PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Fit of a Formula for Discrepant Observations

W. F. Sheppard

Phil. Trans. R. Soc. Lond. A 1929 **228**, 115-150
doi: 10.1098/rsta.1929.0003

III. *The Fit of a Formula for Discrepant Observations.*

By W. F. SHEPPARD, *Sc.D., LL.M.*

(*Communicated by* E. T. WHITTAKER, *F.R.S.*)

(Received October 13, 1927—Revised May 23, 1928—Read December 8, 1927.)

CONTENTS.

INTRODUCTORY.

1. *The Problem.*—When we have a set of quantities $u_1, u_2, \ldots u_m$, which may be regarded as the observed values of certain unknown quantities $U_1, U_2, \ldots U_m$, and we form the hypothesis that these U's can be represented by a formula which either is completely determined by *a priori* considerations or involves constants which have to be determined from the data, how are we to decide whether the hypothesis is justified?

The classic method of testing a hypothesis of this kind is Prof. KARL PEARSON'S "χ^2" method, published in 1900 (ref. 6). But in one important class of cases the accuracy of his formula is open to doubt. The problem has not yet been finally solved: the object of the present paper is to take a step towards its solution.

2. *Points to be met.*—The nature of the problem can be seen by considering an imperfect treatment of it, for a particular class of cases. Suppose that we are concerned with a frequency-distribution, say, of men's heights, and that our data, relating to N men, are the total numbers whose heights Y are respectively below $Y_1, Y_2, \ldots Y_m$. These numbers may be called $Na_1, Na_2, \ldots Na_m$. How are we to test whether this observed distribution can reasonably be regarded as the result of random sampling of N individuals from an indefinitely great population in which the proportion for which Y is less than Y_f ($f = 1, 2, \ldots m$) is a certain hypothetical function of Y_f and of k unknown constants $C_1, C_2, \ldots C_k$?

A crude method would be as follows. We calculate values for the C's, by some method, from the data. We then calculate what the proportions would be if these were the true C's; call these calculated proportions $A_1, A_2, \ldots A_m$. The differences between these "calculated" A's and the "observed" a's are the "discrepancies" of the a's. Now if we took a random sample of N from a large population falling into two categories in which the numbers were proportional to A_f and $1 - A_f$ respectively, and if the numbers actually taken were Na_f and $N(1 - a_f)$, the mean value of $Na_f - NA_f$ would be 0, and its s.d.* would be $\{NA_f(1 - A_f)\}^{\frac{1}{2}}$. The ratio of the actual value of $Na_f - NA_f$ to its s.d. can be found for each of the m values of f; and we can see whether the distribution of these ratios is such as might reasonably be due to random sampling from a set of ratios with mean value 0 and standard deviation 1.

The imperfections of this method are as follows:—

(A) We take no account of the fact that by finding the C's from the a's we have already to some extent fitted the hypothetical formula to the data, and that an approximate agreement between the deduced values and the data must, to this extent, be fallacious. If, to take an extreme case, k were $= m$, so that we used a formula with just as many C's as a's, the discrepancies would all be 0: but this would prove nothing as to the suitability of the formula. Another way of stating the matter is that we compare each discrepancy, not with its own s.d., but with the s.d. of the true error.

* *Abbreviations.*—m.p. = mean product; m.p.e. = mean product of errors; m.s. = mean square; m.s.e. = mean square of error; s.d. = standard deviation.

(B) In considering the distribution of the ratios of discrepancy to s.d. of discrepancy, we are treating these ratios as statistically independent, whereas in reality they are correlated. (Another point, which is not so obvious, is that the ratios are not algebraically independent, but are connected by k relations.)

(C) While the method is tolerably correct when NA_f and $N(1 - A_f)$ are both large, it is not sufficiently correct when either of these numbers is rather small.

In a paper published in 1898 (ref. 7, §§ 24, 32), one object of which was to obtain formulæ for testing the hypothesis of normal distribution or normal correlation, I met (A), so far as these particular cases were concerned, by using the true s.d. of the discrepancy in place of the s.d. of the error. But the method was not really satisfactory, as I did not realise the importance of (B). In view of (C), I limited my method to cases in which the numbers involved were fairly large; *i.e.*, I grouped small frequencies at the extremities of the range.

3. PEARSON'S χ^2 *method.*—In 1900 PEARSON published his χ^2 method, mentioned above. The method may be stated briefly, with some alteration of notation, as follows.

(I) First, take the class of cases in which the hypothetical formula is completely determined by *a priori* considerations. Let the observed quantities be $u_1, u_2 \dots u_m$, and let the corresponding quantities, on the hypothesis we are considering, be $U_1, U_2, \dots U_m$. On this hypothesis, the differences between the u's and the U's are errors, which we can denote by $\varepsilon_1, \varepsilon_2, \varepsilon \dots \varepsilon_m$, so that

$$(\lambda = 1, 2 \dots m) \quad u_\lambda = U_\lambda + \varepsilon_\lambda. \qquad (3.1$$

Let the m.p. of errors ε_f and ε_g be E_{fg}; and let the set reciprocal* to $E_{\lambda\mu}$ be $E^{\lambda\mu}$. Then the frequency of cases in which the errors respectively lie within limits

$$\varepsilon_1 \pm \tfrac{1}{2}d\varepsilon_1,\ \varepsilon_2 \pm \tfrac{1}{2}d\varepsilon_2,\ \dots\ \varepsilon_m \pm \tfrac{1}{2}d\varepsilon_m$$

is proportional to $e^{-\frac{1}{2}S}\, d\varepsilon_1\, d\varepsilon_2 \dots d\varepsilon_m$, where†

$$S = E^{\lambda\mu}\varepsilon_\lambda\varepsilon_\mu. \qquad (3.2)$$

Now let P denote the proportion of cases in which the errors are such that S has the above or a greater value. Then P is taken as giving the grade of frequency of

* The set $E_{\lambda\mu}$ consists of $E_{11}, E_{12} \dots E_{1m}, E_{21} (= E_{12}), E_{22} \dots$. Let Δ be the determinant of which these are the elements. Then E^{fg} is defined as (co-factor of E_{fg} in Δ) $\div$ (value of Δ). Alternatively, we can define $E^{\lambda\mu}$ as the mean-product set of the set "conjugate" to ε_λ.

† I use the shortened sum-notation, in which the duplication of a greek suffix means summation for each value of the suffix. As a rule, $\alpha, \beta \dots$ will refer to summation for α (or $\beta \dots$) $= 1, 2 \dots k$; $\lambda, \mu \dots$ for $1, 2 \dots m$; and $\pi, \rho, \sigma \dots$ for $k+1, k+2 \dots m$. Thus $p_\mu q_\mu$ will mean $p_1q_1 + p_2q_2 + \dots + p_mq_m$. Also I use $\Big|_g^f$ to denote the quantity which is 1 if $f = g$ and 0 if $f \neq g$.

this particular set of errors; the grade varying from 1, when all the errors are 0, to 0 when any of them are indefinitely great. By a change of variables, PEARSON shows that

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}{\int_0^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}, \qquad (3.3)$$

where $\chi^2 = S$, as given by (3.2), the ε's having the actual values given by (3.1).

(II) In the case of a definite frequency-distribution, from which N individuals are supposed to have been taken at random and sorted into $m+1$ categories, let N_0, N_1, $N_2 \ldots N_m$ be the numbers which would come into these categories if the distribution were representative, and let n_0, n_1, $n_2 \ldots n_m$ be the actual numbers, where

$$n_0 + n_1 + n_2 + \ldots + n_m = N.$$

Then, denoting the differences $n_0 - N_0$, $n_1 - N_1 \ldots n_m - N_m$ by ε_0, $\varepsilon_1 \ldots \varepsilon_m$, it is shown that χ in (3.3) is given by

$$\chi^2 = \frac{\varepsilon_0^2}{N_0} + \frac{\varepsilon_1^2}{N_1} + \frac{\varepsilon_2^2}{N_2} + \ldots + \frac{\varepsilon_m^2}{N_m}. \qquad (3.4)$$

(III) Next take the case of a frequency-distribution for which the hypothetical formula is of a given type but contains unknown constants which have to be determined from the data. The "best" values of the constants having been so determined, we calculate the corresponding N's in accordance with them, and find the discrepancies between these calculated N's and the observed n's. It is stated that the previous results then apply, the ε's being replaced by the discrepancies; *i.e.*, that if the discrepancies are θ_0, θ_1, $\theta_2 \ldots \theta_m$, then

$$P = \frac{\int_{\chi_s}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}{\int_0^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}, \qquad (3.5)$$

where

$$\chi_s^2 = \frac{\theta_0^2}{N_0} + \frac{\theta_1^2}{N_1} + \ldots + \frac{\theta_m^2}{N_m}. \qquad (3.6)$$

4. *Observations on the method.*—The method involves the assumption that S is a quadratic function, which means that the errors of observation are small in comparison with the observed quantities. If, therefore, it is applied to an ordinary frequency-distribution, where some of the numbers are small, the difficulty (C) arises. The difficulty is avoided (*cf.* §2) by grouping small frequencies at the extremities of the range. I shall assume throughout this paper that the errors of observation are relatively small, so that (C) can be left out of account. On this understanding, (I) and (II) are sound; and the method has the great merit of considering the errors as a whole, so that, for these cases, (B) is fully met. But does (III) meet (A)? PEARSON

recognised the existence of the difficulty, and discussed it in some detail; but the validity of his reasoning is open to doubt.

The apparent defect of the method can, as already indicated, be expressed in either of two ways. The ε's are the true errors; and we can say that we are wrong in replacing these true errors by the discrepancies between the calculated and the observed values. Or we can say that, since we are dealing with discrepancies, we are wrong in using the m.ss. and m.pp. of the errors. Both these statements are true. But, if we try to remedy the defect by using the m.ss. and m.pp. of the discrepancies in place of the m.ss. and m.pp. of the errors, we shall find that (in consequence of the algebraical relations mentioned in §2) the determinant of the mean-product set is 0, and that the reciprocal set, therefore, does not exist. The problem really needs independent consideration.

There has in recent years (see references at end of this paper) been some discussion of the applicability of PEARSON's method to particular classes of cases, but the general theory does not seem to have been fully investigated. Its treatment necessarily involves some repetition of results obtained by other writers as regards frequency-distributions.

GENERAL FORMULÆ.

5. *Assumption as to frequency of errors.*—It is assumed that the errors with which we are concerned are so distributed that the frequency of joint occurrence of errors lying within the limits $\varepsilon_1 \pm \frac{1}{2}d\varepsilon_1, \varepsilon_2 \pm \frac{1}{2}d\varepsilon_2 \ldots \varepsilon_m \pm \frac{1}{2}d\varepsilon_m$ (these errors being algebraically independent) is proportional to $e^{-\frac{1}{2}S}\,d\varepsilon_1\,d\varepsilon_2 \ldots d\varepsilon_m$, where S is a quadratic function of $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_m$. It is also assumed that the m.ss. and m.pp. of these errors are known; or, at any rate, that our results may be regarded as final when they are expressed in terms of these m.ss. and m.pp.

The former assumption leads to the following results:—

(i) The mean value of any error is 0.

(ii) Let the m.p. of ε_f and ε_g be denoted by E_{fg}; and let $E^{\lambda\mu}$ be the set reciprocal to $E_{\lambda\mu}$. Then $S = E^{\lambda\mu}\varepsilon_\lambda\varepsilon_\mu$.

