FREQUENCY-CURVES

AND

CORRELATION.

BY

W. PALIN ELDERTON.

8/6m

FREQUENCY-CURVES

AND

CORRELATION.

BY

W. PALIN ELDERTON.

Published for the Institute of Actuaries

BY

CHARLES AND EDWIN LAYTON,
FARRINGDON STREET,
LONDON, E.C.

TABLE OF CONTENTS.

PART I.

PART II.

APPENDICES.

PREFACE.

The main object of the present Volume may be regarded as being to give a detailed description of the basis and practical application of those modern statistical methods that are associated with the name of Professor Karl Pearson.

The history of the work is briefly as follows:—In January, 1903, Mr. W. Palin Elderton read before the Institute of Actuaries an interesting paper dealing with the application of the Pearsonian frequency-curves to the graduation of a mortality experience, and it was then felt that the discussion of that paper suffered considerably from the fact that Professor Pearson's methods, which had attracted so much attention in purely statistical circles, were comparatively unfamiliar to the actuarial profession.

It was therefore suggested by more than one member of the Council that it would be exceedingly useful to the profession if Mr. Elderton would contribute to the *Journal* of the Institute "an explanatory paper dealing with frequency-curves, and giving illustrations of their use" based upon actuarial data. To this invitation Mr. Elderton replied in the most public-spirited manner by preparing a lengthy paper which forms the nucleus of the present volume. On consideration it was, however, felt that the work would be more generally useful if published separately instead of in the form of a paper in the *Journal*, and Mr. Elderton was good enough to undertake the necessary alterations and additions.

It is hoped that in its present form the volume may be of use not only to the actuarial profession, but also to other classes of statistical students who may be glad to have a connected account of Professor Pearson's methods.

Actuarial work, on its purely technical side, depends so largely upon the results of statistical enquiries that developments and improvements in statistical methods must always be of great interest to Actuaries, and the profession is much indebted to Mr. Elderton (who has had exceptional opportunities of becoming familiar with Professor Pearson's work) for the preparation of this volume, which must have involved a great expenditure of thought and labour. Time alone can show whether, and to what extent, the methods which he has so ably expounded will prove to be of practical value in actuarial work, and it would as yet be premature to express any opinion on this point. It may, therefore, be well to state that the illustrations based on actuarial data must, for the present, be regarded rather as examples of method than as indications of any official view as to the applicability of the methods to actuarial problems. The fact that no such expression of opinion by the Institute is yet possible, however, does not, in the slightest degree, lessen the indebtedness of all actuarial students to such work as Mr. Elderton's.

F. B. W.

INTRODUCTION BY THE AUTHOR.

By the preparation of the following pages an attempt is made to bring before Actuaries the more practical methods of modern statistical work. It is difficult to tell how far such methods may prove useful in direct application to actuarial problems, but even if they should happen to be only a slight assistance it seems advisable for Actuaries to have some knowledge of the contemporary study of a subject connected with their own work on the theoretical side. It has been necessary to exclude some recent work, and in making a selection it has been decided that the subject should be dealt with more fully from its practical than its theoretical aspect, and that, for the present, at any rate, it is best to omit the more difficult proofs, namely, those of the probable errors of the coefficients of correlation and of the formula for testing goodness of fit, while the proof that the method of moments will probably give very satisfactory results has been omitted because it is not necessary to an appreciation of moments as a practical method (as is obvious when we remember that it was not published till the method had been in use for some years), and also because it seems to the present writer that until adjustments have been found for all possible cases its practical value is somewhat discounted. The mean square contingency is dealt with, but the mean contingency is neglected because the mathematics leading to it are more awkward, and, though the numerical work is less, the results are not so satisfactory.

Some readers may be surprised that the method of least squares finds no mention, but they should bear in mind that the range of its applicability is so limited that there is a growing tendency to put it aside in curve fitting, and it seems best to concentrate attention on those parts of the subject more likely to be of permanent value to those for whom this book is intended. The median is not considered because it has no important bearing on the matter dealt with; and if it should happen to be required the only thing wanted is its definition, that it is the position such that the number of cases before it is equal to the number after it.

The coefficient of variation has not been dealt with, but here again it is well for the reader to know its meaning. In comparing the way two things vary it is necessary to remember that relative size influences not only the mean but the deviation from it, and in discussing variation this is taken into account by using one hundred times the ratio which the standard deviation bears to the mean $\left(\frac{100\sigma}{\text{mean}}\right)$.

Part II. will probably give Actuaries more difficulty than Part I. because it deals with a type of problem that has at present received little attention from actuarial students, and it is because the direct bearing of Part II. on actuarial work is somewhat uncertain that it is dealt with more in outline than Part I. Here, as in all other statistical study, examples must be worked out if the methods and principles are to be mastered. The reader who goes through a book on a practical subject and does not work out examples is as certain to encounter imaginary and miss real difficulties, as he is to fail to obtain any satisfactory knowledge of the subject.

The work may appear to some to demand more mathematical knowledge than most Actuaries possess, and it may therefore be well to point out that a practical man can use frequency curves and correlation reasonably without such knowledge, for the fact that a curve he has found agrees with the statistics from which the moments were obtained is a proof that in the particular case he has obtained proper values for the constants even though he has not followed the mathematical reasoning leading to the equations. It must not of course be inferred that belief without proof is considered advisable, but that it is unwise for a practical man to put aside a practical subject which he can test practically, merely because he cannot follow some of the proofs.

Frequency-curves and correlation form a subject in which there is still much to be done in spite of the rapid progress that has been made recently. There are few subjects which offer a richer field for original work than statistical mathematics and its applications. In this field the reader will find that in recent years we are indebted to Professor Karl Pearson for the majority of the work that has proved a success in practice, and anyone writing on the subject for practical men is bound to follow in his footsteps. Those who become interested in the subject are strongly recommended to study Professor Pearson's papers; it is not until they have done so that they will fully appreciate the great extent of his contribution to statistical science. The present Author has merely tried to bring together some of Professor Pearson's results and give them to members of his profession with examples that tend to show that actuarial statistics can be examined by his methods in the same way as the statistics of biology and anthropology. May not the continuation of such work add some links to the chain of continuity and indicate a wider law than an actuary studying his own subject exclusively might be led to suspect?

As will be readily appreciated, the Author is chiefly indebted to Professor PEARSON, but his indebtedness is of the kind for which it is impossible to offer formal thanks; such thanks would, at their best, fail to express the sense of gratitude which prompted them. He has also to acknowledge much kind help from Messrs. G. J. LIDSTONE and JOHN SPENCER, both of whom read the work in a somewhat different form in MS. and gave him many suggestions in connection with the arrangement of the matter, and the former has also helped him in many ways at the later stages, while Messrs. S. ADLARD and R. L. ELDERTON have devoted a large amount of time to reading the proofs, and have suggested difficulties that would probably arise and ways of removing them. Miss ETHEL M. ELDERTON has rendered assistance in some of the calculations and in other ways.

W. P. E.

LONDON, *August*, 1906.

KEY TO THE ACTUARIAL TERMS AND SYMBOLS USED.

The following explanation of certain technical terms and symbols that are used in this book is given as an assistance to non-actuarial readers. For a fuller account of the functions and notation reference can be made to the *Text-Book* for Actuaries (Parts I. and II.), and to the "Account of the Principles and Methods adopted in constructing the British Offices' Life Tables."—London: C. & E. Layton. 1903.

When an investigation is made into the mortality experienced among lives assured, the number of persons entering at each age, and the numbers passing out of observation at each age owing to (1) *death*, (2) *withdrawal*, by the policies lapsing, being surrendered, or terminating from some other cause, are recorded. The *exposed to risk* of death at age 25 (E_{25}) means the number who had the chance of dying between ages 25 and 26, and were on the average at risk for the whole year.

The number who die between ages 25 and 26, divided by the exposed to risk at age 25, gives *the probability of dying in a year* at age 25 (q_{25}).

When an experience ends in any year, say 1900, there will be a large number of persons who have been at risk, but whose policies are still in force; these are called *existing at the close of the observation.*

When a graduation has been made the *expected deaths* are found by multiplying the exposed to risk by the graduated values of q_x. The result is then compared with the actual deaths.

$O^{NM(5)}$ is the name given to the table of mortality obtained from the male lives assured by ordinary whole-life without profit policies between 1863 and 1893, excluding the first five years of assurance.

O^M is a table constructed from the similar with-profit assurances for all durations, and H^M (healthy males) is the name given to the older experience which ended in 1863.

q_x is (*see* above) the probability of dying in a year.

p_x is the probability of a person aged x living one year.

So if we imagine a stationary community, which a person can only leave by death, and consider l_x to be the number living at exact age x, then $l_{x+1}=p_x\times l_x$ and $l_{x+2}=p_{x+1}\times l_{x+1}$, and so on.

The value at a rate of interest i per unit of a sum of 1 payable if a person aged x be alive at the end of n years, is therefore $\frac{v^n l_{x+n}}{l_x}$ where $v=(1+i)^{-1}$, and the value of an annuity of 1 would be

$$\frac{vl_{x+1}+v^2l_{x+2}+\cdots}{l_x}.$$

Now for convenience in making tables it is well to multiply numerator and denominator of this expression by v^x, and we have as the value of the annuity

$$\frac{v^{x+1}l_{x+1}+v^{x+2}l_{x+2}+\ldots}{v^xl_x}=\frac{D_{x+1}+D_{x+2}+\ldots}{D_x}=\frac{N_{x+1}}{D_x}$$

where $\qquad D_x=v^xl_x$ and $N_x=D_x+D_{x+1}+\ldots$

Similarly $\qquad S_x=N_x+N_{x+1}+\ldots$

Tables of D, N, and S are called *commutation* columns.

$a_{\overline{n}|}$ is written for the value of an annuity of 1 payable for n years certain, independent of any life, so its value is $v+v^2+\ldots+v^n$.

$\bar{a}_{\overline{n}|}$ is the value of a similar annuity of 1 per annum payable m times a year when m takes the limit of ∞, so that its value is

$$\int_0^n v^m dm.$$

Colog p_x is the logarithm of the reciprocal of p_x, and Makeham's hypothesis assumes that its value is $A+Bc^x$; but colog $p_x = \log l_x - \log l_{x+1}$; therefore an alternative way of stating the hypothesis is $l_x = ks^x g^{c^x}$.

When valuing the policies in an assurance office the actuary groups cases together to save labour. When assurances are payable at death they are grouped according to the year of birth, but when they are payable at a certain *maturity age* or previous death *(Endowment Assurances)* they can be grouped either according to year of birth or according to the number of years to run (*unexpired term*). In the latter case they have to be valued by finding an average age at maturity; formerly this was done by taking the mean of the ages, but Mr. G. J. Lidstone has recently shown that a much more accurate result is reached by weighting the ages in Geometrical Progression. The constants used for this purpose are called Z.

A *model office* is an imaginary specimen office which is used for making approximate valuations.

FREQUENCY-CURVES AND CORRELATION.

PART I.

CHAPTER I.

INTRODUCTORY.

1. THE ordinary treatment of probability begins with the assumption that the chance that a certain event will occur is known, and proceeds to solve the problems that arise from the combination of events or the repetition of a particular experiment; it proves that a certain result is more likely to occur from experiment than any other, that a result based on a limited number of trials is unlikely to differ greatly from the expected result, and that the proportional deviation from the most probable result will generally decrease as the number of trials is increased.

Experiments can easily be made to show that the theoretical method leads to results which can be realised in practice when the probabilities can be estimated accurately beforehand; for example, various trials have been made with coin tossing, in which it has been found that if five coins are tossed together and the number of them coming down "heads" is recorded, then the distribution of the cases will agree with the binomial expansion $(\frac{1}{2}+\frac{1}{2})^5$ as the ordinary theory leads us to expect. Sequences of "heads" or "tails" form a series approximating to the Geometrical Progression with a common ratio of $\frac{1}{2}$, and the drawing of cards from a pack gives a result closely agreeing with the numbers that theoretical work suggests.

2. It frequently happens, however, that the probabilities are not known, and it is impossible to tell whether we are dealing with an experiment like coin tossing or sequences or card-drawing; in fact, the only thing known is the distribution of the number of cases into certain groups, and in these circumstances the inverse problem of tracing the theoretical series to which the statistics approximate may become an important matter. The difficulty of the subject is increased because statistics do not give the theoretical distribution exactly, and it is impossible to tell where the differences between the actual and theoretical results lie. To make the position clearer it will be well to re-state the problem and ask whether it is possible to find the theoretical series to which a series resulting from a statistical experiment approximates. It may be difficult, perhaps impossible, to trace the simple probabilities corresponding to a given case, but yet practicable to form a reasonable opinion of the series of numbers that might be reached if the experiment could be repeated an infinite number of times. On turning to the reasons which make it advisable to find this ideal result to which statistics approach, it will be seen that the elementary probabilities are not so important as they seem to be, and a reasonable representation of the series is of far greater practical value. We notice that one of the first objects of a statistician or an actuary dealing with statistical work is to express the observations in a simple form so that practical conclusions can be easily drawn from the figures that have been collected. If the available statistics fall naturally into fifty or sixty groups he has to decide how they can be arranged to bring out the important features of the problem on which he is working, and if he can find four or five numbers closely connected with the original series which can be used as an index to the whole, he can then give the result in a way that might assist comparison with similar statistics, and enable others who have to deal with the facts to appreciate the whole distribution more readily than they could do if it remained in its original form. The statistician has also to supply approximate values for intermediate terms when only a few can be obtained from his experience, or complete or continue a series when only a part of it is known. In many cases he has to keep the same terms as his original series, but remove the roughnesses of material due to limitations in the number of cases available for his investigation; that is, he has to graduate his data.

3. In reality these objects are much alike, for if the statistical tables can be represented by an algebraic or transcendental formula, we can replace the whole series of numbers by a few values (the constants in the formula) which, if we deal systematically with the distributions we meet, facilitate comparison or enable us to supply missing terms, while the roughness of the original material can be removed by making a suitable formula represent the original statistics as nearly as possible. If a formula is based on the theoretical considerations, it should also give a solution of the problem in probabilities mentioned at the outset, and we see that both the practical and theoretical requirements can be dealt with at the same time, for the smooth series sought by the theoretical student is the same thing as the formula required for practical work.

4. The advantages of any system of curves depend on the simplicity of the formulæ and the number of classes of observations that can be dealt with satisfactorily, for a complicated expression is very little improvement on the original groups of statistics, and a system which is not capable of general application leaves the statistician in difficulties whenever it breaks down. One other thing is necessary; if a formula is known to be a suitable one there must be some method of finding the arithmetical constants that will give a good agreement in the particular case. Such a method, if it is to be of practical use, must be simple, reliable and capable of general and systematic application.

A broad idea of the objects to be accomplished ought to be kept clearly before the mind; they are likely to be forgotten because of the large amount of detail necessarily connected with the subject. It is also important because the advantages of systematic treatment are often overlooked, and short cuts and rough and ready methods are adopted to the detriment of the work, and formulæ having no scientific basis and having no connection with others suitable to similar cases are sometimes used in rather haphazard fashion by statisticians. The consequence is that generalisation is impossible, and where a law might be found one can see little but a great variety of attempts by energetic workers to reach their own conclusions regardless of the value of comparative statistics.

CHAPTER II.

Frequency Distribution.

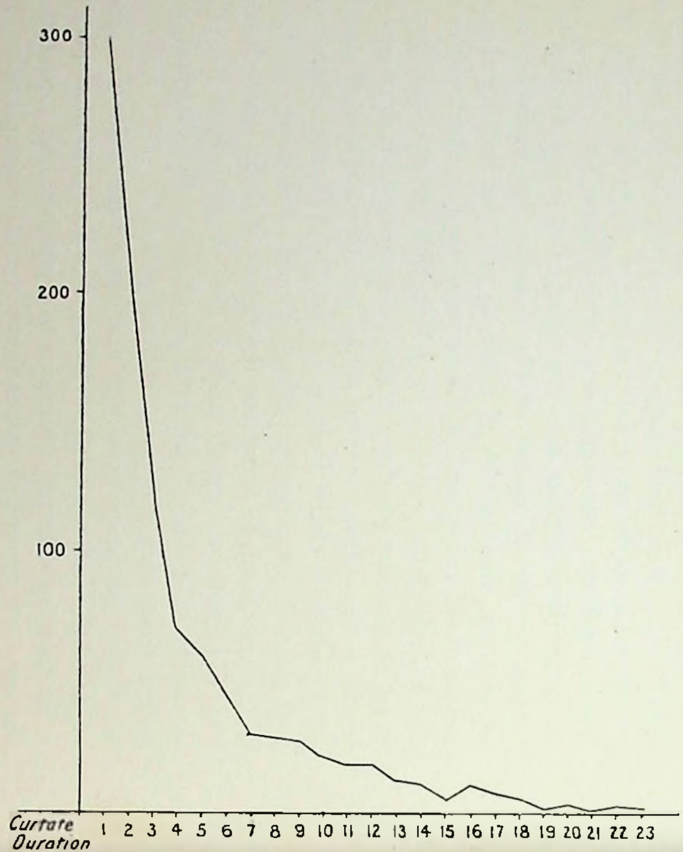
1. If statistics are arranged so as to show the number of times, or frequency with which, an event happens in a particular way, then the arrangement is a frequency distribution. Although some of our results will be of wider applicability, we shall generally confine our attention to these distributions.

2. It is necessary to have a name for the formula used to describe such distributions, and the term frequency-curve has been adopted for the purpose. The geometrical progression which describes the number of sequences in any direct experiment, such as coin tossing or dice throwing, is a frequency-curve, the equation to which is $y = Na^x$.

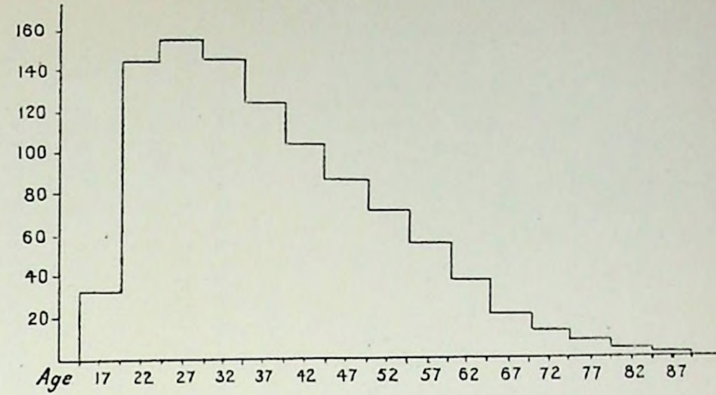
3. Some distributions give the number of cases falling in a certain group of values of the independent variable, while others (*e.g.*, Example V. of Table I.) give the number of cases for an exact value, and in the former case the exact values of the independent variable to which the groups correspond must be considered; for instance, "exposed to risk at age x" includes those from $x-\frac{1}{2}$ to $x+\frac{1}{2}$, but the number of deaths at duration n those from n to $n+1$. When statistics are represented graphically, effect should be given to these differences, and, to bring out the points a little more clearly, the diagrams on pp. 6 and 7 have been prepared. The drawings of distributions, such as those in the diagram, are called frequency polygons or histograms.

4. When statistics give the number of cases for an exact value of the independent variable, it is simple to plot them in a diagram by drawing ordinates and joining their tops, but in the case of groups of values there is a little complication, for

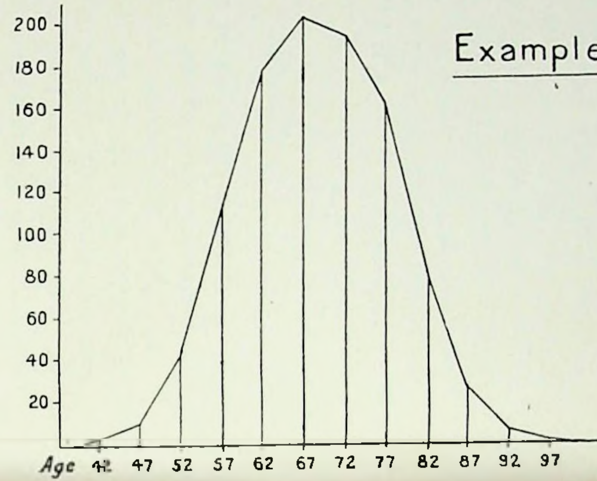
Example I.



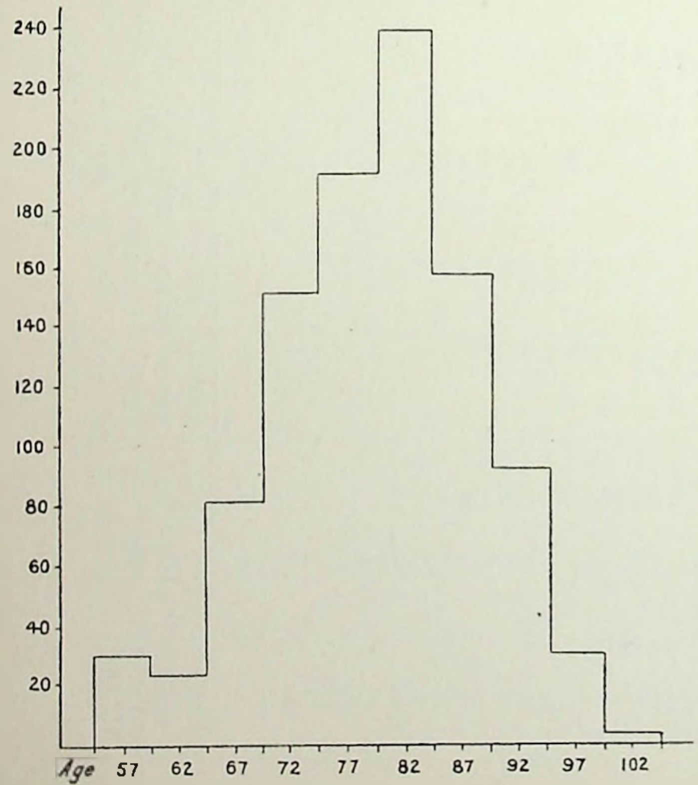
Example II.



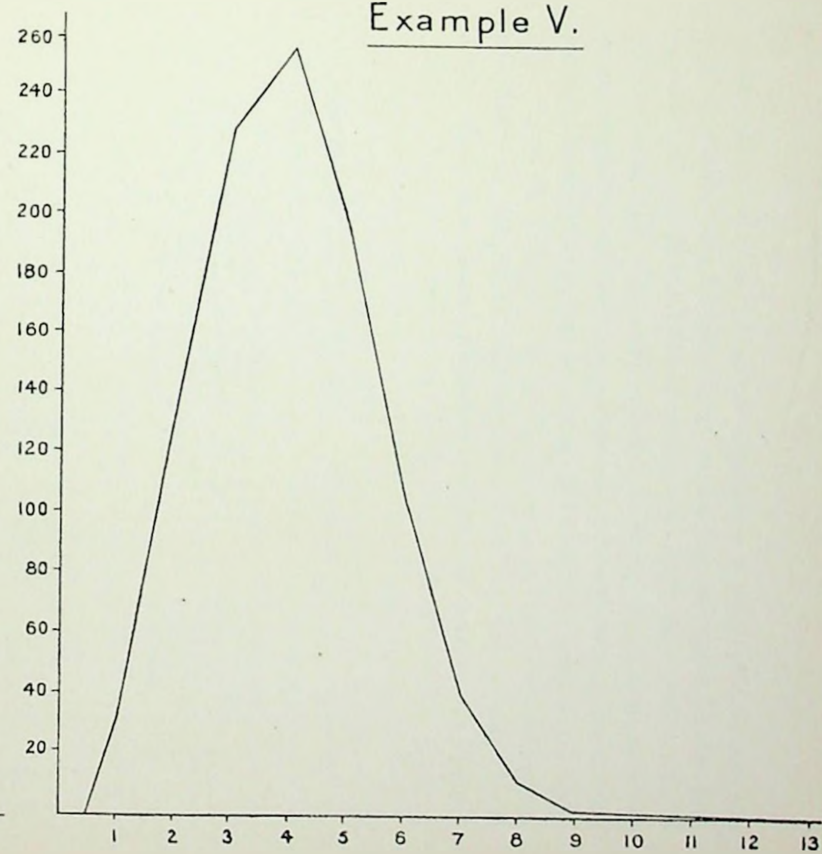
Example III.



Example IV.



Example V.



we can either draw a rectangle standing on the entire base (Ex. II. of diagram) or put in ordinates at the middle points of the bases and then join their tops (Ex. III.). The former method seems to give a better idea of the amount of information conveyed by the statistics, but, for some purposes (*e.g.*, for seeing the possible shape of the curve), the latter is more convenient.

5. If the reader will now examine the examples in Table I., he will notice that the statistics tend towards a smooth series

TABLE I.

Example I.		Example II.		Example III.	Example IV.	Example V.	
Curtate Durations.	Withdrawals with monthly incidence "0" in year of exit (p. 92, *Principles & Methods*).	Ages.	Exposed to risk of Sickness (Watson, *M. U. Tables*, p. 19).	Existing at close of observations Without Profit "Old" Assurances.	Existing at close of observations "Old" Annuities (Females).	Terms of the expansion of $1000(\frac{1}{4}+\frac{3}{4})^{12}$	No. of term.
1	308	–19	34	...	...	32	1
2	200	20–24	145	...	...	127	2
3	118	25–29	156	...	...	232	3
4	69	30–34	145	...	...	258	4
5	59	35–39	123	...	...	194	5
6	41	40–44	103	3	...	103	6
7	29	45–49	86	9	...	40	7
8	28	50–54	71	42	...	11	8
9	26	55–59	55	111	29	2	9
10	21	60–64	37	176	23	1	10
11	18	65–69	21	200	81	...	11
12	18	70–74	13	193	151	...	...
13	12	75–79	7	160	192	...	...
14	11	80–84	3	73	239	...	...
15	5	85–89	1	26	157	...	...
16	11	90–94	...	6	93	...	...
17	7	95–99	...	1	29	...	...
18	6	100–	...	...	6	...	...
19	1	...	...	...	...	...	...
20	3	...	...	...	...	...	...
21	1	...	...	...	...	...	...
22	3	...	...	...	...	...	...
23	2	...	...	...	...	...	...
...	1,000	...	1,000	1,000	1,000	1,000	...
True Total	1,308	...	2,995,724	2,674	172	...	...
Mean	4·182	...	37·8750	68·485	79·400	3·998	...
Standard Deviation	4·1996	...	2·76810	1·771288	1·774894	1·46215	...
Type	I	...	I	II	VII		...

as the total number of cases is increased, and from this it can be seen how naturally practical statistics lead to the conception of a frequency-curve to describe the smooth distribution that would be obtained if an infinite supply of homogeneous material were available for investigation. In other words, such curves would give an approximation to the total "population" of which the particular case investigated is a sample.

6. It may be noticed that a frequency-curve will give a frequency corresponding to every value of the independent variable along the whole range of the distribution, and will not restrict us to a few more or less arbitrary groups as is necessarily done by the actual statistics. The binomial series and geometrical progression do the same when we imagine we are dealing with something that can be divided into a very large number of groups. Thus, if we mix a large quantity of sand of two colours and take out a fixed quantity of the mixture and record the number of grains of sand of either colour in each drawing, we should obtain a continuous curve from a large number of trials.

7. We will now define some important functions. When a distribution is arranged according to the progressive values of a variable characteristic, *e.g.*, duration, age, &c., the average value of that characteristic (not the average of the frequencies) is called the *mean* of the distribution, and is given by

$$\frac{f_a \times a + f_b \times b + f_c \times c + \ldots + f_n \times n}{f_a + f_b + f_c + \ldots + f_n}$$

where f_r is the frequency corresponding to r; thus, in Example I., 200 is the frequency corresponding to 2. If we assume infinitesimal increments, the mean is given by

$$\frac{\int f_x \times x dx}{\int f_x dx},$$

where the limits of the integral will be such as to cover the whole distribution. The mean could also be described as the position of the ordinate through the centre of gravity of the distribution (centroid vertical), and this may be of help to some readers.

8. The *mode* is the characteristic that occurs most frequently, or, in other words, is the position of the maximum ordinate, and its calculation can therefore only be made approximately, unless we know the law connecting the various groups. We cannot find the mode exactly until we know the frequency-curve, because it is the position of an ordinate, and we cannot tell from the rough statistics which ordinate is greatest.

9. Now since one equation or curve might be used for several distributions, one given according to age, a second of a different subject according to duration, a third according to sums assured, and so on, we must have a standard of reference based on the distribution itself. For this purpose a function known as the *standard deviation* is used. It is given by

$$\sqrt{\left\{\frac{f_a a'^2+f_b b'^2+\ .\ .\ .\ +f_n n'^2}{f_a+f_b+\ .\ .\ .\ +f_n}\right\}}$$

where $a', b' \ldots n'$ are the distances from the mean. In the form of integrals the standard deviation is

$$\sqrt{\left\{\frac{\int f_x \times x^2 dx}{\int f_x dx}\right\}}$$

where the distances x are measured from the mean.

The standard deviation measures the way the frequencies are distributed in terms of the unit of measurement. Since the frequencies furthest from the mean are multiplied by the largest values of x, a large standard deviation shows that the frequency distribution spreads out from the mean, while a small standard deviation shows that the frequency is closely concentrated about the mean. In considering the relative sizes of standard deviations, it is necessary to bear in mind the unit of measurement, because, if a given distribution is arranged in two series, first, according to years of age, and then in quinquennial age groups, the standard deviation will be five times as large in the latter case as it is in the former. This can be seen at once by comparing the two expressions

$$\sqrt{\left\{\frac{\int f_x x^2 dx}{\int f_x dx}\right\}} \quad \text{and} \quad \sqrt{\left\{\frac{\int f_x (5x)^2 dx}{\int f_x dx}\right\}}$$

The latter is obviously five times the former. The values of the standard deviations are given in Table I. for each case.

The diagram, on p. 12, shows two curves having the same mean B and approximately the same area, but the dotted curve has the larger standard deviation because it spreads out more on each side of the mean.

The reader will notice from the algebraic expressions given above that the standard deviation is not dependent on the number of cases (*i.e.*, on the absolute size of the curve), but merely on the way they are distributed (*i.e.*, on the proportionate numbers or the shape of the curve); it measures the "spread" or "scatter" of the statistics from the mean.

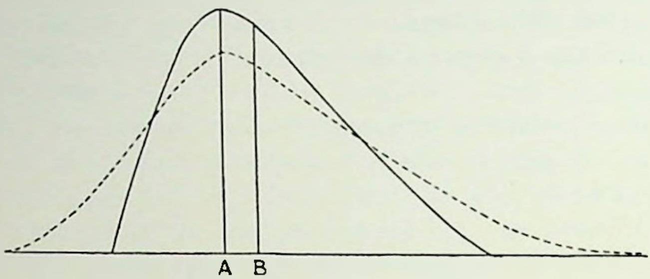
10. An examination of frequency distributions (*see* Table I. and pp. 6 and 7) shows that most of them start at zero, gradually rise to a maximum, and then fall sometimes at a very different rate. If the rise and fall are at the same rate, distribution will be symmetrical about its mean, which will obviously coincide with the mode. The difference between the mean and mode is therefore a function of the *skewness* or deviation from symmetry. In order to get a satisfactory measure, the way the material is grouped must be taken into account, and this leads us to measure skewness by (distance between mean and mode) ÷ standard deviation. If the mean is on the left-hand side of the mode when the statistics are plotted out in diagram, this function will be negative, and to remember the sign it is convenient to write:

$$\text{Skewness} = \frac{\text{Mean}-\text{Mode}}{\text{SD}}$$

The diagram will help to show the rationale of the measure for skewness. It gives two curves having the same mean B and the same mode A, but with different standard deviations, and it is clear that the dotted curve, with its larger standard deviation, is more nearly symmetrical than the other curve.

11. We may summarise these functions by saying that the mean and mode fix the position of the curve on the axis; the standard deviation shows how the material is distributed about the mean, and the skewness shows the amount of the deviation from symmetry exhibited by the material.

These preliminary definitions will be sufficient for our present purpose, but the functions defined will be more easily understood when their actual connection with the practical work of curve fitting has been studied. A student working at the subject for the first time should plot out several distributions on cross-ruled paper, in order to familiarise himself with their nature and appearance. He should calculate and insert the means in the diagrams, but should not attempt to calculate standard deviations until he knows something of the method of moments.



CHAPTER III.

METHOD OF MOMENTS.

1. BEFORE we proceed to deal with suitable forms for use as frequency-curves, it will be well to see if some method of applying them to statistical examples can be found, for it is clearly useless to suggest a curve and have no way of using it. We require, therefore, a general method by which a given formula can be fitted to a particular statistical experience, and may be applied to any expression (for instance, Makeham's formula for the force of mortality) on which we may have decided as the basis of graduation. The first point to be noticed in searching for such a method is that if there are n constants in the formula, we must form n equations between the formula and the statistics. Thus, if we have three terms, say, $y=20$, 40 and 88, when $x=1$, 2, and 3 respectively, and wish to use the curve $y=a+bx+cx^2$ to describe them, we can, of course, find values of a, b and c so that each item is exactly reproduced by equating as follows :—

$$a+b+c=20$$
$$a+2b+2^2c=40$$
$$a+3b+3^2c=88$$

But if we have a fourth term $y=96$ when $x=4$, and use the values of a, b and c found from the three equations just given, we should find that when $x=4$ $y=164$. This suggests that when there are more terms in the statistics than there are constants, the equations must be formed by using all the terms, not by selecting from them. The graduating curve will not necessarily reproduce exactly any of the observations but will run evenly through the roughnesses of the observed facts so as to represent their general trend.

2. Let $a_1, a_2, a_3 \ldots a_n$ be n terms to be graduated; then, if the series were perfectly smooth and followed a known law, each term could be reproduced exactly by, say, $b_1, b_2, b_3 \ldots b_n$ where $a_1=b_1$, $a_2=b_2$, $a_3=b_3 \ldots$ and $a_n=b_n$. Now, if we consider the two series (the a's and the b's), we see that since each term is reproduced exactly

$$\sum_{r=1}^{r=n} a_r \qquad = \sum_{r=1}^{r=n} b_r$$

and

$$\sum_{r=1}^{r=n} c_r a_r = \sum_{r=1}^{r=n} c_r b_r$$

where c_r is a numerical coefficient.

This suggests a possible method to apply when each term cannot be reproduced exactly. The total of the graduated figures must be made equal to the total of the ungraduated, and the further equations necessary for finding the unknown constants must be formed by multiplying the various terms by different factors and similarly equating the sums of the graduated and ungraduated products, *i.e.*, $\Sigma c_r a_r = \Sigma c_r b_r$. It still remains to decide the best form to be given to c_r, and the mean being equal to

$$\frac{a_1 + 2a_2 + \ldots + na_n}{a_1 + a_2 + \ldots + a_n}$$

suggests that $c_r = r$ should give one reasonable equation. Again, since we shall have to use some function of r which, when applied to the graduation formula, will give an integrable form (otherwise, we cannot make an equation between $\Sigma c_r b_r$ and $\Sigma c_r a_r$), the powers of r suggest themselves as convenient when integration by parts is attempted. If, therefore, we write $c_r = r^t$ and give t successively the values 0, 1, 2 ... we can obtain as many equations as we require, and the first two of them give the area and mean of the distribution, which will be the same in the graduated and ungraduated figures.

This method is known as the Method of Moments, (*cf.*, moments of inertia), and Professor Karl Pearson has recently shown (*Biometrika*, vol. i., pp. 267, &c.), that it can be expected to give very good results.

3. Applying the method to solve the three equations given above, we have

$$(a+b+c)+(a+2b+2^2c)+(a+3b+3^2c)=20+40+88$$
$$(a+b+c)+2(a+2b+2^2c)+3(a+3b+3^2c)=20+2\times 40+3\times 88$$
$$(a+b+c)+2^2(a+2b+2^2c)+3^2(a+3b+3^2c)=20+2^2\times 40+3^2\times 88$$

or
$$3a+\ \ 6b+14c=148$$
$$6a+14b+36c=364$$
$$14a+36b+98c=972$$

These equations will give the same result as those from which they were formed, because each of the three terms can be graduated exactly; but if we introduce the fourth term, $x=4$, $y=96$, we can modify the moment method by adding a fourth term to each equation given above and obtain

$$4a+10b+30c \quad = \quad 244$$
$$10a+30b+100c \ = \ 748$$
$$30a+100b+354c=2,508$$

The solution of these equations gives

$$a=-25{\cdot}5$$
$$b=\quad 42{\cdot}6$$
$$c=-3$$

or
$$x=1 \qquad y=14{\cdot}1$$
$$x=2 \qquad y=47{\cdot}7$$
$$x=3 \qquad y=75{\cdot}3$$
$$x=4 \qquad y=96{\cdot}9$$

This is a very simple example, but it will probably help to show the way results are reached, and will serve as a foundation for what follows.

4. The nth moment of a particular frequency is defined as the product of the frequency and the nth power of the distance of the frequency from the vertical about which moments are being taken; or the nth moment of any *ordinate* y of a frequency-curve about the vertical through a point distance x from it is yx^n, and the nth moment of the whole distribution treated as a series of ordinates is $y_1x_1^n+y_2x_2^n+\ldots$ where $y_1+y_2+\ldots$ is the total frequency. Thus, in Example IV., the third moment of the frequency 81 about the vertical through age 77 is $81\times(-2)^3$ where 5 years is the unit distance.

5. If the ordinates are known, we can calculate the moment for them immediately by multiplying the frequencies by the powers of the distances between them and the vertical about which the moments are required and then add the results, care being taken to give the distances their proper signs. If areas are given, an approximation is made by assuming them to be concentrated about the ordinates at the middle points of the bases on which they stand—the moments thus obtained are sometimes said to be based on "loaded ordinates." The columns after the third in Table II. show the calculation of the moments about the vertical through age 77 for Example IV. of Table I., on the assumption that the frequencies are concentrated at the middle points of the bases.

TABLE II.

Central Age of Group x (1)	Frequency f (2)	$\frac{x-77}{5}=s$ (3)	$f \times s$ (4)	$f \times s^2$ (5)	$f \times s^3$ (6)	$f \times s^4$ (7)
57	29	-4	116	464	1,856	7,424
62	23	-3	69	207	621	1,863
67	81	-2	162	324	648	1,296
72	151	-1	151	151	151	151
77	192	0	-498	...	$-3,276$	...
82	239	1	239	239	239	239
87	157	2	314	628	1,256	2,512
92	93	3	279	837	2,511	7,533
97	29	4	116	464	1,856	7,424
102	6	5	30	150	750	3,750
Totals	1,000	...	$+978$ $+480$	3,464	$+6,612$ $+3,336$	32,192

NOTATION FOR MOMENTS.

N = total frequency.

ν_n = nth unadjusted statistical moment about mean.

ν'_n = nth unadjusted statistical moment about any other point.

μ_n = nth moment from curve about mean.

= nth adjusted statistical moment about mean.

μ'_n = nth moment from curve about other point.

= nth adjusted statistical moment about other point.

NOTE.—ν, ν', μ and μ' always refer to a total frequency of unity.

The unit of grouping has been taken as 5 years, and if, as is often convenient, we assume the total frequency to be unity, the totals will have to be divided by 1,000. We should generally deal with the actual numbers that occur, but as they have been given in Table I. as the distribution of 1,000 cases, it will be better to use them in that way in the present case. The numbers -4, -3 . . . in column (3) show the distances from age 77 in terms of the unit of grouping. The centre of any other group would have done almost as well as 77; it is convenient to choose the arbitrary origin so that it is near the mean of the distribution. This makes easier the calculation of the moments about the mean (a result frequently required), and enables the calculator to get a rough check on these moments by comparing them with those about the arbitrary origin. The columns (4) to (7) are sufficiently explained by their headings; they are formed successively and checked by multiplying f by s^1, the values of s^1 being taken from a table of the powers of the natural numbers.

6. It has so far been assumed that moments can be calculated about any point, but it is frequently inconvenient to do so; for if we had required them about age 79·4, we should have had to multiply by the powers of $\frac{77-79{\cdot}4}{5}$, of $\frac{82-79{\cdot}4}{5}$ and so on, and it is quite clear that the labour would have been very great. In such a case we can, however, take the moments about any other more convenient point, and then modify them in the following way:—

Let the distance between A, about which the moments are known, and B, about which they are required, be $+d$; thus, if we want moments about 25·7 and have found them about 25, d is ·7; if we had found them about 26, d would have been $-{\cdot}3$.

Then, if the distance of any ordinate y_r from A is X_r, and from B is x_r, then

$$x_r = X_r - d$$

and

$$x_r^n = (X_r - d)^n.$$

Now, the nth moment of the whole distribution treated as a series of ordinates is $\Sigma y_r X_r^n$ about A, and $\Sigma y_r x_r^n$ about B; so we have

$$\begin{aligned} \nu''_n = \Sigma y_r x_r^n &= \Sigma y_r (X_r - d)^n \\ &= \Sigma[y_r(X_r^n - ndX_r^{n-1} + \ldots + \overline{-1}^n d^n)] \\ &= \nu'_n - nd\nu'_{n-1} + \frac{n(n-1)}{2!} d^2 \nu'_{n-2} - \quad . \quad . \qquad (1) \end{aligned}$$

where ν''_n is written for the nth moment about B, and ν'_n the nth moment about A.

Instead of (1) we may proceed as follows :—

$$\nu'_n = \Sigma y_r X_r = \Sigma[y_r(x_r + d)^n]$$

$$= \nu''_n + nd\nu''_{n-1} + \frac{n(n-1)}{2!} d^2\nu''_{n-2} + \ldots$$

$$\therefore \qquad \nu''_n = \nu'_n - nd\nu''_{n-1} - \frac{n(n-1)}{2!} d^2\nu''_{n-2} - \ \ldots \qquad (2)$$

There is little to choose between these two formulæ, and of course they give identical results.

7. We will now apply formula (2) to work out the moments about the centroid vertical (*i.e.*, vertical through the mean) for the example in Table II. The distance of the mean from any point is

$$\frac{\Sigma(X_r y_r)}{\Sigma y_r} = \frac{\Sigma(X_r y_r)}{N}$$

where N is the total frequency; or we may say that the distance of the mean from any point is the first moment of the distribution about the vertical through that point. It follows that the first moment about the centroid vertical is zero, and this leads me to prefer formula (2) to formula (1) when moments are required about that vertical. When we come to deal with frequency-curves, we shall see that this is generally the case.

8. The arithmetical work is as follows :—

The totals in cols. (4) to (7) are divided by the number of observations (total of col. (2)), and the quotients are the moments (ν') about 77. The moments are dealt with as having reference to a case where unity is the total frequency, *i.e.*, proportional, not actual, frequencies are dealt with.

$$\nu'_1 = \cdot480 \qquad \nu'_2 = 3\cdot464$$

$$\nu'_3 = 3\cdot336 \qquad \nu'_4 = 32\cdot192$$

The value of ν'_1 gives the mean age $= 77 + 5 \times \cdot480 = 79\cdot4$. In order to use formula (1) or (2), the value of d is required, and when the calculation of moments has to be made about the centroid vertical its value is, as we have seen above, the same as ν'_1; in the present case it is the first moment about the vertical through age 77. The powers of d are next calculated by logarithms; as it happens d is a comparatively simple number; if it had been ·48327, say, the propriety of using logarithms would have been more obvious—

$$d^2 = \cdot 2304 \qquad d^3 = +\cdot 110592 \qquad d^4 = \cdot 0530842.$$

In modifying formula (2) for the particular case in which moments about the centroid vertical are required, it should be remembered that ν_1 is zero and ν_0 is unity, because it is merely the total frequency divided by the total frequency.

$$\nu_2 = \nu'_2 - d^2 \qquad = \quad 3\cdot 2336$$

$$\nu_3 = \nu'_3 - 3d\nu_2 - d^3 \qquad = -1\cdot 430976$$

$$\nu_4 = \nu'_4 - 4d\nu_3 - 6d^2\nu_2 - d^4 = \ 30\cdot 416289.$$

A seven-place logarithm table and antilogarithm table (such as Filipowski's), an Arithmometer or Brunsviga should be used. It will be noticed that $6d^2\nu_2$ can be formed very easily from $3d\nu_2$ when logarithms are used, as $\log d$ is known. It is useful to keep a note of this value ($\log d$) in a conspicuous place when the moments are being calculated.

9. Although the above is the most direct and obvious way by which moments can be calculated, another method was suggested by Mr. G. F. Hardy and used by him in his recent graduation of the British Offices Life Tables. He pointed out that by summing the statistical numbers and forming a new series in the same way as the N_x column is formed from the D_x column and then summing these results (*cf.* the column), and so on, equations can be formed. So far as I can trace, Mr. Hardy has not shown the connection between the summation method and the direct calculation of the moments, though he has pointed out that the same results can be obtained.

The arrangement on p. 20 shows both the method of calculation and the form of the expression obtained by the process.

Considering the line opposite the first term, we notice that the sum of the series is given, and that the second summation, which we will call S_2 when the total frequency is taken as unity gives the first moment of the whole distribution about a vertical situated at unit distance before the point corresponding to $f(1)$. Still considering only the first line, we see that S_3 gives each function multiplied by coefficients of the form

TABLE III.

No. of Term.	Function.	Sum of Function.	Second Sum of Function.	Third Sum of Function.	Fourth Sum of Function.	Fifth Sum of Function.
1	$f(1)$	$f(1)+f(2)+\ldots+f(n)$	$f(1)+2f(2)+\ldots+nf(n)$	$f(1)+3f(2)+6f(3)+\ldots+\frac{n\,.\,n+1}{2}f(n)$	$f(1)+4f(2)+\ldots+\frac{n\,.\,n+1\,.\,n+2}{3\,!}f(n)$	$f(1)+5f(2)+\ldots+\frac{n\,.\,n+1\,.\,n+2\,.\,n+3}{4\,!}f(n)$
2	$f(2)$	$f(2)+f(3)+\ldots+f(n)$	$f(2)+2f(3)+\ldots+(n-1)f(n)$	$f(2)+3f(3)+\ldots+\frac{n-1\,.\,n}{2}f(n)$	$f(2)+4f(3)+\ldots+\frac{n-1\,.\,n\,.\,n+1}{3\,!}f(n)$	$f(2)+5f(3)+\ldots+\frac{n-1\,.\,n\,.\,n+1\,.\,n+2}{4\,!}f(n)$
3	$f(3)$	$f(3)+\ldots+f(n)$	$f(3)+2f(4)+\ldots+(n-2)f(n)$	$f(3)+3f(4)+\ldots+\frac{n-2\,.\,n-1}{2}f(n)$	$f(3)+4f(4)+\ldots+\frac{n-2\,.\,n-1\,.\,n}{3\,!}f(n)$	$f(3)+5f(4)+\ldots+\frac{n-2\,.\,n-1\,.\,n\,.\,n+1}{4\,!}f(n)$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$n-2$	$f(n-2)$	$f(n-2)+f(n-1)+f(n)$	$f(n-2)+2f(n-1)+3f(n)$	$f(n-2)+3f(n-1)+6f(n)$	$f(n-2)+4f(n-1)+10f(n)$	$f(n-2)+5f(n-1)+15f(n)$
$n-1$	$f(n-1)$	$f(n-1)+f(n)$	$f(n-1)+2f(n)$	$f(n-1)+3f(n)$	$f(n-1)+4f(n)$	$f(n-1)+5f(n)$
n	$f(n)$	$f(n)$	$f(n)$	$f(n)$	$f(n)$	$f(n)$
General form of the sums on line 1.		$\Sigma f(n)$ $=\nu'_0$	$\Sigma nf(n)$ $=\nu'_1$	$\Sigma\frac{n^2+n}{2}f(n)$ $=\frac{1}{2}(\nu'_2+\nu'_1)$	$\Sigma\frac{n^3+3n^2+2n}{6}f(n)$ $=\frac{1}{6}(\nu'_3+3\nu'_2+2\nu'_1)$	$\Sigma\frac{n^4+6n^3+11n^2+6n}{24}f(n)$ $=\frac{1}{24}(\nu'_4+6\nu'_3+11\nu'_2+6\nu'_1)$

$\frac{n(n+1)}{2!}$ or $\frac{n^2+n}{2!}$, *i.e.*, it gives $\frac{\nu'_2+\nu'_1}{2}$ where ν' is written for the moment, because by definition the tth moment (ν'_t) of the whole distribution is given by the sum of $n^t f(n)$ for all values of n. S_4 and S_5 give each function multiplied by $\frac{n^3+3n^2+2n}{6}$ and $\frac{n^4+6n^3+11n^2+6n}{24}$ respectively.

