would produce similar values for the constants.

28. Advantages of curve-fitting.

The greatest advantage is that the results are ideally smooth, a very important point in rates of mortality to be used for valuation purposes and the construction of premium rates.

The use of some mathematical curves, notably Makeham's and closely allied curves, enables the calculation of complicated functions to be simplified greatly.

The method often throws considerable light on the way mortality is changing from one generation to another and may help in the search for a law of mortality.

In finding the constants of the curve it is usually an easy matter to allow for the weight of the exposed to risk at each age or age-group. Thus, although only rough weights are used in finding the Makeham constant c, the exposed to risk and deaths can be used in finding the other constants A and B.

Usually the total deviation (actual deaths − expected deaths) and the accumulated deviation are automatically made zero or very small, so that in the majority of instances a satisfactory fit should be attained.

29. Disadvantages of curve-fitting.

It is very difficult, particularly with modern data and the heterogeneity involved, to find a suitable curve. Once this has been done subsequent work is largely routine and automatic. In practice, many attempts usually have to be made and quick results can be produced only if one of the earliest proves to be successful. It was partly for this reason, the desire for quick results, that the A 1924–29 data were graduated by a summation formula.

It is doubtful whether a single curve can ever be fitted successfully to heterogeneous data.

30. Conclusion.

The whole subject affords a vast field for research and many functions such as d_x, the number of deaths in the crude mortality table, have only recently been the subject of experiment. For many years, owing to the usefulness of Makeham's formula, attention was focused almost exclusively on μ_x and the allied function $\operatorname{colog} p_x$, but it is possible that other expressions may be used successfully in the future.

BIBLIOGRAPHY

Frequency Curves and Correlation. Sir W. P. Elderton. London, 1938.

"Graduation of the O^{NM} Table." Sir George F. Hardy. *J.I.A.* Vol. xxxviii.

"Re-graduation of the O^{M} Table." G. J. Lidstone. *J.I.A.* Vol. xxxviii.

The Mortality of Annuitants, 1900–1920. Sir W. P. Elderton and H. J. P. Oakley.

"Government Life Annuitants' Mortality, 1900–1920." *J.I.A.* Vol. lv.

"National Health Insurance Life Tables." *J.I.A.* Vol. xlvii.

EXAMPLES 9

1. Graduate the following data by means of Makcham's first formula:

$$\mu_x = A + Bc^x.$$

Age-group	Exposed to risk (in central form)	Actual deaths
20–25	15,750	30
25–30	15,100	40
30–35	12,750	50
35–40	14,700	60
40–45	11,750	70
45–50	11,750	80
50–55	13,900	90

The data relate to the active service of employees of a firm which has a fixed entry age of 20 and a fixed retiring age of 55.

2. Tables of graduated values of q_x have been arrived at by several different methods, all of which produce results which are satisfactory from the point of view of smoothness. It is stated that the most satisfactory graduation is that in which the sum of the squares of the differences between the graduated and ungraduated values is a minimum. Show clearly on what assumptions this statement is based and how the test is derived from these assumptions.

Discuss the applicability of the test to mortality statistics generally.

3. Blend the two following series of F_y between the values of $y=7$ and $y=13$ by means of (*a*) an interpolation formula of the third degree and (*b*) the curve of sines:

y	F_y^A	F_y^B	y	F_y^A	F_y^B
6	167	—	10	311	312
7	200	—	11	325	364
8	235	220	12	395	420
9	272	264	13	—	480
			14	—	544

Why is the usual blending method not suitable in such a case?

4. The following table shows certain values of q_x extracted from English Life Table No. 10 (Males) and the Government Annuitants' Tables 1900–20 (Males):

Age	E.L. No. 10	G.A.	Age	E.L. No. 10	G.A.
39	·00531	·00635	46	·00861	·00832
40	·00562	·00658	47	·00925	·00869
41	·00598	·00683	48	·00990	·00910
42	·00639	·00710	49	·01057	·00957
43	·00687	·00738	50	·01128	·01010
44	·00741	·00768	51	·01206	·01070
45	·00799	·00799			

Blend the two series to obtain values of q_x passing from the E.L. No. 10 values at ages 42 and under to the G.A. values at ages 48 and over.

Criticize the junction effected and state reasons for any unsatisfactory feature. Employ an alternative algebraic process to produce a more suitable series fulfilling the same conditions.

5. In a mortality investigation the data are presented in the following form:

Age-group	Exposed to risk	Deaths
20–26	...	...
27–33	...	...
34–40	...	...
...	...	...
...	...	...
90–96	...	...

The exposed to risk (E_x) and deaths (θ_x) have each been obtained at each individual age x in such a form that $q_x = \theta_x/E_x$ and have then been summed for the age-groups shown. It is desired to obtain graduated values of μ_x by use of the formula $\mu_x = Ac^x + Bc^{2x}$.

State how you would proceed and what difference should in theory be made in the procedure if it were desired to obtain graduated values of m_x by use of the formula $m_x = Ac^x + Bc^{2x}$.

6. You are required to graduate a mortality experience by the formula

$$\mu_x = A + Bc^x + Dc^{-x}.$$

The following is part of the experience:

Age last birthday	Exposed to risk	Deaths
11–15	460	7
16–20	994	11
21–25	2,312	32
26–30	4,824	54
31–35	7,684	117
...	...	...
...	...	...
86–90	19,000	4
91 and over	Nil	Nil

How would you proceed with the graduation?