(iii) Let η_λ be a set of m expressions which are linear functions of the m ε's and are algebraically independent; and let the m.p. of η_f and η_g be H_{fg}. Then the frequency of joint occurrence of values lying within the limits

$$\eta_1 \pm \tfrac{1}{2}d\eta_1,\ \eta_2 \pm \tfrac{1}{2}d\eta_2 \ldots \eta_m \pm \tfrac{1}{2}d\eta_m$$

is proportional to $e^{-\frac{1}{2}S}\,d\eta_1, d\eta_2 \ldots d\eta_m$, where S has the same value as before. Also, by (ii), $S = H^{\lambda\mu}\eta_\lambda\eta_\mu$; and therefore $H^{\lambda\mu}\eta_\lambda\eta_\mu = E^{\lambda\mu}\varepsilon_\lambda\varepsilon_\mu$.

(iv) Let $\eta_1, \eta_2 \ldots \eta_l$ be a smaller number of (algebraically independent) expressions, which are linear functions of the ε's. Then there are similar properties, S being replaced by a quadratic function of $\eta_1, \eta_2 \ldots \eta_l$. This, of course, includes the case in which $\eta_1, \eta_2 \ldots \eta_l$ are l selected ε's.

6. *Notation.*—As a rule, true (or hypothetical) values are denoted by capital letters, observed values by small letters, and errors by small greek letters; values deduced from the data are denoted by small letters or by accented capitals.

$u_1, u_2 \ldots u_m$ are the observed quantities.

$U_1, U_2 \ldots U_m$ are the corresponding quantities for a certain hypothetical formula.

$C_1, C_2 \ldots C_k$ are the constants of this hypothetical formula.

$\epsilon_1, \epsilon_2 \ldots \epsilon_m$ are the "errors" of the observed quantities, *i.e.*,

$$(\lambda = 1, 2 \ldots m) \; u_\lambda = U_\lambda + \epsilon_\lambda. \qquad (6.1)$$

The U's are related to the C's by a definite set of formulæ

$$(\lambda = 1, 2 \ldots m) \; U_\lambda = f_\lambda (C_1, C_2 \ldots C_k); \qquad (6.2)$$

e.g., in the case of a frequency-distribution the U's might be the integrals of the hypothetical function of the C's, the upper limits being magnitudes $Y_1, Y_2 \ldots Y_m$ which are the basis of the classification.

The C's are deducible from the U's by a set of k equations

$$(\alpha = 1, 2 \ldots k) \; C_\alpha = F_\alpha (U_1, U_2 \ldots U_m). \qquad (6.3)$$

These equations must be consistent with (6.2), and must be algebraically independent, but otherwise can be chosen to suit our convenience (see §11).

The values of the constants as deduced from the data are therefore

$$(\alpha = 1, 2 \ldots k) \; c_\alpha = F_\alpha (u_1, u_2 \ldots u_m). \qquad (6.4)$$

With these deduced values of the constants we find the "calculated" U's, namely,

$$(\lambda = 1, 2 \ldots m) \; U'_\lambda = f_\lambda (c_1, c_2 \ldots c_k). \qquad (6.5)$$

The differences between these calculated U's and the observed u's are the "discrepancies," which it is more convenient to denote by θ than by ϵ'. Thus the discrepancies are

$$(\lambda = 1, 2 \ldots m) \; \theta_\lambda = u_\lambda - U'_\lambda. \qquad (6.6)$$

It should be observed, throughout, that the assignment of the suffixes is quite arbitrary. In the case of a correlation-table, for example, the entries might be numbered consecutively by columns or by rows or diagonally or in any other way.

The relations (6.2) and (6.3) may be given implicitly, provided there is no ambiguity as to the result.

By (6.1) and (6.6)

$$\theta_\lambda - \epsilon_\lambda = -(U'_\lambda - U_\lambda); \qquad (6.7)$$

i.e., the error involved in treating the discrepancy as the error is equal but opposite to the error of the calculated U.

7. *Formula for discrepancy.*—The errors being assumed to be small, we are only working to their first powers.

The errors ε_λ of the u's, when c_α is found from (6.4), produce in c_α an error

$$\underset{(6.3)}{\frac{\partial C_\alpha}{\partial U_\lambda}}\,\varepsilon_\lambda\,;$$

" (6.3) " being written underneath to indicate the formula to which the differentiation relates. These errors of the c's, when (6.5) is used, produce in U'_λ an error

$$\underset{(6.2)}{\frac{\partial U_\lambda}{\partial C_\alpha}}\,\underset{(6.3)}{\frac{\partial C_\alpha}{\partial U_\mu}}\,\varepsilon_\mu.$$

(The dummy λ's in the preceding expression have been altered to μ's, in order to keep λ as the leading letter.) The calculated U_λ is therefore

$$U'_\lambda = U_\lambda + \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\,\varepsilon_\mu\,; \quad \ldots\ldots\ldots\ldots \quad (7.1)$$

and, by (6.6) and (6.1), the discrepancy is

$$(\lambda = 1, 2 \ldots m)\ \theta_\lambda = \varepsilon_\lambda - \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\,\varepsilon_\mu\,; \quad \ldots\ldots\ldots \quad (7.2)$$

the partial derivatives being obtained from (6.2) and (6.3). The m.ss. and m.pp. of the ε's being known, those of the θ's can be calculated.

8. *Relations between partial derivatives.*—The occurrence of the expression $\frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\varepsilon_\mu$ in (7.2) leads us to consider the relations between the partial derivatives.

(i) The values of $\frac{\partial U_\lambda}{\partial C_\alpha}$ in (7.2) are to be obtained from (6.2), and those of $\frac{\partial C_\alpha}{\partial U_\mu}$ from (6.3). The μ-summation in $\frac{\partial C_\alpha}{\partial U_\mu}\varepsilon_\mu$ is to be made for all values of μ from 1 to m. If, however, the formula (6.3) for any particular C does not involve a particular U, say U_r, the corresponding error ε_r will not appear in the summation. But this does not mean that $\varepsilon_r = 0$; the reason is that $\frac{\partial C_\alpha}{\partial U_r} = 0$.

(ii) The formulæ (6.2) give us the U's in terms of the C's; and (6.3) give us the C's in terms of the U's. If the expressions for the C's in (6.3) are substituted in (6.2), we shall get formulæ giving each U in terms of the U's, say,

$$(\lambda = 1, 2 \ldots m)\ U_\lambda = g_\lambda(U_1, U_2 \ldots U_m). \quad \ldots\ldots\ldots \quad (8.1)$$

The existence of these relations is due to the fact that, though (6.2) comprises m formulæ, the formulæ only involve k C's, and therefore there are $m-k$ relations between them, *i.e.*, there are only k formulæ which are algebraically independent. Similarly only k of (8.1) are algebraically independent.

We now have

$$(\lambda = 1, 2 \ldots m)\ \underset{(6.2)}{\frac{\partial U_\lambda}{\partial C_\alpha}}\ \underset{(6.3)}{\frac{\partial C_\alpha}{\partial U_\mu}} = \underset{(8.1)}{\frac{\partial U_\lambda}{\partial U_\mu}}.$$

Hence (7.2) may be written

$$\theta_\lambda = \varepsilon_\lambda - \frac{\partial U_\lambda}{\partial U_\mu}\varepsilon_\mu; \quad \ldots \ldots \quad (8.2)$$

the partial derivatives being obtained from (8.1). In practice, however, (7.2) is probably the more convenient form, the expressions for $\frac{\partial U_\lambda}{\partial C_\alpha}$ and $\frac{\partial C_\alpha}{\partial U_\mu}$ being obtained from the original formulæ (6.2) and (6.3).

If any particular ε, say ε_r, does not occur in (8.2), it is because the corresponding U, namely U_r, does not occur in (8.1), and therefore $\frac{\partial U_\lambda}{\partial U_r}$ is $=0$.

(iii) A class of cases of special importance consists of the cases in which the formulæ (6.3) involve only k U's, which we may call $U_1, U_2 \ldots U_k$. The formulæ then take the form

$$(\alpha = 1, 2 \ldots k) \quad C_\alpha = F_\alpha (U_1, U_2 \ldots U_k); \quad \ldots \ldots \quad (8.3)$$

and they are the converse of the first k of (6.2), namely,

$$(\beta = 1, 2 \ldots k) \quad U_\beta = f_\beta (C_1, C_2 \ldots C_k). \quad \ldots \ldots \quad (8.4)$$

The remaining $m-k$ of (6.2), namely

$$(\rho = k+1, k+2 \ldots m) \quad U_\rho = f_\rho (C_1, C_2 \ldots C_k), \quad \ldots \ldots \quad (8.5)$$

give the remaining U's in terms of the C's. They can, by (8.3), be expressed in terms of the first k U's; we may write the result as

$$(\rho = k+1, k+2 \ldots m) \quad U_\rho = g_\rho (U_1, U_2 \ldots U_k). \quad \ldots \ldots \quad (8.6)$$

Thus we have split up (8.1) into two sets of formulæ, namely

$$\left.\begin{array}{ll} (\lambda = 1, 2 \ldots k) & U_\lambda = U_\lambda \\ (\lambda = k+1, k+2 \ldots m) & U_\lambda = g_\lambda (U_1, U_2 \ldots U_k) \end{array}\right\}. \quad \ldots \ldots \quad (8.7)$$

We have then

$$(\lambda = 1, 2 \ldots k) \quad \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\beta} = \frac{\partial U_\lambda}{\partial U_\beta} = \Big|_\beta^\lambda, \quad \ldots \ldots \quad (8.8)$$

and

$$(\lambda = k+1,\ k+2 \ldots m) \quad \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\beta} = \frac{\partial U_\lambda}{\partial U_\beta}; \quad \ldots \ldots \quad (8.9)$$

$\frac{\partial U_\lambda}{\partial U_\beta}$ in (8.9) being found from (8.6). Thus (7.2), which only involves $\varepsilon_1, \varepsilon_2 \ldots \varepsilon_k$, becomes

$$(\lambda = 1, 2 \ldots m) \quad \theta_\lambda = \varepsilon_\lambda - \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\beta}\varepsilon_\beta = \varepsilon_\lambda - \frac{\partial U_\lambda}{\partial U_\beta}\varepsilon_\beta, \quad \ldots \quad (8.10)$$

where $\frac{\partial U_\lambda}{\partial U_\beta}$ is 1 if $\lambda = \beta$, is 0 if λ has any other of the values $1, 2 \ldots k$, and is given by (8.6) if λ has any of the values $k+1,\ k+2 \ldots m$.

(iv) We have so far been considering $\frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}$. The reverse expression, $\frac{\partial C_\alpha}{\partial U_\lambda}\frac{\partial U_\lambda}{\partial C_\beta}$, presents less difficulty. The equations (6.3) must be consistent with (6.2); and therefore, if we substitute from (6.2) in (6.3), we must get C_α. It follows that, for all cases, whether (6.3) contains only k U's or more than k, we have

$$(\alpha = 1, 2 \ldots k) \quad \frac{\partial C_\alpha}{\partial U_\lambda}\frac{\partial U_\lambda}{\partial C_\beta} = \frac{\partial C_\alpha}{\partial C_\beta} = \Big|{}^{\alpha}_{\beta}. \quad \ldots \ldots \quad (8.11)$$

9. *Relations between discrepancies.*—Returning to the general formula for θ_λ, we come upon a paradox. We are assuming that the u's are the observed values of certain definite U's. We can never know the true values of the U's. But in (7.2) we have m equations giving the m discrepancies in terms of the m ε's. Why cannot we, by solving these equations, find the ε's in terms of the θ's, and then, by subtracting the ε's from the u's, obtain the U's?