The following equations result :—

$$S_2 = \nu'_1$$

$$S_3 = \tfrac{1}{2}(\nu'_2+\nu'_1)$$

$$S_4 = \tfrac{1}{6}(\nu'_3+3\nu'_2+2\nu'_1)$$

$$S_5 = \tfrac{1}{24}(\nu'_4+6\nu'_3+11\nu'_2+6\nu'_1).$$

These equations enable us to calculate the moments about the selected origin, but if it is necessary to find moments about the mean, the following relations are more convenient; they can be reached by substituting in the above the values in formula (2), and remembering that $S_2 = d$.

$$\nu_2 = 2S_3 - d(1+d)$$

$$\nu_3 = 6S_4 - 3\nu_2(1+d) - d(1+d)(2+d)$$

$$\nu_4 = 24S_5 - 2\nu_3\{2(1+d)+1\} - \nu_2\{6(1+d)(2+d)-1\}$$
$$- d(1+d)(2+d)(3+d).$$

10. The following table shows the working in the numerical example already dealt with by the direct method. The fifth sum is unnecessary, as the total of the items in the fourth sum gives the only value required :—

TABLE IV.

Frequency	First Sum.	Second Sum.	Third Sum.	Fourth Sum.
29	1,000	5,480	19,372	54,508
23	971	4,480	13,892	35,136
81	948	3,509	9,412	21,244
151	867	2,561	5,903	11,832
192	716	1,694	3,342	5,929
239	524	978	1,648	2,587
157	285	454	670	939
93	128	169	216	269
29	35	41	47	53
6	6	6	6	6
Total (for check) 1,000	5,480	19,372	54,508	132,503

From the totals of the columns we have—

$$S_2 = d = 5{\cdot}48,\ S_3 = 19{\cdot}372,\ S_4 = 54{\cdot}508, \text{ and } S_5 = 132{\cdot}503.$$

The first value S_2 or d shows that the mean is at age $52 + 5{\cdot}48 \times 5 = 79{\cdot}4$. The age 52 is used because it is the centre of the group before that in which numbers occur and, as has been already remarked, the summation method assumes the work to be done with reference to this position. The application of the formula for ν_2, ν_3, and ν_4, given above, enables us to find

$$\nu_2 = \quad 3{\cdot}2336$$
$$\nu_3 = -\ 1{\cdot}43099$$
$$\nu_4 = \quad 30{\cdot}4164$$

11. This is the most obvious way of using the summation method, but if the series contains a great number of terms, it is more convenient to use a central term instead of the first term as the starting point for the summation.* A slight adjustment is then needed because, though there is no difficulty about the calculation of the sum for the terms on the positive side of the selected point, the moments for the terms on the negative side are formed by multiplying by the powers of negative quantities. In order to use the formulæ given above, we require $\Sigma nf(n)$; $\Sigma \frac{n(n+1)}{2} f(n)$, and so on, or when n is negative, $\Sigma - nf(-n)$; $\Sigma \frac{-n(-n+1)}{2} f(-n)$ or $\Sigma \frac{n(n-1)}{2} f(-n)$; $\Sigma \frac{-n(-n+1)(-n+2)}{6} f(-n)$ or $\Sigma - \frac{n(n-1)(n-2)}{6} f(-n)$; and

$$\Sigma \frac{-n(-n+1)(-n+2)(-n+3)}{24} f(-n) \text{ or } \Sigma \frac{n(n-1)(n-2)(n-3)}{24} f(-n).$$

The first of these is given by the *last term* in the ordinary second summation taken negatively; the second is seen from Table III. to come from the *term before the last* in the ordinary third sum; the third is the *second term before the last* in the fourth sum taken negatively; and the fourth is the *third term before the last* in the fifth sum; the sums in each case being begun from the

* I have to thank Mr. G. J. Lidstone for the suggestion of this improvement in the method.

central term but in the reverse direction from the sums on the positive side. To make the method clearer the following table has been prepared, showing the calculation of the summation about the centre of the group of which the frequency is 192.

TABLE IV.(A).

Frequency.	First Sum.	Second Sum.	Third Sum.	Fourth Sum.	Fifth Sum.
29	29	29	29	29	29
23	52	81	110	139	...
81	133	214	324	...	...
151	284	498	...	...	...
192					
239	524	978	1,648	2,587	3,854
157	285	454	670	939	...
93	128	169	216	269	...
29	35	41	47	53	...
6	6	6	6	6	...
1,000					

$$S_2 = \cdot 978 - \cdot 498 = \cdot 48$$

$$S_3 = 1\cdot 648 + \cdot 324 = 1\cdot 972$$

$$S_4 = 2\cdot 587 - \cdot 139 = 2\cdot 448$$

and $$S_5 = 3\cdot 854 + \cdot 029 = 3\cdot 883$$

Hence $$\nu_2 = 2 \times 1\cdot 972 - \cdot 48 \times 1\cdot 48 = 3\cdot 2336$$

and similarly $$\nu_3 = -1\cdot 43097$$

and $$\nu_4 = 30\cdot 41621$$

agreeing with the previous results.

12. A comparison of Table IV.(A) with Table IV. will show that a saving of numerical work is effected by using a central point as the starting point for the summation, for the sums are numerically smaller and the value of S_2 or d, which enters into the formulæ on p. 21, is much smaller. It will be readily appreciated that whenever there is a large number of terms the summation method, and especially the form of it given in Table IV.(A), is a very great improvement on the product method of calculating moments. By means of an adding machine, such as Burroughes' adding machine, the summations can be obtained mechanically with little trouble, even for series containing as many as a hundred terms.

13. It is now necessary to consider the calculation of moments from the curve, for until this has been done it is impossible to form equations for finding the constants.

Let $y_x = f(x, a, b, c \ldots)$ where $a, b, c \ldots$ are constants to be determined.

We have seen, on pp. 13 and 14, that one way of working would be to find

$$f(1, a, b, c \ldots) \times 1^n + f(2, a, b, c \ldots) \times 2^n + \ldots$$

say, $$\sum_{x=1}^{x=r} f(x, a, b, c \ldots) \times x^n,$$

and this would give a result which might be used in forming equations if it were not for the fact that it is often impossible to find an algebraic expression for the sum of such a series in terms of the constants. It is, however, generally possible to find such an expression for the integral, and as we have defined the nth moment of an ordinate y_x as $y_x x^n$, the nth moment of the whole distribution from $x=h$ to $x=k$ is

$$\int_h^k y_x x^n dx \text{ or } \int_h^k f(x, a, b, c \ldots) x^n dx.$$

The total frequency (*i.e.*, total number of cases investigated) is $\int_h^k y_x dx$, and the mean is $\int_h^k y_x x dx \div \int_h^k y_x dx$, as we have already noticed.

14. If the moments from the equation to the curve are calculated in this way and equated to the moments calculated from statistics by assuming that the latter consist of a series of ordinates, an inaccuracy is introduced.

Let us consider the two cases :—

(1) When the statistics are a system of isolated terms or ordinates* and we wish to pass a curve very closely through them.

(2) When they are a system of areas but the moments are calculated by assuming the areas to be concentrated at the middle points of the bases.

* Strictly speaking, not a frequency distribution but a series of values requiring graduation. The distributions referred to on p. 5 have to be dealt with as areas for frequency-curve work because they tell the way the whole number of cases is divided in groups, and the whole area between the curve and the axis of x must therefore be used.

15. (1) In this case the terms $y_0, y_1, y_2 \ldots y_{n-1}$ are given by the statistics, and since $\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx$ is approximately equal to y_0, it is simplest* to assume that $\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y_x dx$ is given by the equation to the curve, and we have to find adjustments to counteract the error caused by equating $\sum_{x=0}^{x=n-1} Xy_x$ to $\int_{-\frac{1}{2}}^{n-\frac{1}{2}} Xy_x dx$ (the error is analogous to that introduced by assuming $(1+i)^{\frac{1}{2}} a_{\overline{n}} = \bar{a}_{\overline{n}|}$). The most practical way of overcoming the difficulty is by calculating the true area corresponding to the ordinates $y_0, y_1 \ldots y_{n-1}$ by means of a quadrature formula (formula of approximate summation). Many formulæ are well known, and some have been given for use in rather special circumstances in *Text-Book*, Part II., pp. 480–491; but for the present purpose it is convenient to have expressions which give approximate values of an area in terms of ordinates lying both within and without the base on which the area to be valued stands. Symbolically, these formulæ express $\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx$ in terms of $y_{-\frac{1}{2}}$, $y_{\frac{1}{2}}$, $y_{-1\frac{1}{2}}$, $y_{1\frac{1}{2}}$, &c., or $y_0, y_1, y_{-1}, y_2, y_{-2}$, &c.

I.—Let

$$y_x = a + bx + cx^2 + dx^3 + ex^4,$$

then

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx = a + \frac{c}{12} + \frac{e}{80},$$

and

$$y_0 = a$$

$$y_{-1} + y_1 = 2(a + c + e)$$

$$y_{-2} + y_2 = 2(a + 4c + 16e).$$

Now, assume the required integral can be equated to

$$hy_0 + k(y_{-1} + y_1) + l(y_{-2} + y_2),$$

substitute the values given just above and equate the coefficients of a, c and e respectively to 1, $\frac{1}{12}$ and $\frac{1}{80}$, and

* It is generally possible to use these limits in case (1), but if other limits have to be taken, such as 0 to n, different quadrature formulæ must be used.

we have
$$h+2k+2l=1$$
$$2k+8l=\frac{1}{12}$$
$$2k+32l=\frac{1}{80}.$$

The solution of these equations gives

$$h=\frac{5178}{5760},\ k=\frac{308}{5760},\ \text{and}\ l=-\frac{17}{5760},$$

and we obtain

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx=\frac{1}{5760}\{5178y_0+308(y_{-1}+y_1)-17(y_{-2}+y_2)\}.$$

II.—If

$$y_x=a+bx+cx^2$$

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx=\frac{1}{24}\{y_{-1}+22y_0+y_1\}.$$

III.—If

$$y_x=a+bx+cx^2+dx^3+ex^4$$

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx=\frac{1}{1440}\{802(y_{\frac{1}{2}}+y_{-\frac{1}{2}})-93(y_{1\frac{1}{2}}+y_{-1\frac{1}{2}})+11(y_{2\frac{1}{2}}+y_{-2\frac{1}{2}})\}.$$

IV.—If

$$y_x=a+bx+cx^2+dx^3$$

$$\int_{-\frac{1}{2}}^{1\frac{1}{2}} y_x dx=\frac{1}{24}\{27y_0+17y_1+5y_2-y_3\}.$$

16. We can now take the calculation of the moments, where $\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y_x dx$ is required in terms of $y_0, y_1 \ldots y_n$.

Now,

$$\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y_x dx=\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx+\int_{\frac{1}{2}}^{1\frac{1}{2}} y_x dx+\ \ldots+\int_{n-1\frac{1}{2}}^{n-\frac{1}{2}} y_x dx.$$

If formula I. be applied it can be used for all the integrals on the right-hand side of this equation except the first two and the last two, and the values of these are given by IV. Summing the values obtained and writing IV. with the denominator 5760, we obtain

$$\int_{-\frac{1}{2}}^{n-\frac{1}{2}} y_x dx = \frac{1}{5760}\Big\{6463y_0 + 4371y_1 + 6669y_2 + 5537y_3 + 5760(y_4 + y_5 + \ldots + y_{n-6} + y_{n-5}) + 5537y_{n-4} + 6669y_{n-3} + 4371y_{n-2} + 6463y_{n-1}\Big\} \quad . \quad . \quad . \quad \text{V.}$$

which means that we can multiply the first and last ordinates by $\frac{6463}{5760}$ (=1·1220485), the second and last but one by $\frac{4371}{5760}$ (=·7588541), the third and last but two by $\frac{6669}{5760}$ (=1·1578127), the fourth and last but three by $\frac{5537}{5760}$ (=·9612847), leave all the other ordinates unaltered, and work out the moments in the usual way from this modified series of ordinates. Of course, if there are less than eight ordinates another formula must be evolved.

17. In the following table the original series and the modified one are set out in the first two columns, and in the other columns the calculations of the first four moments about the middle of the range by the direct method are shown :—

TABLE V.

y_x	Modified by Formula V. y'_x	$y'_x \times x$	$y'_x \times x^2$	$y'_x \times x^3$	$y'_x \times x^4$
51·81	58·13	232·52	930·08	3,720·32	14,881·28
43·74	33·19	99·57	288·71	866·13	2,598·39
35·58	41·01	82·02	164·04	328·08	656·16
27·80	26·72	26·72	26·72	26·72	26·72
20·42	20·42	−440·83	...	−4,941·25	...
13·79	13·26	13·26	13·26	13·26	13·26
8·26	9·52	19·04	38·08	76·16	152·32
4·29	3·26	9·78	29·34	88·02	264·06
1·69	1·90	7·60	30·40	121·60	486·40
208·38	207·41	+ 49·68 −391·15	1,520·63	+ 299·04 −4,642·21	19,078·59

207·41 is then treated as the total frequency, and the moments for unit frequency (μ'_n) would be obtained by dividing $-391·15$, $1520·63$, &c., by 207·41, and not by 208·38, which is not the "total frequency", but merely gives the uncorrected sum of certain equidistant values.

18. The work can sometimes be simplified considerably, for if the values at the ends of the experience are very small and have a tendency to keep close to the axis of x before they finally vanish (*i.e.*, if there is high contact; most actuarial functions l_x, a_x, D_x, &c., have high contact at the old age end of the table), then it is reasonable to suppose that ordinates before the first and after the last exist, but are insignificant in value. Thus the integral corresponding to the whole series of ordinates can be legitimately extended beyond the limits $-\frac{1}{2}$ and $n-\frac{1}{2}$ previously used, because the additional area thus introduced will be evanescent. Now if the area be so extended, the effect will be that in equation V the significant ordinates from y_0 to y_{n-1} will all have the coefficient unity, and the ordinates with weighted coefficients will all vanish. The practical result is, that if there is high contact at one end of the statistics the adjustment need only be made at the other end, while if there is high contact at both ends no adjustment is necessary. Mathematically, high contact means that the first few differential coefficients vanish at the point of contact. The diagrams on pp. 73 and 90 show high contact at both ends of the curves, and the diagram on p. 67 shows high contact at the longer durations.

19. (2) The second case, namely, that in which mid-ordinates are used instead of areas, may now be examined. By concentrating areas about the middle points of their bases, we assume that the distances by which the areas $\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx$; $\int_{\frac{1}{2}}^{1\frac{1}{2}} y_x dx$, &c., must be multiplied, are the same as the distances from y_0 y_1, &c.; that is, the tth moment from the statistics is

$$\int_{-\frac{1}{2}}^{+\frac{1}{2}} y_x dx X^t + \int_{\frac{1}{2}}^{1\frac{1}{2}} y_x dx (X+1)^t + \ldots + \int_{n-1\frac{1}{2}}^{n-\frac{1}{2}} y_x dx (X+n-1)^t$$

and we require $\int_{-\frac{1}{2}}^{n-\frac{1}{2}} (X+x)^t y_x dx$ where X is the distance of y_0 from the ordinate about which moments are calculated.

By formula I. the series of integrals can be written

$$\frac{1}{5760}\Big\{ \dots +[5178h^t+308\{(h-1)^t+(h+1)^t\}$$
$$-17\{(h-2)^t+(h+2)^t\}]y+ \dots \Big\}$$

where h is written for $X+x$ in order to simplify the expression, and working out this general coefficient we have

$$\frac{1}{5760}\Big\{5760h^t+240t(t-1)h^{t-2}+3t(t-1)(t-2)(t-3)h^{t-4}+ \dots \Big\}$$

If $t=1$ this becomes h

" $t=2$ " " $h^2+\frac{1}{12}$

" $t=3$ " " $h^3+\frac{1}{4}h$

" $t=4$ " " $h^4+\frac{1}{2}h^2+\frac{1}{80}$

It has already been noticed that if there is high contact, the value of $\int_{-\frac{1}{2}}^{n-\frac{1}{2}}(X+x)^t y dx$ is found by using the unadjusted ordinates; that is, the second moment is given by a series, the general term of which is h^2y; the third by a series, the general term of which is h^3y, and so on; hence, if μ be written for the true adjusted moment about the mean and ν for the unadjusted moment, the relations between μ and ν are given by

$$\mu_2+\tfrac{1}{12}=\nu_2 \text{ or } \mu_2=\nu_2-\tfrac{1}{12}$$

$$\mu_4+\tfrac{1}{2}\mu_2+\tfrac{1}{80}=\nu_4 \text{ or } \mu_4=\nu_4-\tfrac{1}{2}\nu_2+\tfrac{7}{240}$$

The mean needs no adjustment, for if $t=1$ the general term has the correct coefficient h, and the third moment has to be adjusted by $\frac{1}{4}$ of the first moment, which is zero where the moments are taken about the mean. These adjustments were first given by Mr. W. F. Sheppard in *Proceedings of the London Mathematical Society*, vol. xxix., pp. 353-380. In order to demonstrate the correction for the nth moment by the above method, a parabola of at least the nth order is necessary. If we apply these adjustments to the moments found on p. 19, for Example IV. of Table I., we have $\mu_2=3{\cdot}1503$, $\mu_3=-1{\cdot}430976$, and $\mu_4=28{\cdot}828322$. These adjustments are found to make a considerable difference in the constants obtained from the moments especially when there is a small number of terms.

20. When there is not high contact at both ends of the curve, the adjustments become more difficult to value; suggestions have been made for finding the corrections, but they are not altogether satisfactory, and it is probably best in such cases to use the unadjusted figures. A few particular cases are, however, dealt with in Arts. 15 to 21 of Chap. V.

A student should calculate the moments for one or two distributions, and make the necessary adjustments; he can also find the standard deviations of distributions, for the $SD = \sqrt{\mu_2}$ where the μ_2 has been adjusted in accordance with the above rules. In Examples III. and IV. there is clearly high contact, in II. and V. there is more doubt but the adjustment is advisable, while in I. the rough moment should be used.

21. Before proceeding to deal with fitting more complicated curves it is advisable to consider the application of the method of moments to a simple case, namely, when $y = a + bx + cx^2 + \&c.$

Let the range be $2l$, and let the origin be at the middle point of the range, and m_0 stand for the area and m_n for the nth moment of the whole distribution about the middle of the range.

Then
$$m_{2s} = \int_{-l}^{+l} (a + bx + cx^2 + \ldots) x^{2s} dx$$

$$= 2l \times l^{2s} \left(\frac{a}{2s+1} + \frac{cl^2}{2s+3} + \ldots \right)$$

and similarly
$$m_{2s+1} = 2l \times l^{2s+1} \left(\frac{bl}{2s+3} + \frac{dl^3}{2s+5} + \ldots \right)$$

These equations show that the even moments give the constants a, c, e, &c., and the odd moments give the constants b, d, f, &c. This is, of course, the result of using moments about the middle of the range, and makes the solution of the equations less laborious than they would otherwise have been. The solution can also be simplified a little by writing

$$\frac{1}{2l} \cdot \frac{m_{2s}}{l^{2s}} = \frac{a}{2s+1} + \frac{cl^2}{2s+3} + \ldots$$

so that

$$\left.\begin{aligned} \frac{1}{2l} \cdot m_0 &= a + \frac{cl^2}{3} + \frac{el^4}{5} + \ldots \\ \frac{1}{2l} \cdot \frac{m_2}{l^2} &= \frac{a}{3} + \frac{cl^2}{5} + \frac{el^4}{7} + \ldots \\ \frac{1}{2l} \cdot \frac{m_4}{l^4} &= \frac{a}{5} + \frac{cl^2}{7} + \frac{el^4}{9} + \ldots \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots \quad \text{I.}$$

and similarly

$$\left.\begin{aligned}
\frac{1}{2l}\cdot\frac{m_1}{l} &= \frac{bl}{3} + \frac{dl^3}{5} + \frac{fl^5}{7} + \dots \\
\frac{1}{2l}\cdot\frac{m_3}{l^3} &= \frac{bl}{5} + \frac{dl^3}{7} + \frac{fl^7}{9} + \dots \\
\frac{1}{2l}\cdot\frac{m_5}{l^5} &= \frac{bl}{7} + \frac{dl^3}{9} + \frac{fl^9}{11} + \dots
\end{aligned}\right\} \quad \dots\dots\dots \quad \text{II.}$$

The solution of these equations gives the constants required, for example—

(i.) if $y=a+bx$, we have—

$$a=\frac{1}{2l}m_0$$

$$b=\frac{3}{l}\frac{1}{2l}\cdot\frac{m_1}{l}$$

(ii.) if $y=a+bx+cx^2$

$$a=\frac{3}{4}\left\{\frac{3}{2l}\cdot m_0-\frac{5}{2l}\cdot\frac{m_2}{l^2}\right\}$$

$$b=\frac{3}{l}\cdot\frac{1}{2l}\cdot\frac{m_1}{l}$$

$$c=\frac{15}{4l^2}\left\{-\frac{1}{2l}m_0+\frac{3}{2l}\cdot\frac{m_2}{l^2}\right\}$$

(iii.) if $y=a+bx+cx^2+dx^3$

$$a=\frac{3}{4}\left\{\frac{3}{2l}\cdot m_0-\frac{5}{2l}\cdot\frac{m_2}{l^2}\right\}$$

$$b=\frac{15}{4l}\left\{\frac{5}{2l}\cdot\frac{m_1}{l}-\frac{7}{2l}\cdot\frac{m_3}{l^3}\right\}$$

$$c=\frac{15}{4l^2}\left\{-\frac{1}{2l}\cdot m_0+\frac{3}{2l}\cdot\frac{m_2}{l^2}\right\}$$

$$d=\frac{35}{4l^3}\left\{-\frac{3}{2l}\cdot\frac{m_1}{l}+\frac{5}{2l}\cdot\frac{m_3}{l^3}\right\}$$

The above results, which can easily be extended if it is wished, may now be applied to one or two numerical examples.

22. As a first example, we shall graduate the statistics in Table V., Art. 17, for which the moments about the middle of the range have been calculated. Taking the curve $y=a+bx+cx^2$, the following values from Table V. will be required:

$$2l=9 \qquad \text{or} \qquad l=4{\cdot}5$$

$$m_0 = \quad 207{\cdot}41$$

$$m_1 = - \ 391{\cdot}15$$

$$m_2 = \quad 1520{\cdot}63$$

Hence
$$a=\frac{3}{4}\left\{\frac{622{\cdot}23}{9}-\frac{5}{9}\times\frac{1520{\cdot}63}{(4{\cdot}5)^2}\right\}$$
$$=20{\cdot}563$$
$$b=\frac{3}{4{\cdot}5}\times\frac{1}{9}\times\frac{(-391{\cdot}15)}{4{\cdot}5}$$
$$=-6{\cdot}4387$$
$$c=\frac{15}{4(4{\cdot}5)^2}\left\{-\frac{207{\cdot}41}{9}+\frac{3}{9}\times\frac{1520{\cdot}63}{(4{\cdot}5)^2}\right\}$$
$$={\cdot}36815$$

23. The best way to obtain the ordinates corresponding to this graduation is by calculating $b+c$ the first difference, and $2c$ the second difference, from the middle term; their values are $-6{\cdot}0706$ and ${\cdot}7363$ respectively. Since second differences are constant, the work is done continuously, and is as follows:

	Δ	Δ^2
52·208	−9·016	·736
43·192	−8·279	
34·913	−7·543	
27·370	−6·807	
20·563	−6·071	
14·492	−5·335	
9·157	−4·599	
4·558	−3·862	
·696		

These graduated figures will be found to agree fairly well with those given in the first column of Table V.

24. As a further example the following statistics, taken from a paper by S. H. J. W. Allin (*Journal of the Institute of Actuaries*, xxxix., p. 350), and giving the values of annuities to widows in pension funds according to the age of the member, may be considered :—

Age.	Value of Annuity.	Modified by Formula V. p. 24. a'	Distance from middle of range multiplied by 2. d	$a' \times d$	$a' \times d^2$	$a' \times d^3$
27	21·20	23·79	−7	166·53	1165·71	8159·97
32	19·91	15·11	−5	75·55	377·75	1888·75
37	19·34	22·40	−3	67·20	201·60	604·80
42	18·58	17·86	−1	17·86	17·86	17·86
				−327·14		−10671·38
47	16·74	16·09	+1	16·09	16·09	16·09
52	15·69	18·17	+3	54·51	163·53	490·59
57	14·70	11·15	+5	55·75	278·75	1393·75
62	12·99	14·58	+7	102·06	714·42	5000·94
		139·15		+228·41	2935·71	+6901·37
				− 98·73		−3770·01

In calculating the above moments it has been assumed that the figures to be graduated represent a system of ordinates; if they had represented a system of areas the adjustment by formula V. would have been unsuitable.

When there is an even number of terms the difficulty of calculating the moments about the middle of the range is that the terms have to be multiplied by ·5, 1·5, 2·5, &c., and if the series to be graduated contains only a few terms, it is best to deal with the distance d, in the way shown above, and then divide the totals by 2, 4 and 8, in order to obtain the first, second and third moments respectively. In this way, we have

$$l = 4$$

$$m_0 = 139{\cdot}15$$

$$m_1 = -49{\cdot}36$$

$$m_2 = 733{\cdot}93$$

$$m_3 = -471{\cdot}25$$

We will now fit the statistics with each of the three curves, the formulæ for which have been given, and compare the resulting graduations.

(i.) $y = 17{\cdot}394 - 1{\cdot}157x$

(ii.) $y = 17{\cdot}633 - 1{\cdot}157x - {\cdot}0451x^2$

(iii.) $y = 17{\cdot}633 - 1{\cdot}190x - {\cdot}0451x^2 + {\cdot}0035x^3$

The following table shows the graduations:—

Age.	Ungraduated.	(i.)	(ii.)	(iii.)
27	21·20	21·44	21·13	21·13
32	19·91	20·29	20·24	20·28
37	19·34	19·13	19·27	19·31
42	18·58	17·97	18·20	18·22
47	16·74	16·82	17·04	17·02
52	15·69	15·66	15·80	15·76
57	14·70	14·50	14·46	14·43
62	12·99	13·34	13·03	13·05

Formulæ (ii.) and (iii.) are practically identical, and both are considerably closer to the original figures than (i.).

25. The results obtained so far may be summarized as follows:—

(1) The method of moments is a general method of finding the constants in a formula suitable to a particular statistical example, and it consists of equating the values of $\Sigma f(n) \times n^t$ (which is called the tth moment, and is summed for all values of n that occur) to similar expressions obtained from the graduation formula. These latter expressions will be algebraic, and simultaneous equations have to be solved in order to find the arithmetical constants.

(2) The moments from the statistics can be calculated by multiplying the frequencies by appropriate values of n^t, or by Mr. G. F. Hardy's summation method.

(3) If moments have been obtained about any one vertical, they can be transferred to any other by the formulæ in Art. 6 of Chap. III.

(4) Since the moments from the graduation formula must generally be found by means of the integral

calculus, while those from the statistics are found by summation, the latter have to be adjusted before the equations for obtaining the constants can be correctly formed. The adjustments depend on whether the statistics are a system of ordinates or a system of areas; in the former case adjustment is made by equation V., and in the latter by the formulæ in Art. 19 (Sheppard's adjustments), if there is high contact at both ends of the curve.

CHAPTER IV.

Frequency-Curves.

1. When it becomes necessary in practical work to decide on a system of curves for describing frequency distributions, we have to bear in mind that

(1) Any expression used must be a graduation formula; it must remove the roughness of the material.

(2) There must not be so many constants in the formula that we require a great number of moments, for this means that the accuracy is reduced. The higher the moment the more liable it is to error when deduced from ungraduated observations; this is clear, when we remember that the ends of the experiences are multiplied by the highest numbers and their powers.

(3) There must be a systematic method of approaching frequency distributions.

2. Now, considering the more obvious characteristics of frequency distributions, we find they generally start at zero, rise to a maximum, and then fall sometimes at the same but often at a different rate. At the ends of the distribution there is often high contact. This means, mathematically, that a series of equations $y=f(x)$; $y=\phi(x)$, &c., must be chosen, so that in each equation of the series $\frac{dy}{dx}=0$ in certain cases; at the maximum (for the test of a maximum is that the first differential coefficient is zero and the second negative) and when $y=0$, for there is to be contact at one end, at least, of the range of the distribution, or, in other words, the angle formed by the tangent to the curve at this point must be zero, in order that the tangent of the angle

(*i.e.*, differential coefficient) may be zero. In non-geometrical language, the finite difference between two successive ordinates must be zero, or there will not be contact.

The above suggests that $\frac{dy}{dx}$ may be put equal to $\frac{y \times (x+a)}{F(x)}$, then, if $y=0$, $\frac{dy}{dx}=0$, and if $x=-a$, $\frac{dy}{dx}=0$, and we have the maximum we require. So long as $F(x)$ is general the form assumed for $\frac{dy}{dx}$ is extremely general and includes cases when $\frac{dy}{dx}$ may not be zero when y is zero. $F(x)$ is expanded by Maclaurin's theorem in ascending powers of x, and we have

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

We shall return to this equation and show how it can be put in the form $y=f(x)$, so as to express y as a direct function of x; but as the matter has up to the present been approached from an experimental point of view, it will be interesting to see how equation I. can be obtained up to the x^2 term in the denominator from elementary propositions in the theory of probabilities.

3. If p be the probability of an event happening and q the probability of its failing, then the probabilities of its happening once, twice, and so on out of n trials are given by the terms of the expansion $(p+q)^n$; or if we have N cases, the terms of N $(p+q)^n$ give the frequency distribution of the N cases into n groups. The binomial series does not represent nearly all the probabilities that arise, and another series that occurs is the hypergeometrical. Thus the chances of getting $r, r-1, \ldots 0$ black balls from a bag containing pn black and qn white balls when r balls are drawn, are given by the successive terms of the series

$$\frac{pn(pn-1)\ldots(pn-r+1)}{n(n-1)\ldots(n-r+1)}\left\{1+\frac{rqn}{pn-r+1}\right.$$
$$\left.+\frac{r.r-1}{2\,!}\,\frac{qn(qn-1)}{(pn-r+1)(pn-r+2)}+\ldots\right\}$$

A numerical example may help to make the way the series arises clear. A bag contains seven balls, of which four are black and three white; then if three balls are drawn the probability that

all will be black is $\frac{4.3.2}{7.6.5}$

two will be black is $\frac{4.3.2}{7.6.5} \times {}_3C_1$

one will be black is $\frac{4.3.2}{7.6.5} \times {}_3C_2$

none will be black is $\frac{3.2.1}{7.6.5}$

The sum of these four expressions is unity. The terms can be seen to agree with the series by putting $n=7$, $pn=4$, $qn=3$, and $r=3$.

Other series may arise, but those given will be sufficient for the present purpose, and we shall proceed to consider how they can be put in the form of equation I. The inconvenience of the expressions as they now stand becomes fairly obvious when an attempt is made to calculate numerical values for a large number of groups, and besides this, they are not continuous, while the statistics of practical work often are.

Considering the hypergeometrical series, and remembering that the function required for equation I. is $\frac{1}{y}\frac{dy}{dx}$, and as the series is discontinuous, finite differences must be used, we have

$$y_x = \frac{pn(pn-1)\ldots(pn-r+1)}{n(n-1)\ldots(n-r+1)} \cdot \frac{r(r-1)\ldots(r-x+2)}{(x-1)!} \cdot \frac{qn(qn-1)\ldots(qn-x+2)}{(pn-r+1)(pn-r+2)\ldots(pn-r+x-1)}$$

$$\Delta y_x = y_{x+1} - y_x = y_x \left\{ \frac{r-x+1}{x} \cdot \frac{qn-x+1}{pn-r+x} - 1 \right\}$$

$$= y_x \left\{ \frac{(r+1)(qn+1) - x(n+2)}{x(pn-r+x)} \right\} \text{ for } p+q=1$$

and

$$y_{x+\frac{1}{2}} = \tfrac{1}{2}(y_{x+1} + y_x) = \tfrac{1}{2} y_x \left\{ \frac{(r+1)(qn+1) - x[2(r+1) + n(q-p)] + 2x^2}{x(pn-r+x)} \right\}$$

$$\therefore \quad \frac{\Delta y_x}{y_{x+\frac{1}{2}}} = \frac{2\{(r+1)(qn+1) - x(n+2)\}}{(r+1)(qn+1) - x\{2(r+1) + n(q-p)\} + 2x^2},$$

which may be put in the form of equation I.,

$$\frac{1}{y}\frac{dy}{dx}=\frac{a+x}{b_0+b_1x+b_2x^2}.$$

4. Returning to equation I., we see that it can be written in the form

$$(b_0+b_1x+b_2x^2+\ .\ .\ .)\frac{dy}{dx}=y(x+a)\ ;$$

multiplying each side by x^n, and integrating with respect to x, we have

$$\int x^n(b_0+b_1x+b_2x^2+\ .\ .\ .)\frac{dy}{dx}dx=\int y(x+a)x^n dx.$$

Integrate the left-hand side by parts treating $\frac{dy}{dx}$ as one part, and the right-hand side as the sum of two functions, and then

$$x^n(b_0+b_1x+b_2x^2+\ .\ .\ .)y-\int(nb_0x^{n-1}+(n+1)b_1x^n$$
$$+(n+2)b_2x^{n+1}+\ .\ .\ .)ydx$$
$$=\int yx^{n+1}dx+\int yax^n dx,$$

or since $y=0$ at the ends of the range of the curve the expression $x^n(b_0+b_1x+b_2x^2+\ .\ .\ .)\,y$ vanishes, and using the notation we have already adopted, namely,

$$\mu'_n=\int yx^n dx,$$

we have

$$-nb_0\mu'_{n-1}-(n+1)b_1\mu'_n-(n+2)b_2\mu'_{n+1}-\ .\ .\ .=\mu'_{n+1}+a\mu'_n.$$

If we put $n=0, 1, 2\ .\ .\ .\ s$ respectively, we get $s+1$ equations to enable us to find a, b_0, b_1 . . . &c., in terms of the moments (μ') as shown by the following equations, which have been obtained by writing the equation in the form

$$a\mu'_n+nb_0\mu'_{n-1}+(n+1)b_1\mu'_n+(n+2)b_2\mu'_{n+1}+\ .\ .\ .=-\mu'_{n+1},$$

and then putting $n=0$, 1, 2, &c.

$$\left.\begin{array}{l} a\mu'_0+0\times b_0+b_1\mu'_0+2b_2\mu'_1+\ .\ .\ .\ .\ =-\mu'_1 \\ a\mu'_1+b_0\mu'_0+2b_1\mu'_1+3b_2\mu'_2+\ .\ .\ .\ .\ =-\mu'_2 \\ a\mu'_2+2b_0\mu'_1+3b_1\mu'_2+4b_2\mu'_3+\ .\ .\ .\ =-\mu'_3 \\ a\mu'_3+3b_0\mu'_2+4b_1\mu'_3+5b_2\mu'_4+\ .\ .\ .\ =-\mu'_4 \end{array}\right\}\quad .\ .\quad \text{II.}$$

&c., &c.

Let us now make $\mu'_1=0$, and alter the other moments in the way indicated in Chap. II., for the result of making $\mu'_1=0$ is to change the origin of the system to the mean of the distribution. We can also treat μ'_0 as 1, and these simplications lead to the following results :—

(1) Keeping b_0 only, we have

$$\frac{1}{y}\frac{dy}{dx}=-\frac{x}{\mu_2}.$$

(2) Keeping b_0 and b_1, the first three equations in the system II. above give

$$a+b_1=0$$

$$b_0=-\mu_2$$

and $$a\mu_2+3b_1\mu_2=-\mu_3$$

or $$b_1=-\frac{\mu_3}{2\mu_2}$$

and $$a=\frac{\mu_3}{2\mu_2}$$

and the differential equation becomes

$$\frac{1}{y}\frac{dy}{dx}=\frac{x+\dfrac{\mu_3}{2\mu_2}}{\mu_2+\dfrac{\mu_3}{2\mu_2}x}.$$

(3) Keeping b_0, b_1, b_2, the system gives

$$a+b_1=0$$

$$b_0+3b_2\mu_2=-\mu_2$$

$$a\mu_2+3b_1\mu_2+4b_2\mu_3=-\mu_3$$

$$a\mu_3+3b_0\mu_2+4b_1\mu_3+5b_2\mu_4=-\mu_4.$$

The solution of these simultaneous equations is perfectly straightforward, and leads to

$$\frac{1}{y}\frac{dy}{dx}=-\frac{x+\dfrac{\mu_3(\mu_4+3\mu_2^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}}{\dfrac{\mu_2(4\mu_2\mu_4-3\mu_3^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}+\dfrac{\mu_3(\mu_4+3\mu_2^2)}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}x+\dfrac{2\mu_2\mu_4-3\mu_3^2-6\mu_2^3}{10\mu_2\mu_4-18\mu_2^3-12\mu_3^2}x^2}$$

In this last form put $\beta_1=\dfrac{\mu_3^2}{\mu_2^3}$ and $\beta_2=\dfrac{\mu_4}{\mu_2^2}$ and

$$\frac{1}{y}\frac{dy}{dx}=-\frac{x+\dfrac{\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}}{\dfrac{\mu_2(4\beta_2-3\beta_1)+\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2+3)x+(2\beta_2-3\beta_1-6)x^2}{2(5\beta_2-6\beta_1-9)}}\quad . \qquad \text{III.}$$

5. The reasoning by which equation I. was first obtained showed that a is the distance between the origin and the mode, or as the origin has now been transferred to the mean by putting $\mu'_1=0$, a is the distance between the mean and the mode. This distance in terms of the moment is, therefore,

$$\frac{\sigma\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}$$

where σ is the standard deviation $\sqrt{\mu_2}$.

Since the skewness is the distance between the mean and mode divided by the standard deviation.

$$\text{skewness}=\frac{\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}.$$

6. It would be possible to obtain constants in the differential equation I. by using a greater number of terms and retaining b_3, b_4, &c., but there are strong practical objections to such a course. Besides the large increase in arithmetical work, the gain in introducing additional constants is not great because the higher moments become untrustworthy, owing, as we have already noticed, to their probable errors being very large. Professor Pearson has shown* that "we might easily on a "random sample reach a 7th or 8th moment having half or "double the value it actually has in the general population. "Constants based on these high moments will be practically "idle. They may enable us to describe closely an individual "random sample, but no safe argument can be drawn from this "individual sample as to the general population at large, at "any rate so far as the argument is based on the constants "depending on these high moments." In some actuarial statistics where there are as many as 100,000 cases, it might be worth while to go as far as the next term of the series, but even here the value of the work is discounted because any other smaller body of statistics on the same subject could not be compared satisfactorily with the result. For practical purposes it is probable that the equation taken as far as b_2 will be sufficient, and we shall confine our attention to the forms thus obtained, merely remarking that in some extreme cases in graduation another term might be required.

* "Skew Correlation and non-linear Regression," *Drapers' Company Research Memoir*, 1905, p. 9.

7. Turning to the particular form of equation I. given in equation III. it will be seen that it is possible to obtain a formula representing the statistics by inserting in that equation the values of the moments found from the statistics, but this would not give a graduation in the same form as that in which the original data appeared, for in the latter we have y, while the former gives $\frac{1}{y}\frac{dy}{dx}$ or $\frac{d \log y}{dx}$. It would, therefore, be necessary to integrate the expression we obtain in order to get terms comparable with the original data, and it is better in practical work to deal with the equations in the forms in which we require them for comparison, rather than by using the differential equations and then integrating the result. The latter method could only give proportional not actual frequencies.

8. The next step is, therefore, to replace the equation

$$\frac{d \log y}{dx} = \frac{x+a}{b_0+b_1x+b_2x^2}$$

by one of the form $y=f(x)$, and to do this $\frac{x+a}{b_0+b_1x+b_2x^2}$ must be integrated.

Let us consider equation III. as a general expression for integration, then we notice that the form the integral takes depends on the particular values of the coefficients of x in the denominator. The problem is, in fact, merely a consideration of the forms taken by the denominator for

$$b_0+b_1x+b_2x^2 = b_2\left[x - \frac{-b_1+\sqrt{\{b_1^2-4b_0b_2\}}}{2b_2}\right]\left[x - \frac{-b_1-\sqrt{\{b_1^2-4b_0b_2\}}}{2b_2}\right]$$

and the criterion for fixing the form in a particular case is, obviously, the same as that for the nature of the roots of the equation $b_0+b_1x+b_2x^2=0$, viz., $\frac{b_1^2}{4b_0b_2}$, which, by substituting in formula III., gives—

$$\frac{\beta_1(\beta_2+3)^2}{4(2\beta_2-3\beta_1-6)(4\beta_2-3\beta_1)}.$$

9. If this is negative the roots are real and of different sign (Type I.), if positive and less than unity they are complex

(Type IV.), and if positive and greater than unity they are real and of the same sign (Type VI.). This really covers all the cases, but just at the point where one type changes into another we can use a slightly simpler transition curve. Thus when the criterion is ∞, one root is ∞ (Type III.), when it is unity the two roots are equal (Type V.), while when it is zero the roots are equal in magnitude but of opposite sign (Type II.). The only other transition curve arises when $b_1 = b_2 = 0$, and the criterion is again zero (Normal Curve of Error, Type VII.).

10. The actual integration can now be considered.

Type I.—The factors in the denominator, when the roots of $b_0 + b_1 x + b_2 x^2 = 0$ are real and of different signs, take the form

$$b_2\left[x - \frac{-b_1 + \sqrt{\text{a positive quantity}}}{2b_2}\right]\left[x - \frac{-b_1 - \sqrt{\text{a positive quantity}}}{2b_2}\right]$$

and the expression to be integrated is therefore of the form

$$\frac{x+a}{(x+A_1)(x-A_2)} = \frac{A_1 - a}{A_1 + A_2} \cdot \frac{1}{x + A_1} + \frac{A_2 + a}{A_1 + A_2} \cdot \frac{1}{x - A_2}$$

by partial fractions.

The integration is now simple, and gives

$$\log y = \frac{A_1 - a}{A_1 + A_2} \log (x + A_1) + \frac{A_2 + a}{A_1 + A_2} \log (x - A_2) + \text{a constant.}$$

$$\therefore \qquad y = y'(x + A_1)^{\frac{A_1 - a}{A_1 + A_2}} (x - A_2)^{\frac{A_2 + a}{A_1 + A_2}}$$

where y' results from the constant introduced by integration.

If the origin is now transferred to the mode (*i.e.*, put x for $x + a$), we have

$$y = y_0\left(1 + \frac{x}{a_1}\right)^{\nu a_1}\left(1 - \frac{x}{a_2}\right)^{\nu a}$$

the form given in Table VI.

Type II.—In this type $a_1 = a_2$ in Type I., and it is, therefore, unnecessary to give the working.

Type III.—This type is reached when the criterion is ∞, which happens when $b_2=0$,

$$\log y=\int\frac{x+a}{b_0+b_1x}\,dx$$

$$=\int\left(\frac{1}{b_1}+\frac{a-\frac{b_0}{b_1}}{b_1x+b_0}\right)dx$$

$$=\frac{x}{b_1}+\left(a-\frac{b_0}{b_1}\right)\frac{1}{b_1}\log\left(b_1x+b_0\right)+\mathrm{C}$$

and $$y=y'e^{\frac{x}{b_1}}\left(b_1x+b_0\right)^{\frac{1}{b_1}\left(a-\frac{b_0}{b_1}\right)}$$

or, by changing the origin,

$$y=y_0e^{-\gamma x}\left(1+\frac{x}{a}\right)^{\gamma a}$$

where a has a meaning different from that implied in Equation I. This type can be seen to be a particular case of Type I. when a_2 becomes infinite.

Type IV.—If the roots of the equation $b_0+b_1x+b_2x^2=0$ are complex, it is impossible to throw the denominator into real factors; and when this occurs, we have to integrate by putting the expression on the right-hand side of the fundamental differential equation in the form

$$\frac{\mathrm{X}+c}{b_2(\mathrm{X}^2+\mathrm{A}^2)}$$

where $$\mathrm{X}=x+\frac{b_1}{2b_2},\quad c=a-\frac{b_1}{2b_2},\ \text{and}\ \mathrm{A}^2=\frac{b_0}{b_2}-\frac{b_1^2}{4b_2^2}$$

Then

$$\log y=\int\frac{\mathrm{X}+c}{b_2(\mathrm{X}^2+\mathrm{A}^2)}\,d\mathrm{X}$$

$$=\int\frac{\mathrm{X}}{b_2(\mathrm{X}^2+\mathrm{A}^2)}\,d\mathrm{X}+\int\frac{c}{\mathrm{X}^2+\mathrm{A}^2}\,d\mathrm{X}$$

$$=\frac{1}{2b_2}\log(\mathrm{X}^2+\mathrm{A}^2)+\frac{c}{\mathrm{A}}\tan^{-1}\frac{\mathrm{X}}{\mathrm{A}}+\text{constant}.$$

$$\therefore\qquad y=y'\left(\mathrm{X}^2+\mathrm{A}^2\right)^{\frac{1}{2b_2}}e^{\frac{c}{\mathrm{A}}\tan^{-1}\frac{\mathrm{X}}{\mathrm{A}}}$$

or $$y=\frac{y_0}{\left(1+\frac{x^2}{a^2}\right)^m}e^{-\nu\tan^{-1}\frac{x}{a}}$$

where a has a meaning different from that implied in equation I. The relation between this type and Type I. can be seen by factorising the denominator of the right-hand side of the differential equation, $b_2(X-iA)(x+iA)$, and then obtaining an expression for y having the same form as Type I., but containing complex expressions.

Type V.—In this case, when the roots are real and equal,

$$\log y = \int \frac{1}{b_2} \frac{x+a}{\left(x+\frac{b_1}{2b_2}\right)^2} dx$$

$$= \int \frac{1}{b_2} \frac{\left(x+\frac{b_1}{2b_2}\right)+\left(a-\frac{b_1}{2b_2}\right)}{\left(x+\frac{b_1}{2b_2}\right)^2} dx$$

$$= \int \frac{dx}{b_2\left(x+\frac{b_1}{2b_2}\right)} + \int \frac{a-\frac{b_1}{2b_2}}{b_2\left(x+\frac{b_1}{2b_2}\right)^2} dx$$

$$= \frac{1}{b_2}\log\left(x+\frac{b_1}{2b_2}\right) + \frac{a-\frac{b_1}{2b_2}}{b_2\left(x+\frac{b_1}{2b_2}\right)} + \text{constant}$$

$$\therefore \quad y = y'\left(x+\frac{b_1}{2b_2}\right)^{\frac{1}{b_2}} e^{\frac{a-\frac{b_1}{2b_2}}{b_2\left(x+\frac{b_1}{b_2}\right)}}$$

$$= y_0 x^p e^{-\gamma/x}$$

Type VI.—The factorising is the same as Type I., but the roots of the equation being of like sign, the factors of the denominator take the form $(x+A_1)(x+A_2)$. The work is then the same, but at the end the origin is put not at the mode but so that one of the expressions $x+A_1$ or $x+A_2$ can be written as x. The form is then

$$y = y_0(x-a)^{m_1} x^{-m_2}.$$

Type VII.—Putting

$$b_1 = b_2 = 0$$

$$\log y = \int \frac{x+a}{b_0} dx$$

$$= \frac{x^2}{2b_0} + \frac{ax}{b_0} + \text{constant}$$

$$= \frac{(x+a)^2}{2b_0} + \text{constant}$$

$$\therefore \quad y = y' e^{(x+a)^2/2b_0}$$

or, by changing the origin and remembering that the sign of the expression in equation III. is negative,

$$y = y_0 e^{-x^2/c}.$$

11. The table on p. 47 gives a list of the curves, a description of their appearance and range, the position of the mode and the criteria. The values of β_1 and β_2 in the cases of Type II. and Type VII. can be seen to be required by examining equation III. The third moment about the mean must be very small (theoretically, zero) if the curve is symmetrical, and therefore $\beta_1 = 0$, and it is only when $\beta_2 = 3$ and $\beta_1 = 0$ that both the coefficients of x and x^2 in equation III. vanish; this being the condition for Type VII.