The part of the table given should be used for examples of the first steps; no arithmetical work on the equations formed is required but only a description of the methods used to find the constants.

7. Obtain graduated withdrawal rates from the following data:

(*a*) by a graphic process,
(*b*) by fitting the formula: withdrawal rate at duration $t = a + bt$,

and compare the results.

Duration	Exposed to risk	Withdrawals	Duration	Exposed to risk	Withdrawals
0	1200	120	5	600	27
1	1000	70	6	500	20
2	900	45	7	400	12
3	800	48	8	350	7
4	700	30	9	300	7

CHAPTER X

PIVOTAL VALUES AND OSCULATORY INTERPOLATION

1. Perhaps the method of least general application is that of pivotal values and osculatory interpolation, used for the English Life Tables Nos. 7 to 10. This was devised by George King for the first two of these tables and has been used with slight modifications ever since. A good deal of criticism has been directed to the method, which has been described by one leading statistician as "the highest of high-class cookery". It should, however, be remembered that it was produced to deal with very special problems and these must be borne in mind in assessing its merits or failings.

Chief of these special problems were the following:

(*a*) The data were the population of England and Wales (males and females separately) and the deaths of the years 1901–10 for E.L. No. 7 and of the three years 1910–12 for E.L. No. 8. Consequently very large numbers were involved and apart from local disturbances the progression was already fairly smooth before graduation commenced.

(*b*) Tables were required for spinsters, married women, and widows, as well as for all females combined.

(*c*) King was required to produce not only rates of mortality for the construction of a life table, but also a "Graduation of Ages", as it was called, i.e. a tabulation of the population age by age with inaccuracies removed as far as possible.

(*d*) The chief problem to be solved was how to eliminate "local" mis-statements of age due to a preference for even numbers and numbers ending in 5.

(*e*) Medical Officers of Health had been in the habit of comparing local mortality with that of other districts or of the country as a whole by means of the crude death rates, i.e. the number of deaths per thousand, irrespective of the age distribution. It was therefore suggested that some simple and easily applied method should be devised for their use in constructing mortality tables for districts.

The problem is admirably summed up in King's own words:

"In constructing the tables it was desirable that a method should be employed, simple in theory, easy in application, and which would produce curves of smooth graduation, and curves which would adhere closely to the original data"; and

"The table is not intended to forecast the future, but merely to give accurately the present populations."

Even in 1911 the data exhibited irregularities due to such factors as changing rates of birth, migration and mortality. Such irregularities were inherent in the data and could not be removed in arriving at an accurate picture of the numbers at each age, as they would have been by a powerful method such as was used for the N.H.I. table. It should be remembered that for the latter table the object in view was a table of smooth rates suitable for the calculation of reserve values.

The separate tables for spinsters, widows, and married women also had a restrictive effect on the choice of method, because, as King says: "Under my instructions it was also necessary that the graduation should reproduce the total population exactly, and it was evidently also desirable that the method should be such that when applied separately to each of the sections of the population which make up the whole, the sum of the populations of the several sections at each year of age should be identical with the corresponding total population."

A summation formula would have achieved this last result but would not have dealt satisfactorily with the minor mis-statements of age which have always been the distinctive feature of census data.

The most effective way of meeting this difficulty was to group the data quinquennially. Experiments were carried out at each census since 1911 to determine which grouping was most successful. For instance, 29–33, 34–38, etc., were adopted on one occasion and 33–37, 38–42, etc., on another.

In Chapter I we proved King's formula

$$u_0 = \cdot 2w_0 - \cdot 008\Delta^2 w_{-1},$$

giving the central term of a quinquennial group in terms of the totals for that group and the two neighbouring groups.

By means of this formula King derived what he called "graduated quinquennial pivotal values" of E_x^c, the exposed to risk at age x last birthday, and θ_x, the deaths at age x last birthday in a calendar year.

It should perhaps be emphasized that the graduation was produced

(*a*) by grouping the data quinquennially, and

(*b*) by deducing pivotal values on the assumption that fourth and higher differences of the grouped values were zero.

Of these (*a*) was of course by far the more important.

2. Osculatory interpolation.

The pivotal values were not subsequently altered and the intermediate values were inserted by a process of osculatory interpolation. Any of the well-known formulae such as Everett's might have been used, but the objection to these was that for each new interval the quinquennial values involved were different from those used in the previous interval, so that although the curve of values was continuous the gradient was discontinuous on passing from one interval to the next. To overcome this, King used a formula specially designed to make the gradient continuous. His method of deriving this formula, although not the simplest, is instructive, and as it usually causes difficulty the method is dealt with in some detail in the next paragraph.

3. King's formula for osculatory interpolation.

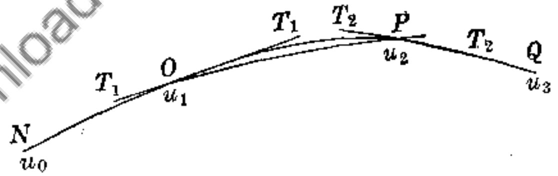


Fig. 9.

King decided to use a third degree curve for the purpose of interpolation over each range of six values consisting of two consecutive pivotal values and four interpolated values. As he had four available constants he made this curve not only pass through the pivotal values at the ends of the range but also touch certain lines at those points.