The explanation is that the m discrepancies are not algebraically independent, but are connected by k relations (*cf.* FISHER, ref. 3, p. 93), so that there would really be only $m - k$ equations for finding the ε's. Suppose, for instance, that our method of finding the constants of a frequency-distribution includes the equating of 1st moments. By doing this, we arrive at a hypothetical distribution whose 1st moment is equal to that of the u's. We then find the calculated U″s. But the 1st moment of these must also be equal to the 1st moment of the hypothetical distribution, and therefore equal to the 1st moment of the u's. The 1st moment of the discrepancies will therefore be 0. Similarly, if we use the mean-and-mean-square method, the 2nd moment of the discrepancies will be 0. Or, again, suppose we determine c_1 and c_2 from u_1 and u_2. Then, when we calculate the U″s from these values of c_1 and c_2, U'_1 and U'_2 will necessarily be equal to u_1 and u_2, *i.e.*, θ_1 and θ_2 will both necessarily be 0.

In the 1st-moment case, we equate c_1 to $p_\lambda u_\lambda$, where $p_1, p_2 \ldots$ are certain coefficients depending on the way in which the data are set out; and this gives $p_\lambda \theta_\lambda = 0$. This suggests

$$(\alpha = 1, 2 \ldots k) \quad \frac{\partial C_\alpha}{\partial U_\lambda}\theta_\lambda = 0 \quad \ldots \ldots \quad (9.1)$$

as the general formula for the relations between the θ's. This is easily verified. For, substituting from (7.2), we find, by (8.11),

$$\begin{aligned}\frac{\partial C_\alpha}{\partial U_\lambda}\theta_\lambda &= \frac{\partial C_\alpha}{\partial U_\lambda}\left(\varepsilon_\lambda - \frac{\partial U_\lambda}{\partial C_\beta}\frac{\partial C_\beta}{\partial U_\mu}\varepsilon_\mu\right)\\ &= \frac{\partial C_\alpha}{\partial U_\lambda}\varepsilon_\lambda - \Big|^{\alpha}_{\beta}\frac{\partial C_\beta}{\partial U_\mu}\varepsilon_\mu\\ &= \frac{\partial C_\alpha}{\partial U_\lambda}\varepsilon_\lambda - \frac{\partial C_\alpha}{\partial U_\mu}\varepsilon_\mu\\ &= 0.\end{aligned}$$

10. *Method of procedure.*—The procedure is therefore as follows. Having settled the form of the functions in (6.3), we obtain $c_1, c_2 \ldots c_k$, and thence find the actual values of $\theta_1, \theta_2 \ldots \theta_m$. We select $m-k$ of these, which we can call θ_ρ. These $m-k$ may be any we please, provided we do not include any which are necessarily zero; if, for instance, the c's are found from $u_1, u_2 \ldots u_k$, we must use $\theta_{k+1}, \theta_{k+2} \ldots \theta_m$. In any case, since the order in which we number the u's is immaterial, we can call the θ's $\theta_{k+1}, \theta_{k+2} \ldots \theta_m$. The expressions for the θ's in terms of the ε's are given by (7.2); and therefore, knowing the m.pp. of the ε's, we can calculate those of the θ's. Let Θ_{ps} be the m.p. of θ_p and θ_s; and let $\Theta^{\pi\sigma}$ be the set reciprocal to $\Theta_{\pi\sigma}$ (§3). Then the frequency of joint occurrence of discrepancies lying between $\theta_{k+1} \pm \frac{1}{2}d\theta_{k+1} \ldots \theta_m \pm \frac{1}{2}d\theta_m$ is proportional to $e^{-\frac{1}{2}S}\,d\theta_{k+1}\ldots d\theta_m$, where $S = \Theta^{\pi\sigma}\theta_\pi\theta_\sigma$. We then follow PEARSON's general method. Writing

$$\chi'^2 \equiv S = \Theta^{\pi\sigma}\theta_\pi\theta_\sigma, \qquad (10.1)$$

the θ's being the actual discrepancies in the particular case, we find that the proportion of cases in which S has this or a greater value is

$$P' = \frac{\int_{\chi'}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-k-1}\,d\chi}{\int_0^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-k-1}\,d\chi}. \qquad (10.2)$$

It will be noticed that this differs from PEARSON's formula not only in replacing m by $m-k$ but also in the form of the expression for χ'^2. The necessity for the former alteration in the case of a frequency-distribution was pointed out by R. A. FISHER (ref. 3, p. 93; *cf.*, YULE, ref. 8, p. 96).

It will be seen from §5 (iii) that, instead of calculating $\Theta^{\pi\sigma}\theta_\pi\theta_\sigma$, we can take any linear functions of the θ's, say $\eta_{k+1}, \eta_{k+2} \ldots \eta_m$, and calculate $H^{\pi\sigma}\eta_\pi\eta_\sigma$, where H_{fg} is the m.p. of η_f and η_g. This will give us the same value of χ'^2, and the same value of P'.

11. *Discrepancies under different systems.*—The expression for S, given in (10.1), is supposed to be obtained by finding the c's from the data by means of the equations (6.4). But, as has already been pointed out (§ 6), there is an infinite number of ways of forming these equations; *i.e.*, there is an infinite number of routes by which we can proceed from any particular data to a value of χ'^2. In the case, *e.g.*, of a supposed Gaussian distribution we might, as the mean and mean square of the distribution, take the actual average and average square given by the distribution as a whole; or we could deduce the mean and the mean square from the average cube and average fourth power; or we might make a selection from the data in any arbitrary manner, and deduce the mean and mean square from these. What method are we to choose? It is, *e.g.*, necessary to use the method which gives the most probable values of the constants, on some definite assumption as to relative probability? Or, if not, what other criterion is to be adopted?

The answer is that it does not matter what method we use, provided it is a "correct" method, *i.e.*, provided that it would give the correct* values of the constants if there were no errors in the u's, and provided also, of course, that when we have found the constants we take account of all the discrepancies, subject to the algebraical relations between them. Different methods will give different discrepancies; but they will also give different m.ss. and m.pp. of the discrepancies, and the resulting value of S will be the same whatever method is adopted.

To prove the above statement, suppose that by one method we obtain a set of $m-k$ discrepancies θ_ρ, the mean product of θ_p and θ_s being Θ_{ps}; and suppose that from the same data, by another method, we obtain a set of discrepancies ϕ_ρ, the mean product of ϕ_p and ϕ_s being Φ_{ps}. (It will be seen from § 10 that the θ's and the ϕ's will not necessarily be discrepancies of the same set of u's; indeed, the sets will usually be different.) Then in the one case we get $S = \Theta^{\pi\sigma}\,\theta_\pi\theta_\sigma$, and in the other case we get $S = \Phi^{\pi\sigma}\,\phi_\pi\,\phi_\sigma$. We want to prove that these two expressions are equal; it will then follow that the two deduced values of P′ are equal.

It will be seen from § 5 that in order to prove this it is sufficient to prove that the ϕ's are linear functions of the θ's, or, which comes to the same thing, that the θ's are linear functions of the ϕ's. Here we are up against the difficulty mentioned in § 9. If we were dealing with m θ's, not algebraically connected, and with m ϕ's, also not algebraically connected, each of these would be a linear function of the m ε's, and they would therefore all be linear functions of one another. But there are only $m-k$ θ's and $m-k$ ϕ's. The matter therefore requires further consideration.

12. *Types of variation of method.*—To make the enquiry complete, let us consider separately the different kinds of variation that we might make in the method; since we are only dealing with first powers of the errors, any change of system could be made up of these separate variations.

* Strictly speaking, this proviso excludes the method of moments, if the formulæ for the moments are only approximate.

(i) First, suppose that instead of dealing with the m u's we deal with m other quantities derived from them in a definite way. In the case of a frequency-distribution, for example, tabulated on the ordinary system, we might have $m+1$ categories determined by values $Y_0, Y_1, Y_2 \ldots Y_{m+1}$ of a variable Y; and the u's would then be the proportions falling into the first m of these categories, the proportion in the remaining category being $1 - u_1 - u_2 \ldots - u_m$. We might find it more convenient to deal with the successive sums $u_1, u_1 + u_2, u_2 + u_2 + u_3 \ldots$, which would be the proportions for which Y was less than Y_1, less than Y_2, and so on. Or (see § 15) we might wish, for simplicity of calculation, to replace the u's by functions of the u's whose mean products of error would all be 0. Or, in the case of a supposed Gaussian distribution, we might, having obtained these sums, deduce from them the corresponding deviations from the mean in the standard Gaussian figure, *i.e.*, in the standard tables, the values of x for which the area $\frac{1}{2}(1+\alpha)$ has the values $u_1, u_1 + u_2, u_1 + u_2 + u_3 \ldots$

Let these new quantities, obtained from the u's, be

$$v_\lambda \equiv v_1, v_2 \ldots v_m;$$

and let their true values be V_λ. Then the U's and the V's are connected by definite equations of the form

$$V_\lambda = \text{function}_\lambda (U_1, U_2 \ldots U_m); \quad \ldots \ldots \quad (12.1)$$

and the errors ε_λ in the u's will produce in the v's errors

$$\eta_\lambda = \frac{\partial V_\lambda}{\partial U_\nu} \varepsilon_\nu. \quad \ldots \ldots \quad (12.2)$$

If ϕ_s is the discrepancy in v_s, we see from (7.2) that the formula for ϕ_s in terms of the η's is

$$\phi_s = \eta_s - \frac{\partial V_s}{\partial C_\alpha} \frac{\partial C_\alpha}{\partial V_\mu} \eta_\mu. \quad \ldots \ldots \quad (12.3)$$

Substituting from (12.2) in (12.3), we have

$$\begin{aligned}\phi_s &= \frac{\partial V_s}{\partial U_\nu} \varepsilon_\nu - \frac{\partial V_s}{\partial C_\alpha} \frac{\partial C_\alpha}{\partial V_\mu} \frac{\partial V_\mu}{\partial U_\nu} \varepsilon_\nu \\ &= \frac{\partial V_s}{\partial U_\nu} \varepsilon_\nu - \frac{\partial V_s}{\partial C_\alpha} \frac{\partial C_\alpha}{\partial U_\nu} \varepsilon_\nu.\end{aligned}$$

But (7.2) gives

$$\begin{aligned}\frac{\partial V_s}{\partial U_\nu} \theta_\nu &= \frac{\partial V_s}{\partial U_\nu} \varepsilon_\nu - \frac{\partial V_s}{\partial U_\nu} \frac{\partial U_\nu}{\partial C_\alpha} \frac{\partial C_\alpha}{\partial U_\mu} \varepsilon_\mu \\ &= \frac{\partial V_s}{\partial U_\nu} \varepsilon_\nu - \frac{\partial V_s}{\partial C_\alpha} \frac{\partial C_\alpha}{\partial U_\mu} \varepsilon_\mu.\end{aligned}$$

Hence

$$\phi_s = \frac{\partial V_s}{\partial U_\nu} \theta_\nu; \quad \ldots \ldots \quad (12.4)$$

i.e., ϕ_s is a linear function of θ_λ. By reason of the k relations between the θ's, the k that we are not using can be expressed in terms of the remainder ; and therefore ϕ_s is a linear function of θ_ρ.