12. It is now necessary to recapitulate the method, and see the steps that have to be taken to fit a frequency-curve to statistics.

1. Arrange the statistics in sequence.
2. Calculate the moments about a convenient vertical.
3. Transfer the moments to the centroid vertical (vertical through the mean).
4. If there is high contact at both ends of the curve, apply Sheppard's adjustments to the moments (*i.e.*, deduct $\frac{1}{12}$ and $\frac{1}{2}\nu_2 - \frac{7}{240}$ from the second and fourth moments respectively).
5. Calculate the criterion.
6. By means of Table VI. decide which curve should be used.

Table VI. gives a reference to the page on which the formulæ for the constants of each curve in terms of the moments are to be found.

Table VI.

Frequency-Curves.

No. of Type.	Description of Curve.	Equation to Curve.	For calculation of Constants, *see* page	Mode.	Criterion.		
					κ	β_1	β_2
I.	Limited range in both directions (skew) ...	$y=y_0\left(1+\frac{x}{a_1}\right)^{\nu a_1}\left(1-\frac{x}{a_2}\right)^{\nu a_2}$	53	Origin	<0	...	...
II.	Limited range in both directions (symmetrical)	$y=y_0\left(1-\frac{x^2}{a^2}\right)^m$	61	Origin	$=0$	$=0$	not $=3$
III.	Limited range in one direction (skew) ...	$y=y_0\left(1+\frac{x}{a}\right)^{\gamma a}e^{-\gamma x}$	65	Origin	∞ *	...	...
IV.	Unlimited range (skew)	$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m}e^{-\nu\tan^{-1}\frac{x}{a}}$	69	$-\frac{\nu a}{2m}$	>0 & <1	...	...
V.	Limited range in one direction (skew) ...	$y=y_0x^{-p}e^{-\gamma/x}$	78	$\frac{\gamma}{p}$	$=1$	...	...
VI.	Limited range in one direction (skew) ...	$y=y_0(x-a)^{q_2}x^{-q_1}$	83	$-\frac{aq_1}{q_1-q_2}$	>1 & $<\infty$	...	...
VII.	Unlimited range (symmetrical)	$y=y_0e^{-x^2/2\sigma^2}$	87	Origin	$=0$	$=0$	$=3$

$\kappa=\frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}$, $\beta_1=\frac{\mu_3^2}{\mu_2^3}$, $\beta_2=\frac{\mu_4}{\mu_2^2}$, and the moments are calculated about the mean.

* Strictly, when $2\beta_2-3\beta_1-6=0$; can also be used in cases of slight skewness, even when the criterion is not very large.

CHAPTER V.

Calculation.

1. The next point to be considered is the calculation of the constants for any particular distribution, when the moments have been calculated and the type to be used has been decided. The formulæ required for the numerical work will be given for each type, a numerical example, including the calculation of the graduated figures, will follow, and finally the proofs of the formulæ.

2. Some general points relating to the calculation of the curves when the constants have been found may be conveniently considered here. When the constants are known, we can calculate the ordinate for any value of x by substituting that value in the expression for the frequency-curve; and if areas are required, some method of proceeding from ordinates to areas must be found. The most simple is probably to calculate mid-ordinates, and then by the quadrature formula I. or II. find the areas. It is occasionally more convenient to calculate the ordinates at the beginning of each group, and then formula III. should be used. These formulæ can be best applied in the form of differences; thus, from II. we have

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx = y_0 - \frac{1}{24}\{\Delta y_{-1} - \Delta y_0\}$$

from I.

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx = y_0 - \frac{291}{5760}\{\Delta y_{-1} - \Delta y_0\} + \frac{17}{5760}\{\Delta y_{-2} - \Delta y_1\}$$

from III.

$$\int_{-\frac{1}{2}}^{\frac{1}{2}} y_x dx = \tfrac{1}{2}\{y_{\frac{1}{2}} + y_{-\frac{1}{2}}\} + \frac{82}{1440}\{\Delta y_{-1\frac{1}{2}} - \Delta y_{\frac{1}{2}}\} - \frac{11}{1440}\{\Delta y_{-2\frac{1}{2}} - \Delta y_{1\frac{1}{2}}\}$$

Formula II. is generally sufficiently accurate, while the others will be found to give a result true to five figures in ordinary cases—exceptional cases will be referred to in the numerical examples that follow.

3. It is sometimes a help to see the graduation expressed graphically, and this has been done with some of the examples. The best method is to insert a vertical height y_0 at the mode; note the ends of the curve, and the heights of the ordinates that have been calculated. These heights give points on the curve, which can be drawn through them fairly easily. In drawing the curve, as well as in calculating the constants, the sign of the skewness must be borne in mind, for it is possible to draw the curve with the skewness on the wrong side of the mode, and if the distribution is nearly symmetrical, it is not so easy to notice the mistake as it seems to be. The tangent to the curve at the mode is parallel to the axis of x.

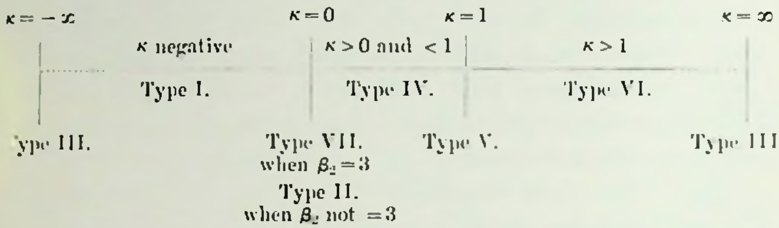
4. It is best to draw on a rather large scale in order to gain distinctness, and the curves given here were drawn larger than their present size; the reduction being, of course, made in the process of reproduction.

The base elements should also be fairly large in proportion to the height, so that the curve may not ascend too steeply; otherwise small horizontal differences between the graduated and ungraduated curves are apt to conceal large vertical differences when the curve is rising or falling rapidly, but it is the latter differences that are of importance. It is sometimes necessary to use more closely-ruled paper than that generally favoured by actuaries, and it can be procured in very convenient rulings from Messrs. W. G. Pye & Co., Granta Works, Cambridge.

5. The reader should notice that all the cases considered in the following pages assume complete distributions, and it is in general only possible to find the curve from part of a distribution by means of successive approximation which is extremely laborious. Another point, to which reference will again be made, is with regard to grouping statistics; it is sometimes impossible to obtain many groups, but for accuracy in finding moments the greater the number of groups the better, unless the total number of cases is small. A little discretion is needed in this respect, but in actuarial statistics which are sometimes based on as many as 200,000 cases,

seventy or eighty groups would not be excessive. In our examples we have grouped merely to save work, space and printing, and the grouping does not alter the method.

6. Another matter with which it seems advisable to deal here is connected with the criterion, κ. This may have any value from $-\infty$ to $+\infty$, and from the following diagram it will be seen how the types cover all the possible values of the criterion and do not overlap.



Just before $\kappa=0$ Type I. becomes nearly symmetrical, and after that value is passed we have a skew curve of unlimited range, and so on. At each critical point there is a "transition" curve, as it is sometimes called; so Types II., III., V. and VII., are the transition types. If by a mistake a student should use the wrong type he will necessarily find his mistake by reaching an imaginary in one of the square roots which occur in the equations for the constants, but transition types can be used when the values of the criterion approximate to the theoretical values; they can, in fact, be viewed as approximations which give an accurate result in a limiting case. It is impossible to say within what limits one is justified in using a transition type; theoretically the justification depends on the size of the probable error of the function dealt with, but in practice one can be guided to a great extent by the size of the experience; if there are few cases a larger deviation in the criterion will arise than if there are many. It would probably be sufficiently accurate to use Type III., provided κ was arithmetically greater than 4; individual cases must be considered on their merits, but if the student finds himself in doubt he should avoid using the transition type as he will then be on the safe side in the matter of accuracy.

7. In the formulæ that are given for the various types, the choice of sign for a square root depends on the sign of μ_3. If the frequency is concentrated more closely before the mean than after it, the mode is on the left-hand side of the mean and μ_3 is positive ; the signs of certain constants in each type must therefore depend on the signs of μ_3 in order that the mode and mean may lie in their correct relative positions. Where, however, no remark is made as to the sign of the expression in which a square root is given the positive root is implied, and the reader will find that these rules become easier to follow when he has worked out two examples, one giving a positive and the other a negative value for μ_3. Thus, if we imagine the frequencies in the example for Type I. to be written in the opposite order 1, 3, 7, 13, &c., all the numerical work would be the same, but m_1 would be 2·776978, m_2=·409833, a_1=13·52728, and a_2=1·99638, and the graduation would be the same, but the numbers in the columns of the table on p. 56 would run in the opposite order.

FORMULÆ FOR MOMENTS.

These Formulæ apply to all the Types of Curves.

$$\nu'_1 = d$$

$$\nu_2 = \nu'_2 - d^2$$

$$\nu_3 = \nu'_3 - 3d\nu_2 - d^3$$

$$\nu_4 = \nu'_4 - 4d\nu_2 - 6d^2\nu_2 - d^4$$

or

$$S_2 = d$$

$$\nu_2 = 2S_3 - d(1+d)$$

$$\nu_3 = 6S_4 - 3\nu_2(1+d) - d(1+d)(2+d)$$

$$\nu_4 = 24S_5 - 2\nu_3\{2(1+d)+1\} - \nu_2\{6(1+d)(2+d)-1\} - d(1+d)(2+d)(3+d)$$

$$\left.\begin{aligned} \mu_2 &= \nu_2 - \tfrac{1}{12} \\ \mu_3 &= \nu_3 \\ \mu_4 &= \nu_4 - \tfrac{1}{2}\nu_2 + \tfrac{7}{240} \end{aligned}\right\} \text{Sheppard's adjustments when the curve has high contact}$$

$$\sigma\text{(standard deviation)} = \sqrt{\mu_2}$$

$$\beta_1 = \mu_3^2 \div \mu_2^3$$

$$\beta_2 = \mu_4 \div \mu_2^2$$

$$\kappa = \frac{\beta_1(\beta_2+3)^2}{4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)}$$

TYPE I.

$$y=y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2}$$

$$\frac{m_1}{a_1}=\frac{m_2}{a_2}$$

FORMULÆ.

The values to be calculated in order are

$$r=\frac{6(\beta_2-\beta_1-1)}{6+3\beta_1-2\beta_2}$$

$$b=\tfrac{1}{2}\sqrt{\mu_2}\sqrt{\{\beta_1(r+2)^2+16(r+1)\}}$$

m_2 and m_1 are given by

$$\frac{1}{2}\left\{r-2\pm r(r+2)\sqrt{\frac{\beta_1}{\beta_1(r+2)^2+16(r+1)}}\right\}$$

and $a_1+a_2=b$, and $a_1\div m_1=a_2\div m_2$

$$y_0=\frac{N}{b}\,\frac{m_1^{m_1}m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}}\,\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}$$

A table of Γ functions is required (*see* Appendix II.).

$$\text{Skewness}=\frac{1}{2}\sqrt{\beta_1}\frac{r+2}{r-2}$$

$$\text{Mode}=\text{Mean}-\frac{1}{2}\frac{\mu_3}{\mu_2}\left\{\frac{r+2}{r-2}\right\}.$$

NOTES.

m_1 is taken with the negative root when μ_3 is positive, and as the positive root when μ_3 is negative.

Sometimes m_1 is negative, which means that the curve has a similar shape to that given in the numerical example of Type III.; it starts at infinity, and falls rapidly; so that though the ordinate is infinite, the area is finite. The difference is that in Type I. the curve ends at a fixed point, while in Type III. it continues indefinitely. In this case a little care is needed in taking out the Γ function, for $\Gamma(t)$ is required where $t<1$; the tables give $\log\Gamma(1+t)$, *i.e.*, $\log t+\log\Gamma(t)$. If both m_1 and m_2 are negative, a U-shaped curve is obtained.

EXAMPLE.

As an example of this type the figures given in Table I. (Example II.) may be used. The moments were first found by Mr. Hardy's Summation Method (*see* Chap. III., Art. 9) in the following form:

Central Age of Group.	Exposed to Risk Example II of Table I.	First Sum.	Second Sum.	Third Sum.	Fourth Sum.
17	34	1,000	5,175	19,809	64,389
22	145	966	4,175	14,634	44,580
27	156	821	3,209	10,459	29,946
32	145	665	2,388	7,250	19,487
37	123	520	1,723	4,862	12,237
42	103	397	1,203	3,139	7,375
47	86	294	806	1,936	4,236
52	71	208	512	1,130	2,300
57	55	137	304	618	1,170
62	37	82	167	314	552
67	21	45	85	147	238
72	13	24	40	62	91
77	7	11	16	22	29
82	3	4	5	6	7
87	1	1	1	1	1
Totals	1,000	5,175	19,809	64,389	186,638

The calculation of $\log y_0$ is as follows:—

$$
\begin{aligned}
\log N &= 3{\cdot}00000 \\
\operatorname{colog} b &= \bar{2}{\cdot}80901 \\
m_1 \log m_1 &= \bar{1}{\cdot}84123 \\
m_2 \log m_2 &= 1{\cdot}23179 \\
\operatorname{colog} (r-2)^{r-2} &= \bar{2}{\cdot}39590 \\
\log \Gamma(r) &= 1{\cdot}50406 \\
\operatorname{colog} \Gamma(m_1+1) &= {\cdot}05219 \\
\operatorname{colog} \Gamma(m_2+1) &= \bar{1}{\cdot}34037 \\
\hline
\log y_0 \quad & 2{\cdot}17455
\end{aligned}
$$

where, of course, $\log \Gamma(m_2+1) = \log \Gamma(3{\cdot}776978) = \log 2{\cdot}776978 + \log 1{\cdot}776978 + \log \Gamma(1{\cdot}776978)$, the last value being taken from the table at the end of the book.

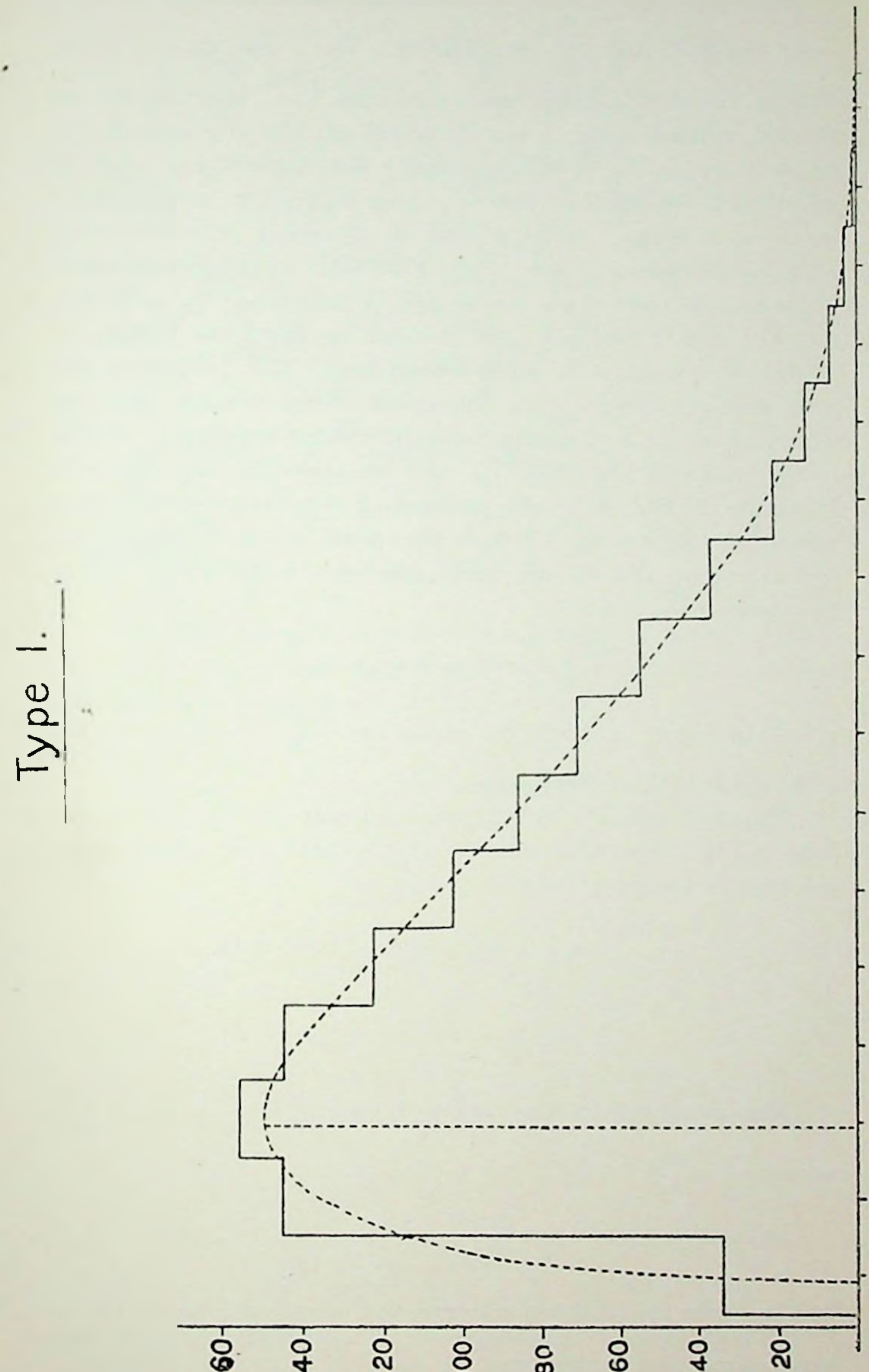
The work to this point gives as the curve for graduating the statistics

$$y = 149{\cdot}47 \left\{ 1 + \frac{x}{1{\cdot}99638} \right\}^{\cdot 409833} \left\{ 1 - \frac{x}{13{\cdot}52728} \right\}^{2{\cdot}776978}$$

where the origin is at age 26·75942 and the unit is five years.

The following table shows the calculation of ordinates of the curve from the equation just given:—

Age	$1+\frac{x}{a_1}$	$1-\frac{x}{a_2}$	log(2)	log(3)	$m_1\times$ col(4)	$m_2\times$ col(5)	col(6) + col(7) $+\log y_0$ $=\log y_x$	y_x	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
17	·02228	1·14429	$\bar{2}$·34792	0·05854	$\bar{1}$·3229	0·1626	$\bar{1}$·6601	45·7	44
22	·52319	1·07037	$\bar{1}$·71866	·02955	$\bar{1}$·8817	·0821	2·1401	138·2	137
27	1·02410	·99614	0·01034	$\bar{1}$·99845	0·0012	$\bar{1}$·9957	2·1745	149·5	149
32	1·52501	·92252	·18327	·96498	·0751	·9027	2·1525	142·1	142
37	2·02592	·84859	·30662	·92870	·1257	·8020	2·1023	126·6	127
42	2·52683	·77466	·40257	·88911	·1650	·6921	2·0317	107·6	108
47	3·02774	·70074	·48111	·84556	·1972	·5711	1·9429	87·7	88
52	3·52865	·62681	·54760	·79714	·2244	·4367	1·8357	68·5	69
57	4·02956	·55289	·60526	·74264	·2481	·2853	1·7080	51·0	51
62	4·53017	·47896	·65615	·68030	·2689	·1122	1·5557	36·0	36
67	5·03136	·40504	·70169	·60750	·2876	$\bar{2}$·9100	1·3722	23·6	24
72	5·53229	·33111	·74291	·51997	·3045	·6670	1·1461	14·0	14
77	6·03320	·25719	·78055	·41025	·3199	·3623	·8568	7·2	7
82	6·53411	·18326	·81519	·26307	·3341	$\bar{3}$·9535	·4622	2·9	3
87	7·03502	·10934	·84726	·03878	·3472	·3307	$\bar{1}$·8525	·7	1
92	7·53593	·03541	·87714	$\bar{2}$·54913	·3595	$\bar{5}$·9709	$\bar{3}$·5050	...	...


Type I.
160
140
120
100
80
60
40
20
17
22
27
32
37
42
47
52
57
62
67
72
77
82
87
92

Cols. (2) and (3) have a constant first difference, viz., $\frac{1}{a_1}$ or ·500907, and $\frac{1}{a_2}$, or ·073925. The value at any point having been calculated and checked, the other items are formed continuously. Cols. (4) to (9) explain themselves, but we may remark that it is generally advisable to use a larger number of figures than five in taking logarithms, especially if m_1 or m_2 is large. A little care is necessary in multiplying such numbers as $\bar{1}$·71866 by m_1(·409833). If an arithmometer is used, m_1 is put on the plate, and is multiplied by −·28134, and the result −·1153 must be put in the form $\bar{1}$·8847, to enable us to add it to other logarithms. Col. (10) gives the area, and was formed by applying one of the formulæ on p. 48. The area of the first group must be treated separately, as the curve starts at age 16·7775, and the base of the group is therefore 2·7225 in length, instead of 5 years as in the other cases. A good way to find the area is to calculate the ordinates for the middle and ends of the base, and apply Simpson's rule, viz. :—

$$\int_0^1 y_x dx = \tfrac{1}{6}\{y_0 + 4y_{\frac{1}{2}} + y_1\},$$

remembering to multiply the result by $\frac{2 \cdot 7225}{5}$ to allow for the different length of the base.

The mid-ordinate is 92·1, the ordinate at the end of the base is 116·5, and the ordinate at the start is of course zero; the area is approximately

$$\frac{2 \cdot 7225}{5} \times \frac{1}{6}\{0 + 4 \times 92 \cdot 1 + 116 \cdot 5\} = 44.$$

PROOF OF FORMULÆ.*

The equation to the curve is $y = y_0\left(1 + \frac{x}{a_1}\right)^{m_1}\left(1 - \frac{x}{a_2}\right)^{m_2}$ where $\frac{m_1}{a_1} = \frac{m_2}{a_2}$.

Let $a_1 + a_2 = b$ and $z = \frac{a_1 + x}{a_1 + a_2}$.

* The reader who has little acquaintance with formulæ of reduction and the Γ and B functions, should consult Appendix II. before reading the proofs of the formulæ for this and the other types.

The area from $x=-a_1$ to $x=+a_2$ is the total frequency N.

$$\therefore \qquad N=\int_{-a_1}^{a_2} y_0\left(1+\frac{x}{a_1}\right)^{m_1}\left(1-\frac{x}{a_2}\right)^{m_2} dx$$

$$=\int_{a}^{a_2} \frac{y_0}{a_1^{m_1}a_2^{m_2}}(a_1+x)^{m_1}(a_2-x)^{m_2}dx$$

$$=\int_0^1 \frac{y_0}{a_1^{m_1}a_2^{m_2}}[z(a_1+a_2)]^{m_1}[(z-1)(a_1+a_2)]^{m_2}(a_1+a_2)dz$$

$$=\int_0^1 \frac{y_0(a_1+a_2)^{m_1+m_2+1}}{a_1^{m_1}a_2^{m_2}} z^{m_1}(z-1)^{m_2}dz$$

$$=\frac{y_0(m_1+m_2)^{m_1+m_2}(a_1+a_2)}{m_1^{m_1}m_2^{m_2}} B(m_1+1,\ m_2+1)\ ;$$

$$\text{Or} \qquad y_0=\frac{N}{b}\cdot\frac{m_1^{m_1}m_2^{m_2}}{(m_1+m_2)^{m_1+m_2}}\cdot\frac{\Gamma(m_1+m_2+2)}{\Gamma(m_1+1)\Gamma(m_2+1)}.$$

Using the same method for the moments as that just given for the area, we see that the nth moment, about the line parallel to the axis of y through $x=-a_1$, is—

$$N\mu'_n=\int_{-a_1}^{a_2} \frac{y_0}{a_1^{m_1}a_2^{m_2}}(a_1+x)^n(a_1+x)^{m_1}(a_2-x)^{m_2}dx$$

$$=\int_0^1 \frac{y_0(a_1+a_2)^{m_1+m_2+n+1}}{a_1^{m_1}a_2^{m_2}} z^{m_1+n}(z-1)^{m_2}dz$$

$$=\frac{y_0(m_1+m_2)^{m_1+m_2}b^{n+1}}{m_1^{m_1}m_2^{m_2}}\cdot\frac{\Gamma(m_1+n+1)\Gamma(m_2+1)}{\Gamma(m_1+m_2+n+2)}.$$

Now, since $\Gamma(p)=(p-1)\Gamma(p-1)$, the moments about the line parallel to the axis of y through $x=-a_1$ are as follows:—

$$\mu'_1=\frac{b(m_1+1)}{m_1+m_2+2}$$

$$\mu'_2=\frac{b^2(m_1+1)(m_1+2)}{(m_1+m_2+2)(m_1+m_2+3)} \text{ and so on.}$$

Changing the origin in order to get moments about the mean and writing $m'_1=m_1+1$ and $m'_2=m_2+1$ and $r=m'_1+m'_2$, we have

$$\mu_2=\frac{b^2m'_1m'_2}{r^2(r+1)}$$

$$\mu_3=\frac{2b^3m'_1m'_2(m'_2-m'_1)}{r^3(r+1)(r+2)}$$

$$\mu_4=\frac{3b^4m'_1m'_2\{m'_1m'_2(r-6)-2r^2\}}{r^4(r+1)(r+2)(r+3)}$$

We can simplify these expressions to obtain the equations on p. 53 by writing $\beta_1=\frac{\mu_3^2}{\mu_2^3}$, $\beta_2=\frac{\mu_4}{\mu_2^2}$, and $\epsilon=m'_1m'_2$; then

$$\beta_1=\frac{4(r^2-4\epsilon)(r+1)}{\epsilon(r+2)^2} \quad \text{or} \quad \frac{\beta_1(r+2)^2}{4(r+1)}=\frac{r^2}{\epsilon}-4,$$

and $$\beta_2=\frac{3(r+1)\{2r^2+\epsilon(r-6)\}}{\epsilon(r+2)(r+3)}$$

$$\text{or} \quad \frac{\beta_2(r+2)(r+3)}{3(r+1)}=\frac{2r^2}{\epsilon}+r-6.$$

Eliminating $\frac{r^2}{\epsilon}$ we find

$$\frac{\beta_1(r+2)^2}{2(r+1)}-\frac{\beta_2(r+2)(r+3)}{3(r+1)}=-8-r+6.$$

Dividing out by $r+2$ we have

$$r=\frac{6(\beta_2-\beta_1-1)}{3b_1-2\beta_2+6}.$$

Using this value in the equation $\frac{\beta_1(r+2)^2}{4(r+1)}=\frac{r^2}{\epsilon}-4$

$$\epsilon=\frac{r^2}{4+\frac{1}{4}\beta_1\frac{(r+2)^2}{r+1}}$$

and from the equation for μ_2

$$b^2=\frac{\mu_2(r+1)r^2}{\epsilon}.$$

The other equations follow at once from $r=m'_1+m'_2$ and $\epsilon=m'_1m'_2$. The distance between the mean and mode is $a_1-\mu'_1=(a_1-bm'_1)\div(m'_1+m'_2)$, which can be easily reduced to the form given. A general value (regardless of type) for the distance was given in Chap. IV., Art. 5.

TYPE II.

$$y = y_0 \left(1 - \frac{x^2}{a^2}\right)^m$$

FORMULÆ.

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)}$$

$$a^2 = \frac{2\mu_2\beta_2}{3 - \beta_2}$$

$$y_0 = \frac{N \times \Gamma(2m + 2)}{a 2^{2m+1} \{\Gamma(m + 1)\}^2}$$

Put $\beta_1=0$ in Type I., for the curve is symmetrical, and therefore $\mu_3=0$. For the same reason it is clear that $m_1=m_2$.

Γ may be approximated to if m is large.

If m is positive, the curve starts at zero, rises to a maximum and falls again to zero; but if m is negative, it starts at infinity, falls, and then rises to infinity again.

EXAMPLE.

In the discussion that followed the reading of Mr. Lidstone's paper on Endowment Assurances, Mr. G. F. Hardy said that "the errors in the successive groups formed a curve very similar to the normal curve of error" (*Journal of the Institute of Actuaries*, xxxiv., p. 87), and the series in question is a rather interesting example of a symmetrical distribution.

Unexpired Term in Years.	Error involved in using "Mean Age" Method.
0– 4	11
5– 9	116
10–14	274
15–19	451
20–24	432
25–29	267
30–34	116
35, &c.	16
...	1,683

Moments were calculated about the centre of the 15–19 group, and ·4985146, 2·161022, 3·104576, and 12·60666 were found for the first four moments; transferring to the mean (17·5+2·492573=19·992573), and using Sheppard's adjustments, the following values result:—

$$\mu_2= 1{\cdot}829172$$
$$\mu_3= {\cdot}120452$$
$$\mu_4= 8{\cdot}52636$$
$$\beta_1= {\cdot}0023706$$
$$\beta_2= 2{\cdot}548313$$
$$\kappa_2=-{\cdot}007492,$$ which shows that

Type II. can be used. The equations for the type give

$$m = 4{\cdot}141766$$
$$a = 4{\cdot}543079$$
$$y_0 = 462{\cdot}57$$

The mean and mode coincide, because the curve is symmetrical.

For calculating a series of values, the following arrangement is convenient :—

$\frac{x}{a}$ (1)	$\log\left(1+\frac{x}{a}\right)$ (2)	$\log\left(1-\frac{x}{a}\right)$ (3)	(2) + (3) (4)	$\log y_x = m \times (4) + y_0$ (5)

It is easier to work in this way than by calculating values of $1-\frac{x^2}{a^2}$. In the particular example, ordinates were calculated at the beginning, middle, and end of each group, and Simpson's quadrature formula was used for finding the areas, viz., $\int_0^1 y\,dx = \frac{1}{6}\{y_0 + 4y_{\frac{1}{2}} + y_1\}$.

Group.	Areas.	Mid-ordinates.
0– 4	14	11
5– 9	109	104
10–14	286	287
15–19	433	440
20–24	433	440
25–29	285	287
30–34	109	104
35, &c.	14	11
...	1,683	...

A comparison of the mid-ordinates with the areas gives an idea of the error involved in using the former for the latter ; the differences are largest at the "tails" and near the mode.

The curve starts at $19{\cdot}992573 - 22{\cdot}71540 = -2{\cdot}72283$, and ends at $42{\cdot}70797$.

It sometimes happens that $\beta_2 > 3$, and so a^2 and m are negative ; if $a'^2 = -a^2$ and $m' = -m$, in such a case the equation to the curve becomes $y = y_0\left(1+\frac{x^2}{a'^2}\right)^{-m'}$ and the value of y_0 can best be found by

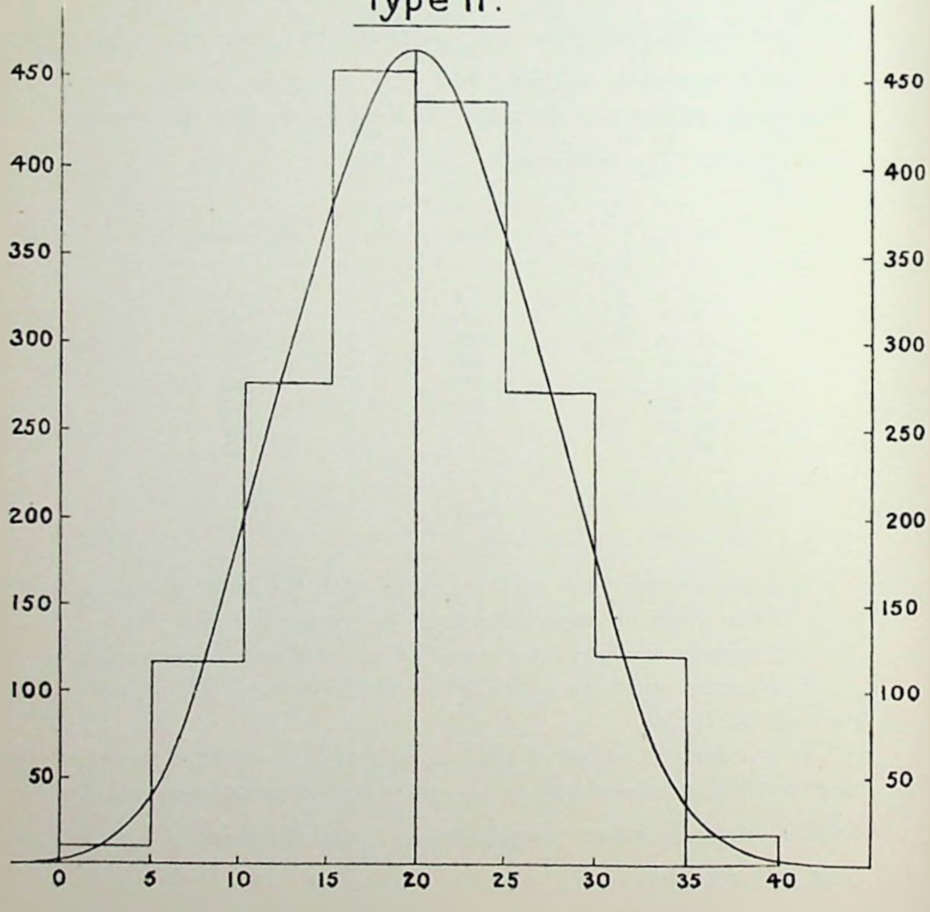
$$N = \int_{-\infty}^{+\infty} y_0\left(1+\frac{x^2}{a'^2}\right)^{-m'} dx = 2\int_0^{\infty} y_0\left(1+\frac{x^2}{a'^2}\right)^{-m'} dx,$$

then putting $1+\frac{x^2}{a'^2}=z^{-1}$ the reader will have no difficulty in showing that $N=\int_0^1 y_0 . a'(1-z)^{-\frac{1}{2}}z^{m'-1\frac{1}{2}}dz = a'y_0 B(m'-\frac{1}{2}, \frac{1}{2})$ by Appendix II., or $y_0=\frac{N}{a'}\cdot\frac{\Gamma(m')}{\Gamma(m'-\frac{1}{2})\Gamma(\frac{1}{2})}$ and $\Gamma(\frac{1}{2})=\sqrt{\pi}$.

In a similar manner we could show that an alternative value for y_0 to that given on the previous page is

$$y_0=\frac{N}{a}\cdot\frac{\Gamma(m+1\frac{1}{2})}{\sqrt{\pi}\Gamma(m+1)}$$

Type II.



TYPE III.

$$y = y_0 e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a}$$

FORMULÆ.

$$\gamma = \frac{2\mu_2}{\mu_3}$$

$$a = \frac{2\mu_2^2}{\mu_3} - \frac{\mu_3}{2\mu_2}$$

$$y_0 = \frac{N}{a}\frac{p^{p+1}}{e^p\Gamma(p+1)}, \text{ where } p = \gamma a$$

$$\text{Mode} = \text{Mean} - \frac{1}{2}\frac{\mu_3}{\mu_2}$$

$$\text{Sk.} = \frac{1}{2}\frac{\mu_3}{\mu_2^{3/2}}$$

NOTES.

If p is positive, the shape of the curve is like that shown in the example of Type I.; but instead of ending at a fixed point, it goes to infinity.

EXAMPLE.

The following statistics are taken from a paper in the *Transactions of the Actuarial Society of Edinburgh*, vol. iv., p. 44, and give the numbers of wives tabulated for the ages of mothers, and according to years since marriage. The mothers' ages for the particular series are 30 to 34.

Year after Marriage.	Number of Wives.	Graduated by Type III. Curve.
1	44	59
2	135	111
3	45	45
4	12	20
5	8	9
6	3	4
7	1	2
8	3	1
Total ...	251	251

The mean is ·3346612 after the middle of the second group, and the moments about the centroid vertical are 1·441787, 3·606622, and 18·93221; so that $\kappa = -8{\cdot}44$.

As this value was large, Type III. was used, and

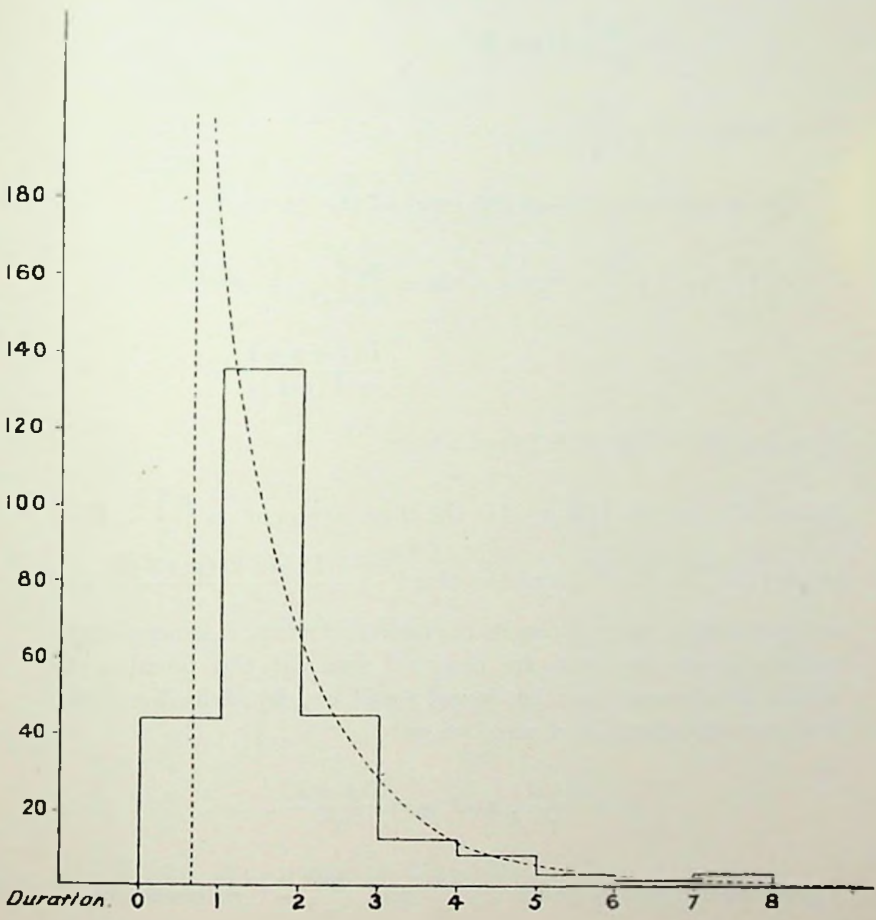
$$\gamma = \quad {\cdot}7995221$$
$$p = - \ {\cdot}0783584$$
$$a = - \ {\cdot}098007$$
$$y_0 = 214{\cdot}8$$

This example is given because it is one which shows a difficulty rather clearly. At first sight, a curve starting at zero, rising to a maximum, and then falling, might be expected. In reality, we find the curve starting at duration ·68192,*

* The mode in ordinary cases of Type III. is given by mean $-\frac{\mu_3}{2\mu_2}$. In this case, $\frac{\mu_3}{2\mu_2} = 1{\cdot}25075$; so the mode would be at ·58391, and the curve would start at $\{\text{"mode"} - a\} = {\cdot}58391 + {\cdot}09801 = {\cdot}68192$.

so that the first group is made up of a strip on a base ·31808 in length, and has a smaller value than the next group, though, of course, any ordinate read off within the first group would be larger than any ordinate in the second group. No adjustment was made to the rough moments.

Type III.



PROOF.

In the equation for the type, viz., $y=y_0\left(1+\dfrac{x}{a}\right)^{\gamma a} e^{-\gamma x}$, put $\gamma a=p$, and substitute z for $\gamma(a+x)$; then, if N be the total frequency,

$$\mathrm{N}=\int_{-a}^{\infty} y_0\left(1+\frac{x}{a}\right)^{p} e^{-\gamma x}dx$$

$$=\int_0^{\infty} y_0 z^p a^{-p} e^{-z+p}\gamma^{-(p+1)}\,dz \text{ for } \frac{dz}{dx}=\gamma$$

$$=y_0\frac{e^p}{\gamma p^p}\int_0^{\infty} z^p e^{-z}dz$$

$$=y_0\frac{ae^p}{p^{p+1}}\Gamma(p+1)$$

This gives $y_0=\dfrac{\mathrm{N}p^{p+1}}{ae^p\Gamma(p+1)}$.

The nth moment about the start of the curve is

$$\frac{1}{\mathrm{N}}\int_{-a}^{\infty} y_0\left(1+\frac{x}{a}\right)^{p} e^{-\gamma x}(x+a)^n dx=\frac{y_0 e^p}{\mathrm{N}p^p\gamma^{n+1}}\int_0^{\infty} z^{p+n}e^{-z}dz$$

$$=\frac{\Gamma(p+n+1)}{\gamma^n\Gamma(p+1)}$$

by using the value of N found above.

Since $\Gamma(p)=(p-1)\Gamma(p-1)$, the first moment is $\dfrac{p+1}{\gamma}$, the second $\dfrac{(p+1)(p+2)}{\gamma^2}$, and the third $\dfrac{(p+1)(p+2)(p+3)}{\gamma^3}$. In order to apply these formulæ to statistical work, it is necessary to have moments about the centroid vertical, the position of which (the mean) can be found; and as, by definition, the first moment about it is zero, we get

$$\mu_2=\frac{p+1}{\gamma^2}, \text{ and } \mu_3=\frac{2(p+1)}{\gamma^3}.$$

These results give γ and p as $\dfrac{2\mu_2}{\mu_3}$ and $\dfrac{4\mu_2^3}{\mu_3^2}-1$ respectively.

TYPE IV.

$$y=y_0\left(1+\frac{x^2}{a^2}\right)^{-m} e^{-\nu\tan^{-1}\frac{x}{a}}$$

FORMULÆ.

$$r=\frac{6(\beta_2-\beta_1-1)}{2\beta_2-3\beta_1-6}$$

$$m=\tfrac{1}{2}(r+2)$$

$$\nu=\frac{r(r-2)\sqrt{\beta_1}}{\sqrt{\{16(r-1)-\beta_1(r-2)^2\}}}$$

$$a=\sqrt{\frac{\mu_2}{16}}\sqrt{\{16(r-1)-\beta_1(r-2)^2\}}$$

$$y_0=\frac{\mathrm{N}e^{\frac{1}{2}\nu\pi}}{a\mathrm{G}(r,\nu)}$$

$$\mathrm{Sk.}=\frac{1}{2}\sqrt{\beta_1}\frac{r-2}{r+2}$$

The origin is not at the mode, but is $\frac{\nu a}{r}$ from the mean, *i.e.*,

$$\text{origin}=\text{mean}+\frac{\nu a}{r}$$

$$\text{mode}=\text{mean}-\frac{1}{2}\frac{\mu_3(r-2)}{\mu_2(r+2)}$$

$$y_0=\frac{\mathrm{N}}{a}\sqrt{\frac{r}{2\pi}}\,\frac{e^{\frac{\cos^2\phi}{3r}-\frac{1}{12r}-\phi\nu}}{(\cos\phi)^{r+1}}$$

is a close approximation where $\tan\phi=\frac{\nu}{r}$.

NOTES.

μ_3 and ν have opposite signs, *i.e.*, when μ_3 is positive ν is negative.

A simple way to calculate the curve is to put it in the form

$$x = a \tan \theta$$

$$y = y_0 \cos^{r+2}\theta e^{-\nu\theta}$$

Then θ is taken as 10°, 20°, 30°, &c., and x and y found; this gives corresponding values of x and y, but the values of y will not be for equidistant values of x. In calculating $e^{\nu\theta}$ the value of θ must be taken in circular measure. If equidistant ordinates are required to be calculated accurately, little is gained by the double form, and if we had good tables of $\log(1+x^2)$ and $\tan^{-1}x$, the calculation of a particular ordinate would be a very simple matter.

The calculation and meaning of $G(r, \nu)$ are dealt with in the proof.

EXAMPLES.

The numbers in the following nearly symmetrical distribution represent the exposed to risk of sickness by Sutton's Sickness Tables (males—all durations) when the number of weeks' sickness is represented by the normal curve of error (Type VII.).

Central Age.	No. Exposed.	Graduated by Type IV.
5	10	6*
10	13	16
15	41	49
20	115	135
25	326	321
30	675	653
35	1,113	1,108
40	1,528	1,535
45	1,692	1,712
50	1,530	1,522
55	1,122	1,074
60	610	601
65	255	274
70	86	102
75	26	32
80	8	8
85	2	2
90	1	1
95	1	...
...	9,154	9,154

* This group has been taken as the area of the rest of the curve.

The following values were obtained :—

$$\begin{aligned} \text{Mean} &= 44\cdot5772339 \\ \mu_2 &= 4\cdot527608 \\ \mu_3 &= -\ \cdot705687 \\ \mu_4 &= 64\cdot98048 \\ \beta_1 &= \cdot0053656 \\ \beta_2 &= 3\cdot169897 \\ \kappa &= \cdot0125 \end{aligned}$$

Type IV. was used because, as there is a large number of cases, the probable error of κ will be small (*see* Chapter VIII.).

$$\begin{aligned} r &= 40\cdot12143 \\ \nu &= 4\cdot450399 \text{ (positive because } \mu_3 \text{ is negative)} \\ a &= 13\cdot39152 \\ m &= 21\cdot06072 \\ \text{Sk.} &= -\ \cdot03313 \end{aligned}$$

When the 5-years unit with which we have been working is changed to one year, a becomes $66\cdot9576$, and $a^2 = 4483\cdot325$.

$$\begin{aligned} \text{The origin } &= \text{mean} + \frac{\nu a}{r} \\ &= 52\cdot504394 \end{aligned}$$

The mode which is wanted if the curve is drawn, is at $44\cdot92989$.

As r is large the approximate form for y_0 was used, $\tan\phi = \dfrac{4\cdot450398}{40\cdot12143}$, or, $\log\tan\phi = \log\tan 6^\circ\ 19' \dfrac{8925}{11537}$; hence $\log\cos\phi = \bar{1}\cdot9973446$, and from this y_0 is found to be $273\cdot3649$.

The value was checked by Dr. Alice Lee's tables (*see* Appendix V).

The calculation of ordinates by the double process is as follows :—

θ	x in years of age.	$4\cdot450398\ \theta \log_{10} e$	$42\cdot1243 \log\cos\theta$	$\log y$	y
0°	0	...	...	2·43675	273·37
1°	1·1687	$\bar{1}$·96637 . . .	$\bar{1}$·99721	2·40033	251·38
2°	2·3382	$\bar{1}$·93253 . . .	$\bar{1}$·98885	2·35813	228·10

The second column is formed directly from the tables of $\tan\theta$ by multiplying by a, and as x is required in years, $13{\cdot}39152 \times 5 = 66{\cdot}9576$ should be used for a. The fourth column is formed by multiplying L $\cos\theta$, and the third continuously by addition. When θ is negative, the fourth column has to be subtracted from the fifth: *i.e.,* it ceases to be negative and becomes positive. In each case the sixth is formed from the fourth + the fifth + $\log y_0$.