For instance, if the points N, O, P, Q in Fig. 9 represented pivotal values of u_0, u_1, u_2 and u_3, the values $u_{1\cdot 2}$, $u_{1\cdot 4}$, $u_{1\cdot 6}$ and $u_{1\cdot 8}$ were inserted between O and P by means of a third degree curve which passed through O and P and touched certain lines T_1OT_1 and T_2PT_2.

Similarly, the third degree curve used for the interval PQ passed through P and Q and touched T_2PT_2 and a similar line at Q.

Since both these curves touched the same line T_2PT_2 they touched one another and the gradient was therefore continuous at P.

The method by which the lines T_1OT_1 and T_2PT_2 were found is apt to cause the student some difficulty. The lines were fixed as the tangents to second degree curves which were used for this purpose only and not for the actual interpolation.

We shall examine the problems in detail.

The second degree curve passing through N, O and P is clearly

$$u_x = u_0 + x\Delta u_0 + \frac{x(x-1)}{2}\Delta^2 u_0,$$

$$\therefore \quad \frac{du_x}{dx} = \Delta u_0 + \frac{2x-1}{2}\Delta^2 u_0.$$

The slope of the tangent T_1OT_1 is found by putting $x=1$; this gives

$$\Delta u_0 + \tfrac{1}{2}\Delta^2 u_0. \qquad \ldots\ldots(1)$$

Similarly, the second degree curve passing through OPQ is

$$u_x = u_1 + x\Delta u_1 + \frac{x(x-1)}{2}\Delta^2 u_1,$$

and the slope of the tangent T_2PT_2 is therefore

$$\Delta u_1 + \tfrac{1}{2}\Delta^2 u_1 = \Delta u_0 + \tfrac{3}{2}\Delta^2 u_0 + \tfrac{1}{2}\Delta^3 u_0. \qquad \ldots\ldots(2)$$

Having obtained these tangents we dispense with the second degree curves. We now have to find a third degree curve passing through O and P and having the above gradients at those points. This curve is to be used for the actual interpolation.

Let its equation be

$$u_{1+x} = u_1 + bx + cx^2 + dx^3.$$

Since the curve passes through P,

$$u_2 = u_1 + b + c + d.$$

I.e. $$b + c + d = u_2 - u_1 = \Delta u_1 = \Delta u_0 + \Delta^2 u_0. \qquad \ldots\ldots(3)$$

Since the curve also touches T_1OT_1 and T_2PT_2,

$$\left(\frac{du_{1+x}}{dx}\right)_{x=0}=b=\Delta u_0+\tfrac{1}{2}\Delta^2 u_0, \text{ from (1)} \qquad \ldots\ldots(4)$$

$$\left(\frac{du_{1+x}}{dx}\right)_{x=1}=b+2c+3d=\Delta u_0+\tfrac{3}{2}\Delta^2 u_0+\tfrac{1}{2}\Delta^3 u_0. \qquad \ldots\ldots(5)$$

Solving these equations we obtain King's osculatory interpolation formula:

$$u_{1+x}=u_1+x\Delta u_0+\frac{x+x^2}{2}\Delta^2 u_0-\frac{x^2-x^3}{2}\Delta^3 u_0. \qquad \ldots\ldots(6)$$

Note: the formula gives u_{1+x} (not u_x) in terms of u_1, the value at the beginning of the interval, and the differences of u_0.

4. Osculatory form of Everett's formula.

Everett's well-known formula can be readily modified so as to produce osculatory interpolation and the student is referred to *Mathematics for Actuarial Students*, Part II, pp. 147 *et seq.* for details.

Formula (6) may be written in the form

$$\begin{aligned}u_{1+x}&=u_1+x(u_2-u_1-\Delta^2 u_0)+\frac{x+x^2}{2}\Delta^2 u_0-\frac{x^2-x^3}{2}(\Delta^2 u_1-\Delta^2 u_0)\\&=xu_2+(1-x)u_1-\frac{x-2x^2+x^3}{2}\Delta^2 u_0-\frac{x^2-x^3}{2}\Delta^2 u_1\\&=xu_2+\frac{x^2(x-1)}{2}\Delta^2 u_1+\xi u_1+\frac{\xi^2(\xi-1)}{2}\Delta^2 u_0, \qquad \ldots\ldots(7)\end{aligned}$$

where $\xi=1-x$.

This differs from the usual formula in having coefficients

$$\frac{x^2(x-1)}{2}, \quad \frac{\xi^2(\xi-1)}{2}$$

instead of

$$\frac{x(x^2-1)}{3!}, \quad \frac{\xi(\xi^2-1)}{3!},$$

An example of the use of this formula will be given later, but King's formula can be applied as follows to give the interpolated values

$$u_{1\cdot2},\ u_{1\cdot4},\ u_{1\cdot6} \quad\text{and}\quad u_{1\cdot8}.$$

Differencing formula (6) at intervals of $1/t$ we obtain the following results, where Δ as before relates to differences of grouped values and δ relates to differences at intervals of $1/t$.

$$\left.\begin{aligned} \delta u_1 &= \frac{\Delta u_0}{t} + \frac{t+1}{2}\frac{\Delta^2 u_0}{t^2} - \frac{t-1}{2}\frac{\Delta^3 u_0}{t^3} \\ \delta^2 u_1 &= \qquad \frac{\Delta^2 u_0}{t^2} - (t-3)\frac{\Delta^3 u_0}{t^3} \\ \delta^3 u_1 &= \qquad\qquad 3\frac{\Delta^3 u_0}{t^3} \end{aligned}\right\}.$$

When $t=5$, these become

$$\left.\begin{aligned} \delta u_1 &= \cdot 2\Delta u_0 + \cdot 12\Delta^2 u_0 - \cdot 016\Delta^3 u_0 \\ \delta^2 u_1 &= \qquad\quad \cdot 04\Delta^2 u_0 - \cdot 016\Delta^3 u_0 \\ \delta^3 u_1 &= \qquad\qquad\qquad \cdot 024\Delta^3 u_0 \end{aligned}\right\}.$$

The four interpolated values can then readily be fitted in, but in order to check the work by reproducing u_2 it is necessary to retain three extra decimal places in the δ's. Moreover, these differences have to be re-calculated for every interval, so that the osculatory form of Everett's formula is usually quicker, particularly if many values have to be inserted.