(ii) Next, suppose that in place of the C's we use as our constants certain other constants $D_1, D_2 \dots D_k$, but that these are so related to the C's that the equations (6.3) still hold. Then it is obvious that we make no alteration in the discrepancies ; for all that we do is to calculate the U''s from (6.5) by means of the *d*'s instead of calculating them directly from the *c*'s. This, indeed, is what we are constantly doing. If, for instance, we use the mean-and-mean-square method, (6.3) might be the formulæ for C_1, the mean, and C_2, the mean square, whereas what we should really do would be to deduce from these the standard deviation, and then use the mean and the standard deviation for calculating the U''s. This case, therefore, presents no difficulty.

(iii) We come finally to the case in which, whether we calculate the U''s directly from the *c*'s or from definite functions of the *c*'s, we replace the equations (6.3) by a different set of equations for the C's. The problem is to prove that the ϕ's derived from these new equations are linear functions of the θ's derived from the original equations. As I found this rather difficult, I will give the steps by which I arrived at a proof.

13. *Proof of linear relation for different sets of equations.*—The problem may be restated. We have, as in (6.2), the definite set of relations

$$(\lambda = 1, 2 \dots m) \quad U_\lambda = f_\lambda (C_1, C_2 \dots C_k). \qquad (13.1)$$

Using one set of equations, consistent with (13.1), namely,

$$(\alpha = 1, 2 \dots k) \quad C_\alpha = F_\alpha (U_1, U_2 \dots U_m), \qquad (13.2)$$

we find the *c*'s, construct the " calculated " U''s, and obtain discrepancies

$$(\lambda = 1, 2 \dots m) \quad \theta_\lambda = \varepsilon_\lambda - a_{\lambda\mu}\varepsilon_\mu, \qquad (13.3)$$

where

$$a_{\lambda\mu} \equiv \frac{\partial U_\lambda}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}, \qquad (13.4)$$

the partial derivatives being obtained from (13.1) and (13.2). Using another set of equations consistent with (13.1), namely,

$$(\beta = 1, 2 \dots k) \quad C_\beta = G_\beta (U_1, U_2 \dots U_m), \qquad (13.5)$$

we obtain discrepancies

$$(\lambda = 1, 2 \dots m) \quad \phi_\lambda = \varepsilon_\lambda - b_{\lambda\mu}\varepsilon_\mu, \qquad (13.6)$$

where

$$b_{\lambda\mu} \equiv \frac{\partial U_\lambda}{\partial C_\beta}\frac{\partial C_\beta}{\partial U_\mu}, \qquad (13.7)$$

the partial derivatives being obtained from (13.1) and (13.5). It is required to prove that the ϕ's are linear functions of the θ's, and the θ's of the ϕ's.

(i) To verify the theorem for a simple case, I considered a Gaussian distribution, with 4 categories: so that $k = 2$, $m = 3$. This left only one discrepancy to be considered. Taking the U's to be the summed frequencies up to values Y_1, Y_2, Y_3 (not necessarily at equal intervals) of the variable Y, I first used u_1 and u_2 for determining the constants, and found (see (17.10)) that

$$p_{12}\theta_3 = p_{12}\varepsilon_3 + p_{23}\varepsilon_1 + p_{31}\varepsilon_2, \quad \ldots \ldots \ldots \quad (13.8)$$

where p_{fg} is a definite function of U_f and U_g. The expression on the right-hand side is symmetrical, and we should get the same expression for $p_{23}\theta_1$ if we used u_2 and u_3, or for $p_{31}\theta_2$ if we used u_3 and u_1. Thus the discrepancies by the three methods are multiples of one another.

(ii) It will be found that, if the mean and the standard deviation of the above distribution are denoted by M and D, p_{fg} is proportional to $\dfrac{\partial(U_f, U_g)}{\partial(M, D)}$. Hence, if we write

$$J_{fg} \equiv \frac{\partial(U_f, U_g)}{\partial(M, D)},$$

(13.8) becomes

$$J_{12}\theta_3 = J_{12}\varepsilon_3 + J_{23}\varepsilon_1 + J_{31}\varepsilon_2 \quad \ldots \ldots \ldots \quad (13.9)$$

This may also be written

$$\frac{\partial(U_1, U_2, U_3)}{\partial(M, D, U_3)}\theta_3 = \frac{\partial(U_1, U_2, U_3)}{\partial(M, D, U_3)}\varepsilon_3 + \frac{\partial(U_1, U_2, U_3)}{\partial(M, D, U_1)}\varepsilon_1 + \frac{\partial(U_1, U_2, U_3)}{\partial(M, D, U_2)}\varepsilon_2; \quad (13.10)$$

in which it is to be understood that in developing the Jacobians $\partial U_g/\partial U_h$ is to be taken* to be $\left|{}^g_h\right.$.

This formula can be extended to the more general case in which k is > 2, $m - k$ being $= 1$. The value of the discrepancy θ_f, obtained by finding $c_1, c_2 \ldots c_k$ from the u's other than u_f, is then given by

$$\frac{\partial(U_1, U_2 \ldots U_k, U_{k+1})}{\partial(C_1, C_2 \ldots C_k, U_f)}\theta_f = \frac{\partial(U_1, U_2 \ldots U_k, U_{k+1})}{\partial(C_1, C_2 \ldots C_k, U_\lambda)}\varepsilon_\lambda; \quad \ldots \quad (13.11)$$

where, as in the preceding paragraph, $\partial U_g/\partial U_h$ is to be taken to be $\left|{}^g_h\right.$. The expression on the right-hand side is the same for all values of f; the discrepancies under the different systems are therefore multiples of one another.

* The U's are given by equations $U_1 = f_1(M, D) \ldots (1)$, $U_2 = f_2(M, D) \ldots (2)$, $U_3 = f_3(M, D) \ldots (3)$. For determining M and D we use equations $M = F_1(U_1, U_2)$, $D = F_2(U_1, U_2)$, which are the inverse of (1) and (2). If we substitute from these equations in (3), we get U_3 in terms of U_1 and U_2, say $U_3 = g_3(U_1, U_2)$. What is stated above is that this last relation is ignored, U_3 being taken as independent of U_1 and U_2.

(iii) Next take the case in which there are only 2 constants, but m is 4 or more. Suppose that we first determine the constants from u_1 and u_2, obtaining discrepancies θ_λ, and then determine them from u_3 and u_4, obtaining discrepancies ϕ_λ. With the notation of (13.3) and (13.6), the formulæ for the discrepancies are

$$\theta_1 = 0,\ \theta_2 = 0,\ \theta_g = \varepsilon_g - a_{g1}\,\varepsilon_1 - a_{g2}\varepsilon_2\ (g = 3, 4 \ldots m), \quad . \quad . \quad . \quad (13.12)$$

$$\phi_3 = 0,\ \phi_4 = 0,\ \phi_h = \varepsilon_h - b_{h3}\,\varepsilon_3 - b_{h4}\,\varepsilon_4\ (h = 1, 2, 5 \ldots m)\ ; \quad . \quad . \quad (13.13)$$

and we want to prove that θ_g is a linear function of the $m - 2$ ϕ's (excluding ϕ_3 and ϕ_4), and ϕ_h a linear function of the $m - 2$ θ's (excluding θ_1 and θ_2).

Supposing that this really is the case, let us see what the form of the expression for θ_g in terms of the ϕ's will be. Since θ_g involves ε_g, ε_1, and ε_2, we should expect the expression to contain ϕ_g, ϕ_1, and ϕ_2; and it will not contain ϕ_3 or ϕ_4. But, if h is any suffix other than g, 1, or 2, and ϕ_h occurs in the expression for θ_g, it can only occur once; and there will be nothing to cancel out ε_h in order to produce the expression in (13.12) when the ϕ's in θ_g are converted into ε's by (13.13). It follows that, if θ_g is a linear function of the ϕ's, it can only involve ϕ_g, ϕ_1, and ϕ_2. And, since the coefficients of ε_g, ε_1, and ε_2 in ϕ_g, ϕ_1, and ϕ_2 are unity, the coefficients of the latter in the expression for θ_g in terms of the ϕ's must be the same as the coefficients of ε_g, ε_1, and ε_2 in the expression for θ_g in terms of the ε's, *i.e.*, in (13.12). What we should have to prove, therefore, is that

$$\theta_g = \phi_g - a_{g1}\phi_1 - a_{g2}\phi_2,$$

and similarly that

$$\phi_h = \theta_h - b_{h3}\theta_3 - b_{h4}\theta_4.$$

(iv) Without stopping to prove this, let us pass on to the general formulæ suggested by it and see whether they are true. Substituting from (13.4) in (13.3), we see that the formula to be established for θ_g is

$$\theta_g = \phi_g - \frac{\partial U_g}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\,\phi_\mu\,;$$

the partial derivatives being obtained from (13.1) and (13.2).

Substituting from (13.6) and (13.7) in the right-hand side of this formula, and using (8.11), we get

$$\varepsilon_g - \frac{\partial U_g}{\partial C_\beta}\frac{\partial C_\beta}{\partial U_\mu}\,\varepsilon_\mu - \frac{\partial U_g}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\left(\varepsilon_\mu - \frac{\partial U_\mu}{\partial C_\beta}\frac{\partial C_\beta}{\partial U_\nu}\,\varepsilon_\nu\right)$$

$$= \left(\varepsilon_g - \frac{\partial U_g}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\,\varepsilon_\mu\right) - \left(\frac{\partial U_g}{\partial C_\beta} - \frac{\partial U_g}{\partial C_\alpha}\frac{\partial C_\alpha}{\partial U_\mu}\frac{\partial U_\mu}{\partial C_\beta}\right)\frac{\partial C_\beta}{\partial U_\nu}\,\varepsilon_\nu$$

$$= \theta_g - \left(\frac{\partial U_g}{\partial C_\beta} - \frac{\partial U_g}{\partial C_\alpha}\Big|_\beta^\alpha\right)\frac{\partial C_\beta}{\partial U_\nu}\,\varepsilon_\nu$$

$$= \theta_g - \left(\frac{\partial U_g}{\partial C_\beta} - \frac{\partial U_g}{\partial C_\beta}\right)\frac{\partial C_\beta}{\partial U_\nu}\,\varepsilon_\nu$$

$$= \theta_g\,;$$

so that

$$\theta_g = \phi_g - \frac{\partial U_g}{\partial C_a}\frac{\partial C_a}{\partial U_\mu}\phi_\mu. \quad \ldots \quad (13.14)$$

This gives θ_g as a linear function of the ϕ's. In the same way we can show that

$$\phi_h = \theta_h - \frac{\partial U_h}{\partial C_\beta}\frac{\partial C_\beta}{\partial U_\mu}\theta_\mu; \quad \ldots \quad (13.15)$$

so that ϕ_h is a linear function of the θ's. Thus the proposition is proved.

(v) We can state this result briefly as follows. Suppose that by one method we obtain discrepancies

$$\theta_\lambda = A_{\lambda\mu}\varepsilon_\mu,$$

and that by another method we obtain discrepancies

$$\phi_\lambda = B_{\lambda\mu}\varepsilon_\mu;$$

the set of m u's being the same in the two cases. Then the θ's and the ϕ's are connected by the relations

$$\theta_\lambda = A_{\lambda\mu}\phi_\mu, \qquad \phi_\lambda = B_{\lambda\mu}\theta_\mu. \quad \ldots \quad (13.16)$$

For a verification of the theorem by a numerical example, and comparison of the value of P′ with that of P given by PEARSON's formula, see §§ 19, 20.

14. *Unknown quantities in the formulæ.*—The formulæ at which we have arrived contain expressions of the form $\partial U/\partial C$ and $\partial C/\partial U$; and we have proceeded as if we knew, or could find, the true values of these quantities. This, of course, we cannot do: we must adopt the usual practice of using values given approximately by the data. The most satisfactory procedure is to find the most probable values of the constants, on the usual assumption as to relative probabilities of different values. A method of doing this, for frequency-constants, is given in Appendix I.