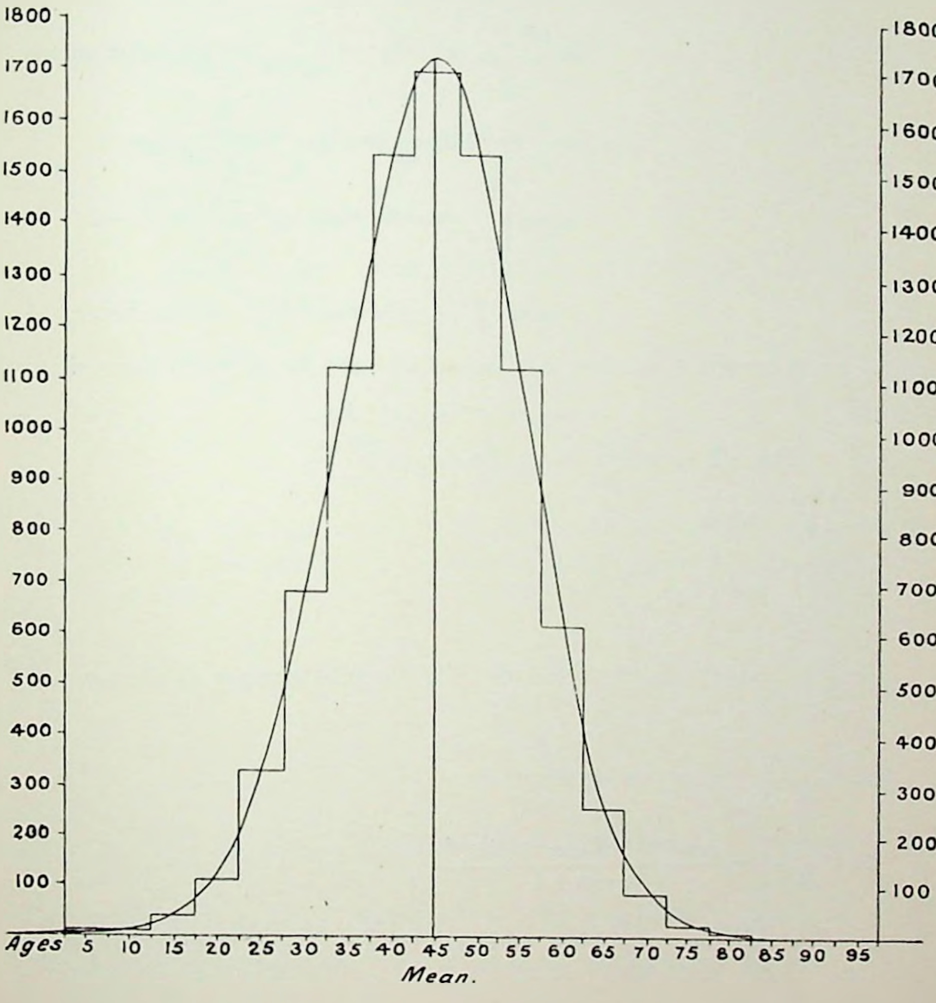
If the calculation is made directly, the following columns would be required:—

$\frac{x}{a}$	$1+\frac{x^2}{a^2}$	$\log\left(1+\frac{x^2}{a^2}\right)$	$\tan^{-1}\frac{x}{a}$ in degrees, &c.	col (4) in circular measure	col (5) $\times(-\nu\log_{10}e)$	$m\times$ col (3)	$\log y_0$ + (6) + (7)	y antilog (8)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)

Col. (2) can be formed best by differences since $\Delta(1+X^2)=2X+1$, $\tan^{-1}\frac{x}{a}$ has to be found by using a table of the tangents of angles inversely. A table helpful for obtaining col. (5) from col. (4) will be found on pp. 251 to 262 of *Chambers' Mathematical Tables* (1897 edition).

When drawing a curve of this type the position and height of the mode can be noted and then corresponding points inserted, *e.g.*, $x=+1{\cdot}1687$ and $y=251{\cdot}38$. Care must be taken to give the curve its maximum at the right point.

Type IV.


1800
1700
1600
1500
1400
1300
1200
1100
1000
900
800
700
600
500
400
300
200
100
Ages 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
Mean.

PROOF.

In $$y = y_0\left\{1+\frac{x^2}{a^2}\right\}^{-m} e^{-\nu\tan^{-1}\frac{x}{a}} \text{ put } \tan\theta = \frac{x}{a},$$

$$\therefore \quad \theta = \tan^{-1}\frac{x}{a} \text{ and}$$

$$\left\{1+\left(\frac{x}{a}\right)^2\right\}^{-m} = \{1+\tan^2\theta\}^{-m} = (\sec^2\theta)^{-m} = \cos^{2m}\theta,$$

$$\therefore \quad y = y_0 \cos^{2m}\theta e^{-\nu\theta}.$$

Now $$\mathrm{N} = \int_{-\infty}^{+\infty} y_0\left\{1+\frac{x^2}{a^2}\right\}^{-m} e^{-\nu\tan^{-1}\frac{x}{a}} dx$$

$$= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} y_0 \cos^{2m}\theta e^{-\nu\theta} \frac{a}{\cos^2\theta} d\theta, \text{ by substituting}$$

$$\tan\theta = \frac{x}{a} \text{ so that } \frac{dx}{d\theta} = a\sec^2\theta = \frac{a}{\cos^2\theta}$$

$$= y_0 a \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^r\theta e^{-\nu\theta} d\theta \text{ where } r = 2m-2$$

$$= y_0 a e^{-\frac{1}{2}\nu\pi} \int_0^{\pi} \sin^r\phi e^{\nu\phi} d\phi, \quad \text{substituting}$$

$\sin\phi$ for $\cos\theta$ so that $\phi + \frac{1}{2}\pi = \theta$ and the limits are changed,

$$= y_0 a e^{-\frac{1}{2}\nu\pi} \mathrm{G}(r, \nu), \text{ say.}$$

The nth moment about the origin is

$$\mu'_n = \frac{1}{\mathrm{N}}\int_{-\infty}^{\infty} y x^n dx$$

$$= \frac{1}{\mathrm{N}}\int_{-\infty}^{\infty} y_0 x^n \left\{1+\frac{x^2}{a^2}\right\}^{-m} e^{-\nu\tan^{-1}\frac{x}{a}} dx$$

$$= \frac{1}{\mathrm{N}}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} y_0 a^{n+1} \cos^{2m-2}\theta \tan^n\theta e^{-\nu\theta} d\theta \text{ by substituting as above}$$

$$= \frac{y_0 a^{n+1}}{\mathrm{N}} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos^{r-n}\theta \sin^n\theta e^{-\nu\theta} d\theta$$

$$= \frac{y_0 a^{n+1}}{\mathrm{N}} \left[-\frac{\cos^{r+n+1}\theta \sin^{n-1}\theta e^{-\nu\theta}}{r-n+1} \right.$$

$$\left. + \int \left\{ \frac{\cos^{r-n+1}\theta}{r-n+1} (\sin^{n-2}\theta \cos\theta e^{-\nu\theta}(n-1) - \nu e^{-\nu\theta} \sin^{n-1}\theta) \right\} d\theta \right]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}$$

by integrating by parts and treating $\sin^{n-1}\theta e^{-\nu\theta}$ as one part and $\cos^{r-n}\theta\ \sin\theta$ as the other, and remembering that $\int\cos^{r-n}\theta\sin\theta d\theta=-\frac{\cos^{r-n+1}\theta}{r-n+1}$.

Now, since $\cos^{r-n+1}\theta\sin^{n-1}\theta e^{-\nu\theta}=0$ when θ becomes $\frac{\pi}{2}$ or $-\frac{\pi}{2}$, we have

$$\mu'_n=\frac{y_0a^{n+1}}{N(r-n+1)}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\begin{aligned}&(n-1)\cos^{r-n+2}\theta\sin^{n-2}\theta e^{-\nu\theta}\\&-\nu\cos^{r-n+1}\theta\sin^{n-1}\theta e^{-\nu\theta}d\theta\end{aligned}$$

$$=\frac{a}{r-n+1}\{(n-1)a\mu'_{n-2}-\nu\mu'_{n-1}\}$$

Further,

$$\mu'_1=\frac{y_0a^2}{N}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^r\theta\tan\theta e^{-\nu\theta}d\theta$$

$$=\frac{y_0a^2}{Nr}\left[-\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\nu\cos^r\theta e^{-\nu\theta}d\theta\right]$$ by putting $n=1$ in the above equation for μ'_n

$$=-\frac{a\nu}{r},\ \text{because } N=y_0a\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos^r\theta e^{-\nu\theta}d\theta$$

Using the last result with the formula for the nth in terms of the two previous moments, and remembering that μ'_0 is unity,

$$\mu'_2=\frac{a^2}{r(r-1)}(r+\nu^2)$$

$$\mu'_3=-\frac{a^3\nu}{r(r-1)(r-2)}(3r-2+\nu^2)$$

$$\mu'_4=\frac{a^4}{r(r-1)(r-2)(r-3)}\{3r(r-2)+\nu^2(6r-8)+\nu^4\}$$

Referring these moments to the centroid vertical, we have, by putting $d=\mu'_1=-\frac{a\nu}{r}$ in the formulæ on p. 19,

$$\mu_2 = \frac{a^2}{r^2(r-1)}(r^2+\nu^2)$$

$$\mu_3 = -\frac{4a^3\nu(r^2+\nu^2)}{r^3(r-1)(r-2)}$$

$$\mu_4 = \frac{3a^4(r^2+\nu^2)\{(r+6)(r^2+\nu^2)-8r^2\}}{r^4(r-1)(r-2)(r-3)}$$

If now, we put z for $r^2+\nu^2$, and write as before,

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2},$$

we have,

$$-\frac{\beta_1(r-2)^2}{2(r-1)} = 8\frac{r^2}{z} - 8$$

and

$$\frac{\beta_2(r-2)(r-3)}{3(r-1)} = r + 6 - \frac{8r}{z}.$$

Adding and dividing out by $r-2$, we have

$$r = \frac{6(\beta_2-\beta_1-1)}{2\beta_2-3\beta_1-6},$$

and

$$z = \frac{r^2}{1-\dfrac{\beta_1(r-2)^2}{16(r-1)}}.$$

Finally, since $\nu^2 = z - r^2$, the other formulæ on p. 69 follow at once.

Since the tangent at the top of the maximum ordinate is parallel to the axis of x, the position of the mode is such that $\frac{dy}{dx}$ is zero at that point, *i.e.*,

$$y_0\left(1+\frac{x^2}{a^2}\right)^{-(m+1)} e^{-\nu\tan^{-1}\frac{x}{a}}\left[-\frac{2mx}{a^2}-\frac{\nu}{a}\right]$$

is zero. There are three cases, $x=-\infty$, $x=+\infty$, and a value of x such that $\frac{2mx}{a^2}+\frac{\nu}{a}$ is zero, or $x=-\frac{\nu a}{2m}$. The distance of the mean from the origin is μ'_1 or $-\frac{\nu a}{r}$, and, therefore, the distance between the mean and mode is $-\frac{2\nu a}{r(r+2)}$, which reduces to the expression given on p. 69, when the values for ν and a, on the same page, are inserted.

It will be useful to give another example of the calculation of y_0 for curves of this type, and may take a curve in which $r=29{\cdot}590$, $\nu=19{\cdot}886$, $a=13{\cdot}650$, $N=2162$. $\therefore \tan\phi = {\cdot}67205$, $\therefore\ \phi=33^\circ\ 54'\frac{8}{42}$, $\cos\phi={\cdot}82998$, $\log\cos\phi=\bar{1}{\cdot}91907$, and ϕ in circular measure is ${\cdot}59172$.

$$
\begin{array}{rrll}
 & & \log N & =3{\cdot}33486 \\
 & & \operatorname{colog} a & =\bar{2}{\cdot}86486 \\
 & & \frac{1}{2}\log r & = \ {\cdot}73557 \\
 & & \log \dfrac{1}{\sqrt{2\pi}} & =\bar{1}{\cdot}60091 \\
\dfrac{\cos^2\phi}{3r} = & {\cdot}00776 & & \\
-\dfrac{1}{12r} = - & {\cdot}00282 & & \\
-\phi\nu = & -11{\cdot}76700 & & \\
\hline
 & -11{\cdot}762 & \times \log_{10} e & =\bar{6}{\cdot}89183 \\
 & & \operatorname{colog}(\cos\phi)^{r+1} & =2{\cdot}47564 \\
\hline
 & & & \bar{1}{\cdot}90367 \\
\hline
 & & y_0 = & {\cdot}80107
\end{array}
$$

The form just considered is sufficiently accurate for all practical purposes provided ν is not very small. If, however, ν is less than 2, $G(r,\nu)$ should be calculated by

$$\frac{2\nu\pi e^{-\frac{1}{2}\pi\nu}\Gamma(r+1)}{\displaystyle\operatorname*{Product}_{n=1}^{n=\infty}\left(1+\frac{r^2+\nu^2}{4n^2\left(1+\frac{r}{n}\right)}\right)}$$

TYPE V.

$$y = y_0 x^{-p} e^{-\gamma/x}$$

FORMULÆ.

$$p = 4 + \frac{8 + 4\sqrt{\{4 + \beta_1\}}}{\beta_1}$$

$$\gamma = (p-2)\sqrt{\mu_2(p-3)}$$

$$y_0 = \frac{N\gamma^{p-1}}{\Gamma(p-1)}$$

$$\text{Sk.} = \frac{2\sqrt{p-3}}{p}$$

$$\text{Origin} = \text{Mean} - \frac{\gamma}{p-2}$$

$$\text{Mode} = \text{Mean} - \frac{2\gamma}{p(p-2)}$$

The sign of γ is the same as that of μ_3.

EXAMPLE.

The following series of deaths is taken from Mr. King's paper "On the rate of Mortality amongst Female Nominees, &c." (*Journal of the Institute of Actuaries*, xxxiii., pp. 262–8):

Ages.	Deaths.	Graduated by Type V.
30–34	1	1
35–39	5	3
40–44	8	6
45–49	12	14
50–54	28	32
55–59	82	68
60–64	128	137
65–69	253	247
70–74	342	381
75–79	525	480
80–84	438	441
85–89	265	261
90–94	53	80
95–99	18	10
100, &c.	4	1
...	2,162	2,162

The mean is at age 75·9782605, and the moments (adjusted), &c., are

$$\mu_2 = 3{\cdot}573346$$
$$\mu_3 = -\ 4{\cdot}752613$$
$$\mu_4 = 51{\cdot}02583$$
$$\beta_1 = {\cdot}4950399$$
$$\beta_2 = 3{\cdot}996134$$
$$\kappa = {\cdot}85$$

Strictly speaking, Type IV. should be used, but the value is not very far from unity, and the following Type V. constants were found :

$$p = 37{\cdot}29145$$
$$\gamma = -390{\cdot}6609 \text{ (negative, because } \mu_3 \text{ is)}$$
$$\log y_0 = 56{\cdot}930518$$

The approximation to the value of $\log \Gamma(p-1)$ was used. The origin is at age 131·32606, and the mode at 78·9467.

The columns used for calculating the ordinates were :

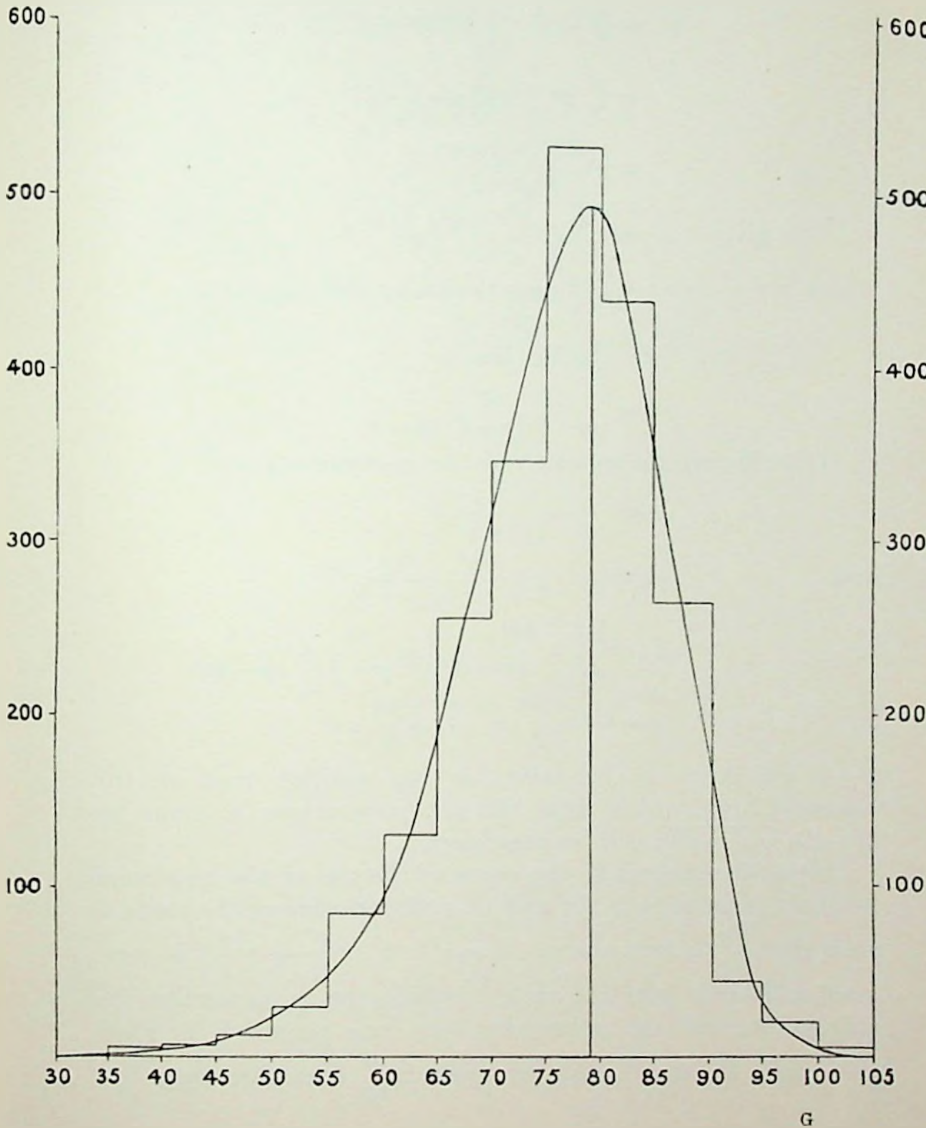
x (1)	$\log x$ (2)	$-p \log x$ (3)	$\frac{1}{x}(-\gamma \log_{10} e)$ (4)	$\log y = \log y_0 + (3) + (4)$ (5)	$y = \text{antilog}\ (3)$ (6)

Col. (4) is best formed by putting $\gamma \log_{10} e$ on the plate of he arithmometer, and multiplying it by $\frac{1}{x}$, obtained, of course, from a table of reciprocals.

The point to be borne in mind in drawing a curve of this type is that as the mode and origin are not at the same place, care must be taken to give the maximum ordinate its right position and magnitude (*cf.* Type IV.).

The graduated figures agree fairly closely with the original statistics below the 90–94 group, but are unsuitable for that and the two later groups. The reason is that Type IV. should be used, and curves of Type V. have a range limited in one direction, while Type IV. curves have an unlimited range. The particular case was chosen partly because an example in which μ_3 is negative is rather more awkward than when μ_3 is positive. In such cases it is a good check to imagine the statistics written in inverse order (in this case 4, 18, 53, &c.), and so avoid the negative signs.

Type V.


600
500
400
300
200
100
600
500
400
300
200
100
30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105
G

PROOF.

Putting $\frac{\gamma}{x}=z$ in $y=y_0e^{-\gamma/x}x^{-p}$, and integrating from 0 to ∞, we have

$$N=y_0\gamma^{1-p}\Gamma(p-1)$$

or
$$y_0=\frac{N\gamma^{p-1}}{\Gamma(p-1)}$$

Using the same substitution, the nth moment about the origin is

$$\mu'_n=\frac{y_0}{N}\gamma^{n-p+1}\int_0^\infty e^{-z}z^{p-n}dz$$

$$=\frac{y_0}{N}\gamma^{n-p+1}\Gamma(p-n-1)$$

$$=\gamma^n\frac{\Gamma(p-n-1)}{\Gamma(p-1)}$$

This gives $\mu'_1=\frac{\gamma}{p-2}$

which is the distance between the mean and origin,

$$\mu'_2=\frac{\gamma^2}{(p-2)(p-3)}$$

$$\mu'_3=\frac{\gamma^3}{(p-2)(p-3)(p-4)}.$$

Transferring the moments to the centroid vertical,

$$\mu_2=\frac{\gamma^2}{(p-2)^2(p-3)}$$

and
$$\mu_3=\frac{4\gamma^3}{(p-2)^3(p-3)(p-4)}$$

$$\therefore\qquad \beta_1=\frac{\mu_3^2}{\mu_2^3}=\frac{16(p-3)}{(p-4)^2}=\frac{16}{p-4}+\frac{16}{(p-4)^2}$$

$$\therefore\qquad (p-4)^2-\frac{16}{\beta_1}(p-4)-\frac{16}{\beta_1}=0.$$

$p-4$ will have to be taken as the positive root of the equation, or γ, which from the above equations is given by $(p-2)\sqrt{\mu_2(p-3)}$, will be imaginary.

Since the tangent to the curve at the top of the maximum ordinate is parallel to the axis of x, the position of the mode is such that $\frac{dy}{dx}$ is zero there, *i.e.*, $y_0x^{-p-1}e^{-\gamma/x}\left\{-p+\frac{\gamma}{x}\right\}$ is zero. $x=0$ and $x=\infty$ give the cases in which the curve touches the axis of x, and the other case, the one required, is when $p-\frac{\gamma}{x}=0$, or $x=\frac{\gamma}{p}$, *i.e.*, the mode is $\frac{\gamma}{p}$ from the origin.

TYPE VI.

$$y = y_0 (x-a)^{q_2} x^{-q_1}$$

FORMULÆ.

$$r = \frac{6(\beta_2 - \beta_1 - 1)}{6 + 3\beta_1 - 2\beta_2}$$

$$1 - q_1 = -\frac{r}{2} + \frac{r(r+2)}{2}\sqrt{\frac{\beta_1}{\beta_1(r+2)^2 + 16(r+1)}}$$

$$1 + q_2 = -\frac{r}{2} - \frac{r(r+2)}{2}\sqrt{\frac{\beta_1}{\beta_1(r+2)^2 + 16(r+1)}}$$

$$a = \frac{1}{2}\sqrt{\mu_2}\sqrt{\beta_1(r+2)^2 + 16(r+1)}$$

$$y_0 = \frac{N a^{q_1 - q_2 - 1}\Gamma(q_1)}{\Gamma(q_1 - q_2 - 1)\Gamma(q_2 + 1)}$$

$$\text{Sk.} = \frac{1}{2}\sqrt{\beta_1}\,\frac{r+2}{r+2}$$

$$\text{Origin} = \text{Mean} - \frac{a(q_1 - 1)}{q_1 - q_2 - 2}$$

$$\text{Mode} = \text{Mean} - \frac{1}{2}\cdot\frac{\mu_3}{\mu_2}\cdot\frac{r+2}{r+2}$$

The range is from a to ∞, and the method is like that of Type I. r and ϵ are found exactly as in Type I., and $1-q_1$ and $1+q_2$ are the roots of $z^2-rz+\epsilon=0$, just as $1+m_1$ and $1+m_2$ were in Type I. The origin is before the beginning of the curve. $1-q_1$ is taken with the negative root and $1+q_2$ with the positive root when μ_3 is negative, and *vice versâ*.

EXAMPLE.

The number of entrants in the recent limited payment policies experience were summed in groups of ten years of age and divided by 100, and the following series was obtained—

No. of Entrants ÷ 100.	Graduated by Type VI. curve.
1	1
56	50
167	168
98	100
34	36
9	10
2	2
1	·5
368	368

The moments, &c., were—

mean at ·402174 after the centre of 167 group—

$$\mu_2 = \cdot 928835$$
$$\mu_3 = \cdot 893096$$
$$\mu_4 = 4\cdot 088800$$
$$\beta_1 = \cdot 9953605$$
$$\beta_2 = 4\cdot 739349$$
$$\kappa_2 = 1\cdot 895$$
$$r = -\ 33\cdot 42129$$
$$1-q_1 = -\ 41\cdot 03080$$
$$1+q_2 = 7\cdot 60950$$
$$q_1 = 42\cdot 03080$$
$$q_2 = 6\cdot 60950$$
$$a = 10\cdot 37947$$
$$\log y_0 = 46\cdot 1821$$

The origin is 12·74270 before the mean or 12·34053 before the centre of the 167 group, and the curve starts at 12·34053 − 10·37947 = 1·96106 before the centre of the largest group. This makes the start of the curve at about age 10, which is reasonable.

The curve was calculated as follows—

x	$\log x$	$\log(x-a)$	$-q_1 \log x$	$q_2 \log(x-a)$	$\log y$	y
(1)	(2)	(3)	(4)	(5)	(6)	(7)

There is no difficulty in writing down the values for columns (2) and (3) without using column (1), as only the whole numbers in x and $x-a$ change, the decimal remaining constant so long as equidistant ordinates are required. Columns (4) and (5) are obtained directly, and column (6) by adding columns (4) and (5) to $\log y_0$.

The mode which is useful for drawing the curve is ·02429 before the centre of the largest group.

The skewness is ·443.

PROOF.

$$N = \int_a^\infty y_0 (x-a)^{q_2} x^{-q_1} dx$$

$$= \int_a^\infty y_0 a^{q_2-q_1} \left(\frac{x}{a}-1\right)^{q_2} \left(\frac{x}{a}\right)^{-q_1} dx$$

$$= \int_1^0 y_0 a^{q_2-q_1} \left(\frac{1}{z}-1\right)^{q_2} z^{q_1} (-a z^{-2}) dz$$

by substituting $\frac{1}{z}$ for $\frac{x}{a}$

$$= \int_0^1 y_0 a^{q_2-q_1+1} (1-z)^{q_2} z^{q_1-q_2-2} dz$$

$$\therefore \quad y_0 = \frac{N}{a^{q_2-q_1+1} B(q_2+1\,;\ q_1-q_2-1)}$$

$$= \frac{N\Gamma(q_1) a^{q_1-q_2-1}}{\Gamma(q_2+1)\Gamma(q_1-q_2-1)}.$$

The nth moment about the origin is

$$\mu'_n = \frac{1}{N}\int_a^\infty y_0 x^n (x-a)^{q_2} x^{-q_1} dx$$

$$= \frac{y_0}{N a^{q_1-q_2-n-1}} \frac{\Gamma(q_1-q_2-n-1)\Gamma(q_2+1)}{\Gamma(q_1-n)}$$

by the same substitution as that used above.

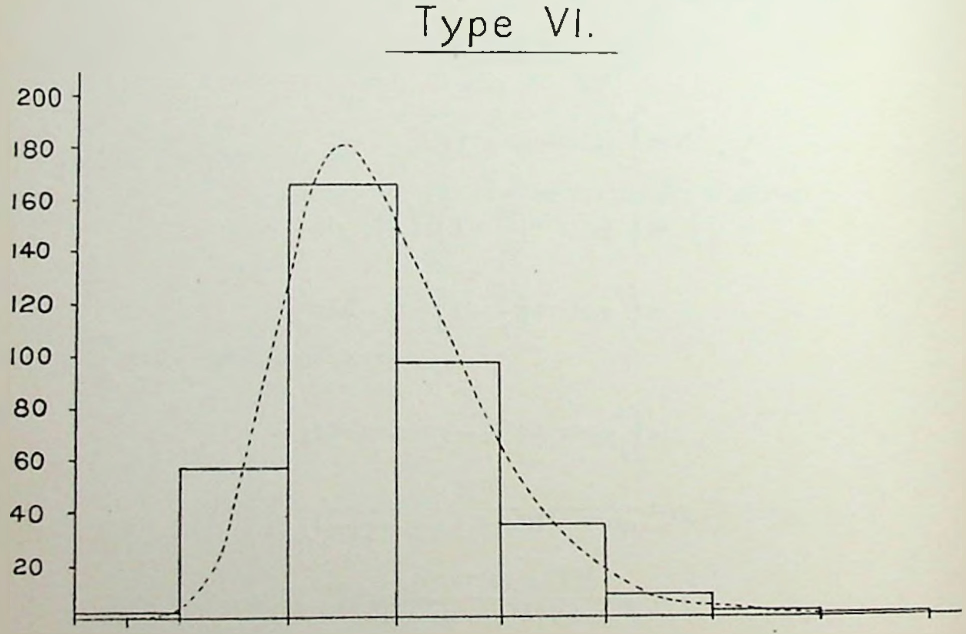
From this last result we obtain, by inserting the value of y_0, and remembering the relationship between $\Gamma(q_1)$ and $\Gamma(q_1-1)$, &c.,

$$\mu'_1 = \frac{a(q_1-1)}{q_1-q_2-2}$$

$$\mu'_2 = \frac{a^2(q_1-1)(q_1-2)}{(q_1-q_2-2)(q_1-q_2-3)}$$

&c.

It will be noticed that these equations are the same as those already obtained for Type I. if $m_1 = -q_1$ and $m_2 = q_2$. Thus, we can use the whole of the Type I. solution, provided we bear in mind that the range is from $x=a$ to $x=\infty$.



TYPE VII.

"NORMAL CURVE OF ERROR."

$$y = y_0 e^{-x^2/c}$$

FORMULÆ.

$$c = 2\mu_2$$

$$y_0 = \frac{N}{\sqrt{2\pi\mu_2}}$$

EXAMPLES.

The following table gives, in column (2), the sums assured and bonuses, and in column (4) the reserves resulting from grouping a number of Endowment Assurances according to their office years of birth :—

Central Age for 5 groups of years of birth.	Sum Assured and Bonuses ÷ 1,000.		Reserves ÷ 1,000.	
	Ungraduated.	Graduated.	Ungraduated.	Graduated.
(1)	(2)	(3)	(4)	(5)
17	11	13	·6	·6
22	48	40	2·8	2·7
27	124	104	11·5	10·9
32	213	202	27·7	30·1
37	281	282	59·1	58·4
42	295	288	84·7	80·1
47	185	214	74·1	76·9
52	104	116	50·5	52·2
57	40	44	23·2	25·0
62	15	13	12·2	8·4
67	3	3	1·3	2·4
Total ...	1,319	1,319	347·7	347·7

The following table shows the moments and constants :—

Constant.	Sum Assured and Bonus.	Reserves.
Mean age	39·202426	43·967213
μ_2	3·066840	2·769635
μ_3	·650127	·029805
μ_4	27·02516	22·40663
β_1	·014653	·0000418
β_2	2·873346	2·920997
κ	− ·005	− ·0002
$\sigma(=\sqrt{\mu_2})$	1·751237	1·664222
σ^{-1}	·5710248	·6008813
y_0	300·4760	83·34959

The criteria for the normal curve are $\kappa=0$, $\beta_1=0$, and $\beta_2=3$. The values given above do not differ very greatly from these, but a comparison of the graduated and ungraduated figures shows that the reserve curve agrees better than the sum assured curve; partly because the value of β_2 is closer to 3, and β_1 has a larger value in the case of the sum assured.

For the calculation of y_0 the value of

$$\log \frac{1}{\sqrt{2\pi}} = \bar{1}\cdot6009100657$$

is required.

In finding the areas for the comparison between the graduated and ungraduated figures it is unnecessary to calculate the ordinates, as one of the calculated tables of the probability integral can be used. The best table was recently given by Mr. W. F. Sheppard (*Biometrika*, vol. ii., pp. 174, &c.), and the columns in the following table show how it was used to calculate the areas in one of the cases (the reserves). Mr. Sheppard's tables give the areas and ordinates of the normal curve in terms of the standard deviation; that is, he assumes the standard deviation to be unity, and his tables must be entered by using intervals of σ^{-1}.

Age x.	Distance from origin in calculation units, *i.e.*, 5 years of age.	Previous column $\times \sigma^{-1}$.	Values of α from Sheppard's Tables using differences (area from origin to x).	Difference of previous column = area for age group x to $x+5$.	Area multiplied by 317·7 (total frequency)
14·5	5·893443	3·541258	...	·00144*	·6
19·5	4·893443	2·940377	·99836	·00785	2·7
24·5	3·893443	2·339496	·99049	·03152	10·9
29·5	2·893443	1·738615	·95897	·08659	30·1
34·5	1·893443	1·137734	·87238	·16806	58·4
39·5	·893443	·536853	·70432	·22985†	80·1
44·5	·106557	·064028	·52553	·22141	76·9
49·5	1·106557	·664909	·74694	·15018	52·2
54·5	2·106557	1·265790	·89712	·07190	25·0
59·5	3·106557	1·866671	·96902	·02418	8·4
64·5	4·106557	2·467552	·99320	·00572	2·0
69·5	5·106557	3·068143	·99892	·00108*	·4

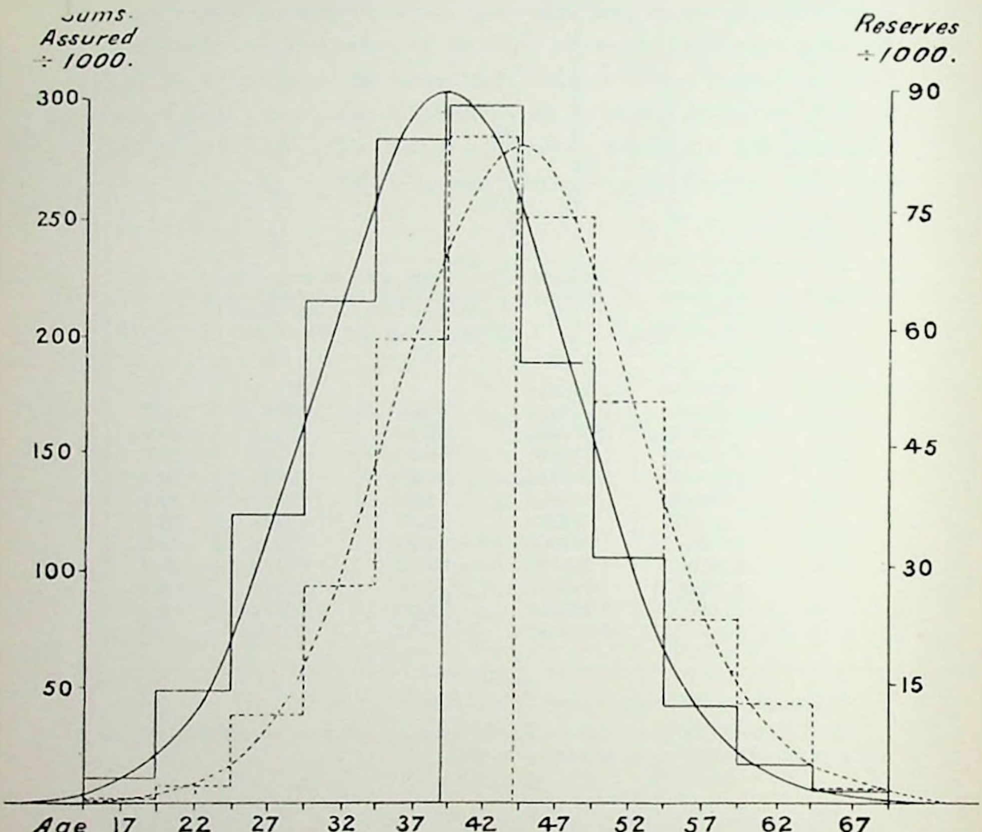
* Remainders of areas beyond 19·5 and 69·5.

† (·70432 − ·50000) + (·52553 − ·50000) because we pass across the origin, and a piece of the group is on each side of it.

The second column can be left out when the method has been grasped. The ages in the first column were taken consistently with the assumptions that 17, 22, etc., were the central ages of the groups.

If ordinates are required, the z column in Mr. Sheppard's tables must be used. It was with its help that the curves in the figure were drawn. The statistics and curve for the reserves are shown by the dotted lines.

Type VII.



An average reserve for any group can be obtained by means of the graduated figures, and it could be used to test the reserves obtained at any future valuation. This is by no means the only rough check that can be applied, but it is interesting because it shows a use to which frequency-curves might be put in practical office routine.

PROOF.

To show that

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

let

$$\int_0^\infty e^{-x^2}dx=\kappa$$

then, substituting ax for x, we have

$$\int_0^\infty e^{-a^2x^2}adx=\kappa$$

$$\therefore \qquad \int_0^\infty e^{-a^2(1+x^2)}adx=\kappa e^{-a^2}.$$

Hence,

$$\int_0^\infty\int_0^\infty e^{-a^2(1+x^2)}adadx=\kappa\int e^{-a^2}da=\kappa^2$$

But

$$\int_0^\infty e^{-a^2(1+x^2)}ada=\frac{1}{2}\cdot\frac{1}{1+x^2}$$

$$\therefore \qquad \frac{1}{2}\int_0^\infty\frac{dx}{1+x^2}=\kappa^2$$

$$\therefore \qquad \kappa^2=\frac{\pi}{4} \text{ or } \kappa=\frac{\sqrt{\pi}}{2}.$$

Hence,

$$\int_{-\infty}^\infty e^{-x^2}dx=\sqrt{\pi}$$

The other constant is obtained as follows :—

$$\int_{-\infty}^\infty y_0e^{-\frac{x^2}{c}}dx=y_0\left[xe^{-\frac{x^2}{c}}+\int\frac{2x}{c}e^{-\frac{x^2}{c}}xdx\right]_{-\infty}^\infty \text{ by parts}$$

$$=\frac{2y_0}{c}\int_{-\infty}^\infty x^2e^{-\frac{x^2}{c}}dx$$

or

$$N=\frac{2N}{c}\mu_2$$

$$\therefore \qquad c=2\mu_2$$

8. Up to the present we have merely considered examples with a view to illustrating the various types of frequency-curves, but it seems advisable to consider one or two practical examples which may help to show the range of applicability of the curves in actuarial work, and give an opportunity of noticing a few difficulties which may arise in applying them.

The function with which actuaries generally wish to deal in practical work is not an exposed to risk or series of deaths or withdrawals, but the ratio between the deaths and the exposed; that is, with the rates of mortality, sickness, marriage, and withdrawal. An actuary studying frequency-curves may therefore naturally ask whether any of these rates can be graduated by means of the curves we have examined, and, if they fail, must they be put aside for some other method? Now the first point to be considered is whether these rates are frequency distributions; if they are not, the use of the frequency-curve is empirical. A rate of mortality gives the proportion of people at each age who die, and if we imagine 1,000 persons exposed to risk at each integral age, the number of deaths would be 1,000 times the rate of mortality, and this seems to show that it is possible to consider the rate of mortality as a distribution, though it is one that could hardly arise in actual experience. It is impossible to describe the rates of mortality or sickness by a single frequency-curve. On the other hand, the rates of marriage are certainly much like frequency-curves, and the rates of withdrawal, whether regarded according to age or duration, might take a form like our example in Type III. There are, however, practical objections to the direct operation on rates, even apart from the very exaggerated idea of frequency distributions in which it is necessary to indulge. The numbers exposed to risk at the end of any table become small, and a single death or marriage there gives a very large rate, while at several ages near there may be a zero rate shown by the ungraduated data. This is extremely awkward, as it tends to make the ratios dealt with far rougher in application than the actual observations are in fact, and we are forced to group the material before using it, which introduces an arbitrary practice which it is well to avoid as far as possible. It must not, of course, be inferred that a small number of say fifty or one hundred deaths must necessarily be grouped according to each year of age, but that even if there are two or three thousand the roughnesses introduced by the use of rates influence the result considerably. The reason is that an equal weight is given to each rate of mortality which is very far from the weight indicated by the exposed to risk.

9. It will be useful to consider a case bearing out these objections and then deal with a practical method of overcoming them. The statistics to be considered have been taken from a paper by Mr. M. Mackenzie Lees "On Rates of Mortality and Marriage among daughters of Peers and Heirs Apparent, &c." (*Transactions of the Faculty of Actuaries*, vol. i., p. 276), and may be summarized as on page 94.

The moments were calculated by Mr. G. F. Hardy's Summation Method, and were found, about the mean 28·77191, to be

$$\mu_2 = 63{\cdot}2092$$
$$\mu_3 = 627{\cdot}101$$
$$\mu_4 = 19{,}103{\cdot}3$$
$$\therefore \quad \beta_1 = 1{\cdot}557153$$
$$\beta_2 = 4{\cdot}781321$$

The criterion was $\kappa = -1{\cdot}5$, but as I had neglected the rate ·00089 at 71 in calculating the moments, I used Type III. The inclusion of the rate at that age would have lengthened the curve and considerably increased the arithmetical value of the criterion.

The constants for Type III. were

$$\gamma = {\cdot}201592$$
$$p = 1{\cdot}56881$$
$$a = 7{\cdot}78189$$
$$\text{Mode} = 23{\cdot}81128$$

The curve starts, therefore, at age 16·02939.

$$y_0 = 890{\cdot}05.$$

The rates resulting from this graduation are given in the table, and while they tend to show that the distribution of rates of marriage is closely allied to a frequency-curve, they do not give a satisfactory graduation, and the

Marriage Rates of Spinsters.

Age x	Exposed to Risk E	No. of Marriages M_x	Rate of Marriage m_x	Rate of Marriage Graduated by Frequency-Curve.	Hypothetical Exposed E'_x	No. of Marriages M'_x	Rate of Marriage Graduated.
15	3,658	3	·0008	...	3,695	3	...
16	3,603	8	·0022	·0027	3,433	7	·0018
17	3,528·5	49	·0139	·0132	3,187	44	·0157
18	3,393·5	114	·0336	·0332	2,957	99	·0350
19	3,187	176	·0552	·0517	2,742	151	·0541
20	2,945	219	·0744	·0667	2,541	189	·0695
21	2,688·5	192	·0714	·0776	2,354	168	·0809
22	2,443	211	·0864	·0846	2,179	188	·0880
23	2,187	212	·0969	·0881	2,016	196	·0917
24	1,956	194	·0992	·0889	1,864	185	·0920
25	1,758	146	·0831	·0875	1,723	143	·0901
26	1,583·5	137	·0865	·0845	1,591	138	·0861
27	1,417	121	·0854	·0803	1,469	126	·0812
28	1,270·5	105	·0826	·0753	1,355	112	·0754
29	1,148·5	75	·0653	·0698	1,249	82	·0693
30	1,068	60	·0562	·0640	1,151	65	·0631
31	984	64	·0650	·0583	1,061	69	·0569
32	904·5	41	·0453	·0528	976	43	·0508
33	848·5	30	·0354	·0475	897	32	·0452
34	802	39	·0486	·0425	825	40	·0400
35	752	20	·0266	·0378	758	20	·0352
36	711	25	·0352	·0335	696	25	·0309
37	672·5	18	·0268	·0295	639	17	·0270
38	638	11	·0172	·0260	586	10	·0235
39	612·5	14	·0229	·0228	537	12	·0205
40	586·5	15	·0256	·0199	492	12	·0176
41	568·5	9	·0158	·0173	451	7	·0151
42	541·5	6	·0111	·0151	412	5	·0130
43	515	8	·0155	·0131	376	6	·0112
44	491·5	2	·0041	·0113	345	1	·0096
45	476	5	·0105	·0097	315	3	·0082
46	454	5	·0110	·0084	288	3	·0070
47	440·5	2	·0045	·0072	262	1	·0060
48	416	5	·0120	·0062	239	3	·0051
49	395	2	·0051	·0054	218	1	·0044
50	378·5	...	...	·0046	199	...	·0037
51	363·5	...	...	·0039	181	...	·0031
52	348·5	1	·0029	·0034	165	...	·0026
53	335·5	2	·0089	·0029	150	1	·0022
54	317·5	...	...	·0024	139	...	·0019
55	304	...	...	·0020	124	...	·0016
56	291	3	·0103	·0018	...	1	...
57	278·5	1	·0036	·0015	...	...	...
58	261	...	...	·0013	...	...	...
59	248·5	...	...	·0011	...	...	...
60	234·5	...	...	·0009	75·7	...	·0007
61	219·5	1	·0046	·0007	...	...	...
62	209·5	...	...	·0006	...	...	...
63	201·5	...	...	·0005	...	...	...
64	191	...	...	·0004	...	...	...
65	177	...	...	·0004	45·7	...	·0003
66	165·5	...	...	·0003	...	...	...
67	154	...	...	·0002	...	...	...
68	147·5	...	...	·0002	...	...	...
69	135·5	...	...	·0001	...	...	...
70	124·5	...	...	·0001	27·2	...	·0001
71	112·5	1	·0089	...	...	...	...
72	105·5	...	...	...	...	...	...
73	95	...	...	...	...	...	...
74	84·5	...	...	...	...	...	...
75	79	...	...	...	...	...	...

failure is due almost entirely to the objections referred to above. Of course, if we were examining the algebraic form taken by rates of marriage, we should begin by work on population data where the roughness of material is avoided by the large numbers of individuals dealt with; as, however, we are seeking for a graduation, we must see how these objections, which of course apply to some extent to any method of graduation, can be overcome. It has been remarked that the cause of the difficulty is that incorrect weights are given to the items used, and the most obvious suggestion is that the actual exposed and marriages should be graduated separately. This, however, entails a large amount of additional work, and a shorter method can be used which avoids the double graduation. This method consists of using a series allied to the exposed, and treating it as a hypothetical exposed to risk from which a new series of marriages can be calculated. The advantages are that we have only to make one graduation, and the weights of the various parts of the table are given approximately. In a similar way q_x can be graduated, and in this connection it may be remarked that as the exposed to risk is generally capable of being represented by a frequency-curve, it is natural to suggest that the hypothetical exposed might be taken as the simplest form assumed by such curves (viz., Type VII.); this is also convenient because the ordinates for such curves have been tabulated.

10. The hypothetical exposed can be fixed by trial or from the values of the exposed. The column E'_x in the table given above is taken from Sheppard's Tables of the Probability Integral, x being taken as 3·06, 3·084, 3·108, 3·132, &c., and the entries were multiplied by 10^6. $M'_x = E'_x \times m_x$ was then formed and graduated. The following values were obtained for the M'_x series:—

$$\text{Mean} = 24{\cdot}85779$$
$$\mu_2 = 29{\cdot}5006$$
$$\mu_3 = 190{\cdot}112$$
$$\mu_4 = 4{,}361{\cdot}12$$
$$\beta_1 = 1{\cdot}40775$$
$$\beta_2 = 5{\cdot}01114$$
$$\kappa = -\ 7{\cdot}102$$

As this is large, Type III. was used, and

$$\gamma = \cdot 310350$$
$$\rho = 1\cdot 841405$$
$$a = 5\cdot 933325$$
$$y_0 = 192\cdot 625$$
$$\text{Mode} = 21\cdot 63562$$

The curve was then worked out and the rates of marriage in the final column were obtained by dividing M′ by E′. They agree closely with the ungraduated figures.

11. A numerical example of the application of the method to the $O^{MX(5)}$ Table may now be given. The normal curve with $\sigma = 10$ and origin at age $52\frac{1}{2}$ was used, and the values were multiplied by q_x with the help of Crelle's tables.

A part of the work was

$q \times E \times 10^5$	Age.	Ordinate from Sheppard's Tables $= E$	Age.	$q \times E \times 10^5$
810	52	·3984439	53	801
597	51	·3944793	54	850
644	50	·3866681	55	875

&c.

Summing these entries ($q \times E \times 10^5$) in fives, I formed the following:—

Age.	$q \times E \times 10^5$
20	13
25	70
30	218
35	594
40	1,394
45	2,460
50	3,702
55	4,519
60	4,385
65	3,602
70	2,249
75	1,197
80	461
85	133
90	31
95	5
100	1
	25,034

The abbreviations (use of Crelle's tables and grouping) were adopted to save labour, and as the figures were required for an example they are sufficiently accurate.

The following values were then found :—

$$\text{Mean Age} = 59{\cdot}439762$$
$$\mu_2 = 4{\cdot}584327$$
$$\mu_3 = -{\cdot}4999871$$
$$\mu_4 = 61{\cdot}17014$$
$$\text{Type of curve.—No. I.}$$
$$m_1 = 32{\cdot}81166$$
$$m_2 = 26{\cdot}57123$$
$$a_1 = 18{\cdot}78553$$
$$a_2 = 15{\cdot}21272$$
$$y_0 = 4609{\cdot}884$$
$$\text{Mode age} = 59{\cdot}730789$$

(The unit is 5 years of age.)

The ordinates were then calculated for every fifth age, and finding that the curve is not very far removed from the normal curve of error, I interpolated in the second differences of the logarithms of the ordinates for those at the other ages.* A quadrature formula was used for finding areas, and q_x was found by dividing by the hypothetical figure already used for the exposed.