Example 1.

Given the following data, calculate the quinquennial pivotal value of q_{39} and interpolate the values q_{35} to q_{43} by an osculatory formula.

Age (1)	Population at 30th June 1935 (2)	Deaths in years 1934–36 (3)	Age (4)	Population at 30th June 1935 (5)	Deaths in years 1934–36 (6)
30	2270	27	40	2346	35
31	2049	26	41	2048	35
32	2198	29	42	2186	38
33	2192	28	43	2073	35
34	2203	29	44	2009	37
35	2226	30	45	2028	38
36	2252	30	46	1912	39
37	2176	32	47	1858	40
38	2324	35	48	1906	41
39	2241	34	49	1762	43

Age x	Quinquennial pivotal values of q_x
24	·00365
29	·00387
34	·00439
44	·00605
49	·00810
54	·01167

To obtain the pivotal value of q_{39} it is first necessary to group the data into the ranges 32–36, 37–41, etc. King's formula, $u_0 = \cdot 2w_0 - \cdot 008\Delta^2 w_{-1}$, then gives the graduated pivotal values of the population and deaths at age 39 as follows:

Central age (1)	Group population (2)	Δ (2) (3)	Δ² (2) (4)	Group deaths (5)	Δ (5) (6)	Δ² (5) (7)
34	11,071			146		
		64			25	
39	11,135		−991	171		−9
		−927			16	
44	10,208			187		

The adjusted population E_x^c at age 39 is therefore

$$\cdot 2(11{,}135) - \cdot 008(-991) = 2235$$

and the adjusted deaths are

$$\cdot 2(171) - \cdot 008(-9) = 34\cdot 27$$ for the three years or 11·42 per annum.

$$\therefore\ m_{39} = \frac{11\cdot 42}{2235}$$

and $$q_{39} = \frac{11\cdot 42}{2235 + 5\cdot 72} = \cdot 00510.$$

To interpolate the required values by the formula

$$u_{1+x} = xu_2 + \frac{x^2(x-1)}{2}\Delta^2 u_1 + \xi u_1 + \frac{\xi^2(\xi-1)}{2}\Delta^2 u_0$$

we proceed as follows:

Value of x	·2	·4	·6	·8
Value of coefficient $\dfrac{x^2(x-1)}{2}$	−·016	−·048	−·072	−·064

As these coefficients are all multiples of ·008 it is convenient to tabulate ·008$\Delta^2 u$ as follows:

Table XXXII

Age y (1)	$q_y \times 10^5$ (2)	Δ (2) (3)	Δ^2 (2) (4)	·008Δ^2 (2) (5)
24	365			
		22		
29	387		30	·240
		52		
34	439		19	·152
		71		
39	510		24	·192
		95		
44	605		110	·880
		205		
49	810		152	1·216
		357		
54	1167			

Table XXXIII

Age (1)	xu_2 (2)	$\frac{x(x-1)}{2}\Delta^2 u_1$ (3)	(2)+(3) (4)	ξ terms (5)	Interpolated value (4)+(5) (6)
29	—	—	—		
30	87·8	− ·304	87·496		
31	175·6	− ·912	174·688		
32	263·4	−1·368	262·032		
33	351·2	−1·216	349·984		
34	—	—	—	—	(439)
35	102	− ·384	101·616	349·984	451·600
36	204	−1·152	202·848	262·032	464·880
37	306	−1·728	304·272	174·688	478·960
38	408	−1·536	406·464	87·496	493·960
39	—	—	—	—	(510)
40	121	−1·760	119·240	406·464	525·704
41	242	−5·280	236·720	304·272	540·992
42	363	−7·920	355·080	202·848	557·928
43	484	−7·040	476·960	101·616	578·576
44	—	—	—	—	(605)

(For illustration the full number of decimal places is retained.)

In performing the actual interpolation only the "x" terms are calculated, since in reverse order they form the "ξ" terms for the succeeding interval. Thus the first four items in column (5) of Table XXXIII are merely the first four items of column (4) in reverse order. This is similar to the ordinary use of Everett's formula: see Example 2 on pp. 73–5 of *Mathematics for Actuarial Students*.

For ages 35–38, $u_1 = q_{34} \times 10^5$, $u_2 = q_{39} \times 10^5$.

For ages 40–43, $u_1 = q_{39} \times 10^5$, $u_2 = q_{44} \times 10^5$.

5. King's short method of constructing abridged Life Tables.

For the benefit of Medical Officers of Health and others interested in vital statistics the following method was devised for constructing an abridged mortality table and values of $\mathring{e}_x$.

First, formulae were required giving the sum of five values for successive ages in terms of the values at quinquennial intervals.