15. *Choice of method: ratios of aggregate frequencies.*—As we get the same result whatever method is adopted, we are at liberty to choose our method. It will usually be found convenient to choose it so as to simplify calculation of the m.ss. and m.pp. of the θ's. In view of (7.2), this means that we should (i) restrict the c-formulæ (6.4) to k of the u's and (ii) choose the u's—or, in the language of § 12 (i), choose v's which are functions of the u's—so as to be statistically independent, *i.e.*, so that their m.pp. e. shall be zero.

In the case of a frequency-distribution, we can do this by dealing with the ratios of aggregate frequencies. Let the numbers in the successive categories (taken in any order we please) be $n_0, n_1, n_2 \ldots n_m$, their sum being N; and let the successive sums be $t_0, t_1, t_2 \ldots t_m$, so that

$$t_0 = 0,\ t_1 = n_0, \ldots t_{s+1} = n_0 + n_1 + \ldots + n_s, \ldots t_{m+1} = n_0 + n_1 + \ldots + n_m = N. \quad (15.1)$$

Then it is obvious that the existence of an error in t_{s+1} will not affect the proportion of the t_{s+1} that fall into t_s, and so on. We therefore write

$$(f = 1, 2 \ldots m)\ r_f \equiv \frac{t_f}{t_{f+1}}; \quad \ldots \ldots \ldots \quad (15.2)$$

and it will be found that (capitals, as before, denoting true values)

$$\text{m.s.e. of } r_f = \frac{R_f(1-R_f)}{T_{f+1}} = R_f^2\left(\frac{1}{T_f} - \frac{1}{T_{f+1}}\right), \quad \ldots \ldots \quad (15.3)$$

$$\text{m.p.e. of } r_f \text{ and } r_g\ (f \neq g) = 0. \quad \ldots \ldots \ldots \quad (15.4)$$

The general method of making use of this property is an extension of the method for a Gaussian distribution, explained in § 18.

16. *Limitations of the method.*—The results obtained in §§ 5–15 are quite general, and the scope of the method is, in theory, very wide. It is applicable, for instance, to such a problem as that of testing whether two or more frequency-distributions can be regarded as independent random samples from the same population, of a specified kind.

In practice, however, the value of the general results is limited by the amount of calculation involved. The expression $\Theta^{\pi\sigma}\theta_\pi\theta_\sigma$ looks simple, but its calculation in the ordinary way involves finding the $\frac{1}{2}(m-k)(m-k+1)$ co-factors of a symmetrical determinant of order $m-k$. This is a troublesome matter, unless $m-k$ is very small. There is therefore a good deal to be done before the problem can be regarded as completely solved. We have already (see § 4) had to limit our solution to the class of cases in which the errors are relatively small; and we have now to make the further general limitation, so far as practical utility is concerned, that $m-k$ is not to be large.

The limitation will not be necessary in any class of cases in which we can reduce $\Theta^{\pi\sigma}\theta_\pi\theta_\sigma$ to an expression which can be easily calculated. The important case is that of frequency-distributions, for which PEARSON has given the expression χ_s^2 of (3.6). The question which has now to be considered is whether, or under what conditions, χ_s^2 can be used in place of χ'^2. I begin by obtaining some general formulæ for the case of a normal distribution, for the purpose of considering a numerical example.

APPLICATION TO NORMAL DISTRIBUTION.

17. *Formulæ for normal distribution.*—For a normal or Gaussian distribution, we use the method and notation of § 15: the categories being determined by values Y_1, $Y_2 \ldots Y_m$ (not necessarily at equal intervals) of a variate Y, in addition to the bounding values $Y_0\ (= -\infty)$ and $Y_{m+1}\ (= +\infty)$.

(i) Let the mean and s.d. of the true distribution be M and D; the values determined from the data being m and d, with errors μ and δ. Then, if the true proportion of cases for which Y is less than Y_f is A_f, we shall have, with the ordinary notation for a Gaussian distribution,

$$\left.\begin{array}{c} X_f \equiv (Y_f - M)/D, \quad Z_f = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X_f^2}, \quad A_f = \int_{-\infty}^{X_f} Z\, dX, \\ R_f = A_f/A_{f+1}, \quad T_f = NA_f. \end{array}\right\} \quad . \; . \quad (17.1)$$

Corresponding to the true values A_f, R_f, T_f, we have the observed values a_f, r_f, t_f, with errors α_f, ρ_f, τ_f.

We know that

$$\text{m.p.e. of } Na_f \text{ and } Na_g = NA_f(1 - A_g)\ (f \leq g)\,; \quad . \; . \; . \; . \; . \quad (17.2)$$

and thence it follows that, as stated in § 15,

$$\text{m.s.e. of } r_f = \frac{R_f(1 - R_f)}{T_{f+1}} = R_f^2\left(\frac{1}{T_f} - \frac{1}{T_{f+1}}\right), \quad . \; . \; . \; . \; . \; . \quad (17.3)$$

$$\text{m.p.e. of } r_f \text{ and } r_g\ (f \neq g) = 0. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.4)$$

(ii) It will be found from (17.1) that errors μ and δ in m and d will produce in the calculated value of A_f an error

$$- Z_f\mu/D - X_fZ_f\,\delta/D, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.5)$$

so that the discrepancy in a_f will be

$$\alpha'_f = \alpha_f + Z_f\mu/D + X_fZ_f\,\delta/D. \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.6)$$

(iii) Now suppose that we determine m and d from the observed values a_g and a_h, which contain errors α_g and α_h. We can find the resulting errors μ and δ by using (17.5); for, if we find m and d from a_g and a_h, and then with these values of m and d calculate A_g and A_h, we shall get back to a_g and a_h, with their errors α_g and α_h. We therefore have equations

$$\left.\begin{array}{l} \alpha_g = - Z_g \,.\, \mu/D - X_gZ_g \,.\, \delta/D \\ \alpha_h = - Z_h \,.\, \mu/D - X_hZ_h \,.\, \delta/D \end{array}\right\} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.7)$$

whence, if we write

$$p_{gh} \equiv (X_h - X_g)\, Z_gZ_h, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.8)$$

we find that

$$\left.\begin{array}{l} p_{gh}\mu/D = - X_hZ_h\alpha_g + X_gZ_g\alpha_h \\ p_{gh}\delta/D = Z_h\alpha_g - Z_g\alpha_h \end{array}\right\}. \quad . \; . \; . \; . \; . \; . \; . \; . \quad (17.9)$$

Substituting in (17.6), the discrepancy α'_f in a_f, due to using m and d as determined by a_g and a_h, is given by

$$p_{gh}\alpha'_f = p_{gh}\alpha_f + p_{hf}\alpha_g + p_{fg}\alpha_h, \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (17.10)$$

which is equivalent to (13.8).

(iv) To deal with the R's, we must choose a_g and a_h. The A's are $A_1, A_2 \ldots A_m$ in addition to A_0 (= 0) and A_{m+1} (= 1); and the R's are $R_1 = A_1/A_2$, $R_2 = A_2/A_3$, $\ldots R_m = A_m/A_{m+1} = A_m$. Suppose we determine m and d from the last two r's, which is equivalent to determining them from a_{m-1} and a_m. Then, ρ_f being the error of r_f, we have

$$\frac{\rho_f}{R_f} = \text{error of } \log r_f = \text{error of } (\log a_f - \log a_{f+1})$$

$$= \frac{\alpha_f}{A_f} - \frac{\alpha_{f+1}}{A_{f+1}}, \quad \ldots \quad (17.11)$$

with, as a special case,

$$\frac{\rho_m}{R_m} = \frac{\alpha_m}{A_m}. \quad \ldots \quad (17.12)$$

It will be found convenient, for purposes of calculation, to deal with ρ/R, which is the error of $\log r$.

If we find m and d from r_{m-1} and r_m, calculate the resulting values of $R'_1, R'_2 \ldots R_{m-2}$, and subtract them from the observed values $r_1, r_2 \ldots r_{m-2}$, we get the discrepancies $\rho'_1, \rho'_2 \ldots \rho'_{m-2}$. We can write

$$\frac{\rho'_f}{R_f} = \frac{\rho_f}{R_f} + P_f \frac{\rho_{m-1}}{R_{m-1}} + Q_f \frac{\rho_m}{R_m}, \quad \ldots \quad (17.13)$$

where P_f and Q_f are coefficients depending on f. By means of (17.10) it may be shown that

$$P_f = \frac{c_{f+1,\,m} - c_{f,\,m}}{c_{m-1,\,m}}, \quad Q_f = P_f + \frac{c_{f+1,\,m-1} - c_{f,\,m-1}}{c_{m,\,m-1}}, \quad (17.14)$$

where

$$c_{gh} \equiv \frac{C\,(X_h - X_g)\,Z_g}{A_g}; \quad \ldots \quad (17.15)$$

C being an arbitrary multiplier which is introduced for convenience of calculation and disappears in P_f and Q_f. The P's and Q's having been calculated, we have

$$\left.\begin{aligned} \text{m.sq. of } \frac{\rho'_f}{R_f} &= \left(\frac{1}{T_f} - \frac{1}{T_{f+1}}\right) + P_f^2\left(\frac{1}{T_{m-1}} - \frac{1}{T_m}\right) + Q_f^2\left(\frac{1}{T_m} - \frac{1}{N}\right) \\ \text{m.p. of } \frac{\rho'_f}{R_f} \text{ and } \frac{\rho'_g}{R_g} &= P_f P_g\left(\frac{1}{T_{m-1}} - \frac{1}{T_m}\right) + Q_f Q_g\left(\frac{1}{T_m} - \frac{1}{N}\right) \end{aligned}\right\}. \quad (17.16)$$

We then proceed as in § 10, θ being replaced by ρ/R. The T's, R's, P's, and Q's occurring in (17.16) must be those given by the most probable values of the constants.

18. *Procedure.*—For comparison with PEARSON's formula, I worked out an example; and, partly to check my calculations, partly to verify that different methods gave the same result, I did this by two methods.

The procedure, briefly, was as follows :—

(i) Construct a typical case as explained in Appendix II, *i.e.*, a case based on an exact Gaussian distribution, but with incorporation of errors such as might be due to random sampling.

(ii) Find the most probable values of the constants, as described in Appendix I.

(iii) With these values for the constants, calculate the values of X, Z, A, T, R, P, Q, on the basis of the method which is to be used in (iv), and thence calculate the mean-product set $\Theta_{\pi\sigma}$ of the discrepancies $\theta \equiv \rho'/\text{R}$. Then find the reciprocal set $\Theta^{\pi\sigma}$.

(iv) Take the distribution as determined by the last two values of t, and find the calculated values R′, the discrepancies ρ', and the ratios of these discrepancies to the "standard" R's of (iii). For this purpose we need not calculate m and d: we find the x's corresponding to t_{m-1} and t_m, and obtain the remaining X′'s by 1st-difference extrapolation.

(v) The values of ρ'/R found in (iv) being the θ's, complete the calculation of $\chi'^2 \equiv \Theta^{\pi\sigma}\theta_\pi\theta_\sigma$, and thence find P′ by (10.2).

(vi) Repeat (v), using the first two values of t in place of the last two. The value of $\Theta^{\pi\sigma}\theta_\pi\theta_\sigma$ should be practically the same as that obtained in (v), *i.e.*, the ratio of their difference to either of them should be of order $1/\sqrt{\text{N}}$ or less.