The expected deaths were as follows :—

Group.	Graduated q_x for Central Age of group.	Actual.	Expected.	Deviation.	
				+	−
15–19	...	...	1·5	1·5	...
20–	·00643	9	8·9	...	·1
25–	·00731	69	61·0	...	8·0
30–	·00850	205	201·6	...	·4
35–	·00991	369	380·7	11·7	...
40–	·01179	588	575·6	...	12·4
45–	·01452	801	811·4	10·4	...
50–	·01866	1,064	1063·8	...	·2
55–	·02505	1,399	1386·6	...	12·4
60–	·03516	1,752	1773·2	21·2	...
65–	·05118	2,164	2136·7	...	27·3
70–	·07682	2,216	2261·2	45·2	...
75–	·11648	1,965	1925·8	...	39·2
80–	·17462	1,237	1241·9	4·9	...
85–	·24870	494	514·4	20·4	...
90–	·33286	129	126·0	...	3·0
95–	·43289	18	17·3	...	·7
100–	...	1	1·5	·5	...
		14,480	14492·1	115·8	103·7
				$\underbrace{\qquad}_{219{\cdot}5}$	

* As $e^{-(x-h)^2/2\sigma^2}$ is the equation to normal curve, the logarithm is $Ax^2 + Bx + C$, say. The criterion of course shows if the curve is nearly normal.

12. It will be interesting to examine a particular case of the method just described, as it is often required by actuaries.

Defining Makeham's hypothesis as colog $p_x = A + Bc^x$, we take a normal curve $(y_0 e^{-(x-h)^2/2\sigma^2})$ to represent the exposed and multiply by the values of colog p_x. This means that we assume that the products can be represented by

$$
\begin{aligned}
y &= (A + Bc^x)\, y_0 e^{-(x-h)^2/2\sigma^2} \\
&= A y_0 e^{-(x-h)^2/2\sigma^2} + B y_0 e^{-(x^2-2hx+h^2-2\sigma^2 x\log_e c)/2\sigma^2} \\
&= A y_0 e^{-(x-h)^2/2\sigma^2} + H B y_0 e^{-(x^2-2[h+\sigma^2\log_e c]x+[h+\sigma^2\log_e c]^2)/2\sigma^2}
\end{aligned}
$$

where $H = e^{(h^2+2\sigma^2 h\log_e c+\sigma^4(\log_e c)^2-h^2)/2\sigma^2} = e^{h\log_e c+\frac{\sigma^2}{2}(\log_e c)^2}$

$$\therefore \qquad y = A y_0 e^{-(x-h)^2/2\sigma^2} + H B y_0 e^{-(x-t)^2/2\sigma^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{I.}$$

i.e., the sum of two normal curves both having the same standard deviation as the exposed curve and one having the same origin.

The difference between the two origins gives $\sigma^2\log_e c$, so $\log_{10} c = \dfrac{t-h}{\sigma^2}\log_{10} e$.

The whole solution is made very simple by taking moments about the known origin (age h), for $\int_{-\infty}^{+\infty} xy\,dx$ and $\int_{-\infty}^{+\infty} x^2 y\,dx$ (the first two moments) give $(t-h)N_2$* and $N_1\sigma^2 + N_2(\sigma^2 + (t-h)^2)$† where $N_1 = Ay_0$ and $N_2 = HBy_0$.

Dividing the values just given by $N_1 + N_2$ (the total frequency), we obtain, as the first moment, about the known origin $\dfrac{(t-h)N_2}{N_1+N_2}$, and, as the second,

$$
\begin{aligned}
&\frac{N_1\sigma^2 + N_2\sigma^2 + N_2(t-h)^2}{N_1+N_2} \\
&= \sigma^2 + (t-h)\mu'_1
\end{aligned}
$$

or
$$t - h = \frac{\mu'_2 - \sigma^2}{\mu'_1}$$

and
$$N_2 = \frac{\mu'_1(N_1+N_2)}{t-h}$$

where μ' is written for moments about h

* Remember that the normal curve is symmetrical, so that the odd moments about the mean of such a curve are zero.

† Can be seen at once as the sum of two integrals; $N_1\sigma^2$ gives the second moment of the first normal curve in I, and $N_2(\sigma^2 + (t-h)^2)$ gives the second moment of the second normal curve.

$$\log_{10}c = \frac{t-h}{\sigma^2}\log_{10}e \quad . \quad . \quad . \quad . \quad . \quad . \quad . \qquad \text{II.}$$

as stated above, and if $y = \frac{10^k}{\sigma\sqrt{2\pi}}$ as is, of course, generally convenient, then,

$$A = N_1 \div 10^k$$

$$\text{and } B = N_2 \div (10^k \times H)$$

$$= \frac{N_2}{10^k} \cdot \frac{1}{e^{h\log_e c + \frac{\sigma^2}{2}(\log_e c)^2}}$$

$$= \frac{N_2}{10^k c^h e^{\frac{(t-h)}{2}\log_e c}} \qquad \textit{(see equation II.)}$$

$$= \frac{N_2}{10^k c^{\frac{t+h}{2}}}$$

13. If we assume, as Mr. Hardy does in his recent graduations of the new experience, that $\log_{10}c$ is known, we only require to calculate one moment which gives us $\frac{(t-h)N_2}{N_1+N_2}$, and this, with the help of equation II., enables us to complete the solution. If c were obtained for the aggregate table we should use this result for the select tables.

14. A numerical example with the O^{NM} Table may be of interest. A normal curve with standard duration 10 and origin $55\frac{1}{2}$ was taken, and the terms multiplied by colog p_x. These were then grouped in fives, and the first two moments calculated about age $55\frac{1}{2}$. One little point should be borne in mind in connection with the grouping; though the centre of the base on which the product $q_x \times$ exposed stands is $x+\frac{1}{2}$, the result colog $p_x \times$ exposed is an ordinate at x; the centre point of five ages 20 to 24 is $22\frac{1}{2}$ when q_x is used and 22 when colog p_x is used.

The figures were

$$N_1 + N_2 = 136387.$$

1st moment about $55\frac{1}{2}$ in 5-years group $= 1{\cdot}416184$.

2nd „ „ „ $= 4{\cdot}1929354$

Deducting (W. F. Sheppard's adjustment) $\frac{1}{12}$ from the second moment and multiplying the first moment and the adjusted second moment by 5 and 25 respectively to make the unit one year instead of 5 years, we have

$$\mu'_1 = 7{\cdot}080920$$
$$\mu'_2 = 164{\cdot}384085$$

then

$$\log(t-h) = {\cdot}9586889$$
$$t-h = 9{\cdot}092617$$
$$\log_{10} c = {\cdot}03948873$$
$$A = {\cdot}00301749$$
$$B = {\cdot}00004518782$$
$$\log_{10} B = \bar{5}{\cdot}6550214$$

q_x was then calculated from the graduated colog p_x obtained from the values of A, B and c, and the following table of expected deaths was worked out. The values of q_x are given in the table showing the frequency-curve graduation :—

Age Group.	Graduated q_x for Central Age of Group.	Expected Deaths.	Actual Deaths.	Deviation. +	Deviation. −
Under 25	...	13·0	9	4·0	...
25–	·00812	67·0	69	...	2·0
30–	·00882	211·6	205	6·6	...
35–	·00991	380·8	369	11·8	...
40–	·01162	566·9	588	...	21·1
45–	·01431	799·7	801	...	1·3
50–	·01854	1057·5	1,064	...	6·5
55–	·02517	1392·7	1,399	...	6·3
60–	·03551	1790·2	1,752	38·2	...
65–	·05160	2153·0	2,164	...	11·0
70–	·07639	2249·3	2,216	33·3	...
75–	·11415	1888·7	1,965	...	76·3
80–	·17053	1213·6	1,237	...	23·4
85–	·23352	519·1	494	25·1	...
90–	·36481	136·6	129	7·6	...
95–	...	20·6	19	1·6	...
...	...	14160·3	14,480	128·2	147·9
				276·1	

This result, which would have been improved by using all the terms instead of grouping, is very like that given by Mr. G. F. Hardy, but avoids having to obtain c by trial. Mr. Hardy's expected and actual deaths balance better than the above, but I do not think the rates have been understated systematically, as the 75–79 group accounts for the disagreement. The total deviation is less than Mr. Hardy's.

15. When dealing with the adjustments for moments it was remarked that it is best to use unadjusted moments except when there is high contact at each end of the curve. In some cases the unadjusted moments are rather far from the truth, and consequently the curve obtained is by no means the best that can be found. This most frequently happens when the curve rises very abruptly, as in our example for Type I., or when it takes the form of the example for Type III., the reason being that, in such cases, the assumption that an area is concentrated at the middle of a base of unit length involves a considerable error. In the Type I. example, for instance, the first group was assumed to be an ordinate at 17, whereas the curve starts at age 16·76, and the central point ought, therefore, to be later. The results can sometimes be improved considerably by basing a second graduation on the first and, in the particular case just referred to, since the first group is too large, we might assume that the curve should start at 17·5. It is unnecessary to find as many as four moments, for by assuming that the start of the curve and the range (b) are known, the equations on p. 59, giving the moments about the start of the curve, afford a very simple solution. We have

$$\mu'_1 = \frac{b(m_1+1)}{m_1+m_2+2} \text{ and } \mu'_2 = \frac{b^2(m_1+1)(m_1+2)}{(m_1+m_2+2)(m_1+m_2+3)},$$

and writing
$$\gamma_1 = \frac{\mu'_1}{b} \text{ and } \gamma_2 = \frac{\mu'_2}{\mu'_1 b},$$

we have
$$m_1+1 = \frac{\gamma_1(\gamma_2-1)}{\gamma_1-\gamma_2}$$

and
$$m_2+1 = \frac{(\gamma_2-1)(1-\gamma_1)}{\gamma_1-\gamma_2}$$

where μ' is written for a moment about the start of the curve.

16. If the range of a curve can be fixed by general considerations, a good deal of labour can thus be saved, while, if the start of the curve is known, the following solution depending on three moments is of use.

Writing
$$\lambda_2 = \frac{\mu'_2}{\mu'^2_1} \text{ and } \lambda_3 = \frac{\mu'_3}{\mu'_2\mu'_1}$$

the values of the constants in the equation to the curve are given by

$$m_1+1=\frac{2(\lambda_2-\lambda_3)}{2\lambda_3-\lambda_2-\lambda_2\lambda_3}$$

$$m_2+1=\frac{2(\lambda_2-\lambda_3)(\lambda_3-1)(1-\lambda_2)}{(2\lambda_3-\lambda_2-\lambda_2\lambda_3)(1+\lambda_3-2\lambda_2)}$$

$$b=\mu'_1\frac{m_1+m_2+2}{m_1+1}$$

and $$a_1:a_2=m_1:m_2$$

17. Returning to the example of Type I., and considering the line for age 22 in the table on p. 54, we see that 4·175 and 14·634 give S_2 and S_3, excluding the first group, and the moments about age 17 are then found to be 4·175 and 29·268; transferring to 17·5, we have 4·075 and 24·268; adding the moments for the first group, $\cdot 034\times\frac{1}{5}$ and $\cdot 034\times(\frac{1}{5})^2$ respectively, $\mu'_1=4\cdot 0818$ and $\mu'_2=24\cdot 26936$. Assuming a range of 15·5, and using the formulæ given above,

$$\begin{aligned} m_1&=\ \ \cdot 3498\\ m_2&=\ 2\cdot 7758\\ a_1&=\ 1\cdot 735\\ a_2&=13\cdot 765\\ y_0&=154\cdot 2 \end{aligned}$$

and the mode is $17\cdot 5+1\cdot 735\times 5=26\cdot 175$.

From these values the graduated figures for the first four groups are 37, 140, 152, 143, which is an improvement on the figures obtained previously. This example is used for convenience, but with regard to adjustments in the particular case the remarks in the footnote on p. 55 should be borne in mind.

18. As a second case, the example for Type III. may be considered. Assuming that the value of $p(-\cdot 0783584)$ is not to be altered, then the first moment about the start of the curve (*see* p. 68) $=\dfrac{p+1}{\gamma}=\dfrac{\cdot 9216416}{\gamma}$. Assuming the curve to start at ·8 the first moment about the point is calculated as follows :—

$$
\begin{array}{rcl}
44 \times \cdot 1^{*} & = & 4\cdot 4 \\
135 \times \cdot 7 & = & 94\cdot 5 \\
45 \times 1\cdot 7 & = & 76\cdot 5 \\
12 \times 2\cdot 7 & = & 32\cdot 4 \\
8 \times 3\cdot 7 & = & 29\cdot 6 \\
3 \times 4\cdot 7 & = & 14\cdot 1 \\
1 \times 5\cdot 7 & = & 5\cdot 7 \\
3 \times 6\cdot 7 & = & 20\cdot 1 \\
\hline
251 & & 277\cdot 3
\end{array}
$$

The first moment is therefore $\dfrac{277\cdot 3}{251} = 1\cdot 0875$, and hence $\gamma = \cdot 84752$. The value of y_0 is 205·0, and the graduation is 47, 123, 48, 19, 8, 3, 2, 1. The equation to the curve is $y = 205\cdot 0 x^{-\cdot 07186} e^{-\cdot 84752x}$ with the origin at the start of the curve, so that in this case the y_0 was calculated by the formula $y_0 = \dfrac{N\gamma^{p+1}}{\Gamma(p+1)}$ which the reader will have no difficulty in reproducing for himself. When moments are calculated about the start of the curve and the form taken is that of the present curve, it is convenient to use the equation in this form rather than in that given on p. 65. It may be mentioned that as the value of p is nearly zero in the particular example, a good result would be obtained by assuming that value, or, which is the same thing, by putting $y = y_0 e^{-\gamma x}$; γ now becomes $1\cdot 0875^{-1} = \cdot 91958$, and $y_0 = 251 \times \cdot 91958 = 230\cdot 8$, and the graduation is 43, 125, 50, 20, 8, 3, 1, 1.

19. It sometimes happens that the error involved in the calculations of the moment tends to balance that resulting from the curve not starting at the beginning of the unit base assumed for the first group. An instance of this is the first example of Table I. for which the mean is at duration 5·182, and the moments and constants are—

$$
\begin{array}{llcll}
\mu_2 = & 17\cdot 63688 & \qquad & \beta_1 = & 3\cdot 34846 \\
\mu_3 = & 135\cdot 5361 & & \beta_2 = & 6\cdot 18392 \\
\mu_4 = & 1923\cdot 565 & & \kappa = & -1\cdot 307
\end{array}
$$

* Strictly should be rather earlier, because the ordinates are decreasing very rapidly.

so that the curve will be of Type I. and equation to it is

$$y = \cdot 89082 x^{-\cdot 629685}(25\cdot 49729 - x)^{1\cdot 624275}$$

where the origin is at 1·02897 where the curve starts.

The graduation by this curve is shown in the following table—

Duration.	Withdrawals.	Graduated by Type I. curve.
1	308	312
2	200	198
3	118	101
4	69	76
5	59	58
6	44	45
7	29	37
8	28	30
9	26	25
10	21	21
11	18	18
12	18	15
13	12	13
14	11	11
15	5	9
16	11	7
17	7	6
18	6	5
19	1	4
20	3	3
21	1	2
22	3	2
23	2	1
24	...	1
	1,000	1,000

20. The calculation of the graduated area of the first group may present a difficulty, as a quadrature formula cannot be applied, and the following method gives the best way of obtaining a correct value—

$$\int_0^x y_0 x^{m_1}(b-x)^{m_2}dx = \int_0^x y_0 x^{m_1}\left(b^{m_2} - m_2 b^{m_2-1}x + \frac{m_2(m_2-1)}{2}b^{m_2-2}x^2 - \ldots\right)dx$$

$$= y_0 x^{m_1+1} b^{m_2}\left(\frac{1}{m_1+1} - \frac{m_2 x}{b(m_1+2)} + \ldots\right)$$

which is a rapidly convergent series when x is small. In the last example where x is $1\cdot 5 - 1\cdot 02897 = \cdot 47103$ the second term barely affects the result. y_0 must, of course, be calculated by the formula

$$\frac{N}{b^{m_1+m_2+1}} \cdot \frac{\Gamma(r)}{\Gamma(m_1+1)\Gamma(m_2+1)}$$

which is an analogous form to that given for similar Type III. curves in Art. 18.

21. The expression for finding the area of the first group in Type III. curves is

$$\int_0^x y_0 e^{-\gamma x} x^p dx = y_0 x^{p+1}\left(\frac{1}{p+1} - \frac{\gamma x}{p+2} + \ldots\right)$$

PART II.

CHAPTER VI.

Correlation.

1. Two measurable characteristics, A and B, are said to be correlated, when, with different values, x of A, we do not find the same value, y of B, equally likely to be associated. In other words, certain values of B are relatively more likely to occur with the value x than others.

2. In practice, as one characteristic increases, the other generally either steadily increases or decreases, and it is exceptional to find that while one increases steadily the other increases for a time and then decreases. Put in a rough-and-ready way, the definition can in particular cases, with which actuaries are familiar, be stated: "The mean ages at maturity in Endowment Assurances increase with the unexpired term when the policies are grouped according to the unexpired term"; or, "the older a bachelor, the less likely is he to marry and have children." There is correlation between ages at maturity and unexpired term, and between the age of a bachelor and the number of children, and it is required to find a method of measuring the amount of correlation statistically. The easiest way to appreciate the nature of the problem is with the help of a table of double entry, such as the following, which gives particulars of 2,870 endowment assurances grouped according to their unexpired term. A little examination of the table shows that there is a connection between the two functions, but does not give any measure of the correlation suitable for comparison with

the experiences of other offices or with that of the same office at a later date :—

Unexpired term of Endowment Assurances.	Central Age at Maturity.										Total.	Mean Maturity Age for the row.
	30	35	40	45	50	55	60	65	70	75		
0–4	2	...	...	2	26	6	14	6	...	...	56	53·75
	24	20	16	12	8	4	0	4				
5–9	1	1	2	6	62	36	40	22	2	...	172	55·03
	18	15	12	9	6	3	0	3	6			
10–14	...	2	9	17	117	99	127	52	8	1	432	55·85
		10	8	6	4	2	0	2	1	6		
15–19	3	...	6	24	145	155	237	84	11	...	665	56·59
	6	5	1	3	2	1	0	1	2	3		
20–24	...	1	...	3	133	167	271	78	20	1	674	57·58
	0	0	0	0	0	0	0	0	0	0		
25–29	...	...	...	9	90	123	231	71	11	3	538	57·88
				3	2	1	0	1	2	3		
30–34	...	...	...	1	11	49	127	49	8	2	247	59·91
				6	4	2	0	2	4	6		
35–39	...	...	...	...	...	6	49	22	...	...	77	61·04
						3	0	3				
40–44	...	...	...	...	...	2	2	3	...	1	8	62·50
						4	0	4	8	12		
45–49	...	...	...		...	...	...	1	...	...	1	65·00
							0	5				
Total	6	4	17	62	584	613	1,098	388	60	8	2,870	

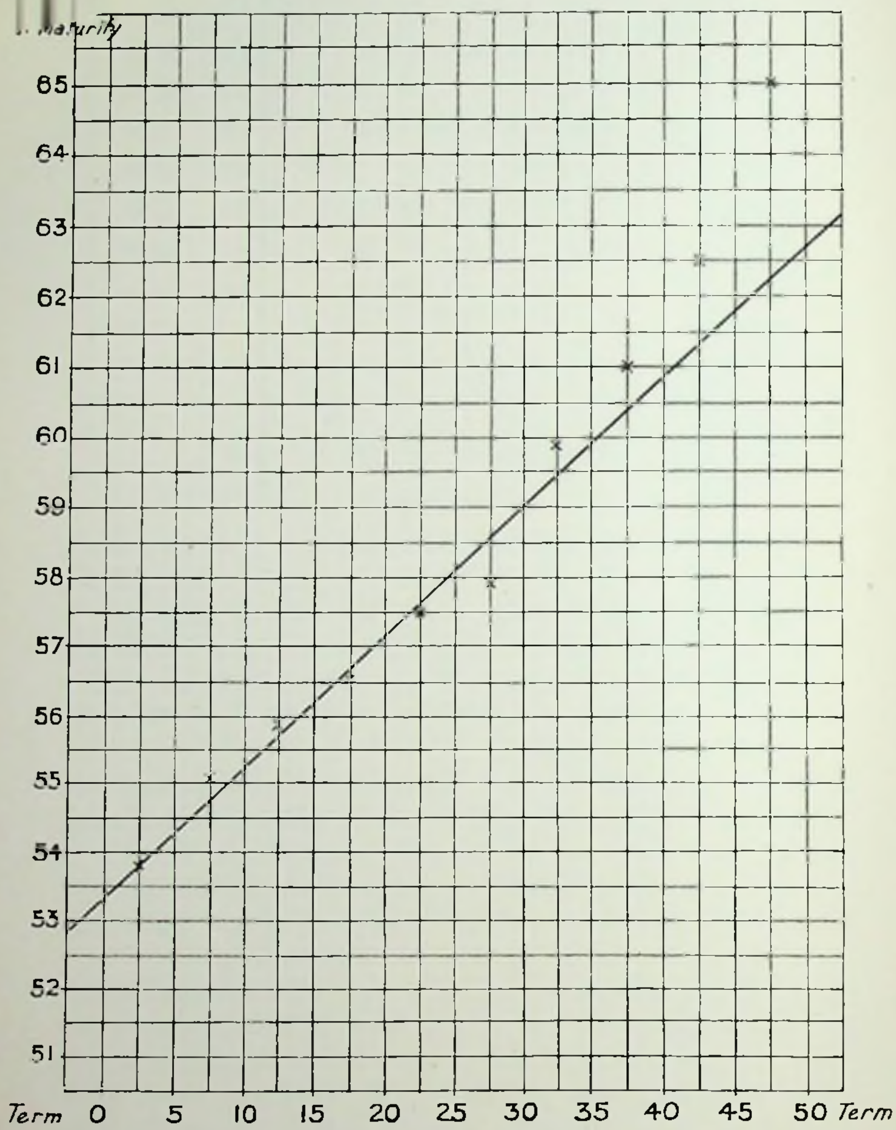
Note.—For explanation of small numbers, see Art. 16.

The above table is called a correlation or frequency table, and a column or row in it is called an array. The middle value of the variable with which the row is associated is called its type, so that the third column (*i.e.*, that headed 40) would be called the y-array of type 40, and the fourth row would be called the x-array of type 17·5, because 17·5 is the middle of the 15–19 group.

3. Now, returning to our definition of correlation, and examining the last column of the table, which gives the mean age in each row, we see that the numbers in it tend to increase as we go down the column, and the age at maturity is therefore correlated with the unexpired term. The figure on p. 108 shows the series of mean values clearly.

4. A little consideration will lead to further information, for if we imagine a case in which there is no correlation, the series of the means of the rows will be independent of the other function; that is, they will run horizontally when plotted out as a diagram. Another point to be noted is that there are two kinds of correlation (positive and negative), for the functions may increase together, or one may increase

FIGURE


Maturity
65
64
63
62
61
60
59
58
57
56
55
54
53
52
51
Term 0 5 10 15 20 25 30 35 40 45 50 Term

and the other decrease; in the case of the endowment assurances the correlation is positive, because as the term increases the mean maturity age also increases.

5. These introductory remarks will give an indication of the nature of the problem to be solved, and may help to render the following proof easier to follow. It should be remembered that the proof deals with a function of n variables.* The table given above has only two variables, but it is easy to see how more variables may be introduced in similar tables; for instance, endowment assurances by limited payments give three variables (term of premium, term of assurance and age of life), while an increase in the number of lives—say, joint-life endowment assurances by limited payments—gives four variables. We may now consider what equation will represent the numbers in the body of the correlation table and how this equation can be utilized to express the relationship between the two functions with which the correlation table is concerned.

[6.] Let $\eta_1, \eta_2, \eta_3 \ldots \eta_n$ be deviations from their respective means of a complex of measurable characteristics. The sizes of the functions measured, or organs, are determined by a large number of *independent* contributory causes. Let there be m of these causes, and let their deviations from their means be $\epsilon_1, \epsilon_2, \epsilon_3 \ldots \epsilon_m$, then $\eta_1, \eta_2, \eta_3 \ldots \eta_n$ will be functions of $\epsilon_1, \epsilon_2, \epsilon_3 \ldots \epsilon_m$. Further, if $m > n$ certain of the ϵ's will appear only in certain of the η's, and the ϵ's will not be fully determined for a given η complex. We also assume that the variations in intensity of the contributory causes are small as compared with their absolute intensity, and that these variations follow the normal law of distribution; that is, we assume that the deviations from the mean value can be graduated by the normal curve of error (Type VII.). The mean complex being reached with the mean intensities of contributory causes, we have, by the principle of the super position of small quantities,

$$\left.\begin{array}{l} \eta_1 = a_{11}\epsilon_1 + a_{12}\epsilon_2 + a_{13}\epsilon_3 + \ldots + a_{1m}\epsilon_m \\ \eta_2 = a_{21}\epsilon_1 + a_{22}\epsilon_2 + a_{23}\epsilon_3 + \ldots + a_{2m}\epsilon_m \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ \eta_n = a_{n1}\epsilon_1 + a_{n2}\epsilon_2 + a_{n3}\epsilon_3 + \ldots + a_{nm}\epsilon_m \end{array}\right\} \quad . \quad . \quad . \quad . \quad (i.)$$

* A student reading the subject for the first time would do well to omit the paragraphs indicated by brackets. After the statistical idea underlying correlation has been understood, it will be found easier to follow the theoretical work.

The a's are coefficients whose values have to be determined, and any of the system of a's may be zero, for a particular contributory cause may have no effect on a particular result. Further, the chance that we have a conjunction of contributory causes lying between ϵ_1 and $\epsilon_1+\delta\epsilon_1$, ϵ_2 and $\epsilon_2+\delta\epsilon_2$ . . . and between ϵ_m and $\epsilon_m+\delta\epsilon_m$ will be given by

$$P = Ce^{-\left(\frac{\epsilon_1^2}{2\kappa_1^2}+\frac{\epsilon_2^2}{2\kappa_2^2}+\ldots+\frac{\epsilon_m^2}{2\kappa_m^2}\right)} \times \delta\epsilon_1\delta\epsilon_2 \ldots \delta\epsilon_m \quad . \quad . \quad . \quad \text{(ii.)}$$

where the standard deviations of the distributions are $\kappa_1\kappa_2 \ldots \kappa_m$ and C is constant.*

Now, by (i.) let n of the variables ϵ, say the first n, be replaced by the variables η, then the probability that we have a complex with organs lying between η_1 and $\eta_1+\delta\eta_1$, η_2 and $\eta_2+\delta\eta_2$ . . . η_n and $\eta_n+\delta\eta_n$, together with a series of contributory causes lying between ϵ_{n+1} and $\epsilon_{n+1}+\delta\epsilon_{n+1}$, ϵ_{n+2} and $\delta\epsilon_{n+2}$ . . . ϵ_m and $\epsilon_m+\delta\epsilon_m$, will be

$$P' = C'e^{-\frac{1}{2}\phi^2}\delta\eta_1\delta\eta_2 \ldots \delta\eta_n\delta\epsilon_{n+1}\delta\epsilon_{n+2} \ldots \delta\epsilon_m$$

where C′ is a constant, a function of C and the a's, and ϕ^2 consists of the following parts:—

(i.) A quadratic function of the η's from η_1 to η_n.

(ii.) A quadratic function of the ϵ's from ϵ_{n+1} to ϵ_m.

(iii.) A series of functions of the type

$$\epsilon_{n+1}(b_{1,\,n+1}\eta_1 + b_{2,\,n+1}\eta_2 + \ldots + b_{n,\,n+1}\eta_n)$$
$$\epsilon_{n+2}(b_{1,\,n+2}\eta_1 + b_{2,\,n+1}\eta_2 + \ldots + b_{n,\,n+1}\eta_n)$$
$$. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad .$$
$$\epsilon_m(b_{1,\,m}\eta_1 + b_{2,\,m}\eta_2 + \ldots + b_{n,\,m}\eta_n)$$

where some of the b's may be zero.

Now, if P′ be integrated for all values from $-\infty$ to $+\infty$ of all the contributory causes ϵ_{n+1}, ϵ_{n+2} . . . ϵ_m, we shall have the whole chance of a complex with organs falling between η_1 and $\eta_1+\delta\eta_1$, η_2 and $\eta_2+\delta\eta_2$ . . . η_n and $\eta_n+\delta\eta_n$. But every time we integrate with regard to an ϵ, say ϵ_{n+1}, we alter the constants of each contributory part of ϕ^2, but do not alter the

* Consider any particular case of the normal curves, the chance of getting a result between ϵ_1 and $\epsilon_1+\delta\epsilon_1$ when the distribution is Type VII. is $y_0e^{-\epsilon_1^2/2\kappa_1^2}\delta\epsilon_1$ where κ_1 is the standard deviation; similarly with each of the other causes. As the causes are independent, the product of the various chances gives the required chance.

triple constitution of ϕ^2 except to cause one ϵ to disappear from its (ii.) and (iii.) constituents. At the same time we alter C′ without introducing into it any terms in η. Thus, finally, after $m-n$ integrations, ϕ^2 is reduced to its first constituent, or we conclude that the chance of a complex of organs between η_1 and $\eta_1+\delta\eta_1$, η_2, and $\eta_2+\delta\eta_2 \ldots \eta_n$ and $\eta_n+\delta\eta_n$ occurring is given by

$$P = Ce^{-\frac{1}{2}\chi^2}\delta\eta_1, \delta\eta_2, \delta\eta_3, \ldots \delta\eta_n \quad \ldots \ldots \ldots \quad \text{(iii.)}$$

where χ^2 is a quadratic function of the η's. This is the law of frequency for the complex.

Consider the expression (iii.) but replace χ^2 by a quadratic function, then

$$P = Ce^{-\frac{1}{2}\{c_{11}\eta_1^2+c_{22}\eta_2^2+\ldots+2c_{12}\eta_1\eta_2+2c_{13}\eta_1\eta_3+\ldots\}}$$

$$= Ce^{-\frac{1}{2}\{S_1(c_{pp}\eta_p^2)+2S_2(c_{pq}\eta_p\eta_q)\}}.$$

Here C, c_{pp}, c_{pq} are constants, and S_1 denotes a summation for every value of p, and S_2 for every pair of value of p and q in the series.

Taking the simplest case, when there are two variables, this becomes

$$P = Ce^{-\frac{1}{2}(c_1\eta_1^2+c_2\eta_2^2+2c_{12}\eta_1\eta_2)}.*$$

Integrate P for all values of η_1 from $-\infty$ to $+\infty$, and we must have the normal curve of η_2 variation.

$$\therefore \qquad \frac{1}{2\sigma_1^2} = c_2\left(1-\frac{c_{12}^2}{c_1c_2}\right).$$

Similarly integrating for all values of η_2,

$$\frac{1}{2\sigma_2^2} = c_1\left(1-\frac{c_{12}^2}{c_1c_2}\right).$$

Integrating for all values of η_1 and η_2 to obtain the total frequency, we have—

$$N = \frac{C\pi}{\sqrt{c_1c_2-(c_{12})^2}}.$$

* In Appendix III., some integrals connected with the normal curve of error are dealt with. The result $\frac{1}{2\sigma_1^2}=c_2\left(1-\frac{c_{12}^2}{c_1c_2}\right)$ can be reached at once by rearranging the index in the expression for P as a perfect square $-\frac{1}{2}c_2\left(1-\frac{c_{12}^2}{c_1c_2}\right)\eta_2^2$ as is done in No. (v.) of Appendix III.

Now put $r = -\frac{c_{12}}{c_1 c_2}$ and write x and y for η_1 and η_2, and we have—

$$z = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{-\frac{1}{2}\left\{\frac{x^2}{\sigma_1^2(1-r^2)} - \frac{2xyr}{\sigma_1\sigma_2(1-r^2)} + \frac{y^2}{\sigma_2^2(1-r^2)}\right\}} \dots \text{(iv.)}$$

The equation just given is a graduation formula capable of representing tables like that on p. 107. It has been obtained on certain assumptions which may not all be realised in practice, but it has a far larger scope in practical work than the analogous normal curve of error has in frequency-curve operations.

[7.] Since $e^{-\frac{x^2}{2\sigma_1^2}}$ represents one series of variations and $e^{-\frac{y^2}{2\sigma_2^2}}$ the other, it follows that if there were no correlation at all the frequency of any particular result, x, y would be the product of the two chances, *i.e.*, proportional to $e^{-\left(\frac{x^2}{2\sigma_1^2}+\frac{y^2}{2\sigma_2^2}\right)}$, which means that the size of the term $\frac{2xyr}{\sigma_1\sigma_2(1-r^2)}$, or r, is a measure of the correlation. r is called *the coefficient of correlation*.

[8.] Perhaps the easiest way tosee how its value can be obtained from an actual experience is by looking at the matter from the curve-fitting point of view, and dealing with the expression for z by moments.

It will be remembered that the moments were obtained by summing for all values of the frequencies multiplied by the powers of the independent variable, but as we now have two variables we can take n powers of the one and m of the other. Thus, if we take the second powers of the x distances and the zero power of the y distances (*i.e.*, neglect y), we obtain the ordinary second moment of the frequencies reckoned only in the x direction. Similarly, with the second powers of the y distances and the zero power of the x distances. These calculations give two of the unknown constants for $\sigma = \sqrt{\mu_2}$ (*see* Type VII.). There is, however, another second-order term to be considered, namely, that obtained by taking the first powers of both the x distances and the y distances, *i.e.*, multiplying the frequencies by xy. This may be written :—

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z\,xy\,dx\,dy, \text{ where } z \text{ has the value given above.}$$

This double integral reduces to $Nr\sigma_1\sigma_2$ (*see* Appendix III.) or the (xy) moment of the total distribution $= Nr\sigma_1\sigma_2$, or $r = \frac{(xy)\text{ moment}}{N\sigma_1\sigma_2}$.

To calculate the coefficient of correlation we have, therefore, to find the x^2 moment, the y^2 moment and the xy moment about the centroid vertical.

[9.] Now we have seen that equation (iv.) gives an expression for describing a correlation table such as the table of endowment assurances on p. 107, and from that equation it is clear that the distribution of an array of any type t is

$$z = z_0 e^{-(g_1x^2 - 2htx + g_2t^2)}$$

where g_1, h and g_2 are written for the longer expressions in σ_1, σ_2, and r. If we make the index into a perfect square, we have

$$z = z_0 e^{-g_1\left\{x - \frac{ht}{g_1}\right\}^2} e^{-g_2t^2 + \frac{h^2t^2}{g_1}}$$

$$= z'_0 e^{-g_1\left\{x - \frac{ht}{g_1}\right\}^2}$$

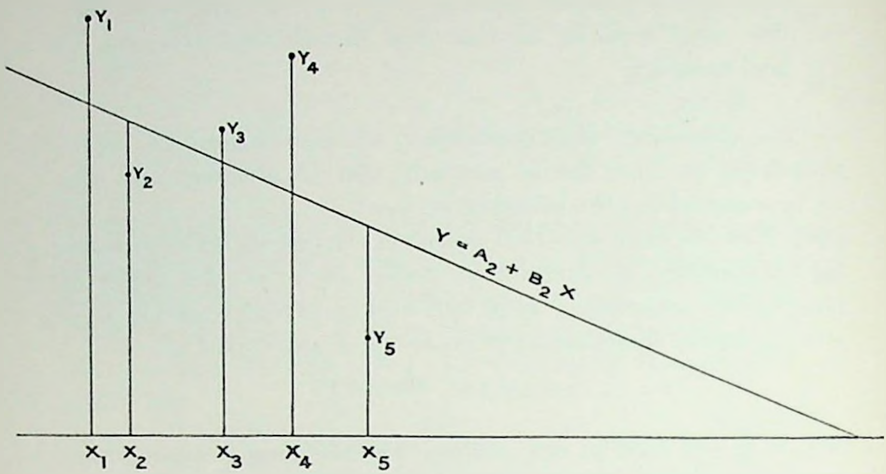
This last expression is a normal distribution having the same standard deviation as that of the whole surface; but its mean differs from that of the whole surface by $\frac{ht}{g_1}$; and it follows that

(1) The deviation of the mean of the array is directly proportional to the type; or the means of the arrays increase or decrease in arithmetical progression or lie on a straight line (called the regression line).

(2) The standard deviations of all parallel arrays are equal and independent of their types.

10. Before returning to the statistical example it will be well to consider the following proof,* which proceeds on the principle that we require to fit a straight line ($y = a_2 + b_2x$) to the correlation table.

* This proof is a modification of one given by Mr. G. U. Yule, in the *J.S.S.*, vol. lx., pp. 812, &c., and *Proc. Roy. Soc.*, 1897, vol. lx., pp. 477, &c. It has been altered to avoid the introduction of the method of least squares.



Let $x_1\,y_1$, $x_2\,y_2$, &c., be associated deviations, and let $y=a_2+b_2x$ be the straight line used in the graduation; then the graduated figure corresponding to x_1 is $a_2+b_2x_1$.

Now, if we proceed as we did in fitting frequency-curves by the method of moments, we make the graduated and ungraduated areas, means, &c., equal, or

$$(a_2+b_2x_1)+(a_2+b_2x_2)+\ .\ .\ .=y_1+y_2+\ .\ .\ .\ ;$$

or $$Na_2+b_2S'(x)=S'(y).$$

And $$(a_2+b_2x_1)x_1+(a_2+b_2x_2)x_2+\ .\ .\ .=x_1y_1+x_2y_2+\ .\ .\ .\ ;$$

or $$a_2S'(x)+b_2S'(x^2)=S'(xy),$$

where $S'(x)$ gives the first moment of the x's, $S'(y)$ the first moment of the y's, $S'(x^2)$ the second moment for the x's, and $S'(xy)$ a moment in which any frequency is multiplied by the product of the distances in the x and y directions.

If these moments are now transferred to the mean, as was done in fitting the frequency-curves, we have

$$Na_2=0,$$

or $$a_2=0;$$

and $$b_2S'(x^2)=S'(xy),$$

or $$b_2=\frac{S'(xy)}{S'(x^2)}.$$

But we have already seen that the second moment of the whole frequency (N) is $N\sigma_1^2$;

$$\therefore \qquad b_2 = \frac{S'(xy)}{N\sigma_1^2};$$

$$\therefore \qquad y = \frac{S'(xy)}{N\sigma_1^2}x.$$

If we now write $S(xy) = N\sigma_1\sigma_2 r$, we have

$$\left.\begin{aligned} y &= r\frac{\sigma_2}{\sigma_1}x \\ x &= r\frac{\sigma_1}{\sigma_2}y \end{aligned}\right\}$$

where r will represent the statistical measure of correlation (coefficient of correlation) between the x's and y's.

11. At first sight it may appear that the two equations just given, showing the relationship between x and y, are not consistent. It must, however, be remembered that the first, $y = r\frac{\sigma_2}{\sigma_1}x$, gives the mean values of y corresponding to particular values of x, while the second gives the mean values of x corresponding to particular values of y. To take a simple case as an example, assume that $\sigma_1 = \sigma_2 = 1$ and that $r = \cdot 1$, then if $x = 0$ the mean of the y's corresponding to this value of x is 0, and if $x = 20$ the mean of the y's will be 2. When we turn the matter round, however, we cannot, of course, assert that the mean of the x's corresponding to $y = 2$ is 20; it will be ·2.

12. After this preliminary remark we may return to the two equations and consider how it is that r is a measure of correlation and whether it can always be treated as a satisfactory measure. We can best see that r is a measure of correlation by rewriting the equation $y = r\frac{\sigma_2}{\sigma_1}x$ in the form $\frac{y}{\sigma_2} = r\frac{x}{\sigma_1}$ or $Y = Xr$, and we can then interpret it as giving one characteristic in terms of the other where the mean is the origin (this is due to referring moments to the mean in the proof) and the unit of measurement is the standard deviation in each case. In this form we see at once that as one characteristic (X) increases the mean (Y) of the corresponding series of the other characteristic increases to an extent which depends on the value of r, while if r is negative Y decreases. It is only if r is unity that the increments of X and Y become equal and absolute correlation is reached. If Y remains constant as the value of X increases the definition at the beginning of this chapter tells us that there is no correlation, and r in this case is zero as can easily be seen from the equation $Y=Xr$. The value of r lies between -1 and $+1$ (*see* Chap. X., Art. 4), and its sign has no influence on its numerical value. In other words a large negative value does not mean that the two characteristics do not vary together but only that increases in the one correspond with decreases in the other; the numerical value of r indicates the extent to which variations in the two characteristics correspond. This indication is satisfactory provided the means, when plotted in a diagram such as that on p. 108, fall approximately in a straight line (*i.e.*, "regression"* is linear). Distinct deviations from linearity are not so common as might be supposed, but if they are very marked in any case, r ceases to be an entirely satisfactory measure of the correlation.

13. We may take this opportunity of removing another difficulty that is sometimes met. Some students have a loubt which is best shown by the question, "How can there be perfect correlation when one thing is always smaller than another"? As an example we may take the correlation between the lengths of a man's right arm and his left arm; here the coefficient of correlation would be practically unity, and since each characteristic is measured from it own mean, and in terms of its own standard deviation, the coefficient would not be decreased if every left arm was a certain number of inches shorter than the right or if it bore a fixed relation in length, say $\frac{99}{100}$, to the right arm.

14. When dealing with moments on p. 17, we noticed that though we required to find them about the mean, it was best in practice to take them about some point fixed arbitrarily so as to avoid fractions and then adjust the results afterwards. The values of the σ_1 and σ_2 can, of course, be found with the help of the formula on p. 19, viz., $\nu_2=\nu'_2-d^2$. The

* The term "regression" was invented by Mr. Francis Galton in connection with the study of heredity; it indicates the way the children of particular parents tend to "step back" to the ordinary population mean.

deduction of $\frac{1}{12}$ from the second moment should be made for the same reason and in the same cases as in frequency-curve fitting.

With regard to the product moment we have—

$$S(x'y') = S(x+d_1)(y+d_2)$$

$$= S(xy) + d_1S(y) + d_2S(x) + Nd_1d_2;$$

or since $\quad S(x) = S(y) = 0$

$$S(xy) = S(x'y') - Nd_1d_2$$

where $S(x'y')$ is calculated about a point distant d_1 from the mean of the x's and d_2 from the mean of the y's.

15. The statistical example on p. 107 can now be worked through. It will be found to make the proofs and methods given above much easier to grasp.

A point about which moments are to be calculated is first fixed, say, the middle of the group corresponding to maturity age 60 and unexpired terms 20–24 years, and for the present the calculations are made about this point. The following table shows the calculation of the mean and second moment of the totals of the y-arrays, *i.e.*, the totals at the bottom of the table, because columns are y-arrays and rows x-arrays:—

Frequency.	x'	Frequency × x'	Frequency × $(x')^2$
6	−6	36	216
4	−5	20	100
17	−4	68	272
62	−3	186	558
584	−2	1,168	2,336
643	−1	643	643
1,098	0	−2,121	...
388	1	388	388
60	2	120	240
8	3	24	72
2,870 = N		+ 532	4,825
		−1,589	

$$d_1 = -\frac{1589}{2870} = -\cdot553659$$

Hence, the mean age $=60-2{\cdot}76830=57{\cdot}23170$, because the unit of grouping is 5 years.

$$\sigma_1^2=\frac{4825}{2870}-d_1^2-\tfrac{1}{12}\text{ (Sheppard's adjustment)}$$
$$=1{\cdot}37465-{\cdot}08\dot{3}$$
$$=1{\cdot}29132$$
$$\therefore \quad \sigma_1=1{\cdot}13637$$

Treating the rows in the same way, the following table was formed :—

Frequency.	y'	Frequency $\times y'$	Frequency $\times y'^2$
56	−4	224	896
172	−3	516	1,548
432	−2	864	1,728
665	−1	665	665
674	0	−2,269	...
538	1	538	538
247	2	494	988
77	3	231	693
8	4	32	128
1	5	5	25
2,870 = N		+1,300	7,209
		− 969	

$$d_2=-\frac{969}{2870}=-{\cdot}337631$$

Mean unexpired term $=22-1{\cdot}68815=20{\cdot}31185$

$$\sigma_2^2=\frac{7209}{2870}-d_2^2-\tfrac{1}{12}$$
$$=2{\cdot}31453$$

and $\qquad \sigma_2=1{\cdot}52135$

16. The value of $S(xy)$ is formed with the help of the numbers in very small type appearing under the frequencies in the correlation table. The frequency 62 in the 50 column, for instance, is distanced three spaces upwards and two sideways from the arbitrary origin, so the value of $x'y'$ by which it has to be multiplied is $3\times2=6$, as shown in the small type. The other figures are obtained in like manner, but the sign must be borne in mind. Any value from the left-hand upper division of the table, or in the lower right-hand division, will be positive, because the frequency will be multiplied by a product of an x and y having like signs; while any value from the other divisions will be negative, because the x and y by which the frequencies are multiplied are of opposite signs.

The calculation of the product moment is as follows :—

Frequencies.	$y'x'$	Total of frequencies (f)	$f \times x'y'$
155 + 71 − 84 − 123	1	+ 19	+ 19
145 + 99 + 11 + 49 − 11 − 52 − 49 − 90 .	2	102	204
24 + 36 + 3 + 22 − 22 − 6 − 9 . . .	3	48	144
6 + 6 + 8 + 3 − 6 − 8 − 11 − 2 + 117. .	4	113	452
1	5	1	5
3 + 17 + 62 + 2 − 1 − 1 − 2	6	80	480
9 + 26	8	35	280
6	9	6	54
2	10	2	20
2 + 2 + 1	12	5	60
1	15	1	15
1	18	1	18
2	24	2	48
			1,799

$$S(xy) = S(x'y') - Nd_1d_2$$
$$= 1799 - Nd_1d_2$$
$$= 1262{\cdot}51$$
$$r = \frac{S(xy)}{N\sigma_1\sigma_2} = \frac{1262{\cdot}51}{2870 \times 1{\cdot}13637 \times 1{\cdot}52135}$$
$$= {\cdot}25445.$$

The coefficient of correlation between age at maturity and the unexpired term of endowment assurances is ·25445.

The equation representing the one function in terms of the other is

$$x = r\frac{\sigma_1}{\sigma_2}y$$
$$= {\cdot}19007y$$

where all measurements are made from the mean and the unit is 5 years. The line drawn in the figure gives this result.

17. An alternative method similar to the summation method given in Art. 9, Chap. III. for moments can be conveniently used in connection with correlation tables.

Taking the same example, we obtain from the given table another in the same form, giving the y sum of it by summing each column continuously, and then form a third table by summing the second table across continuously.

TABLE *of the y-sum of Correlation Table.*

Unexpired term of Endowment Assurances.	Central Age at Maturity.										
	30	35	40	45	50	55	60	65	70	75	Totals.
0–4	6	4	17	62	584	643	1,098	388	60	8	2,870
5–9	4	4	17	60	558	637	1,084	382	60	8	2,814
10–14	3	3	15	54	496	601	1,014	360	58	8	2,642
15–19	3	1	6	37	379	502	917	308	50	7	2,210
20–24	0	1	0	13	234	347	680	224	39	7	1,545
25–29	0	0	0	10	101	180	409	146	19	6	871
30–34	0	0	0	1	11	57	178	75	8	3	333
35–39	0	0	0	0	0	8	51	26	0	1	86
40–44	0	0	0	0	0	2	2	4	0	1	9
45–49	0	0	0	0	0	0	0	1	0	0	1
Totals	16	13	55	237	2,363	2,977	5,463	1,914	294	49	13,381

The totals in the right-hand column of the second table give the first sum of the total in the right-hand column of the correlation table, and are the same as the column $x = 30$ in the third table. The total of the y sum, or of the first column in the xy table, gives the mean of the y's (13,381÷2,870), and similarly the sum of the first row gives the mean of the x's (18,501÷2,870).