From the formula

$$u_x = u_0 + x\Delta u_0 + \frac{x(x-1)}{2}\Delta^2 u_0 + \frac{x(x-1)(x-2)}{6}\Delta^3 u_0$$

the following formulae were deduced:

$$u_1 + u_{1\cdot2} + u_{1\cdot4} + u_{1\cdot6} + u_{1\cdot8} = 5u_0 + 7\Delta u_0 + 1\cdot6\Delta^2 u_0 - \cdot2\Delta^3 u_0 \quad \ldots\ldots(8)$$

and

$$u_{1\cdot2} + u_{1\cdot4} + u_{1\cdot6} + u_{1\cdot8} + u_2 = 5u_0 + 8\Delta u_0 + 2\cdot6\Delta^2 u_0 - \cdot2\Delta^3 u_0. \quad \ldots(9)$$

For the initial group these formulae fail and were replaced by

$$u_0 + u_{\cdot2} + u_{\cdot4} + u_{\cdot6} + u_{\cdot8} = 5u_0 + 2\Delta u_0 - \cdot4\Delta^2 u_0 + \cdot2\Delta^3 u_0 \quad \ldots\ldots(10)$$

and

$$u_{\cdot2} + u_{\cdot4} + u_{\cdot6} + u_{\cdot8} + u_1 = 5u_0 + 3\Delta u_0 - \cdot4\Delta^2 u_0 + \cdot2\Delta^3 u_0. \quad \ldots\ldots(11)$$

The following gives the subsequent process in detail:

(i) Calculate quinquennial pivotal values of population and deaths and hence q_x at quinquennial intervals.

(ii) Deduce the values of $\log p_x$ at quinquennial intervals.

(iii) Since $\log_5 p_x = \log p_x + \log p_{x+1} + \ldots + \log p_{x+4}$, calculate the values of $\log_5 p_x$ from the values in (ii).
(Formula (10) has to be used for the first interval.)

(iv) Take a suitable radix for l_x and, using the values in (iii), find $\log l_x$ at quinquennial intervals. These values will not extend to the end of the life table and the last values will have to be inserted by some arbitrary but reasonable method.

(v) Taking antilogs obtain l_x at quinquennial intervals. Formula (9) (formula (11) for the first interval) then gives

$$l_{x+1}+l_{x+2}+l_{x+3}+l_{x+4}+l_{x+5}$$

in terms of l_x and the differences taken at quinquennial intervals. (This sum of five consecutive values of l_x at unit intervals was called by King $N'_{x:\overline{5}|}$.)

(vi) By summing the values of $N'_{x:\overline{5}|}$ from the bottom upwards obtain

$$\sum_{z=x}^{\omega} l_x$$

at quinquennial intervals and hence

$$e_x=\sum_{z=x}^{\omega} l_z/l_x.$$

The addition of ·5 gives the complete expectation $\mathring{e}_x$.

Table XXXIV is an example of the use of the method and is taken from King's report on English Life Tables Nos. 7 and 8.

The initial processes were straightforward and are not shown. They were as follows. From the data grouped in the ranges 4–8, 9–13, ... 104–108 quinquennial pivotal values were obtained for ages 11 to 101 inclusive and the values of $-\log p_x$ shown in column (2) calculated.

Formula (8) then enabled $\log_5 p_x$ to be deduced as far as age 91, but $\log_5 p_{96}$ and $\log_5 p_{101}$ were also needed. These were inserted as shown by assuming a constant fourth difference for $\log p_x$. Any other reasonable assumption would have given much the same result.

Column (7) was obtained by the application of formulae (8) and (10). It should be noted that, throughout the work, values at quinquennial intervals only were used, the intervening ages being allowed for in the construction of formulae (8) to (11).

To derive the values of $\mathring{e}_x$ at quinquennial intervals the work was as follows:

Column (13) was obtained by means of formulae (9) and (11) from the values of l_x and the differences. Column (14) was then derived by summing from the bottom upwards and the final column obtained by dividing the entries in column (14) by the corresponding values of l_x and adding ·5.

Table XXXIV

Age (1)	$-\log p_x \times 10^5$ (2)	Δ (2) (3)	Δ^2 (2) (4)	Δ^3 (2) (5)	Δ^4 (2) (6)	$-\log {}_5p_x \times 10^5$ (7)	$\log l_x$ (8)	l_x (9)
11	79					451	5·00000	100,000
		34						
16	113		12			660	4·99549	98,967
		46		−38				
21	159		−26			836	·98889	97,474
		20		46				
26	179		20			965	·98053	95,616
		40		8				
31	219		28			1,222	·97088	93,515
		68		−9				
36	287		19			1,596	·95866	90,920
		87		26				
41	374		45			2,114	·94270	87,640
		132		11				
46	506		56			2,872	·92156	83,476
		188		60				
51	694		116			4,026	·89284	78,134
		304		28				
56	998		144			5,816	·85258	71,216
		448		62				
61	1,446		206			8,416	·79442	62,290
		654		198				
66	2,100		404			12,410	·71026	51,317
		1,058		224				
71	3,158		628			18,895	·58616	38,562
		1,686		77				
76	4,844		705			28,717	·39721	24,958
		2,391		14				
81	7,235		719			42,001	·11004	12,884
		3,110		533				
86	10,345		1,252			60,640	3·69003	4,898
		4,362		−3,461				
91	14,707		−2,209		14,116	76,594	·08363	1,212
		2,153		10,655				
96	16,860		8,446		14,116	97,165	2·31769	208
		10,599		24,771				
101	27,459		33,217		14,116	203,863	1·34604	22
				38,887				
106							$\bar{1}$·30741	0