I began by constructing a distribution of the ordinary type, by equal intervals, taking $m = 6$ (*i.e.*, 7 cells). But the results of the two methods did not agree. The reason, no doubt, was the smallness of the numbers in the extreme cells, which made them quite unsuitable for determining a theoretical distribution to cover the whole range. I therefore constructed a second example, as stated in the next paragraph.

19. *Numerical example.*—(i) The example was constructed as described in Appendix II. Original (hypothetical) distribution with mean $= 2/3$, s.d. $= 20/3$; N $= 300{,}000$; $m = 5$, the determining values of Y being -6, -2, $+1$, $+4$, $+8$. The data, after introducing errors of random sampling, were* :—

f =	0	1	2	3	4	5	6
Y_f =	$-\infty$	-6	-2	$+1$	$+4$	$+8$	$+\infty$
t =	0	47494	103459	155641	207099	259240	300000

(ii) A third approximation to the constants of the distribution gave (see Appendix I)

$$\text{Mean of Y} = 0\cdot67658\ \ 99206,$$
$$\text{s.d. of Y} = 6\cdot66950\ \ 00707.$$

(I mostly worked to 10 figures, but 7 would probably have been sufficient.)

* The calculations were done with a machine lent me by the Royal Society.

(iii) This gave (for determining the distribution by means of a_5 and a_4)—

Mean-product set Θ_{fg}.

	$f = 3$.	$f = 2$.	$f = 1$.
$g = 3$	·00000 43876 64877	0·00000 46137 73786	0·00000 94403 32876
$g = 2$		0·00001 09772 78683	0·00001 58917 70403
$g = 1$			0·00004 42762 28166

Reciprocal set Θ^{fg}.

	$f = 3$.	$f = 2$.	$f = 1$.
$g = 3$	+ 4 78448·91683	− 1 11181·66494	− 62106·54477
$g = 2$		+ 2 15469·03706	− 53631·31951
$g = 1$			+ 55077·02831

(iv) The observed values a_5 and a_4 gave

$$x_5 = +1{\cdot}09907\ \ 96317\ \ 1, \qquad x_4 = +0{\cdot}49678\ \ 59557\ \ 9,$$

whence we get the calculated values

f.	X'_f.	ρ'_f.	$\theta_f \equiv \rho'_f \div$ (standard R_f).
3	+0·04506 56988 5	+0·00120 34887	+0·00160 09754
2	−0·40665 45580 9	+0·00420 91305	+0·00635 28147
1	−1·00894 82340 1	+0·00163 45521	+0·00355 07859

(v) Combining (iii) and (iv), we get

$$\chi'^2 = \Theta^{\pi\sigma}\theta_\pi\theta_\sigma = 5{\cdot}22941\ \ 009, \qquad \chi' = 2{\cdot}28679\ \ 035, \qquad P' = 0{\cdot}15574\ \ 9.$$

(vi) Repeating the process, using a_1 and a_2 instead of a_5 and a_4, I got

$$\chi'^2 = 5{\cdot}23041\ \ 587, \qquad \chi' = 2{\cdot}28701\ \ 025, \qquad P' = 0{\cdot}15568\ \ 2.$$

(vii) The results in (v) and (vi), obtained by starting with different pairs of a's, agree very closely; the two values of χ'^2 are within $\frac{1}{2}$ in 5230, or roughly 1 in 10,000, of their mean, and the values of P′ differ by less than 0·0001. We can safely take the true value of χ'^2 to be 5·230, with an error not exceding 1 or 2 in final figure.

20. *Comparison with* PEARSON's *value.*—Now let us see what we should get by using PEARSON's formula (3.6). His adoption of the formula was based (ref. 6, p. 165) on

the supposition that the "best" values of the constants were to be used; but it is instructive to begin without making this supposition.

(i) Using a_5 and a_4, as in the first method of § 19, I obtained

$$\chi_s^2 = 14\cdot99160\quad 4, \qquad \chi_s = 3\cdot87189\quad 9.$$

From this, using Pearson's formula (3.5), we get

$$P = 0\cdot01039\quad 8.$$

But, if we use Fisher's modification (ref. 3, p. 93) of Pearson's formula, *i.e.* (see § 10 above) replace m in (3.5) by $m - k$ ($= 3$), we get

$$P = 0\cdot00182\quad 4.$$

(ii) Using a_1 and a_2, as in the second method of § 19, we get

$$\chi_s^2 = 14\cdot98679\quad 1, \qquad \chi_s = 3\cdot87127\quad 8;$$

and thence

$$P = 0\cdot01041\quad 9 \quad \text{or} \quad 0\cdot00182\quad 8,$$

according as we use $m = 5$ or replace it by $m - k = 3$.

(iii) The values of χ_s^2 obtained by these two methods happen to agree with one another very well, but they are much too great, and therefore give too small a value of P, even if we retain χ^{m-1} instead of χ^{m-k-1} in the integration. This result is due to the fact that in each case we take account of only two values of a, so that we are using a distribution which is a poor fit to the data. Let us therefore see what happens if we use the best fit we can get, namely, the distribution for which the mean and the s.d. have their most probable values. Since the numbers involved are large, this (as may be seen by applying Stirling's theorem to (I.4) of Appendix I) is practically the same thing as using the minimum value of χ_s^2 (*cf.* Bowley and Connor, ref. 1, p. 7; Fisher, ref. 4, p. 7, and ref. 5, p. 446).

The values of the constants which were used in § 19 for finding the standard values of R, P, Q, etc., were those given by a third approximation, namely

$$M = 0\cdot67658\quad 99206, \qquad D = 6\cdot66950\quad 00707.$$

To make sure of getting the best result I went to a fourth approximation, but found that this only affected the figures in the 10th decimal place. The calculated A's used for § 19 (iii) were therefore sufficient. Using these, I obtained

$$\chi_s^2 = 5\cdot22013\quad 6, \qquad \chi_s = 2\cdot28476\quad 2,$$

whence

$$P = 0\cdot38961\quad 0 \quad \text{or} \quad 0\cdot15636\quad 9$$

for $m = 5$ or $m - k = 3$ respectively. This value of χ_s^2 is within about 1 in 500 of the true value χ'^2, so that the difference between them is relatively of the 1st order of small quantities; and the value of P, using $m - k$ in the integral in place of m, is therefore approximately the same as the true value P′.

21. *Another example.*—We cannot, of course, lay much stress on a single example. The calculations for finding $\Theta^{\pi\sigma}\theta_\pi\theta_\sigma$ are so laborious that I have not worked out another example of the same character. But, as a simple test for a Gaussian distribution, I have taken the extreme case in which $m = 3$, so that there is only one discrepancy to be calculated. I combined the first two groups, and also the last two, in § 19, so that the numbers in the four categories were

$$103459, \quad 52182, \quad 51458, \quad 92901,$$

the data being

Y =	$-\infty$	-2	$+1$	$+4$	$+\infty$
t =	0	103459	155641	207099	300000.

By successive approximations, I found for the most probable values of the constants

$$M = 0\cdot67570 \quad 15490, \quad D = 6\cdot69658 \quad 61344,$$

the last figure in each case being approximately correct. Working in one direction I found

$$\chi'^2 = 1\cdot43566 \quad 476;$$

and in the other direction

$$\chi'^2 = 1\cdot43309 \quad 497.$$

The two values do not agree so closely as in the former example, but they only differ by about 1 in 500, so they may be taken to be correct. The PEARSON formula (3.6) gives

$$\chi_s^2 = 1\cdot43439 \quad 696,$$

which lies between the two values of χ'^2. The PEARSON-FISHER result for P is thus practically identical with the true value.

These two examples definitely support the view that PEARSON'S expression χ_s^2 may be taken as giving the value of χ'^2, provided that the calculated values of the numbers in the different categories are found by using the most probable values of the constants of the distribution.

GENERAL THEORY.

22. *Contingency tables.*—The application of the theory to testing of independence in the case of a contingency table has been considered by G. UDNY YULE (ref. 8)*. Inconsistencies in results obtained by PEARSON'S formula suggested the correction

* I am indebted to a referee for calling my attention to this investigation, which has led me to alter the views expressed in the first draft of my paper.

subsequently put forward by FISHER, namely, the substitution of $m - k$ for m in (3.5). This correction being adopted, the method employed by YULE for testing the accuracy of PEARSON'S χ_s^2 was to make a large number of groups of experiments under conditions securing independence, calculate the value of χ_s^2 for each group, and compare the frequency-distribution of these values of χ_s^2 with the corresponding theoretical distribution as obtained from the corrected formula for P. There was found to be a substantial agreement.

As has already been pointed out, different ways of calculating the discrepancies give different values of χ_s^2. The method adopted by YULE was to calculate the "independence-values" from the observed sub-totals (totals of rows and of columns, "marginal frequencies"). It is shown below that this is equivalent to using the most probable values of the constants. YULE'S results therefore give an experimental support to the view that the resulting value of χ_s^2 is practically equal to χ'^2.

To prove this proposition, I will take first the simple case of an association ("fourfold") table, and then the general case of a contingency table of any number of columns and rows.

(i) The association table may be set out in the form given below; A and B denoting true values, and α β γ true errors.

N (1)	$A + \alpha$ (2)	$N - A - \alpha$ (3)
$B + \beta$ (4)	$\frac{AB}{N} + \gamma$ (5)	$\frac{(N - A)B}{N} + \beta - \gamma$ (6)
$N - B - \beta$ (7)	$\frac{A(N - B)}{N} + \alpha - \gamma$ (8)	$\frac{(N - A)(N - B)}{N} - \alpha - \beta + \gamma$ (9)

There are 3 algebraically independent entries, and 2 unknown constants, so that $m = 3$, $k = 2$. There is therefore only one discrepancy to be considered.

(*a*) To find the true value χ'^2, we may use any discrepancy we like. Let us take the discrepancy between the entry in (5), as calculated from (2) and (4), and its actual value. We have then

$$N\theta_1 = N\gamma - B\alpha - A\beta; \quad \ldots \ldots \quad (22.1)$$

whence we find (*cf.* SHEPPARD, ref. 7, § 21; YULE, ref. 8, p. 96) that

$$\chi'^2 = \frac{\theta_1^2}{\Theta_{11}} = \frac{N(B\alpha + A\beta - N\gamma)^2}{A(N - A)B(N - B)}. \quad \ldots \ldots \quad (22.2)$$

(*b*) To find the most probable values of the constants, say A and B, we use Appendix I. In the notation of the Appendix, $P_0\, P_1\, P_2\, P_3$ and $n_0\, n_1\, n_2\, n_3$ correspond to entries in compartments (5) (6) (8) (9) of the table. Let the unknown constants C_1 and C_2 be the proportions which properly belong to compartments (2) and (4). Then

$$C_1 = P_0 + P_2, \qquad C_2 = P_0 + P_1. \quad . \quad . \quad . \quad (22.3)$$

The condition of independence gives

$$P_0 = C_1 C_2, \;\; P_1 = (1 - C_1) C_2, \;\; P_2 = C_1 (1 - C_2), \;\; P_3 = (1 - C_1)(1 - C_2). \quad . \quad (22.4)$$

By differentiation, we find that

$$\frac{\partial \log P_0}{\partial C_1} = \frac{1}{C_1} \qquad \frac{\partial \log P_1}{\partial C_1} = -\frac{1}{1 - C_1}, \text{ etc.}$$

Hence, by (I.5) of Appendix I,

$$\frac{n_0 + n_2}{C_1} - \frac{n_1 + n_3}{1 - C_1} = 0, \qquad \frac{n_0 + n_1}{C_2} - \frac{n_2 + n_3}{1 - C_2} = 0,$$

so that

$$C_1 = \frac{n_0 + n_2}{N}, \qquad C_2 = \frac{n_0 + n_1}{N}. \quad . \quad . \quad . \quad (22.5)$$

Thus the most probable values of the constants are the values deduced from the observed sub-totals, the entries in the substantive table being ignored (*cf.* BOWLEY and CONNOR, ref. 1, p. 8).