TABLE *of x-sum of above Table, i.e., Table giving all cases for xy group and over in Correlation Table.*

Unexpired term of Endowment Assurances.	Central Age at Maturity.										
	30	35	40	45	50	55	60	65	70	75	Totals.
0–4	2,870	2,861	2,860	2,843	2,781	2,197	1,554	456	68	8	18,501
5–9	2,814	2,810	2,806	2,789	2,729	2,171	1,534	450	68	8	18,179
10–14	2,642	2,639	2,636	2,621	2,567	2,071	1,470	426	66	8	17,146
15–19	2,210	2,207	2,206	2,200	2,163	1,784	1,282	365	57	7	14,481
20–24	1,545	1,545	1,544	1,544	1,531	1,297	950	270	46	7	10,279
25–29	871	871	871	871	861	760	589	171	25	6	5,887
30–34	333	333	333	333	332	321	261	86	11	3	2,349
35–39	86	86	86	86	86	86	78	27	1	1	623
40–44	9	9	9	9	9	9	7	5	1	1	68
45–49	1	1	1	1	1	1	1	1	0	0	8
Totals	13,381	13,365	13,352	13,297	13,060	10,697	7,720	2,257	343	49	87,521

The total of the last table gives the xy moment (87,521), and the x standard deviation is found by forming from the first row the series 18501, 15631, 12767, 9907, 7064, 4283, 2086, 532, 76, 8, and summing it, *i.e.*, 70,855. The second moment about the mean can then be found, the numerical work being as follows :—

$$x-\text{mean}=\frac{18501}{2870}=6{\cdot}4463$$

$$\begin{aligned}\nu_2&=2S_3-d(1+d)\\&=\frac{2\times 70855}{2870}-6{\cdot}4463\times 7{\cdot}4463\\&=1{\cdot}3747\end{aligned}$$

Similarly with the y moments

$$y\text{ mean}=\frac{13381}{2870}=4{\cdot}6624$$

$$\nu_2=\frac{(13381+10511+7697+5055+2845+1300+429+96+10+1)}{2870}-4{\cdot}6624\times 5{\cdot}$$

$$=2{\cdot}2312$$

$$\text{The } xy \text{ moment}\quad =\frac{87521}{2870}-6{\cdot}4463\times 4{\cdot}6624$$

$$={\cdot}4399$$

Remembering that $\nu_2-\frac{1}{12}$ (Sheppard's adjustment) $=\sigma^2$ and that the means are in the above work measured from the centre of the group $x=25$, $y=-5$ to -1 years, the values just given will be found to agree with those previously obtained by the direct method. The xy moment (·4399) is the same as $\frac{1262{\cdot}5}{2870}$, *i.e.*, $S(xy)\div N$.

18. Before dealing with other examples and methods, it may be well to point out a use to which the particular example might be put. The result in the equation form gives the average age corresponding to each unexpired term. Now, we might weight each entry with Mr. Lidstone's Z's,* or with the temporary annuities, then work out an equation in each case, and get new series of average ages. The results used in a valuation would give the *relative* accuracy of the three methods. I have worked out the formula with the Z weights (H^M Table), and found that

Age at maturity $=57{\cdot}595+{\cdot}1200\times$ (unexpired term).

* The method used by me was approximate and can probably be improved, the result is merely given as an indication of a possible line for research.

The results could also be used as a rough check on the average ages at valuations, and there certainly seems a possibility of doing something towards making a simple "model office" for endowment assurances with the help of the method we have been using.

19. When constructing correlation tables a little care is necessary, because in certain arrangements of statistical material it is possible to obtain a significant value for a coefficient of correlation when in reality the two functions are absolutely uncorrelated. Such a result is called "spurious correlation," and as the manner in which it arises is by the use of indices, it is defined as the correlation which will be found between indices, when the absolute values of the functions dealt with have been selected purely at random. As an example of the way spurious correlation might be introduced in actuarial statistics, we may refer to endowment assurances by limited payments on the books of a company doing a large quantity of such business, and consider the term of the original assurance (t_1), the number of premiums to be paid in future (t_2), and the number of years for which the policy has been in force (t_3). If we formed the ratios $\frac{t_2}{t_1}$ and $\frac{t_3}{t_1}$, and worked out the coefficients of correlation, we should not obtain a measure of the correlation between number of premiums payable in future and the number of years in force, because the result of using fractions with the same denominator in each would be to exaggerate correlation —that is, to introduce spurious correlation.

The general propositions of spurious correlation, of which the result just mentioned is a particular case, are as follows:—

I. *To find the mean of an index in terms of the means, standard deviations and coefficients of correlation of the two absolute measurements.*

Let x_1, x_2, x_3, x_4, be the absolute sizes of any four correlated subjects, m_1, m_2, m_3, m_4, their mean values; σ_1, σ_2, σ_3, σ_4, their standard deviations; r_{12}, r_{23}, r_{34}, r_{41}, r_{24}, r_{13}, the six coefficients of correlation; ϵ_1, ϵ_2, ϵ_3, ϵ_4, the deviations of the four subjects from their means, *i.e.*, $x_1 = m_1 + \epsilon_1$ &c.; i_{13} the mean value of the index $\frac{x_1}{x_3}$ and i_{24} the mean value of $\frac{x_2}{x_4}$; Σ_1 and Σ_2 the standard deviations of the indices $\frac{x_1}{x_3}$ and $\frac{x_2}{x_4}$ respectively, and N the total number of groups.

We shall suppose the ratios of the deviations of the mean values of the organs are so small that their cubes may be neglected.

Then $$i_{13}=\frac{1}{N}S\left(\frac{x_1}{x_3}\right)=\frac{1}{N}\cdot\frac{m_1}{m_3}\cdot S\left\{\left(1+\frac{\epsilon_1}{m_1}\right)\left(1+\frac{\epsilon_3}{m_3}\right)^{-1}\right\}$$

$$=\frac{1}{N}\cdot\frac{m_1}{m_3}\left\{N+\frac{S(\epsilon_1)}{m_1}-\frac{S(\epsilon_3)}{m_3}-\frac{S(\epsilon_1\epsilon_3)}{m_1m_3}+\frac{S(\epsilon_3)^2}{m_3^2}\right\}$$

But $S(\epsilon_1)=S(\epsilon_3)=0$ and $S(\epsilon_1\epsilon_3)=N\sigma_1\sigma_2r_{13}$ and $S(\epsilon_3)^2=N\sigma_3^2$

$$\therefore \qquad i_{13}=\frac{m_1}{m_3}\left(1+\frac{\sigma_3^2}{m_3^2}-\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}\cdot r_{13}\right)$$

and $$i_{24}=\frac{m_2}{m_4}\left(1+\frac{\sigma_4^2}{m_4^2}-\frac{\sigma_2}{m_2}\cdot\frac{\sigma_4}{m_4}\cdot r_{24}\right)$$

II. *To find the standard deviation of an index in terms of the standard deviations and coefficient of correlation of the two absolute measurements.*

$$N\times\Sigma_{13}^2=S\left(\frac{x_1}{x_3}-i_{13}\right)^2$$

$$=\frac{m_1^2}{m_3^2}S\left\{\left(1+\frac{\epsilon_1}{m_1}\right)\left(1+\frac{\epsilon_3}{m_3}\right)^{-1}-\left(1+\frac{\sigma_3^2}{m_3^2}-\frac{\sigma_1}{m_1}\cdot\frac{\sigma_3}{m_3}\cdot r_{13}\right)\right\}^2$$

$$=\frac{m_1^2}{m_3^2}S\left\{\frac{\epsilon_1}{m_1}-\frac{\epsilon_3}{m_3}+\text{square terms}\right\}^2$$

$$=i_{13}^2\left(N\frac{\sigma_1^2}{m_1^2}+N\frac{\sigma_3^2}{m_3^2}-2N\frac{\sigma_1}{m_1}\cdot\frac{\sigma_3}{m_3}\cdot r_{13}\right)$$

or $$\Sigma_{13}=i_{13}\sqrt{\left\{\frac{\sigma_1^2}{m_1^2}+\frac{\sigma_3^2}{m_3^2}-2\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}\cdot r_{13}\right\}}$$

III. *To find the coefficient of correlation of two indices in terms of the coefficients of correlation of four absolute measurements and their standard deviations.*

Let $\frac{x_1}{x_3}$ and $\frac{x_2}{x_4}$ be the two indices.

Then, if ρ be the coefficient of correlation of the two indices,

$$N\rho\Sigma_{13}\Sigma_{24}=S\left(\frac{x_1}{x_3}-i_{13}\right)\left(\frac{x_2}{x_4}-i_{24}\right)$$

$$=\frac{m_1m_2}{m_3m_4}S\left(1+\frac{\epsilon_1}{m_1}-\frac{\epsilon_3}{m_3}-\frac{\epsilon_1\epsilon_3}{m_1m_3}+\frac{\epsilon_3^2}{m_3^2}-1-\frac{\sigma_3^2}{m_3^2}+\frac{\sigma_1}{m_1}\cdot\frac{\sigma_3}{m_3}\cdot r_{13}\right)$$

$$\times\left(1+\frac{\epsilon_2}{m_2}-\frac{\epsilon_4}{m_4}-\frac{\epsilon_2\epsilon_4}{m_2m_4}+\frac{\epsilon_4^2}{m_4^2}-1-\frac{\sigma_4^2}{m_4^2}+\frac{\sigma_2}{m_2}\cdot\frac{\sigma_4}{m_4}\cdot r_{24}\right)$$

$$=i_{13}i_{24}S\left(\frac{\epsilon_1}{m_1}-\frac{\epsilon_3}{m_3}\right)\left(\frac{\epsilon_2}{m_2}-\frac{\epsilon_4}{m_4}\right)$$

as we neglect the terms of cubic order.

$$\therefore\quad \rho\Sigma_{13}\Sigma_{24}=i_{13}i_{24}\left(\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}r_{12}-\frac{\sigma_1}{m_1}\cdot\frac{\sigma_4}{m_4}r_{14}-\frac{\sigma_2}{m_2}\cdot\frac{\sigma_3}{m_3}r_{23}+\frac{\sigma_2}{m_2}\cdot\frac{\sigma_4}{m_4}r_{24}\right)$$

Hence,

$$\rho=\frac{\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}r_{12}-\frac{\sigma_1}{m_1}\cdot\frac{\sigma_4}{m_4}r_{14}-\frac{\sigma_2}{m_2}\cdot\frac{\sigma_3}{m_3}r_{23}+\frac{\sigma_2}{m_2}\cdot\frac{\sigma_4}{m_4}r_{24}}{\sqrt{\left\{\frac{\sigma_1^2}{m_1^2}+\frac{\sigma_3^2}{m_3^2}-2\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}r_{13}\right\}}\sqrt{\left\{\frac{\sigma_2^2}{m_2^2}+\frac{\sigma_4^2}{m_4^2}-2\frac{\sigma_2}{m_2}\cdot\frac{\sigma_4}{m_4}r_{24}\right\}}}$$

Proposition I. shows that the mean of an index is not the ratio of the means of the corresponding absolute measurements, and Proposition III. shows that the ρ will vanish when the four subjects forming the indices are quite uncorrelated, while, if two, say, the third and fourth, are identical, so that $r_{34}=1$ and $\frac{\sigma_3}{m_3}=\frac{\sigma_4}{m_4}$, we have

$$\rho=\frac{\frac{\sigma_1}{m_1}\cdot\frac{\sigma_2}{m_2}r_{12}-\frac{\sigma_1}{m_1}\cdot\frac{\sigma_3}{m_3}r_{13}-\frac{\sigma_2}{m_2}\cdot\frac{\sigma_3}{m_3}r_{23}+\frac{\sigma_3^2}{m_3^2}}{\sqrt{\left\{\frac{\sigma_1^2}{m_1^2}+\frac{\sigma_3^2}{m_3^2}-2\frac{\sigma_1}{m_1}\cdot\frac{\sigma_3}{m_3}r_{13}\right\}}\sqrt{\left\{\frac{\sigma_2^2}{m_2^2}+\frac{\sigma_3^2}{m_3^2}-2\frac{\sigma_2}{m_2}\cdot\frac{\sigma_3}{m_3}\cdot r_{23}\right\}}}$$

This would become applicable in the endowment assurances by limited payments to which we referred.

An interesting special case arises when the subjects x_1, x_2, x_3 are not correlated and $\frac{x_1}{x_3}$ and $\frac{x_2}{x_3}$ are formed, then

$$\rho=\frac{\frac{\sigma_3^2}{m_3^2}}{\sqrt{\left\{\frac{\sigma_1^2}{m_1^2}+\frac{\sigma_3^2}{m_3^2}\right\}}\sqrt{\left\{\frac{\sigma_2^2}{m_2^2}+\frac{\sigma_3^2}{m_3^2}\right\}}}$$

CHAPTER VII.

Correlation of Characters not Quantitatively Measurable.

1. Before the theory in this section is dealt with, we will give a table showing the class of problem with which it deals, drawn from vaccination statistics. This subject was brought before the Institute of Actuaries recently by the late A. F. Burridge, but his figures are not in a form convenient for the present purpose, and the table is taken from a paper on the subject by Dr. W. R. Macdonell,* and relates to the Sheffield smallpox outbreak of 1887–1888:—

Degree of effective Vaccination.	Strength to resist Smallpox when incurred.			
	Cicatrix.	Recoveries.	Deaths.	Total.
	Present	3,951	200	4,151
	Absent	278	274	552
	Total	4,229	474	4,703

The functions between which we want to find the correlation are "Strength to resist smallpox when incurred" and "Degree of effective vaccination," and the statistics we have cannot be arranged in a more detailed manner than the above. The characters cannot be measured quantitatively; but as the absence of such measurement does not mean that there is no correlation, we must see how the coefficient can be obtained in such a case.

* *Biometrika*, vol. i., pp. 375, *et seq.* This paper and a supplementary one deal with the subject in a way that shows clearly the strength of the evidence on the side of vaccination. The question of class is investigated, a practical point frequently neglected.

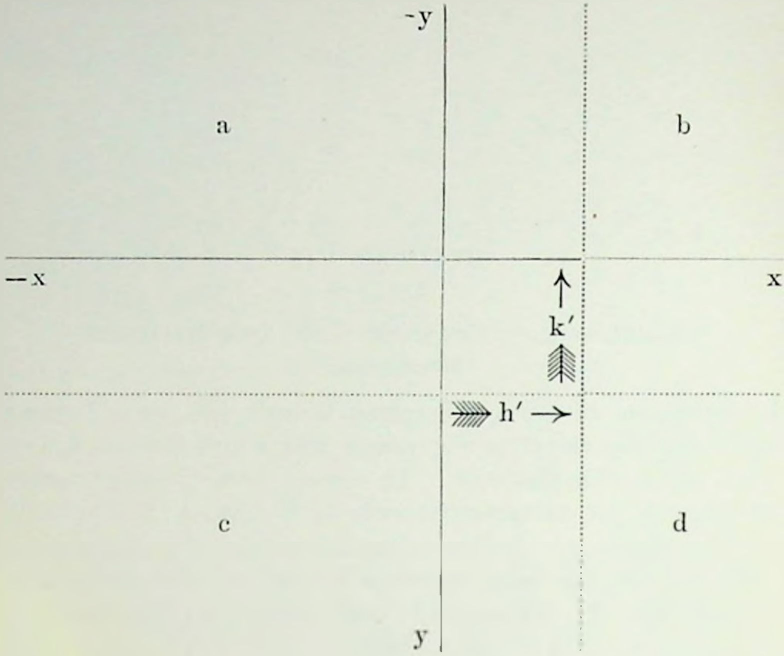


Table of Frequencies.

a	b	a+b
c	d	c+d
a+c	b+d	N

2. Using the same notation as that of the previous chapter, imagine the frequency surface

$$z = \frac{N}{2\pi\sqrt{1-r^2}\,\sigma_1\sigma_2}\,e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)}$$

to be divided into four parts by two planes at right angles to the axes of x and y at distances h' and k' from the origin; as suggested by the figures above.

Then

$$d=\frac{\mathrm{N}}{2\pi\sqrt{1-r^2}\sigma_1\sigma_2}\int_{h'}^{\infty}\int_{k'}^{\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2}+\frac{y^2}{\sigma_2^2}-\frac{2rxy}{\sigma_1\sigma_2}\right)}dxdy$$

$$=\frac{\mathrm{N}}{2\pi\sqrt{1-r^2}}\int_{h}^{\infty}\int_{k}^{\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}(x^2+y^2-2rxy)}dxdy$$

by substituting

$$x^2 \text{ for } \frac{x^2}{\sigma_1^2} \text{ and } y^2 \text{ for } \frac{y^2}{\sigma_2^2}$$

and writing

$$h=\frac{h'}{\sigma_1} \text{ and } k=\frac{k'}{\sigma_2}.$$

Further,

$$b+d=\frac{\mathrm{N}}{\sqrt{2\pi}\sigma_1}\int_{h'}^{\infty}e^{-\frac{1}{2}\frac{x^2}{\sigma_1^2}}dx$$

$$=\frac{\mathrm{N}}{\sqrt{2\pi}}\int_{h}^{\infty}e^{-\frac{1}{2}x^2}dx$$

and

$$c+d=\frac{\mathrm{N}}{\sqrt{2\pi}}\int_{k}^{\infty}e^{-\frac{1}{2}y^2}dy$$

and, remembering that N the total frequency $=a+b+c+d$, we have—

$$\mathrm{N}-2(b+d)=\mathrm{N}-\mathrm{N}\sqrt{\frac{2}{\pi}}\int_{h}^{\infty}e^{-\frac{1}{2}x^2}dx$$

$$\therefore \qquad \frac{(a+c)-(b+d)}{\mathrm{N}}=\sqrt{\frac{2}{\pi}}\int_{0}^{h}e^{-\frac{1}{2}x^2}dx$$

and, similarly,

$$\frac{(a+b)-(c+d)}{\mathrm{N}}=\sqrt{\frac{2}{\pi}}\int_{0}^{k}e^{-\frac{1}{2}y^2}dy$$

Since a, b, c, and d are known, h and k can be found from Sheppard's Tables, and the problem becomes

"To find a value for r from the equation

$$\frac{\mathrm{N}}{2\pi\sqrt{1-r^2}}\int_{h}^{\infty}\int_{k}^{\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}(x^2+y^2-2rxy)}dxdy=d$$

where d, N, h, and k are known."

The solution given by Professor Pearson (*see* Appendix III.) leads to the following equation.

$$\frac{ad-bc}{N^2HK}=r+\frac{r^2}{2}hk+\frac{r^3}{6}(h^2-1)(k^2-1)+\frac{r^4}{24}h(h^2-3)k(k^2-3)$$
$$+\frac{r^5}{120}(h^4-6h^2+3)(k^4-6k^2+3)$$
$$+\frac{r^6}{720}h(h^4-10h^2+15)k(k^4-10k^2+15)$$
$$+\frac{r^7}{5040}(h^6-15h^4+45h^2-15)(k^6-15k^4$$
$$+45k^2-15)+\text{ etc.}$$

where $$H=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}h^2},\text{ and }K=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}k^2}.$$

The numerical solution has to be obtained by approximating to the roots, and Newton's method* is convenient for the purpose.

3. The numerical work of our example is as follows :—

$$\sqrt{\frac{2}{\pi}}\int_0^h e^{-\frac{1}{2}x^2}dx=\frac{(a+c)-(b+d)}{N}=\frac{3755}{4703}$$
$$=\cdot 7984265$$
$$\therefore \qquad h=1\cdot 27716$$

by interpolation in Sheppard's Tables. In using these tables for this purpose remember that the value ·7984265 corresponds to a in his notation, so $\frac{1}{2}$ of $(1+\cdot 7984265)=\cdot 8992132$ must be looked up inversely in his Table I. If his Table III. be used it must be entered with ·7984265.

Similarly,

$$\sqrt{\frac{2}{\pi}}\int_0^k e^{-\frac{1}{2}y^2}dy=\cdot 7652561$$
$$\therefore \qquad k=1\cdot 18833$$

We next require $\dfrac{ad-bc}{N^2HK}$; and we first get from Sheppard's Tables

$$H=\cdot 1764870 \quad \therefore \quad \log H=\bar{1}\cdot 2467127$$
$$K=\cdot 1969111 \quad \therefore \quad \log K=\bar{1}\cdot 2942702$$

* *Newton's method of approximating to the root of an equation.* Let $f(x)=0$ be an equation from which the value of x is to be found and let b be a value near to x so that $x=b+h$ where h is small, then $f(x)=f(b+h)=f(b)+hf'(b)+$ terms involving higher powers of h by Taylor's Theorem, and since $f(x)=0$, we have $h=-\dfrac{f(b)}{f'(b)}$ or $x=b-\dfrac{f(b)}{f'(b)}$. The chief objection to the method is that there may be more than one root near the value b, but this does not hold in the application to correlation. (*Cf.* Approximations to rate of interest from an annuity, *Text-Book*, Part I., p. 110, formula 8.)

Hence $\log\dfrac{ad-bc}{N^2HK}=\cdot1258266$

and $$\frac{ad-bc}{N^2HK}=1\cdot336062$$

Dr. Macdonell gives 56 instead of 62 as the last two figures; the difference is probably due to interpolation.

Turning to the expression for r, we notice that hk is a product in the coefficients of r^2, r^4, r^6, &c., so it is well to work out its value and keep a note of it while the coefficients are being found. It is also advisable to begin the work by writing down the first six or seven powers of h and k.

Dr. Macdonell gives the following series:—

$$\cdot097083r^7+\cdot008170r^6+\cdot119614r^5+\cdot137450r^4 \\ +\cdot043352r^3+\cdot758844r^2+r=1\cdot336056$$

In order to obtain r we must find a value near the true one as a first approximation.

Taking $$\cdot758844r^2+r-1\cdot336056=0$$

we have $$r=\frac{-1+\sqrt{\{1+4\times1\cdot336\times\cdot7588\}}}{1\cdot5177}$$
$$=\cdot79$$

Now, this value will be in excess of the truth, owing to our only using two terms of the series on the left-hand side of the equation for finding r, and we may take $\cdot77$ as a trial rate. Applying Newton's Rule, we have:—

$$r=\cdot77-\frac{\begin{array}{l}-1\cdot336056+(\cdot77)+\cdot7588(\cdot77)^2+\cdot0434(\cdot77)^3\\ \quad+\cdot1375(\cdot77)^4+\cdot1196(\cdot77)^5+\cdot0082(\cdot77)^6+\cdot0971(\cdot77)^7\end{array}}{\begin{array}{l}1+2(\cdot77)(\cdot7588)+3(\cdot77)^2(\cdot0434)+4(\cdot77)^3(\cdot1375)\\ \quad+5(\cdot77)^4(\cdot1196)+6(\cdot77)^5(\cdot0082)+7(\cdot77)^6(\cdot0971)\end{array}}$$

$$=\cdot77-\frac{\cdot0022}{2\cdot861}$$

$$=\cdot7692$$

In work such as this, a table giving the first seven powers of the natural numbers is a help. There is one in the first edition of Barlow's Tables, but this edition is now difficult to get, though a copy of it will be found in the Library of the Institute. A table of the same powers has been given in *Biometrika*, vol. ii., pp. 474, *et seq.*

4. The value we have found for r can be checked in the following way:—

We had

$$d=\frac{N}{2\pi\sqrt{1-r^2}}\int_h^{\infty}\int_k^{\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}(x^2+y^2-2rxy)}dxdy$$

$$=\frac{N}{2\pi\sqrt{1-r^2}}\int_k^{\infty}e^{-\frac{1}{2}y^2}\left\{\int_h^{\infty}e^{-\frac{1}{2}\frac{1}{1-r^2}(x-yr)^2}dx\right\}dy$$

$$=\frac{N}{2\pi\sqrt{1-r^2}}\int_k^{\infty}e^{-\frac{1}{2}y^2}\left\{\int_{\frac{h-yr}{\sqrt{1-r^2}}}^{\infty}e^{-\frac{1}{2}X^2}\sqrt{1-r^2}\,dX\right\}dy$$

$$=\frac{N}{2\pi}\int_k^{\infty}e^{-\frac{1}{2}y^2}\left\{\int_t^{\infty}e^{-\frac{1}{2}X^2}dX\right\}dy$$

where $\quad t=\dfrac{h-yr}{\sqrt{1-r^2}}$.

To approximate to the double integral we can find

$$\int_t^{\infty}e^{-\frac{1}{2}X^2}dX$$

for a few equidistant values of t and apply a quadrature formula. The following table shows the work:—

Values of y	Values of $\frac{h-yr}{\sqrt{1-r^2}}=t$	$\frac{1}{\sqrt{2\pi}}\int_t^{\infty}e^{-\frac{1}{2}X^2}dX$	$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$	Product of two previous cols.	Application of Simpson's Rule.
(k)1·18833	·5682	·2849	·1968	·0561	$\times 1=$ ·0561
$(k+\frac{1}{2})$1·68833	$-$ ·0342	·5133	·0960	·0491	$\times 4=$ ·1964
2·18833	$-$ ·6367	·7370	·0363	·0265	$\times 2=$ ·0530
2·68833	$-$1·2392	·8923	·0107	·0096	$\times 4=$ ·0384
3·18833	$-$1·8417	·9672	·0025	·0024	$\times 2=$ ·0048
3·68833	$-$2·4441	·9795	·0004	·0004	$\times 4=$ ·0016
...	...	...	...	...	·3503 $\div$ 6 = ·584

The final column shows the application of Simpson's first quadrature rule, viz.:—

$$\int_0 y_x dx=\frac{1}{6}\left\{y_0+4y_{\frac{1}{2}}+2y_1+4y_{1\frac{1}{2}}+\ \ldots\ \right\}$$

and gives ·584 as the value of the double integral; multiplying this by N(4703), we have, 274·6 as the value of the group called *d*, which agrees with the figure given in the correlation table.

CHAPTER VIII.

Probable Errors.

1. In the previous chapters we have assumed that the means, standard deviations, moments, constants, and coefficients of correlation obtained from a body of statistics give an exact measure of the constants or of the correlation between two functions. This is not really the case. If it were possible to make an infinite number of trials bearing on a given subject, we could obtain constants or measure correlation accurately, but in practice it is only possible to take a sample from this total "population." The variation that results from using a random sample, instead of the whole "population," to find the value of any particular constant, could be reasonably measured by the standard deviation of that constant, for, as we have already remarked, the standard deviation measures the way statistics are collected round their mean or their "scatter" from it. Custom has, however, led to the use of another function known as the Probable Error, which is ·67449 times the Standard Deviation. The connection between these two functions is due to the theory having been developed from the normal curve of error, and arises in the following way. The probable error gives that value of x (say p) which divides the part of the normal curve representing positive errors into two equal portions; it is therefore given by $\int_0^p \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = \cdot 25$, where the whole area of the curve (positive *plus* negative deviations) is unity. In order to find p in terms of the standard deviation, we have,

therefore, to obtain the value of x, corresponding to $\frac{1}{2}(1+a)=\cdot75$ in Sheppard's Tables, where a is $\int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}dx$. This can be done by interpolating inversely by Lagrange's formula, and p is thus found to be $\cdot67449$ approximately. Probably the best way of viewing the use of the probable error is to regard it as a conventional reduction of the standard deviation.

2. The general rule followed by statisticians when considering probable errors is that unless a result exceeds the expected by two or three times the probable error, it is not safe to assume that the particular case differs from the expected result.

3. We will first consider the most simple case, and find the probable error of an event happening mp times when m trials are made and p is the probability of the event happening, and q of its failing. The whole series is given by $(p+q)^m=p^m+mp^{m-1}q+\ldots+q^m$. Taking moments about the centre of the group represented by p^m the first moment is $mp^{m-1}q+m(m-1)p^{m-2}q+\ldots+mq^m=mq(p+q)^{m-1}=mq$.

The second moment about the same point is

$$mp^{m-1}q+2m(m-1)p^{m-2}q^2+\tfrac{3}{2}m(m-1)(m-2)p^{m-3}q^3+\ldots+m^2q^m$$

$$=mp^{m-1}q+m(m-1)p^{m-2}q^2+\frac{m(m-1)(m-2)}{2\,!}p^{m-3}q^3+\ldots+mq^m$$

$$+m(m-1)p^{m-2}q^2+m(m-1)(m-2)p^{m-3}q^3+\ldots+m(m-1)q^m$$

$$=mq+m(m-1)q^2$$

The second moment about the mean is, therefore,

$$mq+m(m-1)q^2-m^2q^2=mpq$$

$\therefore$ The standard deviation $=\sqrt{\mu_2}=\sqrt{mpq}$

and the probable error is $\cdot67449\sqrt{mpq}$.

4. This value is of considerable use in statistical work, and it will, therefore, be advisable to see its application to a few examples.

It has been remarked that the number of male children born is to the number of female children born as 1,050 : 1,000;

in other words, the probability of a child being male is $\frac{1,050}{2,050}$. If 51,350 out of 100,000 children proved to be males in a certain community, would it be safe to base any theory connected with the variation from the usual probability on the statistics? The expected result is 51,220, and the probable error is $\cdot 67449\sqrt{100,000\frac{1050}{2050}\cdot\frac{950}{2050}} = \pm 103\cdot 9$. The difference between the actual case and the expected result was 130, and as this is only one and a quarter times the probable error, no definite conclusion can be based on the divergence from the result.

If the number of cases had been 10,000,000, and the actual number 5,135,000, then the probable error being 1,039, and the actual difference 13,000, it would have been sufficient evidence for the conclusion that the ratio 1,050 : 1,000 did not fit the particular case.

5. If the probability of death within a year is ·007, the probable error in 200 cases is $\cdot 67449\sqrt{200 \times \cdot 007 \times \cdot 993} = \cdot 80$, and it would, therefore, be possible to approximate to a loading for emergencies if 2·2 was taken instead of 1·4 as the number of deaths expected in a year out of 200 cases on risk for a year. That is, it would not be unreasonable to treat ·0110 as the rate of mortality instead of ·007, in order to obtain some idea of an emergency loading for term assurances on the assumption that the number of cases is about 200 and the average age is such that ·007 might be taken as the probability of death in a year. It has also been assumed that it is correct to treat each class as if it were subject to its own rate of mortality, and had to be treated independently of the rest of the business; this is, however, a debatable point.

6. It will be noticed that if m remains constant, then $\sqrt{mpq}$ has its largest numerical value when $p = q = \frac{1}{2}$, which shows that an office will generally find that if it has two classes of equal size, and one is subject to a higher rate of mortality than the other, the former will have the larger actual deviations from the expected number of claims, because the probability of dying in a year only reaches the value $\frac{1}{2}$ at the end of the mortality table.

7. If we now consider a frequency distribution instead of an individual experiment, it is clear that if y_s is the theoretical frequency in the sth group that would occur in N cases, the probability of a particular one of the N cases falling in the sth group is $\frac{y_s}{N}=p$, and the probability of its falling elsewhere is $1-\frac{y_s}{N}=q$. Then in m trials the distribution of frequency of this group will be given by $(p+q)^m$ and its standard deviation

$$\sigma_{y_s}=\sqrt{mpq}=\sqrt{m\frac{y'_s}{m}\left(1-\frac{y'_s}{m}\right)}$$

where $y'_s=\frac{m}{N}y_s$, and is accordingly the proportion of y_s which we should expect in the typical group of m out of N individuals. In actual practice we have, however, only the sample, but since the sample is only likely to deviate from the theoretical value to a proportionally small extent, we can replace the theoretical value by the observed frequency, and write

$$\sigma^2_{y_s}=y_s\left(1-\frac{y_s}{m}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(i.)}$$

where y_s is now taken as the frequency of the sth group in the sample.

8. If one frequency in any distribution is too large, then the others must on the average be too small, and this shows that the errors between groups are correlated. The next point to be investigated is the amount of the *correlation between deviation y_s and $y_{s'}$, or between deviations in the frequencies of the sth and s'th groups.*

Let δy_s=deviation from y_s the most probable value in the sth group, then, since

$$y_1+y_2+y_3+\ .\ .\ .\ +y_s+\ .\ .\ .\ y_{s'}+\ .\ .\ .\ +y_n=m$$

$$\delta y_1+\delta y_2+\delta y_3+\ .\ .\ .\ +\delta y_s+\ .\ .\ .\ +\delta y_{s'}+\ .\ .\ .\ +\delta y_n=0$$

Now, if the sample has given too large a value in the sth group, it is proper to suppose that this error will be distributed among the other groups in the proportion of their relative frequencies. This assumes that deviations are only due to random sampling, not to defective measurement or classification; if the deviations were due to these latter causes, it might be reasonable to assume that the excess was drawn from adjacent groups, but it is necessary to confine our attention to errors due to random sampling, and we then have

$$\delta y_{s'} = -\delta y_s \times \frac{y_{s'}}{m - y_s}$$

and
$$\delta y_{s'} \delta y_s = -\frac{(\delta y_s)^2}{m} \frac{y_{s'}}{1 - y_s/m}$$

This gives the effect of the error in the sth group on that in the s'th group on the assumption that the error in the sth group is the cause of all deviations; to give effect to the fact that all groups contribute we must sum the expression for all samplings, and obtain

$$\sigma_{y_{s'}} \sigma_{y_s} r_{y_s y_{s'}} = -\frac{\sigma^2_{y_s}}{m} \frac{y_{s'}}{1 - y_s/m} = -\frac{y_s y_{s'}}{m} \text{ by (i.)}$$

Hence
$$\sigma_{y_s} \sigma_{y_{s'}} r_{y_s y_{s'}} = -\frac{y_s y_{s'}}{m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(ii.)}$$

The only step likely to cause difficulty in the above is in summing $\delta y_{s'} \delta y_s$, but if it is borne in mind that the (xy)-moment of a correlation table gives $r\sigma_1\sigma_2$, the difficulty should immediately disappear.

9. *To find the standard deviation σ_h of the mean h of a system of observations.*

Measured from a fixed point we have

$$h = \frac{S(xy)}{S(y)} = \frac{S(xy)}{m}$$

where y is the frequency of size x.

$$\therefore \quad m\delta h = S(x\delta y)$$

Squaring each side, we have

$$m^2(\delta h)^2 = S(x_s^2 \delta y_s^2) + 2S'(x_s x_{s'} \delta y_s \delta y_{s'})$$

where S′ is the sum for all values of s and s', for which s is not equal to s'.

By dividing by the number of samplings after summing for all such samples, this gives

$$m^2\sigma_h^2 = S(x_s^2 \sigma_{y_s}^2) - 2S'(x_s x_{s'} \sigma_{y_s} \sigma_{y_{s'}} r_{y_s y_{s'}})$$

or, using (i.) and (ii.),

$$\begin{aligned} m^2\sigma_h^2 &= S(x_s^2 y) - S\left(x_s^2 \frac{y_s^2}{m}\right) - 2S'\left(x_s x_{s'} \frac{y_s y_{s'}}{m}\right) \\ &= m\mu'_2 - m \frac{S(x_s y_s)}{m} \times \frac{S(x_{s'} y_{s'})}{m} \\ &= m(\mu'_2 - h^2) \end{aligned}$$

where $m\mu'_2$ is the second moment of the whole distribution about the fixed point. But $\mu'_2 - h^2 = \sigma^2 =$ square of standard deviation of sample. Hence

$$\sigma_h = \sigma/\sqrt{m} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(iii.)}$$

10. This last result is of considerable use in statistical work. A large number of cases is recorded and the mean used to compare the particular experiment with another of a like kind. Is an actual difference between the means due to some cause other than random sampling? A practical application would be the comparison of the average profit from various classes of business for a number of years. The standard deviation of the profits in the various years would be obtained by taking the square root of the second moment about the mean and dividing it by the square root of the number of years; the quotient would give σ_h of (iii). It is only by using the standard deviations or probable errors deduced from them that it would be possible to say definitely whether a lower average profit in a certain part of the business was due to chance or to some cause requiring removal.

11. In a similar way it is possible to find the probable errors of the moments and constants, but this leads to the more theoretical parts of the subject, with which it is unadvisable to deal in a book of this character. It is, however, necessary to call attention to the probable error of the coefficient of correlation owing to the importance of that function in statistical work.

Its value has been shown* to be $\cdot 67449 \frac{(1-r^2)}{\sqrt{n}}$. We have already found in Chapter VI. that the coefficient of correlation between the age at maturity and the unexpired term of endowment assurances is $\cdot 25445$, but it is not right to assert definitely, on the strength of this information, that this coefficient represents any real relationship, until we have seen how large a deviation in the value of such a coefficient might arise purely from our having taken a random sample. The probable error of r is found by inserting 2870 for n in the formula given above, and we have $\pm \cdot 00118$ as the probable error. It is customary to show this by writing $r = \cdot 25445 \pm \cdot 00118$. In this case, therefore, the probable error is so small that the result is reliable, but it sometimes happens, especially when only a few cases have been

* The proof is given in *Phil. Trans. A.*, vol. cxci, pp. 231-241. It is, however, of a complicated nature and unsuitable for insertion in the present work. This last remark also applies to the proof for the probable error when the fourfold table is used.

considered, that the probable error is large enough to make it impossible to base any definite conclusion on the result produced. If, for instance, it had been found that $r = \cdot 0827 \pm \cdot 0621$, it would have been impossible to say definitely that the correlation had not arisen merely from chance.

12. In using the fourfold table the probable errors are larger, as would be expected, because the grouping is rougher, and the formula by which they should strictly be calculated becomes complicated. The formula referred to, gives as the probable error of r,

$$\frac{\cdot 67449}{\chi\sqrt{n}}\left\{\frac{(a+d)(c+b)}{4n^2}+\psi_2^2\frac{(a+c)(d+b)}{n^2}+\psi_1^2\frac{(a+b)(d+c)}{n^2}\right.$$
$$\left.+2\psi_1\psi_2\frac{ad-bc}{n^2}-\psi_2\frac{ab-cd}{n^2}-\psi_1\frac{ac-bd}{n^2}\right\}^{\frac{1}{2}}$$

where
$$\chi=\frac{1}{2\pi\sqrt{1-r^2}}e^{-(h^2+k^2-2rhk)/2(1-r^2)}$$

$$\psi_1=\frac{1}{\sqrt{2\pi}}\int_0^{\frac{h-rk}{\sqrt{1-r^2}}}e^{-\frac{1}{2}x^2}dx$$

$$\psi_2=\frac{1}{\sqrt{2\pi}}\int_0^{\frac{k-rh}{\sqrt{1-r^2}}}e^{-\frac{1}{2}x^2}dx$$

and it is assumed that the fourfold table is so arranged that $a+c>b+d$ and $a+b>c+d$, where a, b, c, and d, have the meanings indicated on p. 126. The numerical work for finding the probable error of r for the example in Chapter VII. is as follows—

$$\psi_1=\frac{1}{\sqrt{2\pi}}\int_0^{\frac{h-rk}{\sqrt{1-r^2}}}e^{-\frac{1}{2}x^2}dx=\frac{1}{\sqrt{2\pi}}\int_0^{\cdot 56821}e^{-\frac{1}{2}x^2}dx=\cdot 21505$$

by Sheppard's Tables.

$$\psi_2=\frac{1}{\sqrt{2\pi}}\int_0^{\frac{k-rh}{\sqrt{1-r^2}}}e^{-\frac{1}{2}x^2}dx=\frac{1}{\sqrt{2\pi}}\int_0^{\cdot 32230}e^{-\frac{1}{2}x^2}dx=\cdot 12639$$

$$\chi=\frac{1}{2\pi\sqrt{1-r^2}}e^{-(h^2+k^2+2hkr)/2(1-r^2)}=\frac{1}{2\pi\times\cdot 63900}e^{-\cdot 86741}=\cdot 10462$$

$\therefore\ \log\frac{1}{\chi}=\bar{1}\cdot 98039$, $\log\psi_1=\bar{1}\cdot 33254$, and $\log\psi_2=1\cdot 10171$;

$\therefore$ the probable error of r is—

$$\frac{\cdot 67449}{\cdot 10462\sqrt{4703}}\left\{\cdot 02283+\cdot 00145+\cdot 00479+\cdot 00252-\cdot 00408-\cdot 01015\right\}^{\frac{1}{2}}=\pm\cdot 0124.$$

13. It is often sufficient to assume that the probable error by the above formula will be three times that by the formula $\cdot 67449\,\dfrac{1-r^2}{\sqrt{n}}$. The latter gives $\pm\cdot 0040$ in the above case, which is a good approximation to $\frac{1}{3}$ of $\cdot 0124$.

CHAPTER IX.

The Test of Goodness of Fit.

1. When the values of ordinates and areas were calculated in the examples of the various types of frequency-curves, no systematic attempt was made to test the graduations, in order to ascertain whether the results obtained were reasonable. Actuaries have generally been in the habit of imposing on the graduated values of any table on which they may have been working rough checks which have amounted to a comparison of the totals in various groups and an inspection of the changes of sign in the differences between the graduated and ungraduated figures. The problem of the goodness of fit needs, however, more accurate treatment, for inspection, even when aided by the calculation of a mean error for each group,* can only tell that certain differences are large; and if the mean error be exceeded in two or three cases, it is impossible to say whether the excesses are in any way balanced by equalities in the rest of the graduation. A test is required which will give some measure of the disagreement as judged by the whole graduation.

2. Now, if there be N observations distributed in $n+1$ groups, the numbers in the group being $m'_1, m'_2 \ldots m'_{n+1}$, we have to find a criterion to enable us to decide when the series $m_1, m_2 \ldots m_{n+1}$ will be a legitimate graduation. We may clearly take a legitimate graduation to be one in which the observed values (m') do not differ from the theoretical (m) by more than the deviations that would be expected in random sampling. What we require to know is not the

* Generally calculated as $\frac{4}{5}\sqrt{npq}$, which gives approximately the average magnitude of the deviations irrespective of sign from the mean result. G. F. Hardy, *Journal of the Institute of Actuaries*, xxvii., pp. 214, *et seq.*

probability that the particular series of m''s will occur if the m's represent the theory, but the probability that the m''s, or an *equally likely or less likely* series, will arise. To appreciate the difficulties of the problem we may consider the simplest case, that of a coin-tossing experiment, and suppose that a coin has been tossed six times and come down 4 heads and 2 tails. The "graduation" we make is 3 heads and 3 tails, and to test it we require to find the probability of obtaining a result as unlikely, or more unlikely, than the observed one. This probability is the same as that of getting any one of the following results:—

6	heads	and	0	tails
5	,,	,,	1	,,
4	,,	,,	2	,,
2	,,	,,	4	,,
1	,,	,,	5	,,
0	,,	,,	6	,,

It is obviously impossible to calculate such probabilities directly, even when the simple probabilities leading to the deviations are known, in any but the easiest cases; but when we do not know the simple probabilities, or the case is a complicated one, a further difficulty is introduced, owing to our inability to tell from *a priori* reasoning which of the possible cases are more or less likely than that which has actually arisen. It would, for instance, be impossible to say, without a large amount of arithmetical work, when 20 dice were being thrown, whether the probability of getting ten "sixes" or more was greater than that of getting two "sixes" or fewer; but this is an extremely simple case compared with the general proposition in which deviations over a series of numbers have to be considered.

3. If it is assumed in any measurement on one subject that the deviations from the mean take the form of the normal curve of error (Type VII.), and it is required to estimate the chance of obtaining deviations greater than a certain value (t, say), it will be necessary to sum all values of the normal curve beyond t on each side of the mean, *i.e.*, take

$$\int_{-\infty}^{-t} e^{-x^2}dx + \int_{t}^{\infty} e^{-x^2}dx = 2\int_{t}^{\infty} e^{-x^2}dx$$

and divide the result by the area of the whole curve, *i.e.*, by the total deviations. Assuming that there are two measurements instead of one (the exposed to risk, for instance, at two ages), the deviations are, as it were, in two directions instead of one; and it is necessary to take an expression with two variables instead of one. The expression analogous to the normal curve is the correlation surface

$$z = z_0 e^{-\frac{1}{2}\left\{\frac{x^2}{\sigma_1^2} - \frac{2xyr}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\} \div (1-r^2)}$$

with which we have already dealt. The integrations must be performed for both variables from t and t' onwards, and compared with the total. If there are n measurements, it becomes necessary to deal with a function of n variables, and this will give the reader a slight idea of the problem from the mathematical point of view, and suggest that he will expect the quotient of two n-fold integrals to give the probability. The next step is to reduce these n-fold integrals to the form of ordinary integrals, and it has been shown* that the result

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{1}{2}x^2} x^{n-1} dx}{\int_{0}^{\infty} e^{-\frac{1}{2}x^2} x^{n-1} dx} \dagger$$

is reached. In this expression χ stands for a complex function depending on the n variables from which the expression was evolved, and measures the position that is indicated by the probability of the particular distribution, the test for the graduation of which is required.

4. Before a measure of the probability P can be obtained, a value for χ must be found from the statistics of the particular graduation, and in the paper to which reference has already been made its value is shown to be such that

$$\chi^2 = S\left\{\frac{(m_r - m'_r)^2}{m_r}\right\}$$

It is almost obviously necessary to use the square of the difference in order that negative differences may, equally with

* Professor Karl Pearson: "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from Random Sampling."—*Phil. Mag.*, July, 1900.

† A fairly extensive table of P will be found in *Biometrika*, vol. i., p. 155, &c. It gives values of P for all values of $n+1$ from 3 to 30, corresponding to χ^2 from 1 to 30, with a few additional values and auxiliary tables for the calculation of further values.

positive differences, increase the improbability of the system, while a ratio is required to bring into account the size of the group; for an error of 15 in a group of 20 would be very large, but in a group of 1,000 would be negligible.

5. The practical aspects of the test of fit and its application may now be dealt with.

(1) If the facts representing the graduated and ungraduated figures are only available in groups, then the value of the probability by the test will, as a rule, be lower as the number of groups is increased. This practical point should be borne in mind, as it sometimes happens that graduations are tested in groups of, say, 5 years of age; but the graduated figures for individual ages are then used unreservedly, though, strictly speaking, they may be no better than interpolated values.

(2) The test assumes a distribution, and would not be applicable if the numbers were a series of ordinates, though the application of the test would probably give a fair idea of the goodness of fit if a large number of ordinates had been given in the series.

(3) The tails of the experience will be very small and never fit exactly. "We ought to take our final theoretical "groups to cover as much of the tail area as amounts to at "least a unit of frequency in such cases." (*Phil. Mag.*, footnote, p. 164.)

(4) If the number of observations be multiplied by t, say, and the deviations are also multiplied by t, then the value of χ^2 will be multiplied by the same figure, and the test will show that the fit is worse. This may seem strange at first, but a little consideration will show that it is reasonable, since a large number of cases will give smoother series than a small number; then, if the results are proportionally the same in two examples having the same theoretical distribution but different total frequencies, it follows that the one with greater frequency is less probable than the one with less frequency. The probability of a result as bad, or worse, than three heads and one tail in coin tossing (two heads and two tails being the theoretical result) is ·625; but the probability of a result as bad, or worse, than $3 \times 2 = 6$ heads and $1 \times 2 = 2$ tails is ·289.

(5) I have found, in applying the test, that when the numbers dealt with are very large, the probability is often small, even though the curve appears to fit the statistics very

closely. The explanation is that the statistics with which we deal in practice nearly always contain a certain amount of extraneous matter, and the heterogeneity is concealed in a small experience by the roughness of the data. The increase in the number of cases observed removes the roughness, but the heterogeneity remains. The meaning, from the curve-fitting point of view, is that the experience is really made up of more than one frequency-curve; but a certain curve, approximating to the one calculated, predominates.

(6) It is sometimes thought that the introduction of additional constants must necessarily improve the fit of a curve. It may do so in some cases, but it is quite possible to take a curve with ten constants and find it gives a worse result than another having only three.

(7) It may sometimes be advisable to use a curve giving a slightly worse agreement than another for simplicity, or for reasons such as those which prompt actuaries to employ Makeham's hypothesis; but, as a rule, it is well to use the best-fitting curve in any case, that is, the curve giving the highest value of P.