Table XXXIV

Age	l_x*	$-\Delta$ (9)	$-\Delta^2$ (9)	Δ^3 (9)	$N'_{x:\overline{5}\vert}$	$\sum_x^{\omega} l_x$	$\mathring{e}_x$
	(9)	(10)	(11)	(12)	(13)	(14)	(15)
11	100,000				497,104	5,166,263	52·16
		1,033					
16	98,967		460		490,521	4,669,159	47·68
		1,493		+ 95			
21	97,474		365		481,918	4,178,638	43·37
		1,858		+ 122			
26	95,616		243		471,924	3,696,720	39·16
		2,101		− 251			
31	93,515		494		460,026	3,224,796	34·98
		2,595		− 191			
36	90,920		685		445,074	2,764,770	30·91
		3,280		− 199			
41	87,640		884		426,120	2,319,696	26·97
		4,164		− 294			
46	83,476		1178		401,905	1,893,576	23·18
		5,342		− 398			
51	78,134		1576		370,633	1,491,671	19,59
		6,918		− 432			
56	71,216		2008		330,113	1,121,038	16·24
		8,926		− 39			
61	62,290		2047		279,297	790,925	13·20
		10,973		+ 265			
66	51,317		1782		218,846	511,628	10·47
		12,755		+ 933			
71	38,562		849		151,862	292,782	8·09
		13,604		+2379			
76	24,958		−1530		87,444	140,920	6·15
		12,074		+2558			
81	12,884		−4088		38,784	53,476	4·65
		7,986		+ 212			
86	4,898		−4300		12,036	14,692	3·50
		3,686		−1618			
91	1,212		−2682		2,348	2,656	2·69
		1,004		−1864			
96	208		− 818		286	308	1·99
		186		− 654			
101	22		− 164		22	22	1·50
		22					
106	0				—	—	—

* col. (9) on p. 273.

6. English Life Table No. 7.

These tables (one for each sex) were based on the population of England and Wales at the censuses of 1901 and 1911 and the recorded deaths in the calendar years 1901–10 inclusive.

The populations of 1901 were supplied for each of the first four years of age and then in quinquennial groups 5–9, 10–14, etc., with a final group 100 and over.

The 1911 figures were available for each year and were grouped in the same way as the 1901 figures in order that the mean population could be found by Waters's method. Finally the interpolated figures for 100 and over were split into those for ages 100–104 and those for ages 105 and over in the proportions shown by the 1911 Census figures.

The deaths for the ten years were given for each of the first five years, then in quinquennial groups up to 24 last birthday, and then in decennial groups 25–34, etc. up to 75–84, with a final group 85–99. For 1901 to 1909 the deaths of centenarians were given age by age and the deaths for 1910 were subdivided into the groups 100–104 and 105 and over in the same proportion. Similarly, the decennial groups were split into quinquennial groups by means of the figures for the years 1910–12, which were available for each age up to 99. Using the grouping 5–9, 10–14, ... 100–104 graduated quinquennial pivotal values were obtained for ages 12, 17, ... 97 for populations and deaths and the values of q_x were deduced.

For osculatory interpolation $\log(q_x + \cdot 1)$ was used instead of q_x and values were obtained for ages 17 to 92 inclusive.

For each of the ages 0 to 4 q_x was derived from the records of births and deaths, while q_x for age 12 had already been obtained as a pivotal value. Using the values of q_x (not $\log(q_x + \cdot 1)$) for ages 3, 4, 12, 17 and 18 the intermediate values were found by Lagrange's formula, although divided differences would have given the same result rather more simply.

For old ages $\log p_x$ was the function operated on. The values for 89, 90, 91, 92 and 97 were available and, by assuming a constant fourth difference, the remaining values were easily found.

7. English Life Table No. 8.

The populations at the 1911 Census were available for each age, as were the deaths of the years 1910–12 inclusive as far as age 99, although centenarians were grouped together.

The populations were first grouped as for the 1901 Census so that they could be brought down to 1st July 1911, the mid-point of the three years.

They were then grouped for ages 4–8, 9–13, etc., so as to deal as effectively as possible with local mis-statements of age. The deaths over age 100 were split into the necessary ranges 100–103 and 104 and over by means of the deaths for 1912, which were available age by age.

Having thus obtained the population and deaths for age-groups 4–8, 9–13, ... 99–103, pivotal values were calculated for ages 6, 11, 16, ... 96 and q_x obtained for these ages. Osculatory interpolation produced the values from age 16 to 91 inclusive and $\log p_x$ was calculated for ages 88, 89, 90, 91 and 96 (the last pivotal value). Assuming a constant fourth difference as before the table was completed at the high ages.

q_x was calculated from statistics of births and deaths for each of the first six years of life and, by using the values of q_x for ages 4, 5, 11, 16 and 17, the remaining values were inserted by Lagrange's formula.

8. English Life Table No. 9.

The population enumerated at the 1921 Census and the deaths of the years 1920–22 were available for each age and, as births and deaths were available for each quarter of the calendar years involved, the rates of mortality for infantile ages were calculated more accurately than was possible before. The quinquennial groupings adopted were 2–6, 7–11, ... 92–96. Pivotal values of populations and deaths were obtained and osculatory interpolation produced rates for ages 14 to 84 inclusive.