(*c*) It remains to prove that the resulting value of χ_s^2 is equal to χ'^2. Working from the sub-totals, the discrepancies between the calculated entries in (5) (6) (8) (9) and the actual entries will be θ_1, $-\theta_1$, $-\theta_1$, θ_1; θ_1 being given by (22.1). The value of χ_s^2 will therefore be

$$\theta_1^2 \left\{ \frac{N}{AB} + \frac{N}{(N - A) B} + \frac{N}{A (N - B)} + \frac{N}{(N - A)(N - B)} \right\}, \quad . \quad . \quad . \quad (22.6)$$

which will be found to be equal to χ'^2 as given by (22.2).

(ii) The general case of a contingency table can be dealt with on the same lines.

Suppose that the table has $h + 1$ columns and $k + 1$ rows, so that the number of algebraically independent entries is $(h + 1)(k + 1) - 1 = hk + h + k$. Then the number of unknown constants, on the hypothesis of independence, is $h + k$. Thus the numbers corresponding to the m and k of the general formulæ are $hk + h + k$ and $h + k$.

We adapt the notation of Appendix I by using double suffixes for the cell-entries of n's and P's. For the sub-totals of n's we can write

$$n_{f\omega} \equiv n_{f0} + n_{f1} + \ldots + n_{fk}, \; n_{\omega g} \equiv n_{0g} + n_{1g} + \ldots + n_{hg}.$$

For the sub-totals of P's we can write

$$Q_f \equiv P_{f0} + P_{f1} + \ldots + P_{fk}, \quad R_g \equiv P_{0g} + P_{1g} + \ldots + P_{hg} \quad . \quad . \quad . \quad . \quad (22.7)$$

Thus the scheme of the n's (observed) is

$N \equiv n_{\omega\omega}$	$n_{0\omega}$	$n_{1\omega}$	$n_{2\omega}$	$n_{3\omega}$	$\ldots$	$n_{h\omega}$
$n_{\omega 0}$	n_{00}	n_{10}	n_{20}	n_{30}	$\ldots$	n_{h0}
$n_{\omega 1}$	n_{01}	n_{11}	n_{21}	n_{31}	$\ldots$	n_{h1}
$n_{\omega 2}$	n_{02}	n_{12}	n_{22}	n_{32}	$\ldots$	n_{h2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$n_{\omega k}$	n_{0k}	n_{1k}	n_{2k}	n_{3k}	$\ldots$	n_{hk}

and that of the P's (theoretical) is

1	Q_0	Q_1	Q_2	Q_3	$\ldots$	Q_h
R_0	P_{00}	P_{10}	P_{20}	P_{30}	$\ldots$	P_{h0}
R_1	P_{01}	P_{11}	P_{21}	P_{31}	$\ldots$	P_{h1}
R_2	P_{02}	P_{12}	P_{22}	P_{32}	$\ldots$	P_{h2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
R_k	P_{0k}	P_{1k}	P_{2k}	P_{3k}	$\ldots$	P_{hk}

where Q_0 and R_0 are to be replaced by $1 - Q_1 - Q_2 \ldots - Q_h$ and $1 - R_1 - R_2 - \ldots - R_k$ respectively. The condition of independence gives

$$P_{fg} = Q_f R_g. \qquad (22.8)$$

(a) To find χ'^2, suppose that the true values of the n's are N_{00}, N_{01}, N_{10}, etc., so that

$$N_{fg} = NP_{fg} = NQ_f R_g ; \qquad (22.9)$$

the true errors being ε_{00}, ε_{01}, ε_{10}, etc. The sum of these errors is 0; we shall take ε_{00} to be the dependent error, so that $\varepsilon_{00} = -\varepsilon_{01} - \varepsilon_{02} - \ldots - \varepsilon_{10} - \varepsilon_{11} - \ldots - \varepsilon_{hk}$. As the $h+k$ unknown constants we shall take $Q_1\, Q_2 \ldots Q_h$ and $R_1\, R_2 \ldots R_k$; and for calculating the discrepancies we shall take these constants to have their observed values, *i.e.*, we shall work from the actual sub-totals $n_{1\omega}\, n_{2\omega} \ldots n_{h\omega}$ and $n_{\omega 1}$, $n_{\omega 2} \ldots n_{\omega k}$. The discrepancies which will appear in χ'^2 are therefore those that do not contain a 0 in the suffix, *i.e.*, they are those of the entries that remain when we have struck out the 1st column and the 1st row. The $h+k$ relations between discrepancies, given by (9.1), merely state that the sum of the discrepancies in any complete column or in any complete row is 0, *i.e.*,

$$\left.\begin{array}{ll} (f = 0, 1, 2 \ldots h) & \theta_{f1} + \theta_{f2} + \ldots + \theta_{fk} = -\theta_{f0} \\ (g = 0, 1, 2 \ldots k) & \theta_{1g} + \theta_{2g} + \ldots + \theta_{hg} = -\theta_{0g} \end{array}\right\} \qquad (22.10)$$

Working from the given Q's and R's, we find that the discrepancy of n_{fg} is

$$\begin{aligned}\theta_{fg} &= (N_{fg} + \varepsilon_{fg}) - (NQ_f + \varepsilon_{f\omega})(NR_g + \varepsilon_{\omega g})/N \\ &= \varepsilon_{fg} - Q_f\varepsilon_{\omega g} - R_g\varepsilon_{f\omega}\,; \qquad (22.11)\end{aligned}$$

and thence that

$$\left.\begin{aligned}&\text{m.s. of } \theta_{fg} = NQ_f(1-Q_f)\,R_g(1-R_g)\\ &\text{m.p. of } \theta_{fg} \text{ and } \theta_{fg'} = -NQ_f(1-Q_f)\,R_gR_{g'}\\ &\quad,,\quad \theta_{fg} \quad ,, \quad \theta_{f'g} = -NQ_fQ_{f'}\,R_g(1-R_g)\\ &\quad,,\quad \theta_{fg} \quad ,, \quad \theta_{f'g'} = NQ_fQ_{f'}R_gR_{g'}\end{aligned}\right\} \qquad (22.12)$$

These mean squares and mean products are the entries in the set $\Theta_{\pi\sigma}$. We have then to find the reciprocal set $\Theta^{\pi\sigma}$. This would be rather troublesome, were it not that the entries are indicated by the expression we have to obtain for χ'^2. It will be found (see example below) that in $\chi'^2 \equiv \Theta^{\pi\sigma}\theta_\pi\theta_\sigma$

$$\left.\begin{aligned}\text{co. } \theta_{fg}^2 &= \frac{1}{N_{00}} + \frac{1}{N_{0g}} + \frac{1}{N_{f0}} + \frac{1}{N_{fg}}\\ \text{co. } 2\theta_{fg}\theta_{fg'} &= \frac{1}{N_{00}} + \frac{1}{N_{f0}}\\ \text{co. } 2\theta_{fg}\theta_{f'g} &= \frac{1}{N_{00}} + \frac{1}{N_{0g}}\\ \text{co. } 2\theta_{fg}\theta_{f'g'} &= \frac{1}{N_{00}}\end{aligned}\right\} \qquad (22.13)$$

Writing out χ'^2 accordingly, collecting terms, and taking account of (22.10), we get finally

$$\chi'^2 = \frac{\theta_{00}^2}{N_{00}} + \frac{\theta_{01}^2}{N_{01}} + \ldots + \frac{\theta_{10}^2}{N_{10}} + \ldots + \frac{\theta_{hk}^2}{N_{hk}}\,; \qquad (22.14)$$

in which the θ's are the discrepancies obtained by working from the sub-totals.

To make the work clear, take as an example a table with 3 rows and 3 columns. By (22.12), the mean-product set, with each element divided by N, will be

	θ_{11}	θ_{12}	θ_{21}	θ_{22}
θ_{11}	$Q_1(1-Q_1)\,R_1(1-R_1)$	$-Q_1(1-Q_1)\,R_1R_2$	$-Q_1Q_2R_1(1-R_1)$	$Q_1Q_2R_1R_2$
θ_{12}	$-Q_1(1-Q_1)\,R_1R_2$	$Q_1(1-Q_1)\,R_2(1-R_2)$	$Q_1Q_2R_1R_2$	$-Q_1Q_2R_2(1-R_2)$
θ_{21}	$-Q_1Q_2R_1(1-R_1)$	$Q_1Q_2R_1R_2$	$Q_2(1-Q_2)\,R_1(1-R_1)$	$-Q_2(1-Q_2)\,R_1R_2$
θ_{22}	$Q_1Q_2R_1R_2$	$-Q_1Q_2R_2(1-R_2)$	$-Q_2(1-Q_2)\,R_1R_2$	$Q_2(1-Q_2)\,R_2(1-R_2)$

The set which is to be proved to be the reciprocal set, with each element multiplied by N, is

	θ_{11}	θ_{12}	θ_{21}	θ_{22}
θ_{11}	$\frac{1}{P_{00}}+\frac{1}{P_{01}}+\frac{1}{P_{10}}+\frac{1}{P_{11}}$	$\frac{1}{P_{00}}+\frac{1}{P_{10}}$	$\frac{1}{P_{00}}+\frac{1}{P_{01}}$	$\frac{1}{P_{00}}$
θ_{12}	$\frac{1}{P_{00}}+\frac{1}{P_{10}}$	$\frac{1}{P_{00}}+\frac{1}{P_{02}}+\frac{1}{P_{10}}+\frac{1}{P_{12}}$	$\frac{1}{P_{00}}$	$\frac{1}{P_{00}}+\frac{1}{P_{02}}$
θ_{21}	$\frac{1}{P_{00}}+\frac{1}{P_{01}}$	$\frac{1}{P_{00}}$	$\frac{1}{P_{00}}+\frac{1}{P_{01}}+\frac{1}{P_{20}}+\frac{1}{P_{21}}$	$\frac{1}{P_{00}}+\frac{1}{P_{20}}$
θ_{22}	$\frac{1}{P_{00}}$	$\frac{1}{P_{00}}+\frac{1}{P_{02}}$	$\frac{1}{P_{00}}+\frac{1}{P_{20}}$	$\frac{1}{P_{00}}+\frac{1}{P_{02}}+\frac{1}{P_{20}}+\frac{1}{P_{22}}$

The reciprocal relation can be checked in the usual way by adding products of corresponding terms in pairs of rows, remembering that

$$P_{00}=(1-Q_1-Q_2)(1-R_1-R_2),\quad P_{01}=(1-Q_1-Q_2)R_1,\quad P_{02}=(1-Q_1-Q_2)R_2,$$
$$P_{10}=Q_1(1-R_1-R_2),\quad P_{11}=Q_1R_1,\quad P_{12}=Q_1R_2,$$
$$P_{20}=Q_2(1-R_1-R_2),\quad P_{21}=Q_2R_1,\quad P_{22}=Q_2R_2.$$

(*b*) For the most probable values of the constants, the formula (I.5) of Appendix I becomes

$$\sum_{f=0}^{h}\sum_{g=0}^{k} n_{fg}\frac{\partial \log P_{fg}}{\partial C_\alpha}=0, \quad \ldots\ldots\ldots\ldots \quad (22.15)$$

where C_α is any one of the Q's (Q_1 to Q_h) or of the R's (R_1 to R_k).