6. In a recent paper "On the Comparative Reserves of Life Assurance Companies, &c." (*Journal of the Institute of Actuaries*, xxxvii., pp. 458–9), Mr. King remarked that it is permissible to use the H^M Model Office for the O^M; and it will be interesting to apply the formulæ given above to see what is the probability of the O^M distribution if the H^M be taken as the theoretical distribution :—

Central Age in Group.	Policies issued arranged in Age-Groups.		$O^M - H^M$		Square of $O^M - H^M$ $\div H^M$
	H^M	O^M	+	−	
20	6·97	7·30	·33	...	·02
25	17·75	20·45	2·70	...	·41
30	21·04	23·11	3·07	...	·45
35	18·41	18·40	...	·01	·00
40	13·82	13·05	...	·77	·04
45	9·45	8·44	...	1·01	·11
50	6·23	5·07	...	1·16	·22
55	3·51	2·58	...	1·93	1·60
60	1·97	1·20	...	·77	·30
65	·85	·40	...	·45	·24
...	100·00	100·00	6·10	6·10	$\chi^2=3·39$

There are ten groups, and $\chi^2=3{\cdot}39$, and the table in *Biometrika* gives $P={\cdot}964295$ and ·911413 when $\chi^2=3$ and 4 respectively. There is, however, a further point, for it has to be decided if it is sufficient to test for 100 new policies. 500 would reduce the probability to about ·05, which means that in only one case out of twenty would a random sampling lead to a system of deviations from the H^M as great as that shown by the O^M. This result will remind the student of the great danger of dealing with percentages without considering the actual number of cases investigated. Mr. King's other table, which is of greater importance in his work (policies according to attained age), shows a much closer agreement, as P is ·831051 for 10,000 cases.

In a paper on Makeham's formula (*Journal of the Institute of Actuaries*, xxxv.) Mr. Calderon gave some graduations of the H^F mortality table, and on pp. 188 and 189 his results re summarized in a form which is convenient for applying e test. His methods A, B and C, give 35·11, 34·02 and ·77 as the values of χ^2 for 20 groups. The probability if $=30$ is ·051798, and if $\chi^2=40$ is ·003272. The odds against the best of the three graduations must be 30 to 1, which shows that Makeham's formula is unsuitable or the methods of application unsatisfactory.

In the numerical examples of Chapter V., the value of P for Type I. is about ·98. Type II. gives ·7, Type IV. $\frac{2}{3}$, and the sums assured in Type VII. ·2, and the reserves give a probability greater than ·9.

7. The only other point to which reference is necessary is the actual value of P at which a good fit ends and a bad one begins. It is impossible to fix such a value. We have merely a measure of probability for the whole table, and if the odds against the graduation are twenty or thirty to one the result is unsatisfactory; if they are ten to one the graduation is not unreasonable, but the exact value when a result must be discarded cannot be given. As, however, it is clearly impossible to imagine any test which can fix an absolutely definite standard, there is no reason for objecting to the particular method because it fails to do so.

CHAPTER X.

THE THEORY OF CONTINGENCY.

1. A TABLE showing no correlation can be formed from the totals of any ordinary correlation table, such as that on p. 107, by dividing the total of each column in proportion to the totals of the rows. Thus, the first column would be—

Unexpired Term ...	0—4	5—9	. . .
Frequency with no correlation ...	$6 \times \frac{56}{2870}$	$6 \times \frac{172}{2870}$	. . .

and the remaining part of the table would be formed in a similar way. A moment's consideration of the definition given at the beginning of Chapter VI. will show that such a method of formation must necessarily give the required table, because, since each column is formed in proportion to the total, the means of the columns must all be the same as the mean of the total, which shows at once from the definition that no correlation can exist in such a table.

2. The following table shows the figures exhibiting no correlation in ordinary type, and those exhibiting correlation in small type. Now, if these two sets of figures coincide exactly in any particular case there is clearly no correlation in the table; if they differ slightly there is a slight amount, and if they differ greatly there is a considerable amount of correlation, and we come therefore to the conclusion that an alternative method of finding the correlation between two things is by measuring the difference between the figures in the actual correlation table and those that would have arisen if there had not been any correlation. In the last chapter we discussed a method of measuring the goodness of fit (or amount of agreement) between two sets of figures, and this suggests that we might calculate χ^2 by squaring the difference between each pair of figures in the table and dividing the result by the frequency when there is no correlation. The reason for choosing the figure from the table with no correlation as the divisor is that it always has a value, while the correlation table may give a frequency of zero, which renders it impossible to use the latter as a divisor.

Unexpired term of Endowment Assurances.	Central Age at Maturity.										Total.
	30	35	40	45	50	55	60	65	70	75	
0–4	·1 2	·1 ..	·3 ..	1·2 2	11·4 26	12·5 6	21·4 14	7·6 6	1·2 ..	·2 ..	56
5–9	·4 1	·2 1	1·0 2	3·7 6	35·0 62	38·6 36	65·8 40	23·2 22	3·6 2	·5 ..	172
10–14	·9 ..	·6 2	2·6 9	9·3 17	87·8 117	96·8 99	165·4 127	58·4 52	9·0 8	1·2 1	432
15–19	1·4 3	·9 ..	3·9 6	14·4 24	135·3 145	149·0 155	254·4 237	89·9 84	13·9 11	1·9 ..	665
20–24	1·4 ..	·9 1	4·0 ..	14·6 3	137·2 133	151·0 167	257·8 271	91·1 78	14·1 20	1·9 1	674
25–29	1·1 ..	·8 ..	3·2 ..	11·6 9	109·5 90	120·6 123	205·9 231	72·7 71	11·2 11	1·4 3	538
30–34	·5 ..	·4 ..	1·5 ..	5·3 1	50·3 11	55·3 49	94·6 127	33·4 49	5·2 8	·7 2	247
35–39	·2 ..	·1 ..	·5 ..	1·7 ..	15·7 ..	17·2 6	29·4 49	10·4 22	1·6 ..	·2 ..	77
40–44	·0 ..	·0 ..	·0 ..	·2 ..	1·6 ..	1·8 2	3·1 2	1·1 3	·2 ..	·0 1	8
45–49	·0 ..	·0 ..	·0 ..	·0 ..	·2 ..	·2 ..	·1 ..	·2 1	·0 ..	·0 ..	1
Total	6	4	17	62	584	643	1,098	388	60	8	2,870

3. As it is clear that χ^2 will give a measure of the correlation, it will be interesting to see the connection between it and the coefficient of correlation r; and the following proof shows that

$$r=\sqrt{\frac{\phi^2}{1+\phi^2}}\text{ where }\phi^2=\frac{\chi^2}{N}$$

and the correlation table can be approximately represented by the normal correlation surface.

Using the same notation as that of Chapter VI., the frequency with no correlation is given by

$$z_0 = \frac{N}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_1^2} + \frac{y^2}{\sigma_2^2}\right)}$$

while that with correlation is—

$$z = \frac{N}{2\pi\sqrt{1-r^2}\sigma_1\sigma_2} e^{-\frac{1}{2}\frac{1}{1-r^2}\left(\frac{x^2}{\sigma_1^2} - \frac{2rxy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right)}$$

Then $\phi^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \frac{(z-z_0)^2}{Nz_0} dxdy$

$$= \frac{1}{N}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\left(\frac{z^2}{z_0} - 2z + z_0\right)dxdy$$

$$= \frac{1}{2\pi}\left\{\frac{1}{1-r^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left\{x'^2\frac{1+r^2}{1-r^2} - \frac{4rx'y'}{1-r^2} + y'^2\frac{1+r^2}{1-r^2}\right\}} dx'dy'\right.$$

$$- \frac{2}{\sqrt{1-r^2}}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}\left\{x'^2\frac{1}{1-r^2} - \frac{2rx'y'}{1-r^2} + y'^2\frac{1}{1-r^2}\right\}} dx'dy'$$

$$\left. + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x'^2+y'^2)}dx'dy'\right\}$$

where $x' = \frac{x}{\sigma_1}$ and $y' = \frac{y}{\sigma_2}$

$$= \frac{1}{1-r^2}\frac{1}{\sqrt{\left(\frac{1+r^2}{1-r^2}\right)^2 - \frac{4r^2}{(1-r^2)^2}}} - \frac{2}{\sqrt{1-r^2}}\frac{1}{\sqrt{\frac{1}{(1-r^2)^2} - \frac{r^2}{(1-r^2)^2}}} + 1$$

by No. (vi.) of Appendix III.

$$= \frac{1}{1-r^2} - 2 + 1$$

$$= \frac{r^2}{1-r^2}$$

or $\qquad r = \pm\sqrt{\frac{\phi^2}{1+\phi^2}}$

4. The result just obtained may be considered a little more closely, as it leads to some valuable conclusions:—

(1) It shows clearly that r must lie between -1 and $+1$.

(2) Since the value of ϕ^2 will not be affected by the order of the columns (or rows), it will be seen that it is permissible to interchange them, provided, of course, the whole column (or row) be moved at once.

(3) The proof shows that r will not necessarily be obtained exactly if a very small number of groups is used, because by using the integral calculus an infinite number of groups was assumed.

(4) We also assumed, however, that we were dealing with perfectly smooth series; but since χ^2 is a measure of the goodness of fit between the correlation and no-correlation figures, a very large number of groups gives undue prominence to the chance deviation, due to the use of a random sample, and the value of r found from that of ϕ^2 may differ considerably from the value reached by the xy-moment. Too fine a grouping may give a less accurate result than a less fine one.

5. These conclusions are borne out by practical work, and any student who cares to go into the subject can find the value of r by the two methods from a large table, using various groupings, and he will see that the best agreements are obtained when the grouping is neither very fine nor very rough. It should, however, be borne in mind that unless the correlation table takes the form assumed in the proof, an exact agreement between the two methods cannot be expected; it is, for this reason, well to distinguish the value $\sqrt{\frac{\phi^2}{1+\phi^2}}$ from the value r by calling the former the *coefficient of contingency*. It seems to me that if the difficulty about grouping could be overcome, the coefficient of contingency would be more useful than the coefficient of correlation (r).

6. The following table shows the table given above grouped so as to enable us to obtain the coefficient of contingency more conveniently :—

Unexpired term of Endowment Assurances.	Central Age at Maturity						Total.
	30, 35, 40 & 45	50	55	60	65	70 & 75	
0–9	7·0 14	46·4 88	51·1 42	87·2 54	30·8 28	5·5 2	228
10–14	13·4 28	87·8 117	96·8 99	165·4 127	58·4 52	10·2 9	432
15–19	20·6 33	135·3 145	149·0 155	254·4 237	89·9 84	15·8 11	665
20–24	20·9 4	137·2 133	151·0 167	257·8 271	91·1 78	16·0 21	674
25–29	16·7 9	109·5 90	120·6 123	205·9 231	72·7 71	12·6 14	538
30–34	7·7 1	50·3 11	55·3 49	94·6 127	33·4 49	5·9 10	247
35–49	2·7 ..	17·5 ..	19·2 8	32·9 51	11·7 26	2·0 1	86
Total	89	584	643	1,098	388	68	2,870

Working out χ^2 from this table, the value is found to be 257·9,* which gives $\phi^2 = \frac{\chi^2}{N} = \cdot 0899$, and the coefficient of contingency is ·2872. This differs from the value of r found by the other method by about ·03; but an inspection of the table on p. 107 leads to the conclusion that the totals cannot be considered to be curves of Type VII., and the condition for $r = \sqrt{\frac{\phi^2}{1+\phi^2}}$ is not therefore satisfied.

7. The probable error of the coefficient of contingency may be taken as approximately one and a third times that of r.

8. Though we have dealt with the theory of contingency from the point of view of its particular application to ordinary correlation tables, the reader should bear in mind that in statistical practice its chief use is when characters not capable of quantitative measurement are being examined, such, for instance, as colours, shapes, diseases, &c. The method of application is, of course, exactly the same.

* In case any student may not follow the method easily, we may mention that the contributions to χ^2 from the first column are 7·0, 15·9, 7·5, 13·7, 3·6, 5·8, and 2·7.

9. We may now, in conclusion, refer briefly to some of the practical applications to which the theories of correlation and contingency can be put. Some actuarial applications have already been made, such as the investigation by Professor Pearson and Miss Beeton into the inheritance of duration of life (see *Journal of the Institute of Actuaries*, vol. xxxv., pp. 112, *et seq.*, and 458, *et seq.*; and *Biometrika*, vol. i., part 1.); while the correlations between ages of husband and wife and age at death of a man and the number of his children under age 21 are obvious applications. The last-mentioned case seems to suggest that it might be possible to apply the method in connection with the valuation of pension funds, while we have already noticed that it is possible that it might be of use for checking average ages, &c., in endowment assurance valuations.

APPENDIX I.

USEFUL CONSTANTS.

$$e = 2{\cdot}71828\ 18285$$

$$e^{-1} = {\cdot}36787\ 94417$$

$$\pi = 3{\cdot}14159\ 26536$$

$$\log_{10} e = {\cdot}43429\ 44820$$

$$\log_e 10 = 2{\cdot}30258\ 509$$

$$\log(\log_{10} e) = \bar{1}{\cdot}63778\ 43114$$

$$\log_{10} \pi = {\cdot}49714\ 98728$$

$$\log_{10} \sqrt{\pi} = {\cdot}24857\ 49364$$

$$\log_{10} \frac{1}{\sqrt{2\pi}} = \bar{1}{\cdot}60091\ 00657$$

$$\log_{10} e^{-\frac{1}{12}} = \bar{1}{\cdot}96380\ 87932$$

APPENDIX II.

B AND Γ FUNCTIONS.

$$B(m, n)=\int_0^1 x^{m-1}(1-x)^{n-1}dx$$

$$\Gamma(p)=\int_0^\infty e^{-x}x^{p-1}dx$$

I. $$\int x^{p-1}e^{-x}dx= -e^{-x}x^{p-1}+(p-1)\int x^{p-2}e^{-x}dx$$

by integration by parts.

When $p-1$ is positive, $e^{-x}x^{p-1}$ vanishes when $x=0$, and when $x=\infty$ it can be written $\dfrac{x^{p-1}}{e^{+x}}$, and the rule for evaluation of undetermined forms (Edwards' *Diff. Calc.*, ch. xiv.) can be applied.

$\therefore$ $$\int_0^\infty x^{p-1}e^{-x}dx=(p-1)\int_0^\infty x^{p-2}e^{-x}dx$$

or $$\Gamma(p)=(p-1)\Gamma(p-1).$$

If p be an integer, $\Gamma(p)=\underline{|p-1}$.

II. To prove $B(m, n)=\dfrac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$.

Putting zx for x in the equation for $\Gamma(m)$, we have

$$\int_0^\infty e^{-xz}x^{m-1}dx=\frac{\Gamma(m)}{z^m}$$

$\therefore$ $$\Gamma(m)=\int_0^\infty e^{-zx}z^m x^{m-1}dx$$

and $$\Gamma(m)e^{-z}z^{n-1}=\int_0^\infty e^{-z(1+x)}z^{m+n-1}dx$$

$\therefore$ $$\Gamma(m)\int_0^\infty e^{-z}z^{n-1}dz=\int_0^\infty\left[\int_0^\infty e^{-z(1+x)}z^{m+n-1}dx\right]x^{m+1}dx$$

But if $z(1+x)=y$, we get

$$\int e^{-z(1+x)}z^{m+n-1}dz=\frac{1}{(1+x)^{m+n}}\int_0^{\infty}e^{y}y^{m+n-1}dy$$

$$=\frac{\Gamma(m+n)}{(1+x)^{m+n}}$$

$$\therefore \qquad \Gamma(m)\Gamma(n)=\Gamma(m+n)\int_0^{\infty}\frac{x^{m-1}}{(1+x)^{m+n}}dx$$

But putting $1+x=\frac{1}{1-z}$ in this integral, we obtain

$$\int_0^1\frac{z^{m-1}}{(1-z)^{m-1}}(1-z)^{m+n}\frac{1}{(1-z)^2}dz$$

which reduces at once to $B(m, n)$, and

$$\therefore \qquad B(m, n)=\frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

III.—To prove $\Gamma(\frac{1}{2})=\sqrt{\pi}$.

We have already shown in the proof for Type VII. that $2\int_0^{\infty}e^{-x^2}dx=\sqrt{\pi}$, and by putting $x^2=z$, we have

$$2\int_0^{\infty}e^{-x^2}dx=\int_0^{\infty}e^{-z}z^{-\frac{1}{2}}dz=\Gamma(\tfrac{1}{2})=\sqrt{\pi}.$$

For statistical work a table of $\Gamma(x)$ or $\log\Gamma(x)$ is required, and Legendre has given a table of the latter to twelve figures for values of x between 1 and 2, from which $\log\Gamma(x)$ can be found easily, provided x is small.

When x is large, $\log\Gamma(x)$ can be approximated to. The best known approximation is

$$\Gamma(x+1)=\sqrt{2\pi x}x^xe^{-x}e^{-\frac{1}{12x}} \; *$$

or

$$\log_{10}\Gamma(x+1)=\log_{10}\sqrt{2\pi}+(x+\tfrac{1}{2})\log_{10}x-\left(x+\frac{1}{12x}\right)\log_{10}e$$

and it can be used when x is not less than 8. To show how the table of $\log\Gamma(x)$ is used, and also how the approximation

* A proof of this well-known approximation will be found in *Chrystal's Algebra*, vol. ii., pp. 368, &c., or in *Boole's Finite Differences*, chapter VI.

approaches the true value, the following table has been prepared :—

TABLE *comparing True and Approximate Values of* $\log \Gamma(x)$.

x	Log $\Gamma(x)$		
	True.	Approximate.	Δ
1·372	$\bar{1}$·948975	...	...
2·372	·086329	·086743	·000414
3·372	·461444	·461532	·000088
4·372	·989332	·989369	·000037
5·372	1·630012	1·630025	·000013
6·372	2·360148	2·360156	·000008
7·372	3·164124	3·164130	·000006
8·372	4·032009	4·032010	·000001
9·372	4·954838	4·954837	− ·000001
10·372	5·926670	5·926669	− ·000001

A six-figure table of $\log \Gamma(p)$, obtained by abridging Legendre's, is given on pp. 166 and 167.

APPENDIX III.

The Integration of some Expressions connected with the Normal Curve of Error.

1. On page 91 we showed that

$$\int_{-\infty}^{+\infty} e^{-x^2}dx = \sqrt{\pi} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{i.})$$

$$\int_{-\infty}^{+\infty} e^{-x^2/h^2}dx = h\sqrt{\pi} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{ii.})$$

Since the curve is symmetrical, we have

$$\int_{-\infty}^{+\infty} x^{2n+1}e^{-x^2}dx = \text{zero} \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{iii.})$$

If we integrate $\int x^{2n}e^{-x^2}dx$ by parts, we have

$$\int x^{2n}e^{-x^2}dx = \frac{x^{2n+1}}{2n+1}e^{-x^2} + \int 2x\frac{x^{2n+1}}{2n+1}e^{-x^2}dx$$

and inserting the limits $-\infty$ and $+\infty$ we have

$$\int_{-\infty}^{+\infty} x^{2n}e^{-x^2}dx = \frac{2}{2n+1}\int_{-\infty}^{+\infty} x^{2n+2}e^{-x^2}dx \quad . \quad . \quad . \quad (\text{iv.})$$

This last formula shows the connection between the successive even moments.

2. Referring to p. 112, let

$$z = z_0 e^{-\frac{1}{2(1-r^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2xyr}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}}$$

Then z can be put in the form

$$z_0 e^{-\frac{1}{2(1-r^2)}\left\{\frac{x}{\sigma_1} - \frac{yr}{\sigma_2}\right\}^2} e^{-\frac{1}{2(1-r^2)}\cdot\frac{y^2(1-r^2)}{\sigma_2^2}}$$

$$= z_0 e^{-\frac{1}{2(1-r^2)\sigma_1^2}\left\{x - \frac{yr\sigma_1}{\sigma_2}\right\}^2} e^{-\frac{y^2}{2\sigma_2^2}}$$

Then $$\int_{-\infty}^{+\infty} z\,dx = z_0\sqrt{2\pi(1-r^2)}\,\sigma_1 e^{-y^2/2\sigma_2^2} \text{ by (ii.)}$$

and $$N = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} z\,dx\,dy = \int_{-\infty}^{+\infty} z_0\sqrt{2\pi(1-r^2)}\,\sigma_1 e^{-y^2/2\sigma_2^2}dy$$

$$= 2\pi\sigma_1\sigma_2\sqrt{1-r^2}\,z_0 \quad . \quad . \quad . \quad . \quad . \quad . \quad (\text{v.})$$

3. Using the same method as that just given, it can be shown that if $ac > b^2$

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(ax^2-2bxy+cy^2)}dxdy = \frac{2\pi}{\sqrt{ac-b^2}} \quad . \quad . \quad . \quad \text{(vi.)}$$

for the index can be written

$$-\frac{a}{2}\left\{x-\frac{b}{a}y\right\}^2+\frac{y^2}{2a}(ac-b^2)$$

and if we then integrate with respect to x, we have

$$\sqrt{\frac{2\pi}{a}}e^{-\frac{y^2}{2a}(ac-b^2)}$$

and if $ac-b^2$ is positive, we can integrate this last expression with respect to y and have

$$\sqrt{\frac{2\pi}{a}}\sqrt{\frac{2\pi a}{ac-b^2}}=\frac{2\pi}{\sqrt{ac-b^2}}$$

4. We will now find $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} zxydxdy$.

Proceeding as in (v.), we have

$$\int_{-\infty}^{+\infty} zxydx = \int_{-\infty}^{+\infty} xyz_0 e^{-y^2/2\sigma_2{}^2} e^{-\frac{1}{2(1-r^2)\sigma_1{}^2}\left\{x-\frac{yr\sigma_1}{\sigma_2}\right\}^2} dx$$

$$= z_0 y e^{-y^2/2\sigma_2{}^2}\int_{-\infty}^{+\infty}\left(X+\frac{yr\sigma_1}{\sigma_2}\right)e^{-\frac{X^2}{2(1-r^2)\sigma_1{}^2}}dX$$

where $X = x - \dfrac{yr\sigma_1}{\sigma_2}$

$$= z_0 y e^{-y^2/2\sigma_2{}^2}\int_{-\infty}^{+\infty}\frac{yr\sigma_1}{\sigma_2}e^{-\frac{X^2}{2(1-r^2)\sigma_1{}^2}}dx$$

because by (iii.) $\int_{-\infty}^{+\infty} e^{-X^2/k^2}XdX$ is zero

$$= z_0\frac{\sigma_1{}^2}{\sigma_2}r\sqrt{2\pi(1-r^2)}\,y^2e^{-y^2/2\sigma_2{}^2}$$

But, by putting $n=0$ in (iv.) and using (ii.), we have

$$\int_{-\infty}^{+\infty} y^2e^{-y^2/2\sigma_2{}^2}dy = \sigma_2{}^3\sqrt{2\pi}$$

$$\therefore \quad \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} zxydxdy = z_0\sigma_1{}^2\sigma_2{}^2 2\pi r\sqrt{1-r^2}$$

$$= N\sigma_1\sigma_2 r \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(vii.)}$$

because by (v.)

$$z_0 = \frac{N}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}$$

5. We may now deal with the problem referred to in Chap. VII—
"To find a value of r from the equation

$$\frac{\mathrm{N}}{2\pi\sqrt{1-r^2}}\int_h^\infty\int_k^\infty e^{-\frac{1}{2}\frac{1}{1-r^2}(x^2+y^2-2rxy)}\,dx\,dy=d$$

where d, N, h, and k are known."

Consider the expression $\dfrac{1}{\sqrt{1-r^2}}e^{-(x^2+y^2-2rxy)/2(1-r^2)}=\mathrm{U}$ say . . (α)

and expand it in terms of r by Maclaurin's theorem, then

$$\mathrm{U}=e^{-\frac{1}{2}(x^2+y^2)}\left(u_0+\frac{u_1r}{1!}+\frac{u_2r^2}{2!}+\ldots+\frac{u_nr^n}{n!}+\ldots\right)\ldots\ (\beta)$$

where $\quad u_n=e^{\frac{1}{2}(x^2+y^2)}\left(\dfrac{d^n\mathrm{U}}{dr^n}\right)_{r=0}$ . . . . (γ)

Now take the logarithmic differential coefficient of U with respect to r, and we have

$$\frac{1}{\mathrm{U}}\frac{d\mathrm{U}}{dr}=-\frac{(x^2+y^2-2rxy)}{2}\frac{d}{dr}(1-r^2)^{-1}-\frac{(1-r^2)^{-1}}{2}\frac{d}{dr}(x^2+y^2-2rxy)$$
$$-\frac{1}{2}\frac{d}{dr}\log(1-r^2)$$
$$=-(x^2+y^2-2rxy)(1-r^2)^{-2}+(1-r^2)^{-1}xy+r(1-r^2)^{-1}$$
$$\therefore\ \frac{d\mathrm{U}}{dr}=\mathrm{U}\{1-r^2\}^{-2}\{xy+r(1-x^2-y^2)+r^2xy-r^3\}$$

Differentiate n times by Leibnitz's theorem and put $r=0$ and we have

$$u_{n+1}=n(2n-1-x^2-y^2)u_{n-1}-n(n-1)(n-2)^2u_{n-3}$$
$$+xy(u_n+n(n-1)u_{n-2})$$

Hence

$$\left.\begin{aligned}u_0&=1\\u_1&=xy\\u_2&=(x^2-1)(y^2-1)\\u_3&=x(x^2-3)(y^2-3)y\\u_4&=(x^4-6x^2+3)(y^4-6y^2+3)\end{aligned}\right\}\ldots\ldots\ldots\ (\delta)$$

The laws indicated by (δ) are

$$\left.\begin{aligned}u_n&=v_n\times w_n\\v_n&=xv_{n-1}-(n-1)v_{n-2}\\w_n&=yw_{n-1}-(n-1)w_{n-2}\end{aligned}\right\}\ldots\ldots\ldots\ (\epsilon)$$

and we can, therefore, re-write (β) as

$$\frac{1}{2\pi}\mathrm{U}=\frac{1}{2\pi}e^{-\frac{1}{2}(x^2+y^2)}\left(1+\frac{v_1w_1}{1!}r+\frac{v_2w_2}{2!}r^2+\ldots\right)\ .\ .\ .\ (\zeta)$$

integrating this from κ to ∞ with respect to x, and remembering that w_n does not involve x, we have

$$\frac{1}{2\pi}\int_h^\infty U\,dx = \frac{1}{2\pi}e^{-\frac{1}{2}y^2}\left\{\int_h^\infty e^{-\frac{1}{2}x^2}dx + \frac{rw_1}{1!}\int_h^\infty e^{-\frac{1}{2}x^2}v_1dx + \dots \right.$$

$$\left. + r^n\frac{w_n}{n!}\int_h^\infty e^{-\frac{1}{2}x^2}v_ndx + \dots \right\}$$

$$= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}\left\{V_0 + \frac{rV_1w_1}{1!} + \dots + \frac{r^nV_nw_n}{n!} + \dots \right\}$$

where V_n is written for $\frac{1}{\sqrt{2\pi}}\int_h^\infty v_ne^{-\frac{1}{2}x^2}dx$.

Now integrate this with respect to y from k to ∞, remembering that V_n does not involve y, and writing W_n for $\frac{1}{\sqrt{2\pi}}\int_k^\infty e^{-\frac{1}{2}y^2}w_ndy$, we see that $\frac{1}{2\pi}\int_h^\infty\int_k^\infty U\,dxdy$ can be expressed as a series of which the general term is $\frac{r^n}{n!}V_nW_n$, and we must now evaluate V_n and W_n.

From (ϵ) it can be shown by induction* that the general form of v_n is

$$x^n - \frac{n(n-1)}{1!}\cdot\frac{x^{n-2}}{2} + \frac{n(n-1)(n-2)(n-3)}{2!}\cdot\frac{x^{n-4}}{2^2} - \&c. \quad . \quad . \quad (\eta)$$

Now we notice that

$$\frac{dv_n}{dx} = nv_{n-1} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (\theta)$$

$\therefore$ by (ϵ)
$$v_n = xv_{n-1} - \frac{dv_{n-1}}{dx}.$$

Multiply by $e^{-\frac{1}{2}x^2}$ and integrate

$$\int e^{-\frac{1}{2}x^2}v_ndx = \int xe^{-\frac{1}{2}x^2}v_{n-1}dx - \int e^{-\frac{1}{2}x^2}\frac{dv_{n-1}}{dx}dx$$

* The proof is as follows:—

$$xv_{n-1} - (n-1)v_{n-2} = x^n - \frac{(n-1)(n-2)}{1!}\cdot\frac{x^{n-2}}{2} + \frac{(n-1)(n-2)(n-3)(n-4)}{2!}\cdot\frac{x^{n-4}}{2^2} + \dots$$

$$-(n-1)x^{n-2} + \frac{(n-1)(n-2)(n-3)}{1!}\cdot\frac{x^{n-4}}{2} - \dots$$

$$= x^n - \frac{n(n-1)}{1!}\cdot\frac{x^{n-2}}{2} + \frac{n(n-1)(n-2)(n-3)}{2!}\cdot\frac{x^{n-4}}{2^2} - \dots$$

$$= v^n$$

and integrating the latter integral by parts we have

$$\int e^{-\frac{1}{2}x^2} v_n dx = -e^{-\frac{1}{2}x^2} v_{n-1}$$

or
$$V_n = \frac{1}{\sqrt{2\pi}} \int_h^\infty v_n e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2} (v_{n-1})_{x=h}$$

Now, writing H for $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}h^2}$, and K for $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$, we have from ($a$)

$$\frac{d}{\mathrm{N}} = \frac{1}{2\pi} \int_h^\infty \int_k^\infty \mathrm{U}\, dx\, dy$$

$$= \frac{1}{2\pi} \int_h^\infty \int_k^\infty e^{-\frac{1}{2}(x^2+y^2)} dx\, dy + \overset{\infty}{\underset{1}{\mathrm{S}}} \left(\frac{r^n}{n!} \mathrm{HK} (v_{n-1})_{x=h} (w_{n-1})_{y=k} \right)$$

$$= \frac{(b+d)(c+d)}{\mathrm{N}^2} + \overset{\infty}{\underset{1}{\mathrm{S}}} \left(\frac{r^n}{n!} \mathrm{HK} (v_{n-1})_{x=h} (w_{n-1})_{y=k} \right)$$

or remembering that $\mathrm{N} = a + b + c + d$, we write

$$\frac{ad - bc}{\mathrm{N}^2 \mathrm{HK}} = \overset{\infty}{\underset{1}{\mathrm{S}}} \left(\frac{r^n}{n!} (v_{n-1})^{x=h} \times (w_{n-1})^{y=k} \right)$$

$$= r + \frac{r^2}{2} hk + \frac{r^3}{6} (h^2 - 1)(k^2 - 1) + \frac{r^4}{24} h(h^2 - 3) k(k^2 - 3)$$

$$+ \frac{r^5}{120} (h^4 - 6h^2 + 3)(k^4 - 6k^2 + 3)$$

$$+ \frac{r^6}{720} h(h^4 - 10h^2 + 15) k(k^4 - 10k^2 + 15)$$

$$+ \frac{r^7}{5040} (h^6 - 15h^4 + 45h^2 - 15)(k^6 - 15k^4 + 45k^2 - 15)$$

$$+ \&\mathrm{c.} \ldots \qquad \text{(viii.)}$$

APPENDIX IV.

Alternative Systems of Frequency-Curves.

This Appendix deals with some of the systems of frequency-curves which have been suggested instead of those of Chapter IV. Objection to Pearson's system has not been generally directed against its practical sufficiency, but has rather been founded on the contention that the theoretical basis of the system is insufficient. It is interesting to note in this connection that in a paper read before the Statistical Society this year, Professor F. Y. Edgeworth, who has himself suggested other methods, points out that Pearson's Generalized Probability Curve appears more justifiable in the matter of *à priori* justification the longer its philosophic basis is subjected to criticism. With these prefatory remarks, we may turn to the suggested methods.

I. *Method of Translation.*—As the normal curve has an approved theoretical basis, graduation might be effected by using $y=e^{-[f(x)]^2}$. This merely conceals the use of an absolutely general expression, and one still requires to know what forms are best for $f(x)$. It is hard to see why the normal curve should be held to be anything more than a first approximation to a general result. For a fuller account of this method, the reader may refer to F. Y. Edgeworth, *Journal of Statistical Society*, vol. lxi., pp. 675-689, or J. C. Kapteyn, "Skew Frequency-Curves in Biology and Statistics," Groningen, 1903.

II. *The use of half one normal curve for positive and half another normal curve for negative frequencies.*—Obviously, there is no theoretical basis, for each separate normal curve has its meaning based on certain assumptions, but the two parts become meaningless. The use of a part of a normal curve for a complete series is empirical; our own use of part of one on p. 95 is so too, but we did not adopt it for actual curve fitting, but as a hypothetical series of numbers. The method cannot give suitable curves for graduating the examples of Type II. or Type III., nor a curve rising abruptly from the axis of x.

III. *The use of the series* $y=A_0\phi(x)+A_3\phi'''(x)+A_4\phi^{iv}(x)+\ldots$ *where* $\phi(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-(x-b)^2/2\sigma^2}$.—The curve has been given by Edgeworth, *Camb. Phil. Trans.*, vol. xx., pp. 36–65, 113–141, C. V. L. Charlier, "Researches into the Theory of Probability," Meddelanden från Lunds Astronomiska Observatorium, Lund, 1906, and T. N. Thiele, "Theory of Observations," London, 1903, either in this form, or as

$$y=\phi(x)\{1+a_3[(x-b)^3-3c(x-b)]$$
$$+a_4[(x-b)^4-6c(x-b)^2+3c^2]+\ldots\}^*$$

which might be developed as $y=\phi(x)\{a_0+a_1x+a_2x^2\ldots\}$. Charlier uses the method of moments for fitting the curve and, using our notation, $b=\mu_1'$, $\sigma^2=\mu_2$, $A_3=-\frac{N\mu_3}{3!}$, $A_4=\frac{N\mu_4-3N\mu_2^2}{4!}$ &c.; he gives tables of $\sigma\phi(x)$, $\sigma^4\phi'''(x)$ and $\sigma^5\phi^{iv}(x)$, and writes his formula as

$$y=\frac{N}{\sigma}\left[\sigma\phi(x)-\frac{\mu_3}{\sigma^3 3!}\sigma^4\phi'''(x)+\frac{1}{4!}\left(\frac{\mu_4}{\sigma^4}-3\right)\phi^{iv}(x)+\ldots\right].$$

An example will be of help. Taking our example of Type IV., we find the mean is 44·5772339, $\sigma=2\cdot127818$, $\frac{N}{\sigma}=4302\cdot0$, $-\frac{\mu_3}{\sigma^3 3!}=\cdot012208$ and $\frac{1}{4!}\left(\frac{\mu^4}{\sigma_4}-3\right)=\cdot007079$, and using Charlier's tables, we obtain the following:—

Central Age x	$\frac{x-44\cdot577234}{5\sigma}$	First Term.	Second Term	Third Term	Sum of three previous cols. multiplied by 4302·0.
5	−3·7200	·0004	+·0002	+·0003	4
10	−3·2500	·0020	+·0006	+·0007	14
15	−2·7800	·0084	+·0013	+·0010	46
20	−2·3100	·0277	+·0018	−·0001	126
25	−1·8401	·0734	+·0006	−·0030	306
30	−1·3701	·1561	−·0030	−·0052	637
35	−·9002	·2661	−·0064	−·0023	1,108
40	−·4302	·3637	−·0054	+·0050	1,563
45	+·0397	·3986	+·0006	+·0084	1,753
50	·5097	·3503	+·0060	+·0037	1,548
55	·9797	·2468	+·0060	−·0032	1,075
60	1·4496	·1394	+·0022	−·0047	589
65	1·9196	·0632	−·0010	−·0025	256
70	2·3895	·0229	−·0018	+·0002	92
75	2·8595	·0067	−·0012	+·0010	29
80	3·3295	·0015	−·0005	+·0007	7
85	3·7984	·0003	−·0001	+·0003	2
90	4·3693	·0000	−·0000	+·0001	1
...	...	...	...	...	9,156

* Thiele obtains the equation by writing $e^{-(x-b)^2/2\sigma^2}=e^{-b^2/2\sigma^2}e^{xb/\sigma^2}e^{-\frac{1}{2}x^2/2\sigma^2}$, and then expands the last term by Maclaurin's theorem. Charlier in "Ueber das Fehlergesetz," Arkiv för Matematik, vol. ii., Stockholm, 1905, adopts a method which follows that of Laplace. Edgeworth gives more than one method of reaching the formula.

This graduation, which shows the method in a suitable application, gives a result which is less probable than the Type IV. graduation as judged by the test in Chapter IX., though the difference is entirely due to the bad agreement in the age 5 group. With the Type II. example the equation is

$$y=1214\cdot4\{\sigma\phi(x)-\cdot0081\sigma^4\phi'''(x)-\cdot01882\sigma^5\phi^{\text{iv}}(x)\}$$

which gives a fair result that is improved into an excellent one by omitting the middle term, but negative frequencies arise (*see* below). The formula fails to graduate distributions such as our example for Type I. or Example I. of Table I., and for these cases Charlier suggests the use of the series

$$\text{IV. } y=\mathrm{F}(xw+c)=\mathrm{B}_0\psi(x)+\mathrm{B}_1\Delta\psi(x)+\mathrm{B}_2\Delta^2\psi(x)+\ \ldots$$

where

$$\psi(x)=\frac{e^{-\lambda}\sin\pi x}{\pi}\left[\frac{1}{x}-\frac{\lambda}{1!\,(x-1)}+\frac{\lambda^2}{2!\,(x-2)}-\ \ldots\right]\text{ or }\frac{e^{-\lambda}\lambda^x}{x!}$$

when x is a positive integer (*cf.* Thiele, *Theory of Observations*, p. 21). Charlier's fitting of this curve is arbitrary as he gives four methods of solution according as we assume certain values for w, c, B_0, &c. Thus one solution is based on $w=1$ and $c=0$, while another, by assuming $\mathrm{B}_1=\mathrm{B}_2=\mathrm{B}_3=0$ finds λ, w, and c. He only gives two examples because two points in connection with its application have still to be cleared up; a third point to which he 'oes not refer is that a statistical criterion depending on the moments required to show which series is to be used and which solution f IV. is to be taken.

Apart, however, from these points the use of a series seems open to many objections; for if one of the later coefficients should have a large value the neglect of later terms may involve considerable error while the higher moments which are necessary to find these coefficients are untrustworthy owing to their large probable errors. The use of a limited number of terms of such series as those suggested, may lead to negative frequency which is objectionable from the practical point of view and hard to reconcile with sound theoretical treatment. Edgeworth's well-known curve

$$y=\frac{1}{c\sqrt{\pi}}e^{-x^2/c^2}\left\{1+2j\left(\frac{x}{c}-\frac{2}{3}\cdot\frac{x^3}{c^3}\right)\right\}$$

is merely the first two terms of the series III., and its inability to graduate distributions having considerable skewness is rather accentuated by Charlier's recent work in which he often finds the next term of the series significant.

APPENDIX V.

BOOKS, REFERENCES, &c.

The following list covers the principal papers to which reference has been made, though it is by no means a complete bibliography. It will, however, be found to cover those papers most likely to prove of value or interest to actuarial students.

THEORETICAL PAPERS, &c.

Biometrika (Editorials) :—

"On the Probable Errors of Frequency Constants." *Biom.*, vol. ii., pp. 273, *et seq.*

"Elementary Proof of Sheppard's Formulæ, &c." *Biom.*, vol. iii., pp. 308, *et seq.*

Blakeman, J. :—

"On Tests for Linearity of Regression in Frequency Distributions." *Biom.*, vol. iv., pp. 332, *et seq.*

Blakeman, J., and Pearson, K. :—

"On the Probable Error of Mean Square Contingency." *Biom.*, vol. v., pp. 191, *et seq.*

Davenport, C. B. :—

"Statistical Methods." New York: John Wiley & Sons; London: Chapman & Hall, 1904.

Galton, F. :—

"Correlations and their Measurement." *Proc. Roy. Soc.*, vol. xlv., pp. 136–145.

Pearson, Karl :—

"Skew Variation in Homogeneous Material." *Phil. Trans. A.*, vol. clxxxvi., pp. 343, *et seq.*, and a supplement in *Phil. Trans. A.*, vol. cxcvii., pp. 443–459.

"Regression, Hereditary and Panmixia." *Phil. Trans. A.*, vol. clxxxvii., pp. 253–318.

"On a Form of Spurious Correlation which may arise when Indices are used, &c." *Proc. Roy. Soc.*, vol. lx., pp. 489–498.

"On the criterion that a given system of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling." *Phil. Mag.*, July, 1900.

PEARSON, KARL (*continued*):—

"On the Correlation of Characters not quantitatively measurable." *Phil. Trans. A.*, vol. cxcv., pp. 1–47.

"On the Lines and Planes of Closest Fit to Systems of Points in Space." *Phil. Mag.*, November, 1901.

"On the Mathematical Theory of Errors of Judgment, with special reference to the Personal Equation." *Phil. Trans. A.*, vol. cxcviii., pp. 235–299.

"On the Influence of Natural Selection on the Variability and Correlation of Organs." *Phil. Trans. A.*, vol. cc., pp. 1–66.

"On a General Theory of the Method of False Position." *Phil. Mag.*, June, 1903.

"On the Theory of Contingency and its relation to Association and Normal Correlation." *Drapers' Company Research Memoir:* Dulau & Co., 1904.

"On the General Theory of Skew Correlation and Non-linear Regression." *Drapers' Company Research Memoir:* Dulau & Co., 1905.

"On the Systematic Fitting of Curves to Observations and Measurements." *Biom.*, vol. i., pp. 265, *et seq.*; and vol. ii., pp. 1, *et seq.*

PEARSON, KARL, AND FILON, L. N. G.:—

"On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation." *Phil. Trans. A.*, vol. cxci., pp. 229–311.

SHEPPARD, W. F.:—

"On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation." *Phil. Trans. A.*, vol. cxcii., pp. 101–167.

"On the Calculation of the Most Probable Values of the Frequency Constants for data arranged according to equidistant divisions of a scale." *Proc. Lon. Math. Soc.*, vol. xxix., pp. 353–380.

YULE, G. U.:—

"On the Significance of Bravais' Formulæ for Regression, &c., in the case of Skew Correlation." *Proc. Roy. Soc.*, vol. lx., pp. 477, *et seq.*

"On the Theory of Correlation." *Journal of Statistical Society*, vol. lx., pp. 812, *et seq.*

TABLES.

PROFESSOR PEARSON *informs us that he hopes to publish very shortly a volume of copyright tables for the use of statisticians, and it has therefore been decided not to include any tables in this volume except that of log $\Gamma(p)$.*

BARLOW'S TABLES OF SQUARES, CUBES, &c. E. & F. N. Spon.

ELDERTON, W. PALIN:—

Tables for Testing the Goodness of Fit of Theory to Observation. *Biom.*, vol. i., pp. 155, *et seq.*

Tables of Powers of Natural Numbers and of the Sums of the Powers of the Natural Numbers, from 1 to 100. *Biom.*, vol. ii., pp. 474, *et seq.*

GIBSON, W.:—

Tables for Facilitating the Computation of Probable Errors. *Biom.*, vol. iv., pp. 385, *et seq.*

LEE, A.:—

Tables of $F(r, \nu)$ and $H(r, \nu)$ Functions, $r=1$ to $r=50$, and $\phi^\circ=0^\circ$ to 45°. *Brit. Assocn. for Advancement of Science*, 1899.

SHEPPARD, W. F.:—

New Tables of the Probability Integral. *Biom.*, vol. ii., pp. 174, *et seq.*

There are also tables of the following functions in Davenport's "Statistical Methods":—

Smaller tables than Sheppard's of the Normal Curve of Error.

Table of log $\Gamma(p)$.

Table of first six powers of Natural Numbers from 1 to 30.

Small table of Probable Errors of r.

Squares, Cubes, Square Roots, Cube Roots and Reciprocals of numbers from 1 to 1,054.

Table to six decimal places of $\mathrm{L}\sin\theta$, $\mathrm{L}\tan\theta$, $\mathrm{L}\cos\theta$, $\mathrm{L}\cot\theta$.

Table of log $\Gamma(p)$.

p	0		1	2	3	4	5	6	7	8	9
1·00	...	...	9750	9500	9251	9003	8755	8509	8263	8017	7773
1·01	$\bar{1}$·99	7529	7285	7043	6801	6560	6320	6080	5841	5602	5365
1·02	$\bar{1}$·99	5128	4892	4656	4421	4187	3953	3721	3489	3257	3026
1·03	$\bar{1}$·99	2796	2567	2338	2110	1883	1656	1430	1205	0981	0757
1·04	$\bar{1}$·99	0533	0311	0089	9868	$\bar{9}$647	$\bar{9}$427	9208	8989	8772	8554
1·05	$\bar{1}$·98	8338	8122	7907	7692	7478	7265	7053	6841	6629	6419
1·06	$\bar{1}$·98	6209	6000	5791	5583	5376	5169	4963	4758	4553	4349
1·07	$\bar{1}$·98	4145	3943	3741	3539	3338	3138	2939	2740	2541	2344
1·08	$\bar{1}$·98	2147	1951	1755	1560	1365	1172	0978	0786	0594	0403
1·09	$\bar{1}$·98	0212	0022	$\bar{9}$833	$\bar{9}$644	$\bar{9}$456	$\bar{9}$269	9082	8896	8710	8525
1·10	$\bar{1}$·97	8341	8157	7974	7791	7610	7428	7248	7068	6888	6709
1·11	$\bar{1}$·97	6531	6354	6177	6000	5825	5650	5475	5301	5128	4955
1·12	$\bar{1}$·97	4783	4612	4441	4271	4101	3932	3764	3596	3429	3262
1·13	$\bar{1}$·97	3096	2931	2766	2602	2438	2275	2113	1951	1790	1629
1·14	$\bar{1}$·97	1469	1309	1150	0992	0835	0677	0521	0365	0210	0055
1·15	$\bar{1}$·96	9901	9747	9594	9442	9290	9139	8988	8838	8688	8539
1·16	$\bar{1}$·96	8390	8243	8096	7949	7803	7658	7513	7369	7225	7082
1·17	$\bar{1}$·96	6939	6797	6655	6514	6374	6234	6095	5957	5818	5681
1·18	$\bar{1}$·96	5544	5408	5272	5137	5002	4868	4734	4601	4469	4337
1·19	$\bar{1}$·96	4205	4075	3944	3815	3686	3557	3429	3302	3175	3048
1·20	$\bar{1}$·96	2922	2797	2672	2548	2425	2302	2179	2057	1936	1815
1·21	$\bar{1}$·96	1695	1575	1456	1337	1219	1101	0984	0867	0751	0636
1·22	$\bar{1}$·96	0521	0407	0293	0180	0067	$\bar{9}$955	$\bar{9}$843	$\bar{9}$732	$\bar{9}$621	9511
1·23	$\bar{1}$·95	9401	9292	9184	9076	8968	8861	8755	8649	8544	8439
1·24	$\bar{1}$·95	8335	8231	8128	8025	7923	7821	7720	7620	7520	7420
1·25	$\bar{1}$·95	7321	7223	7125	7027	6930	6834	6738	6642	6547	6453
1·26	$\bar{1}$·95	6359	6267	6173	6081	5989	5898	5807	5716	5627	5537
1·27	$\bar{1}$·95	5449	5360	5273	5185	5099	5013	4927	4842	4757	4673
1·28	$\bar{1}$·95	4589	4506	4423	4341	4259	4178	4097	4017	3938	3858
1·29	$\bar{1}$·95	3780	3702	3624	3547	3470	3394	3318	3243	3168	3094
1·30	$\bar{1}$·95	3020	2947	2874	2802	2730	2659	2588	2518	2448	2379
1·31	$\bar{1}$·95	2310	2242	2174	2106	2040	1973	1907	1842	1777	1712
1·32	$\bar{1}$·95	1648	1585	1522	1459	1397	1336	1275	1214	1154	1094
1·33	$\bar{1}$·95	1035	0977	0918	0861	0803	0747	0690	0634	0579	0524
1·34	$\bar{1}$·95	0470	0416	0362	0309	0257	0205	0153	0102	0051	0001
1·35	$\bar{1}$·94	9951	9902	9853	9805	9757	9710	9663	9617	9571	9525
1·36	$\bar{1}$·94	9480	9435	9391	9348	9304	9262	9219	9178	9136	9095
1·37	$\bar{1}$·94	9054	9015	8975	8936	8898	8859	8822	8785	8748	8711
1·38	$\bar{1}$·94	8676	8640	8605	8571	8537	8503	8470	8437	8405	8373
1·39	$\bar{1}$·94	8342	8311	8280	8250	8221	8192	8163	8135	8107	8080
1·40	$\bar{1}$·94	8053	8026	8000	7975	7950	7925	7901	7877	7854	7831
1·41	$\bar{1}$·94	7808	7786	7765	7744	7723	7703	7683	7664	7645	7626
1·42	$\bar{1}$·94	7608	7590	7573	7556	7540	7524	7509	7494	7479	7465
1·43	$\bar{1}$·94	7451	7438	7425	7413	7401	7389	7378	7368	7358	7348
1·44	$\bar{1}$·94	7338	7329	7321	7312	7305	7298	7291	7284	7278	7273
1·45	$\bar{1}$·94	7268	7263	7259	7255	7251	7248	7246	7244	7242	7241
1·46	$\bar{1}$·94	7240	7239	7239	7240	7241	7242	7243	7245	7248	7251
1·47	$\bar{1}$·94	7254	7258	7262	7266	7271	7277	7282	7289	7295	7302
1·48	$\bar{1}$·94	7310	7317	7326	7334	7343	7353	7363	7373	7384	7395
1·49	$\bar{1}$·94	7407	7419	7431	7444	7457	7471	7485	7499	7514	7529
p	0		1	2	3	4	5	6	7	8	9

Table of log Γ(p)—continued.