As an alternative the crude values of q_x were calculated from the data; quinquennial groups 5–9, 10–14, etc. were adopted to reduce irregularities. The same process as above was then applied to the grouped values of q_x and the rates obtained were very similar to those produced by operating on populations and deaths separately. For the sake of continuity these latter rates were adopted.

The rates for ages 0 to 5 were calculated from the statistics of births and deaths. The values for ages 85 and over were obtained by a Gompertz graduation in which, since $\mu_x = Bc^x$, the values of $\log p_x$ are in geometric progression.

Taking $\dfrac{\log_{10} p_{94}}{\log_{10} p_{84}} = r$ the ratio $r^{1/10}$ was used to construct the successive values $\log_{10} p_{85}$, $\log_{10} p_{86}$, etc. and the values of q_x were deduced.

For ages 6 to 13 it was assumed that

$$q_x = a + bx + \tfrac{1}{2}cx(x-1) + \tfrac{1}{6}dx(x-1)(x-2),$$

and the constants were found from the values of q_5, q_9, q_{14} and q_{15} already obtained.

9. English Life Table No. 10.

After various trials it was found that the groupings 5–9, 10–14, etc. were as good as any, and rates of mortality were found in the usual way for ages 17 to 87 inclusive. As for E.L. No. 9 a Gompertz curve was used to extend the table to the end of life. In this method

$$r = \frac{\operatorname{colog} p_{x+5}}{\operatorname{colog} p_x} = 1{\cdot}40 \text{ for males and } 1{\cdot}42 \text{ for females.}$$

The rates for ages 0 to 5 were calculated from records of births and deaths. Special attention was paid to q_0, for which more detailed data were available.

For ages 6 to 16 special methods had to be used because of the rapid changes in the birth-rates after the war of 1914–18. As a result the values for ages 17 to 22 already obtained had to be modified to produce a smooth progression. The whole question was however one of construction rather than graduation.

10. Advantages and disadvantages of the method.

It is hardly possible to discuss on general lines the merits and demerits of the method of graduated quinquennial pivotal values and osculatory interpolation. The method was devised to meet a special problem and has proved so successful that it has been modified only in detail.

The weak graduating power of the formula renders it unsuitable for many purposes, particularly if a high degree of smoothness is essential. In assessing any method of graduation however it is only fair to take into account the type of data with which it was intended to deal.

BIBLIOGRAPHY

Supplement to the Registrar-General's 75th Annual Report, Part I. Cmd. 7512, 1914. (Reprint can be obtained from the Institute of Actuaries.)

Review of the above. *J.I.A.* Vol. XLIX.

"Short Method of Constructing on Abridged Mortality Table." G. KING. *J.I.A.* Vol. XLVIII.

"New National Life Tables." G. KING. *J.I.A.* Vol. XLIX.

Registrar-General's Decennial Supplement, 1921. Part I. H.M. Stationery Office.

Registrar-General's Decennial Supplement, 1931, Part I. H.M. Stationery Office.

Mathematics for Actuarial Students, Part II, pp. 147–154. H. FREEMAN. Camb. Univ. Press.

EXAMPLES 10

1. From the following data find values of q_x for ages 42 to 67 by means of graduated quinquennial pivotal values and osculatory interpolation:

Age-group last birthday	Period 1st Jan. 1935 to 31st Dec. 1937	
	Mean population	Deaths
30–34	36,466	294
35–39	48,634	474
40–44	55,100	783
45–49	56,623	1098
50–54	49,684	1440
55–59	37,664	1749
60–64	24,139	1839
65–69	16,511	2043
70–74	11,881	2445

2. Without making use of the interpolated rates found in the question above find $a_{47:\overline{10}|}$ at $3\frac{1}{2}$ per cent interest from the data of that question and compare it with the value based on the interpolated rates.

3. From the undernoted data relating to the period 1934–36 calculate an approximate value of $e_{60:\overline{9}|}$ using King's short method.

Age-group	Mean population	Deaths
50–53	68,400	2400
53–56	64,200	2940
56–59	59,600	3360
59–62	51,200	3750
62–65	44,700	4350
65–68	37,000	4650
68–71	30,100	4980
71–74	22,200	5100
74–77	16,400	4800
77–80	10,200	4080

APPENDIX

Table I. *Values of the ordinates and the distribution function for the Normal Curve.*

The ordinate $$y \text{ or } p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The area to the left of the ordinate at the point x (the distribution function) is

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2}\, dt.$$

The values are tabulated for positive values of x only; a change in the sign of x does not affect y while $F(-x) = 1 - F(x)$.