P	0		1	2	3	4	5	6	7	8	9
1·50	$\bar{1}$·94	7545	7561	7577	7594	7612	7629	7647	7666	7685	7704
1·51	$\bar{1}$·94	7724	7744	7764	7785	7806	7828	7850	7873	7896	7919
1·52	$\bar{1}$·94	7943	7967	7991	8016	8041	8067	8093	8120	8146	8174
1·53	$\bar{1}$·94	8201	8229	8258	8287	8316	8346	8376	8406	8437	8468
1·54	$\bar{1}$·94	8500	8532	8564	8597	8630	8664	8698	8732	8767	8802
1·55	$\bar{1}$·94	8837	8873	8910	8946	8983	9021	9059	9097	9135	9174
1·56	$\bar{1}$·94	9214	9254	9294	9334	9375	9417	9458	9500	9543	9586
1·57	$\bar{1}$·94	9629	9672	9716	9761	9806	9851	9896	9942	9989	0035
1·58	$\bar{1}$·95	0082	0130	0177	0225	0274	0323	0372	0422	0172	0522
1·59	$\bar{1}$·95	0573	0624	0676	0728	0780	0833	0886	0939	0993	1047
1·60	$\bar{1}$·95	1102	1157	1212	1268	1324	1380	1437	1494	1552	1610
1·61	$\bar{1}$·95	1668	1727	1786	1845	1905	1965	2025	2086	2147	2209
1·62	$\bar{1}$·95	2271	2333	2396	2459	2522	2586	2650	2715	2780	2845
1·63	$\bar{1}$·95	2911	2977	3043	3110	3177	3244	3312	3380	3449	3517
1·64	$\bar{1}$·95	3587	3656	3726	3797	3867	3938	4010	4081	4154	4226
1·65	$\bar{1}$·95	4299	4372	4446	4519	4594	4668	4743	4819	4894	4970
1·66	$\bar{1}$·95	5047	5124	5201	5278	5356	5434	5513	5592	5671	5740
1·67	$\bar{1}$·95	5830	5911	5991	6072	6154	6235	6317	6400	6482	6566
1·68	$\bar{1}$·95	6649	6733	6817	6901	6986	7072	7157	7243	7322	7416
1·69	$\bar{1}$·95	7503	7590	7678	7766	7854	7943	8032	8122	8211	8301
1·70	$\bar{1}$·95	8391	8482	8573	8664	8756	8848	8941	9034	9127	9220
1·71	$\bar{1}$·95	9314	9409	9502	9598	9693	9788	9884	9980	0077	0174
1·72	$\bar{1}$·96	0271	0369	0467	0565	0664	0763	0862	0961	1061	1162
1·73	$\bar{1}$·96	1262	1363	1464	1566	1668	1770	1873	1976	2079	2183
1·74	$\bar{1}$·96	2287	2391	2496	2601	2706	2812	2918	3024	3131	3238
1·75	$\bar{1}$·96	3345	3453	3561	3669	3778	3887	3996	4105	4215	4326
1·76	$\bar{1}$·96	4436	4547	4659	4770	4882	4994	5107	5220	5333	5447
1·77	$\bar{1}$·96	5561	5675	5789	5904	6019	6135	6251	6367	6484	6600
1·78	$\bar{1}$·96	6718	6835	6953	7071	7189	7308	7127	7547	7666	7787
1·79	$\bar{1}$·96	7907	8028	8149	8270	8392	8514	8636	8759	8882	9005
1·80	$\bar{1}$·96	9129	9253	9377	9501	9626	9751	9877	0003	0129	0255
1·81	$\bar{1}$·97	0383	0509	0637	0765	0893	1021	1150	1279	1408	1538
1·82	$\bar{1}$·97	1668	1798	1929	2060	2191	2322	2454	2586	2719	2852
1·83	$\bar{1}$·97	2985	3118	3252	3386	3520	3655	3790	3925	4061	4197
1·84	$\bar{1}$·97	4333	4470	4606	4744	4881	5019	5157	5295	5434	5573
1·85	$\bar{1}$·97	5712	5852	5992	6132	6273	6414	6555	6697	6838	6980
1·86	$\bar{1}$·97	7123	7266	7408	7552	7696	7840	7984	8128	8273	8419
1·87	$\bar{1}$·97	8564	8710	8856	9002	9149	9296	9443	9591	9739	9887
1·88	$\bar{1}$·98	0036	0184	0333	0483	0633	0783	0933	1084	1234	1386
1·89	$\bar{1}$·98	1537	1689	1841	1994	2147	2299	2453	2607	2761	2915
1·90	$\bar{1}$·98	3069	3224	3379	3535	3690	3846	4003	4159	4316	4474
1·91	$\bar{1}$·98	4631	4789	4947	5105	5264	5423	5582	5742	5902	6062
1·92	$\bar{1}$·98	6223	6383	6544	6706	6867	7029	7192	7354	7517	7680
1·93	$\bar{1}$·98	7844	8007	8171	8336	8500	8665	8830	8996	9161	9327
1·94	$\bar{1}$·98	9494	9660	9827	9995	0162	0330	0498	0666	0835	1004
1·95	$\bar{1}$·99	1173	1343	1512	1683	1853	2024	2195	2366	2537	2709
1·96	$\bar{1}$·99	2881	3054	3227	3399	3573	3746	3920	4094	4269	4443
1·97	$\bar{1}$·99	4618	4794	4969	5145	5321	5498	5674	5851	6029	6206
1·98	$\bar{1}$·99	6384	6562	6740	6919	7098	7277	7457	7637	7817	7997
1·99	$\bar{1}$·99	8178	8359	8540	8722	8903	9085	9268	9450	9633	9816
P	0		1	2	3	4	5	6	7	8	9

INDEX.

FREQUENCY CURVES

AND

CORRELATION.

ADDENDUM (WITH DIAGRAM) AND ERRATA.

BY

W. PALIN ELDERTON, F.I.A.

Printed for the Institute of Actuaries

BY

CHARLES AND EDWIN LAYTON,

FARRINGDON STREET,

LONDON, E.C. 4.

1917.

FREQUENCY CURVES AND CORRELATION.

ADDENDUM (WITH DIAGRAM) AND ERRATA.

SINCE 1906, when "Frequency Curves and Correlation" was published, further work has been done on the subject, and the following pages are put forward with the idea of bringing the book more up to date.

There is a warning in connection with probable errors dealt with in Chapter VIII which should be borne in mind in all statistical work. It is that we are not justified in assuming that the normal curve of error describes the distribution of errors in all cases. When the number of cases is large it is generally sufficiently accurate, but in many cases it is unsuitable and in some cases may lead to wrong conclusions. This may be inferred from the many peculiar shapes of frequency distributions found in general curve fitting and from the fact that actual samples of correlated material give all these extreme forms and show that the criticism is a practical one which must be borne in mind in practical work.

The study of deviations in these cases is still being developed; it is of some importance in actuarial work where the deviations with which we are concerned arise from a distribution which is by no means normal, *e.g.* $(p+q)^n$ where q the probability of dying in a year is a very small quantity.

A list of mistakes in "Frequency Curves and Correlation" is also given. I apologize for these mistakes and thank the many friends who have helped me by calling my attention to them. I am afraid a few mistakes may still lurk undetected in the book; if so, I can only hope that they will cause little inconvenience to readers.

I.—UNCOMMON FREQUENCY TYPES.

In the course of statistical work a distribution is sometimes found which appears different in its algebraic form from the usual types, but can nevertheless be described accurately by those types. An example which will give an indication of the kind of case we have in mind is a distribution arising from recording the number of sequences in coin-tossing or dice-throwing experiments: the distribution is a geometric progression and this, although a well-known result in probability, is not at first sight a necessary form of Pearson's series of frequency curves. The expression, however, for his Type III is $e^{-\gamma x}\left(1+\frac{x}{a}\right)^p$ and if p be put equal to zero, we obtain the exponential $e^{-\gamma x}$ which gives the series we want. This possible limit was casually pointed out by me in "Frequency Curves and Correlation," but I did not investigate generally the circumstances in which this happened, nor did I appreciate that other interesting special cases might arise with other types. A recent paper* by Prof. Pearson deals with various special cases which are so different from the curves to which we are accustomed that he has dealt with them as new types. These types arise in certain limiting cases of Types I, II and VI, and give straight lines, curves starting with an infinite and ending with a finite ordinate, two separated blocks of frequency, and curves starting at a finite ordinate and ending at zero either at a finite point or at infinity: among these last is, of course, the exponential to which we have already referred.

Before turning to the expressions for these new types it may be useful to give a table of various peculiar distributions that have been obtained from insurance and other material.

Examples of uncommon Frequency Types.

469	45	119	4,165	33	68
186	38	100	2,028	53	24
166	46	86	982	65	17
134	53	75	480	81	14
122	43	61	266	101	12
112	38	50	132	131	11
...	49	39	71	186	10
...	41	27	36	350	10
...	44	22	17	...	10
...	52	12	9	...	11
...	...	3	2	...	12
...	...	...	1	...	20
...	...	...	1	...	...
...	...	...	1	...	...
...	...	...	1	...	...
1,189	449	594	8,192	1,000	219

* *Phil. Trans. A.*, vol. ccxvi, pp. 429–457.

This table shows distributions which at first sight have little likeness to the frequency curves with which we are familiar; they can, however, be fitted with reasonable accuracy by Pearson's generalized curve. Some of them are rather like the J-shaped curve, of which examples are given on pp. 66, 104 of "Frequency Curves and Correlation", but others have little resemblance even to this extreme type. The table includes (col. 6) areas of a U-shaped curve which is rare; in fact, I have not succeeded in finding a suitable distribution of this shape among actuarial statistics, but such a distribution might occur among terminations (including withdrawals) in term policies of ten years, say, or similar endowment assurances.

It will now be convenient to take each curve in order and give the formulæ to enable it to be fitted to the statistics, the circumstances in which it arises (*i.e.*, the criterion), and numerical examples. It may be remarked that the types can be easily picked out from a diagram given in Prof. Pearson's paper; this diagram sets out for all possible values of β_1 and β_2 the type that should be used.

Type VIII.

$$y=y_0\left(1+\frac{x}{a}\right)^{-m}$$

Range from an infinite ordinate at $-a$ to a finite ordinate, y_0, at 0.

m is found from the solution of

$$m^3(4-\beta_1)+m^2(9\beta_1-12)-24\beta_1 m+16\beta_1=0$$

and must be not <0 or >1.

$$a=\pm\sigma(2-m)\sqrt{\left(\frac{3-m}{1-m}\right)}$$

$$y_0=N(1-m)\div a$$

The distance of the mean from $x=-a$ is $a(1-m)\div(2-m)$. When μ_3 is positive a is negative.

The curve is a special case of Type I when m_2 is zero, that is, when

$$r-2=r(r+2)\sqrt{\{\beta_1\div[\beta_1(r+2)^2+16(r+1)]\}}$$

where $\qquad r=6(\beta_2-\beta_1-1)\div(6+3\beta_1-2\beta_2)$

that is when

$$\frac{(4\beta_2-3\beta_1)(10\beta_2-12\beta_1-18)^2-\beta_1(\beta_2+3)^2(8\beta_2-9\beta_1-12)}{(3\beta_1-2\beta_2+6)\{\beta_1(\beta_2+3)^2+4\beta_1(4\beta_2-3\beta_1)(3\beta_1-2\beta_2+6)\}}\text{ or }\lambda\text{ say}$$

is zero.

The criteria for Type VIII can be reduced to (1) special case of Type I, (2) $\lambda=0$, (3) $5\beta_2-6\beta_1-9$ is negative. It may be added that $24\beta_2-27\beta_1-38$ is small; theoretically positive.

If $\beta_1=0$ an interesting special case arises, in which $m=0$, and the curve becomes a horizontal line, which is also the limit of Types IX and XII.

The solution of the cubic for m gives trouble. m can also be found from $m=-2(5\beta_2-6\beta_1-9)\div(3\beta_1-2\beta_2+6)$, and though this involves β_2 it should theoretically give the same value of m as the cubic. As the criterion is not exactly reached in practice the two results differ, and it seems preferable to find m from the cubic by using

$$m=\frac{24\beta_1-\sqrt{\{24^2\beta_1^2-64\beta_1(m'(4-\beta_1)+9\beta_1-12)\}}}{2(m'(4-\beta_1)+9\beta_1-12)}$$

where m' is found from the expression in β_1 and β_2 given above or by some other trial method.

An alternative is to find from the criteria or from the diagram at the end of Pearson's paper the value of β_2 which is the consequence of the particular value of β_1 when a Type VIII curve occurs, and use this theoretical value in finding m instead of the β_2 given by the actual statistics.

*Example.**

Frequency	Graduation (1)	Graduation (2)
469	437	436
186	222	209
166	165	161
134	136	141
122	120	127
112	109	115
1,189	1,189	1,189

The mean is ·65518 of an interval after the centre of the 186 group. The constants were

μ_2	2·986
μ_3	3·295
μ_4	18·252
β_1	·408
β_2	2·047
λ	$-$·05

* In the numerical work, with a few exceptions, I used more figures than are shown.

$5\beta_2-6\beta_1-9$ negative. Hence Type VIII can be used.

m ·500

a $-5·797$

y_0 102·6

The curve runs from ·277 before the middle of the first to ·02 after the end of the last group. The graduation is shown (No. 1) above.

The areas can be calculated by the expression

$$y_{-r}\times(a-r)\div(1-m)$$

which gives the area of the remainder from $-r$ to $-a$. In the particular case the range could be fixed as 6 as the data related to six months' experience of maturities among endowment assurances, and remembering that the mean is $a(1-m)\div(2-m)$ we found

m ·439

y_0 111·1

The areas resulting are given in graduation (2). The following table gives the calculation of the areas in this case. The equation to the curve is

$$y=111·1\left(1-\frac{x}{6}\right)^{-·439}$$

with range from 0 to 6, a being negative because μ_3 is positive.

x (1)	$1-x/6$ (2)	Colog (2) (3)	(3) × m (4)	(4) + log 111·1 = log y_x (5)	Log $\frac{y_x}{1-m}$ (6)	Antilog (6) (7)	Remainder of range (8)	(7) × (8) (9)	Area required (10)
0	...	...	...	...	...	...	6	1,189	115
1	·8333	·0792	·0348	2·0804	2·3318	214·7	5	1,074	127
2	·6667	·1761	·0774	2·1230	2·3744	236·8	4	947	141
3	·5000	·3010	·1323	2·1779	2·4293	268·7	3	806	161
4	·3333	·4815	·2118	2·2574	2·5088	322·7	2	645	209
5	·1667	·7781	·3419	2·3875	2·6389	435·5	1	436	436

Some of the columns can be dispensed with: they are shown in detail to make the method clear.

Both graduations are reasonably close to the facts.

An example of the limiting case will be found in the following statistics:

No.	Frequency	Graduation	Theoretical
1	45	42	45
2	38	45	45
3	46	45	45
4	53	45	45
5	43	45	45
6	38	45	45
7	49	45	45
8	41	45	45
9	44	45	45
10	52	48	45
...	449	450	450

The mean is ·57 after the middle of the 5 group ; the moments are

$$\begin{array}{ll} \mu_2 & 8{\cdot}374 \\ \mu_3 & {\cdot}026 \\ \mu_4 & 124{\cdot}46 \\ \beta_1 & {\cdot}011 \\ \beta_2 & 1{\cdot}78 \end{array}$$

Hence
$$y = 449 \div 2\sqrt{3\mu_2}$$
$$= 44{\cdot}8$$

The range is from ·57 to 10·43.

The series was found by summing in tens the last figure of arlisle $3\frac{1}{2}$ per-cent Table of A_x and the mean should be at 5·5 .heoretically instead of 5·57 and y should be 45. The range should be ·5 to 10·5. The example is interesting as showing how the Pearson-curves graduate in an extreme case. The "graduation" and theoretical results are shown. In the "graduation" decimals have been neglected.

Type IX.

$$y = y_0\left(1 + \frac{x}{a}\right)^m$$

Range from $x = -a$ where $y = 0$ to $x = 0$ where $y = y_0$

$$a = \pm\sigma(m+2)\sqrt{\left(\frac{m+3}{m+1}\right)}$$

m is found by solving

$$m^3(\beta_1-4)+m^2(9\beta_1-12)+24m\beta_1+16\beta_1=0$$

$$y=\mathrm{N}(m+1)\div a$$

The distance of the mean from $x=-a$ is $a(m+1)\div(m+2)$.

As in Type VIII the value of m can be found by simplifying the cubic into a quadratic, or by the other method indicated.

The criteria are reached through the same equation as those for Type VIII, and can be reduced to (1) special case of Type I, (2) $\lambda=0$, (3) $5\beta_2-6\beta_1-9$ is positive, (4) $2\beta_2-3\beta_1-6$ is negative.

If $\beta_2=2{\cdot}4$ and $\beta_1={\cdot}32$ the curve becomes a sloping line

$$y=\frac{\sqrt{2}\mathrm{N}}{3\sigma}\left(1+\frac{x}{3\sqrt{2}\sigma}\right)$$

If $\beta_1=0$ we reach a horizontal line as the limit, while if $\beta_2=9$ and $\beta_1=4$, we have the other limit of Type IX, and find the exponential series (Type X).

Example.

Duration	Exposed to Risk in Annuity Experience	Type IX	Frequency Line
0	119	118	108
1	100	98	97
2	86	85	86
3	75	74	76
4	61	63	65
5	50	52	54
6	39	41	43
7	27	30	32
8	22	20	22
9	12	11	11
10	3	2	0
...	594	594	594

The mean is at 2·909 assuming the exposed to risk to be an ordinate at the duration or an area from $n-\frac{1}{2}$ to $n+\frac{1}{2}$.

μ_2	6·27
μ_3	10·99
μ_4	102·50
β_1	·490
β_2	2·606
$5\beta_2-6\beta_1-9$	positive
$2\beta_2-3\beta_1-6$	negative

The curve is not far from Type IX, and if β_1 had been ·32 and β_2 2·4, we should have reached a straight line $y=111{\cdot}8\left(1-\frac{x}{10{\cdot}63}\right)$ with range from $-{\cdot}6$ to 10·0 and obtained the graduation shown. The whole area $-{\cdot}6$ to $+{\cdot}5$ is taken as the frequency for duration 0. Using Type IX, the following constants are reached:

m	1·123
a	$-10{\cdot}913$
y_0	115·54

The curve runs from $-{\cdot}586$ to 10·3275. The 118 in the first group has been taken as the area from $-{\cdot}586$ to $+{\cdot}5$. Theoretically there cannot be an exposed before duration $-{\cdot}5$, but as we are merely giving an example of fitting a curve to a series of numbers this need not concern us. The difficulty could be met by fitting a system of ordinates or by assuming a starting point for the curve.

If m happens to be less than unity the shape of the curve is somewhat different, *e.g.*, if $y=100\left(1+\frac{x}{10}\right)^{\cdot 25}$ we have the following ordinates:

100, 98, 95, 91, 88, 84, 79, 74, 67, 56, 0.

The actual deaths in a select mortality experience may take this form, but the shape of the curve will be less flat at the start, *e.g.*, in the recent American Medico-Actuarial experience age group 30–34.

Type X.

$$y=\frac{\mathrm{N}}{\sigma}e^{-x/\sigma}$$

Range from 0 to ∞.

Distance of origin from the mean is σ.

The curve is a special case of Type III when $\gamma a=0$, that is when $\beta_1=4$.

The condition for Type III is better stated by $2\beta_2 = 6 + 3\beta_1$ than by $\kappa = \infty$ as mentioned in Frequency Curves, p. 50, &c. Hence the exponential form is given by $\beta_1 = 4$ and $\beta_2 = 9$. The curve is also the limit of Types IX and XI.

Example.

	Frequency	Graduation	Theoretical
1	4,165	4,132	4,096
2	2,028	2,016	2,048
3	982	1,015	1,024
4	480	511	512
5	266	257	256
6	132	130	128
7	71	65	64
8	36	33	32
9	17	17	16
10	9	8	8
11	2	4	4
12	1	2	2
13	1	1	1
14	1	1	1
15	1	...	...
...	8,192	8,192	8,192

The unadjusted mean is 2·0087.

μ_2	2·045
μ_3	6·290
μ_4	39·720
β_1	4·629
β_2	9·502
σ	1·43

When the curve is an exponential the moments and mean require adjustment, but the Sheppard high contact adjustments are, of course, unsuitable. If the curve starts at the beginning of the first group, I think that the mean is overstated when μ_3 is positive by $\frac{1}{12\sigma}$ approximately, and the second moment about the true mean is understated by $\frac{1}{12}$ approximately.*

* The adjustments for the moments for J and U-shaped curves call for consideration. I think something might be done by assuming an exponential form. But I have not had an opportunity of going sufficiently into the point. It is clear that unadjusted moments will not always give entirely satisfactory results. The student should also remember that when the β's are large their probable errors are large too.

Making use of the adjustments the mean is now 1·934, μ_2 is 2·123, and σ is 1·457.

The statistics relate to sequences in coin-tossing and the theoretical figures are added. In the statistics as published the sequence of 11, 12, &c., were 2, 1, 0, 1, 0, 2. In calculating the graduated areas of the curve it is useful to remember that the area from a to b is $(y_a - y_b)\sigma$.

It is interesting to notice how the "graduation" keeps closer to the frequency than the theoretical result.

I give as a second example the following series based on cricket scores known to start at the beginning of the first group:

Score ...	0–19	20–	40–	60–	80–	100–	120–	140–	160–
Series ...	64	34	18	9	6	3	3	0	0
Graduation	64	34	18	10	5	3	1	1	1

The ratio of each term to the preceding is ·54, and the graduation is almost exact. Owing, however, to the 3 at the group 120, the moments give a criterion considerably removed from the theoretical $\beta_1 = 4$, $\beta_2 = 9$.

If we had assumed the start of the curve in the previous example, we should have reproduced the theoretical result almost exactly.

Type XI.

$$y = y_0 x^{-m}$$

Range from $x = b$ where $y = y_0 b^{-m}$ to $x = \infty$ where $y = 0$.

m is found from

$$m^3(4 - \beta_1) + m^2(9\beta_1 - 12) - 24\beta_1 m + 16\beta_1 = 0$$

$$b = \pm\sigma(m-2)\sqrt{\left\{\frac{m-3}{m-1}\right\}}$$

$$y_0 = \mathrm{N}b^{m-1}(m-1)$$

The distance of mean from origin is $b(m-1) \div (m-2)$.

As in Type VIII m can be found by simplifying the cubic into a quadratic or by the other method indicated.

m may have any value from 5 to ∞, but in practice its value is not less than 9.

The curve is a special case of Type VI when $q_2=0$.

The criteria can be expressed as (1) special case of Type VI, (2) $\lambda=0$, (3) $2\beta_2-3\beta_1-6$ is positive.

Example.

Duration	Withdrawals	Graduation by XI
0	165	183
1	65	53·9
2	23	32·6
3	32	20·0
4	13	12·4
5	8	7·6
6	1	4·9
7	6	3·1
8	3	1·9
9	3	1·2
10	1	·8
11	3	1·6
...	323	323

I have not come across a distribution really represented by this type, but I give an unsuccessful attempt to apply it to a series of withdrawals. The constants were

β_1	4·97
m	29·69
b	57·14
$\log y_0$	54·3563
distance of mean from origin	59·205

In calculating areas we use $y_0 a^{-(m-1)} \div (m-1)$ as the area from a to ∞.

Type XII.

$$y=y_0\left(\frac{\sigma(\sqrt{3+\beta_1}+\sqrt{\beta_1})+x}{\sigma(\sqrt{3+\beta_1}-\sqrt{\beta_1})-x}\right)^{\sqrt{\frac{\beta_1}{3+\beta_1}}}$$

Range from $x=\sigma(\sqrt{3+\beta_1}-\sqrt{\beta_1})$ to $x=-\sigma(\sqrt{3+\beta_1}+\sqrt{\beta_1})$.

The origin is at the mean.

$$y_0 = \frac{N}{b\Gamma(m+1)\Gamma(1-m)}$$

where $m = \sqrt{\frac{\beta_1}{3+\beta_1}}$ and $b = 2\sigma\sqrt{3+\beta_1}$.

When μ_3 is positive the negative sign is taken for the square roots.

The limit of the curve when $\beta_1 = 0$ is a horizontal line.

The criterion is $5\beta_2 - 6\beta_1 - 9 = 0$.

Example.

Frequency	Graduation	Ordinates
...	2	18·5
33	31	31·6
...	...	40·6
53	49	49·3
...	...	57·0
65	65	65·2
...	...	73·4
81	83	82·4
...	...	92·3
101	103	103·3
...	...	114·6
131	134	132·9
...	...	155·5
186	191	186·0
...	...	244·2
350	342	405·4
1,000	1,000	...

The mean is ·051 after the centre of the 131 group. The constants are

μ_2	4·266
μ_3	−7·688
μ_4	48·154
β_1	·761
β_2	2·646
$5\beta_2 - 6\beta_1 - 9$	·368

$$y = 87{\cdot}2\{(5{\cdot}808 + x) \div (2{\cdot}204 - x)\}^{\cdot 45}$$

In addition to the graduation a number of equidistant ordinates is given. They show that the curve rises abruptly, then less abruptly and then again more abruptly. The withdrawals in select tables are sometimes of this shape (*e.g.*, Japanese experience age 52, females). A somewhat similar twist occurs in a population curve. The twist is found in Type I where both m_1 and m_2 are numerically less than unity and one is negative. Type XII is a particular case.

U-shaped Curve.

This shape arises in Type I when m_1 and m_2 (or m in Type II) are negative. There are difficulties in fitting it to statistics because we do not know how to adjust the rough moments. The figures given in the table of examples (p. 2, col. 6) were found by calculating the areas of the curve

$$y=10\left(1+\frac{x}{8}\right)^{-\cdot 7}\left(1-\frac{x}{4}\right)^{-\cdot 35}$$

The limit of the U-curve is two separate blocks of frequency at the ends of the range. This limit is reached when

$$\beta_2-\beta_1-1=0.$$

The curves with which we have dealt are rare and in practical curve-fitting may be avoided, for they depend on certain definite values of β_1 and β_2, and the chance of reaching these exact values is negligible. In other words, if the object of fitting a curve in any particular case is to obtain the closest agreement between the actual figures given and the graduated figures, then Types I, IV and VI are all that are necessary, for the other types being transition types and depending on specific values of β_1 and β_2 need not arise. If, however, our object is to study probability in a wider sense, the transition types are of importance and they may, of course, be properly used when the values of the β's only differ from those indicated by the criteria to a small extent. This "small extent" means within the limit suggested by the probable errors of the β's.

I may here deal with a little difficulty that students sometimes encounter in connection with the Types I and VI which can be expressed in the same algebraic form, and may more obviously be found with Types VIII, IX and XI, all of which can be written in the form hx^k, and the question may be asked why we should not fit hx^k from a to b and find h, k, a and b from the equations for the moments. The answer is that the criteria afford in effect a simplification of the equations and automatically tell us a good deal about the value of the constants and the range of the curve.

The curves arising from the equation

$$\frac{d \log y}{dx} = \frac{a_0 + a_1 x}{b_0 + b_1 x + b_2 x^2}$$

give us a means of graduating frequency distributions that vary from two separated lumps of frequency, through U-shaped curves to J-shaped curves, then to the "cocked hat" shaped with which we are familiar; incidentally they give straight lines, the exponential and the Gaussian curve.

In the paper dealing with the curves of which we have given examples Pearson indicates that the whole range of these shapes can be obtained from samples from a large population in which each individual carries any number of characteristics which are correlated together. If the correlation be calculated from m samples of equal size we shall not reach the correlation r known to exist in the whole population, but shall have a frequency distribution giving the number of samples that show various values of the correlation between -1 and $+1$. The distributions ill vary with the size of the sample and the value of r in the hole population and these distributions, although they do not sually lend themselves to expression by simple functions, give ne various shapes that are found from the frequency curve system we have discussed.

II.—Calculation of the Coefficient of Correlation in a Two-row Table.

The method of calculating the coefficient of correlation when both the variates are continuous and quantitative is dealt with in Chapter VI of "Frequency Curves and Correlation" and the method to be used when the variates are not quantitatively measurable is given in Chapter VII. There are, however, intermediate cases where one variate is, and the other is not, quantitatively measurable, as, for instance, in the following table relating to the effect of enlarged glands on the weight of children (boys).*

* The method is given by Prof. Pearson in *Biometrika*, Vol. VII, pp. 96 *et seq.*, and the example is taken from that paper.

Weight	Boys with good glands	Boys with bad glands	Total
14	2	...	2
16	3	5	8
18	15	26	41
20	20	40	60
22	28	47	75
24	34	30	64
26	30	31	61
28	29	20	49
30	30	30	60
32	21	14	35
34	18	11	29
36	18	5	23
38	6	7	13
40	5	2	7
42	7	3	10
44	1	...	1
46	...	...	...
48	3	...	3
50	1	...	1
52	1	...	1
...	...	...	...
62	2	...	2
Total	274	271	545

If the reader considers a volume of frequency built out of a complete table such as that for endowment assurances on p. 107 of "Frequency Curves and Correlation," or out of a correlation table giving relative ages of husbands and wives, he will see that he has a complete distribution. Now if a volume of frequency be cut off from such a complete volume by a vertical plane at a given value of one variate, then the vertical through the centroid of this volume cuts the regression line. The vertical plane in the two-row table is at the division of the rows; in our example where the good glands end and the bad ones begin. If $\bar{p}$ and $\bar{q}$ be the co-ordinates of the point of section where the vertical through the centroid of the volume cut off cuts the regression line, then we have, σ_1 and σ_2 being the standard deviations of the two variates and r the correlation

$$\bar{p}=r\frac{\sigma_1}{\sigma_2}\bar{q}$$

or

$$r=\frac{\bar{p}}{\sigma_1}\div\frac{\bar{q}}{\sigma_2}$$

Now $\bar{p}$ is the mean value of the quantitatively measurable variate for all the pairs with a certain one of the alternative variates, in our example, the mean weight of boys with bad glands and σ_1 is the standard deviation of all the boys. We cannot calculate $\bar{q}$ and σ_2 in a similar way because they relate to glands of which no quantitative measure is available. If we assume the non-measurable variate (glands) to follow the normal probability distribution the proportion of the non-measurable variate gives, with the help of tables of the probability integral, the ratio of $\frac{y}{\sigma_2}$ for the distance from the mean at which the division of this variate occurs, and then

$$\frac{\bar{q}}{\sigma_2} = \frac{N}{\sqrt{2\pi\sigma_2^2}}\int_y^\infty ye^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}dy \div \frac{N}{\sqrt{2\pi\sigma_2}}\int_y^\infty e^{-\frac{1}{2}\frac{y^2}{\sigma_2^2}}dy$$

$$= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y/\sigma_2)^2} \div \frac{1}{\sqrt{2\pi}}\int_{y/\sigma_2}^\infty e^{-\frac{1}{2}y^2}dy$$

The numerator is z and the denominator $\frac{1}{2}(1-a)$ in the notation used in Sheppard's tables of the probability integral.

The working of the numerical example may help to make the method clearer; it is as follows:

The mean weight of all the boys is ... ... 27·7522
The standard deviation is ... ... 6·7502
The mean weight of boys with bad glands is ... 27·3737

$\frac{1}{2}(1-a) = 271 \div 545 = \cdot4972$

$\frac{1}{2}(1+a) = \cdot5028$ and this value corresponds with

$z = \cdot3989$ in Sheppard's tables.

The correlation of good glands and weight is

$$\frac{(27{\cdot}7522 - 27{\cdot}3737)}{6{\cdot}7502} \div \frac{\cdot3989}{\cdot4972}$$
$$= \cdot070$$

III.—The Correlation Ratio.

The correlation coefficient which is discussed in Chapters VI and VII of "Frequency Curves and Correlation" is a satisfactory measure of correlation when the regression line is a straight line, that is when the means of the columns (and the means of the rows) are in a straight line; but in other circumstances its use is open to objection. In Chapter X another method was described which is not open to the same objection, and it may now be of interest to add a short note on a function known as the "correlation ratio" (η) which is a useful measure in many cases.

The value of η is given by

$$\eta^2 = \frac{S\{n_x(\bar{y}-\bar{y}_x)^2\}}{N\sigma_y^2}$$

where $\bar{y}_x$ is the mean of the y's corresponding to the particular array x, n_x is the number of cases in the array x and $\bar{y}$ is the mean of all the y's. The summation extends over all the arrays. In a similar way we can work from the y-arrays and have

$$\eta^2 = \frac{S\{n_y(\bar{x}-\bar{x}_y)^2\}}{N\sigma_x^2}$$

These values of η will not be the same except in the limiting case when regression is linear. The correlation ratio can alternatively be expressed as the ratio of the standard deviation of the means of the y-arrays, each array being weighted with the number in it, to the standard deviation of the y's. The values of η vary between 0 and 1 and are equal to those of r when regression is linear.

Taking the example on p. 107 of "Frequency Curves and Correlation" we should find η as follows:

Mean unexpired term in each column $\bar{y}_x$	Deduct mean of whole (20·312) $\bar{y}_x-\bar{y}$	$(\bar{y}_x-\bar{y})^2$	n_x	$n_x(\bar{y}_x-\bar{y})^2$
10·333	−9·979	99·6	6	598
13·250	7·062	49·9	4	200
13·176	7·136	50·9	17	865
16·113	4·199	17·6	62	1,091
17·230	3·082	9·50	584	5,548
20·141	·171	·029	643	19
21·877	+1·565	2·45	1,098	2,690
21·665	1·353	1·83	388	710
21·500	1·188	1·42	60	85
27·625	7·313	53·4	8	427
...	...	...	2,870	12,233

$$\therefore \qquad \eta^2 = \frac{12233}{2870 \times (7{\cdot}6067)^2}$$

$$\text{or} \qquad \eta = {\cdot}271$$

The figure 7·6067 is the value of σ_2 on p. 118, multiplied by 5 the unit of grouping. The probable error is approximately ·012 being calculated as $\cdot 67449 \frac{1-\eta^2}{\sqrt{N}}$, which is an accurate expression when regression is linear and can be used as a guide in cases which do not depart widely from this condition.

Working similarly with the maturity ages we obtain the following :

$\bar{x}_y - \bar{x}$	$(\bar{x}_y - \bar{x})^2$	$n_x(\bar{x}_y - \bar{x})^2$
3·48	12·11	678
2·20	4·84	832
1·38	1·90	821
·64	·41	273
·35	·12	82
·65	·42	227
2·71	7·34	1,813
3·81	14·52	1,117
5·27	27·77	222
7·77	60·37	60
...	...	6,125

$$\therefore \qquad \eta^2 = \frac{6125}{2870 \times (5{\cdot}6818)^2}$$

$$\text{or} \qquad \eta = {\cdot}257 \pm {\cdot}012$$

It may be of interest to set out the various values found as measures of the correlation of the same table :

Correlation coefficient	·254
Correlation ratio (first value)	·271
Correlation ratio (second value)	·257
Contingency method (Chapter X)	·287

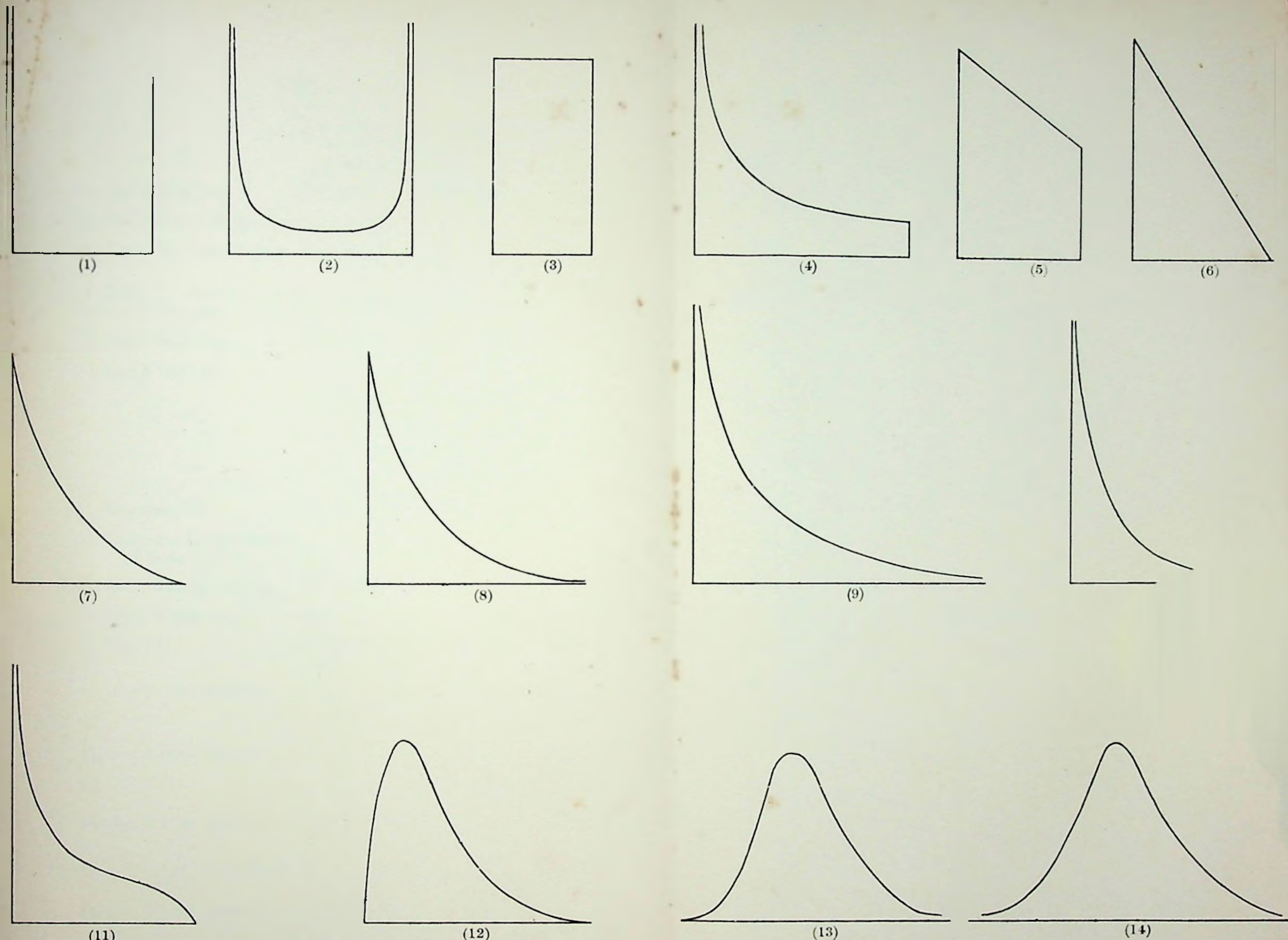


Diagram showing transition from two separated blocks of frequency through U-*shaped and* J-*shaped curves to the more common "cocked hat" shapes.*

(1) *represents two separated blocks of frequency developing into the* U-*shaped curve in* (2). *The horizontal straight line of* (3) *is, as it were, the bottom piece of the* U-*curve and the Type* VIII *curve of* (4) *is like a part of the* U-*curve.* (5) *and* (6) *are limits when straight lines are reached.* (7) *is Type* IX *and* (8) *is the exponential. The next two curves* (9) *and* (10) *are the* J-*shaped curves of Types* III *and* I, *and* (11) *is Type* XII. *From this we proceed to Types* I, III, IV, V *and* VI, *curves of the "cocked hat" shape; three examples being given in* (12), (13) *and* (14).

ERRATA, &c.

PAGE	
15, § 3	The value of a should be -23, and the resulting values of y should be increased by 2·5
18, line 4 from top ...	For $\Sigma y_r X_r$ read $\Sigma y_r X_r^2$
26, line 5 from bottom ...	For y_n read y_{n-1}
27, Table V, first column	For 35·58 read 35·42, and for 8·26 read 8·22
27, Table V, total, and 28, line 3 from top ...	For 208·38 read 207·18
30, line 8 from top ...	After IV add "of Table I"
30, line 9 from top ...	After "contact", delete remainder of sentence, and substitute "and in "I the rough moment should be "used. In II and V there is more "doubt, and in the calculation of "the moments for II (*see* p. 55) no "adjustment was made."
31, Equation II	For $\int l^7$ and $\int l^6$ read $\int l^5$
33, heading of third column of table	For p. 24 read p. 27
37, line 10 from bottom...	For n read $n+1$
39, line 15 from top ...	Delete $y=0$
40, line 16	Insert a *minus* sign before the right-hand side of the equation
43, line 9 from bottom ...	Insert b_2 as a divisor for the right-hand side and as a divisor for the indices in the following line
43, line 4 from bottom ...	The suffix 2 has been dropped in some copies in the index to the second bracket
44, line 6 from bottom ...	Insert $\frac{1}{b_2}$ before second integral
44, lines 4 & 5 from bottom.	For $\frac{c}{A}$ read $\frac{c}{Ab_2}$
44, line 4 from bottom ...	For $2b_2$ read $\frac{1}{2b_2}$
45, line 9 from top ...	The sign after the first bracket should be *minus*, and consequently the sign of the corresponding index in the two following lines should also be negative

PAGE

45, line 10 from top ... For b read b_2; and for $\frac{b_1}{b_2}$ read $\frac{b_1}{2b_2}$

52 ... ... The fourth moment

$$\nu_4 = \nu'_4 - 4d\nu_3 - 6d^2\nu_2 - d^4$$

59, line 3 from top ... In some copies the suffix, 1, to the lower limit of the integral and the *minus* sign before the limit have been dropped

59, lines 4, 5 & 12 from top. For $z - 1$ read $1 - z$

60, line 5 from top ... For $-2r^2$ read $+2r^2$

60, line 10 from bottom... The denominator should be $3\beta_1 - 2\beta_2 + 6$

63, heading to col. (5) ... For y_0 read $\log y_0$

72, heading to col. (7) ... Insert *minus* sign before m

74, line 6 from top ... For $\cos^{2m\theta}$ under integral read $\cos^{2m}\theta$

74, line 9 from bottom ... For $\phi + \frac{1}{2}\pi = \theta$ read $\frac{\pi}{2} = \theta + \phi$

74, line 2 from bottom ... For $\cos^{r+n+1}\theta$ read $\cos^{r-n+1}\theta$

76, line 9 from top ... For $\frac{8r}{z}$ read $\frac{8r^2}{z}$

77, last line ... ... The numerator should be $2^r\pi e^{\frac{1}{2}\pi\nu}$

83 ... ... Instead of lines 5 and 6 read

q_2 and $-q_1$ are given by

$$\frac{1}{2}\left\{r - 2 \pm r(r+2)\sqrt{\frac{\beta_1}{\beta_1(r+2)^2 + 16(r+1)}}\right\}$$

83, last line and line 3 from bottom ... ... For $\frac{r+2}{r+2}$ read $\frac{r+2}{r-2}$

85, heading to col. (5) ... The suffix, 2, has been dropped in some copies

89, heading to fourth col. of table ... ... For α read $\frac{1}{2}(1+\alpha)$

99, line 2 from top ... For y read y_0

99, insert as a note to the formulæ in § 12:

"Care is necessary with regard to the value used for y_0, and consequently with regard to A and B. If Sheppard's tables of ordinates (z) be multiplied by, say, 10^5 and used as the exposed to risk, the values of A and B resulting from the work will be $N_1 \div (10^5\sigma)$ and $N_2 \div (10^5 H\sigma)$. The reason is that his tables are in terms of standard deviation. In the example, pp. 99, 100, § 14, this is obscured because the assumed standard deviation is 10."

PAGE

99, line 15 from bottom... For "duration" read "deviation"

102, line 10 from top ... For 29·268 read 25·093

103, line 12 from top ... For 1·0875 read 1·1048

111, lines 4 & 6 from bottom and foot-note ... $\frac{1}{2\sigma^2}$ should be $\frac{1}{\sigma^2}$, and the suffixes of the σ's should be interchanged

112, top line ... ... For $r = -\frac{c_{12}}{c_1 c_2}$ read $r = -\frac{c_{12}}{\sqrt{c_1 c_2}}$

112, line 3... ... ... For 2π read π

118, second line below table should read

Mean unexpired term $= 22·5 - 1·68815 = 20·81185$

123, line 6 from top ... For σ_2 read σ_3

123, line 7 from top ... For $\frac{\sigma_2}{m_2}$ read $\frac{\sigma_3}{m_3}$

127, line 9 from top ... For $\sqrt{2\pi}\sigma_1$ read $\sqrt{2\pi\sigma_1}$

152, line 2 from bottom ... Insert x^{m-1} under integral

152, line 1 from bottom ... For x^{m+1} read x^{m-1}, and for dx read dz. The upper limit of the integral on the left-hand side is ∞

153, line 2 from top ... The limits of the integral on the left-hand side are 0 to ∞, and the expression under the integral on the right-hand side should be $e^{-y}y^{m+n-1}$

153, line 5 from bottom should be $\Gamma(x+1) = \sqrt{2\pi x}\, x^x e^{-x} e^{\frac{1}{12x}}$

153, line 3 from bottom should be

$$\log_{10}\Gamma(x+1) = \log_{10}\sqrt{2\pi} + \left(x + \frac{1}{2}\right)\log_{10}x - \left(x - \frac{1}{12x}\right)\log_{10}e$$

157, line 13 from top should be $-(x^2+y^2-2rxy)r(1-r^2)^{-2}$ &c.

158, footnote ... ... The last symbol should be r_n not r^n

165... ... ... The volume of tables for statisticians has been recently published by the Cambridge University Press.

Note on Type III, *example on pp.* 66, 67.

As implied at the top of p. 66, the usual shape of a Type III curve is "cocked-hat." This has been over-looked, not unnaturally, by some readers who thought that the example of a J-shaped curve meant that this was the more usual shape for Type III. The particular example was used to bring out the difficulty mentioned at the foot of p. 66. It should have been explained on p. 66 that the value of y_0 was calculated by the expression $N\gamma^{p+1} \div \Gamma(p+1)$, *cf.* p. 103. This expression is used when p is negative to avoid dealing with the logarithms of negative numbers. The value should be 194·2. The method of adjusting the moments in such exceptional cases given on p. 103 is correctly indicated, but the arithmetical values given below the table are inaccurate.

Arithmetical results.—The opportunity may be taken of dealing with two points. The first is with regard to possible arithmetical slips in some of the work. A general check has been put on figures wherever possible, but it could not disclose a slip which did not affect the graduation or the final figures of a constant or coefficient. Thus, if the last three figures of $\log y_0$ on p. 79 were wrong (which I have no reason to suppose) the mistake would be regrettable, but the graduation in the table on that page would be unaffected. The second point relates to certain difficulties found in reproducing exactly the numerical result of another calculator, owing to the usual unreliability of the end figures when many operations have been made. In lengthy arithmetic the two final figures may be unreliable, and two different arithmetical processes may therefore both be correct and yet give considerable divergencies.