x	$y=p(x)$	$F(x)$	x	$y=p(x)$	$F(x)$
0·0	0·39894	0·50000	2·5	0·01753	0·99379
0·1	0·39695	0·53983	2·6	0·01358	0·99534
0·2	0·39104	0·57926	2·7	0·01042	0·99653
0·3	0·38139	0·61791	2·8	0·00792	0·99744
0·4	0·36827	0·65542	2·9	0·00595	0·99813
0·5	0·35207	0·69146	3·0	0·00443	0·99865
0·6	0·33322	0·72575	3·1	0·00327	0·99903
0·7	0·31225	0·75804	3·2	0·00238	0·99931
0·8	0·28969	0·78814	3·3	0·00172	0·99952
0·9	0·26609	0·81594	3·4	0·00123	0·99966
1·0	0·24197	0·84134	3·5	0·00087	0·99977
1·1	0·21785	0·86433	3·6	0·00061	0·99984
1·2	0·19419	0·88493	3·7	0·00042	0·99989
1·3	0·17137	0·90320	3·8	0·00029	0·99993
1·4	0·14973	0·91924	3·9	0·00020	0·99995
1·5	0·12952	0·93319	4·0	0·00013	0·99997
1·6	0·11092	0·94520	4·1	0·00009	0·99998
1·7	0·09405	0·95543	4·2	0·00006	0·99999
1·8	0·07895	0·96407	4·3	0·00004	0·99999
1·9	0·06562	0·97128	4·4	0·00002	0·99999
2·0	0·05399	0·97725	4·5	0·00002	
2·1	0·04398	0·98214	4·6	0·00001	
2·2	0·03547	0·98610	4·7	0·00001	
2·3	0·02833	0·98928	4·8	0·00000	
2·4	0·02239	0·99180			

For intermediate values second difference interpolation is usually sufficient although on occasions third differences should be allowed for if great accuracy is required.

From the above table the following values may be obtained:

$F(x)$	x	$F(x)$	x
·001	−3·090	·999	3·090
·005	−2·576	·995	2·576
·010	−2·326	·990	2·326
·025	−1·960	·975	1·960
·050	−1·645	·950	1·645
·25	−0·674	·75	0·674

Table II. *Table of values of x_0 corresponding to critical values of*

$$P=\frac{1}{2^{\frac{1}{2}f}\Gamma(\frac{1}{2}f)}\int_{x_0}^{\infty}x^{\frac{1}{2}f-1}e^{-\frac{1}{2}x}\,dx$$

Judgement	*P*	Degrees of freedom						
		1	5	10	15	20	25	30
"Much too probable"	·999	—	—	—	—	—	—	—
"Too probable"	·99	·0002	·55	2·56	5·23	8·26	11·52	14·95
"Rather too probable"	·95	·004	1·14	3·94	7·26	10·85	14·61	18·49
	·5	·455	4·35	9·34	14·34	19·34	24·34	29·34
"Of doubtful improbability"	·05	3·84	11·07	18·31	25·00	31·41	37·65	43·77
"Improbable"	·01	6·64	15·09	23·21	30·58	37·57	44·31	50·89
"Very improbable"	·001	10·83	20·52	29·59	37·70	45·32	52·62	59·70

	P	35	40	45	50	55	60	65
"Much too probable"	·999	13·61	16·81	20·12	23·53	27·01	30·56	34·18
"Too probable"	·99	17·88	21·53	25·26	29·06	32·92	36·83	40·78
"Rather too probable"	·95	22·19	26·23	30·34	34·49	38·68	42·91	47·17
	·5	34·50	39·50	44·50	49·50	54·50	59·50	64·50
"Of doubtful improbability"	·05	49·52	55·47	61·37	67·22	73·03	78·80	84·54
"Improbable"	·01	56·53	62·88	69·15	75·35	81·49	87·58	93·63
"Very improbable"	·001	64·94	71·74	78·43	85·02	91·54	97·98	104·37

	P	70	75	80	85	90	95	100
"Much too probable"	·999	37·84	41·55	45·31	49·10	52·93	56·79	60·68
"Too probable"	·99	44·78	48·81	52·87	56·96	61·08	65·22	69·39
"Rather too probable"	·95	51·46	55·77	60·11	64·47	68·85	73·24	77·65
	·5	69·50	74·50	79·50	84·50	89·50	94·50	99·50
"Of doubtful improbability"	·05	90·25	95·93	101·59	107·24	112·86	118·47	124·06
"Improbable"	·01	99·63	105·60	111·54	117·45	123·33	129·19	135·02
"Very improbable"	·001	110·71	117·00	123·24	129·45	135·62	141·76	147·87

	P	105	110	115	120	125	130	135
"Much too probable"	·999	64·60	68·54	72·51	76·50	80·51	84·54	88·59
"Too probable"	·99	73·57	77·78	82·00	86·24	90·50	94·77	99·05
"Rather too probable"	·95	82·07	86·51	90·96	95·42	99·90	104·38	108·87
	·5	104·50	109·50	114·50	119·50	124·50	129·50	134·50
"Of doubtful improbability"	·05	129·63	135·19	140·74	146·28	151·81	157·33	162·83
"Improbable"	·01	140·84	146·63	152·41	158·17	163·91	169·64	175·36
"Very improbable"	·001	153·95	160·00	166·04	172·05	178·04	184·01	189·96

	P	140	145	150	155	160	165	170
"Much too probable"	·999	92·66	96·74	100·84	104·95	109·08	113·22	117·38
"Too probable"	·99	103·35	107·65	111·98	116·31	120·66	125·01	129·37
"Rather too probable"	·95	113·38	117·89	122·41	126·94	131·47	136·02	140·57
	·5	139·50	144·50	149·50	154·50	159·50	164·50	169·50
"Of doubtful improbability"	·05	168·33	173·82	179·30	184·77	190·23	195·69	201·14
"Improbable"	·01	181·06	186·75	192·43	198·10	203·76	209·40	215·04
"Very improbable"	·001	195·89	201·81	207·71	213·60	219·47	225·33	231·17

Note: Seal has pointed out that the values for $f>30$ are only approximate. Accurate χ^2 tables are available for $f=40$, 50, 60, 70, 80, 90 and 100 in *Biometrika*, XXXII, 1941, p. 187.

INDEX