First let C_α be Q_α. Then, since $P_{fg}=Q_fR_g$,

$$\frac{\partial \log P_{fg}}{\partial Q_\alpha}=\frac{\partial \log Q_f}{\partial Q_\alpha}.$$

But, if $f=0$, $Q_f=1-Q_1-Q_2-\ldots-Q_h$. Hence $\partial \log Q_f/\partial Q_\alpha$ is 0 except when $f=0$ or α, in which case it is $-1/Q_0$ or $+1/Q_\alpha$. The equation therefore becomes

$$\frac{n_{0\omega}}{Q_0}=\frac{n_{\alpha\omega}}{Q_\alpha}.$$

This is true for all values of α from 1 to h, so that

$$\frac{n_{0\omega}}{Q_0}=\frac{n_{1\omega}}{Q_1}=\ldots=\frac{n_{h\omega}}{Q_h}=N.$$

We get a similar result with the R's. Thus, as in the case of an association table, the most probable values of the constants are those obtained from the sub-totals.

(c) The resulting value of χ_s^2 is the expression on the right-hand side of (22.14), which we have there shown to be equal to χ'^2.

23. *Summary.*—The present position of the problem, as regards frequency-distributions, may be stated as follows. It is understood that we are dealing with large numbers throughout.

(i) PEARSON's first formula for the ratio which he denotes by P is

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}{\int_0 e^{-\frac{1}{2}\chi^2}\chi^{m-1}\,d\chi}, \qquad \text{(23.1)}$$

where

$$\chi^2 = \frac{\varepsilon_0^2}{N_0} + \frac{\varepsilon_1^2}{N_1} + \ldots + \frac{\varepsilon_m^2}{N_m}; \qquad \text{(23.2)}$$

$m+1$ being the number of cells, N_0, $N_1 \ldots N_m$ (total $=$ N) the true numbers in them for a representative distribution, and ε_0, $\varepsilon_1 \ldots \varepsilon_m$ the differences between the N's and the observed numbers n_0, $n_1 \ldots n_m$. This is the correct formula for the case in which we are enquiring whether the data can reasonably be regarded as a random sample from a hypothetical distribution which is completely known *a priori*.

(ii) For the case in which the hypothetical distribution is not completely known but contains k constants which have to be determined from the data, PEARSON retains the form of P as in (23.1) but replaces the lower limit χ by χ_s, where

$$\chi_s^2 = \frac{\theta_0^2}{N_0'} + \frac{\theta_1^2}{N_1'} + \ldots + \frac{\theta_m^2}{N_m'}; \qquad \text{(23.3)}$$

the N′'s being calculated when the "best" values have been obtained for the k constants, and the θ's being the differences between these N′'s and the observed n's.

On general grounds, it is clear that this cannot be absolutely correct. If for calculating the N′'s we use a distribution which is a good fit to the data—as would be the case for an ordinary Gaussian distribution treated by the method of moments—we shall to some extent be adjusting the theoretical distribution to suit the data, and the value of P will be too great. If, on the other hand, we use a bad fit, P may be too small.

(iii) The correct formula for the cases mentioned in (ii) is

$$P = \frac{\int_{\chi'}^{\infty} e^{-\frac{1}{2}\chi^2}\chi^{m-k-1}\,d\chi}{\int_0^{\infty} e^{\frac{1}{2}-\chi^2}\chi^{m-k-1}\,d\chi}, \qquad \text{(23.4)}$$

where

$$\chi'^2 = \Theta^{\pi\sigma}\theta_\pi\theta_\sigma\,; \qquad \text{(23.5)}$$

Θ_{ps} being the mean product of θ_p and θ_s, and $\Theta^{\pi\sigma}$ being the set reciprocal to $\Theta_{\pi\sigma}$. The summation is for $m-k$ θ's, say θ_{k+1} to θ_m.

The calculation of χ'^2 by means of (23.5) is ordinarily so laborious as to be out of the question. It is therefore important to consider the relation between χ'^2 and χ_s^2.

(iv) Whatever (correct) method we use for finding the constants which determine the N′'s, we shall always arrive at the same value of χ'^2 (differences which are relatively of order $1/\sqrt{N}$ or of a higher order being ignored). But, as pointed out in (ii), differences of method will give different values of χ_s^2.

(v) The truth seems to be that χ_s^2 is actually equal to χ'^2 when the N′'s for (23.3) are calculated from the most probable values of the constants as given by the data; so that PEARSON's formula is correct if the m of (23.1) is replaced by $m-k$. I have not been able to prove this theorem generally; but I have proved it (§ 22) for contingency tables (thus confirming YULE's experimental results) and have verified it arithmetically for a Gaussian distribution in two simple cases (§§ 19–21). As regards the general theorem, we note that χ_s^2 can be expressed in terms of the $m-k$ θ's occurring in χ'^2 by means of the relations $\theta_0 = -\theta_1 - \theta_2 - \ldots - \theta_m$ and, as in (9.1),

$$(\alpha = 1, 2 \ldots k) \quad \frac{\partial C_\alpha}{\partial N_\lambda}\theta_\lambda = 0, \quad \ldots\ldots\ldots\ldots \quad (23.5)$$

where C_α are the k constants of the distribution. To prove the theorem, we should have to prove that, whatever method is adopted for calculating the constants, $\chi_s^2 - \chi'^2$ is equal to an expression which cannot be negative but will be 0 if the conditions (I.5) of Appendix I are satisfied. Since $N_f = NP_f$, and $N_0 + N_1 + \ldots + N_m = N$, these conditions may for our purpose be written in the form

$$(\alpha = 1, 2 \ldots k) \quad \sum_{f=0}^{m} \frac{\partial \log N_f}{\partial C_\alpha}\theta_f = 0. \quad \ldots\ldots\ldots \quad (23.6)$$

(vi) The following are two points of detail.

(*a*) If entries are grouped for the purpose of the test, they must be grouped in the same way for finding the most probable values of the constants, though the original tabulation may of course be used for getting first approximations.

(*b*) The relations (23.6), which are essential to the determination of the best values of the constants, have been obtained on the assumption that all reasonably possible combinations of values of the constants are equally probable *a priori*. Are we to say that when this is not the case the theorem does not hold good, *i.e.*, χ_s^2 is not equal to χ'^2?

I think not. A corresponding assumption has really been made in obtaining the value of χ'^2 itself; and the two assumptions balance one another. There seems no reason to doubt that the theorem, if true on these particular assumptions, is true generally. But this enquiry belongs to a more advanced stage.

APPENDIX I.

Most probable values of frequency-constants.

1. *Problem.*—Let the actual numbers in the $m+1$ categories of a frequency-distribution be $n_0, n_1 \ldots n_m$, the total number being N, so that

$$n_0 + n_1 + \ldots + n_m = \mathrm{N}. \quad \ldots \ldots \ldots \ldots \quad (\mathrm{I.1})$$

On the hypothesis that these numbers can be regarded as the result of random sampling from a source in which the distribution follows a certain law, with constants $C_1, C_2 \ldots C_k$, it is required to find the most probable values of the C's, making the usual assumption as to the relative probabilities of occurrence of different values.

2. *Method.*—On the above hypothesis, let the proportions in the $m+1$ categories of the source, for specified values of the C's, be* $P_0, P_1, P_2 \ldots P_m$. Then each P is a definite function of the C's, *i.e.*,

$$(r = 0, 1, 2 \ldots m) \;\; P_r = f_r(C_1, C_2 \ldots C_k), \quad \ldots \ldots \quad (\mathrm{I.2})$$

these being connected by the relation

$$P_0 + P_1 + P_2 + \ldots + P_m = 1. \quad \ldots \ldots \ldots \quad (\mathrm{I.3})$$

The probability that if N individuals are taken at random the numbers in the $n+1$ categories will be $n_0, n_1, n_2 \ldots n_m$ is

$$p \equiv \frac{\mathrm{N}\,!}{n_0\,!\; n_1\,! \ldots n_m\,!} P_0^{\,n_0} P_1^{\,n_1} \ldots P_m^{\,n_m}. \quad \ldots \ldots \ldots \quad (\mathrm{I.4})$$

To find the most probable values of the C's we make the usual assumption that all combinations of reasonably possible values are equally frequent; and the required values are therefore found (*cf.* FISHER, ref. 2, p. 356) by choosing the C's so as to make p a maximum, the n's, of course, remaining constant.

3. *Resulting equations.*—Taking logarithms of both sides of (I.4), differentiating with regard to each C, and equating the results to zero, we get

$$(\alpha = 1, 2 \ldots k) \sum_{f=0}^{m} n_f \frac{\partial \log P_f}{\partial C_\alpha} = 0 \quad \ldots \ldots \ldots \quad (\mathrm{I.5})$$

as the equations which have to be satisfied in order that p may be a maximum.

* These P's, of course, have nothing to do with the P's of § 17.

As a very simple example, suppose there are only two categories. We can take C to be the proportion in one category, so that $P_0 = C$, $P_1 = 1 - C$. Then (I. 5) gives us

$$\frac{n_0}{P_0} - \frac{n_1}{P_1} = 0\,;$$

i.e., the most probable distribution is that shown by the data.

4. *Approximative equations.*—In theory, the equations (I.5) are sufficient to determine the C's. In practice, the equations cannot usually be solved as they stand, and we have to proceed by successive approximations.

For simplicity, suppose there are only two C's, which we may call M and D. We start with approximate values of M and D, obtained in any convenient way; we denote these, and the corresponding values of the P's obtained from (I.2), by capital letters. We then replace M and D by $M + \psi$ and $D + \omega$, and find the values of ψ and ω that will make the resulting value of p a maximum. Denoting the altered values of the P's, when M and D are replaced by $M + \psi$ and $D + \omega$, by p's, (I.4) becomes

$$\log p = \text{const.} + \sum_{f=0}^{m} n_f \log p_f, \quad \text{(I.6)}$$

where

$$\log p_f = \log P_f + \left(\frac{\partial \log P_f}{\partial M}\psi + \frac{\partial \log P_f}{\partial D}\omega\right) + \tfrac{1}{2}\left(\frac{\partial^2 \log P_f}{\partial M^2}\psi^2 + 2\frac{\partial^2 \log P_f}{\partial M\,\partial D}\psi\omega + \frac{\partial^2 \log P_f}{\partial D^2}\omega^2\right) + \text{etc.} \quad \text{(I.7)}$$

Substituting from (I.7) in (I.6), differentiating with regard to ψ and to ω separately, and equating the results to zero, we get the equations to determine ψ and ω. If we omit terms containing higher powers than the 1st, we get

$$\left.\begin{aligned} \sum_f n_f \frac{\partial^2 \log P_f}{\partial M^2}\psi + \sum_f n_f \frac{\partial^2 \log P_f}{\partial M\,\partial D}\omega &= -\sum_f n_f \frac{\partial \log P_f}{\partial M} \\ \sum_f n_f \frac{\partial^2 \log P_f}{\partial M\,\partial D}\psi + \sum_f n_f \frac{\partial^2 \log P_f}{\partial D^2}\omega &= -\sum_f n_f \frac{\partial \log P_f}{\partial D} \end{aligned}\right\} \quad \text{(I.8)}$$

These equations give approximate values for ψ and ω. The process can be repeated for obtaining further approximations; revised values of $\partial \log P_f/\partial M$, etc., being obtained, at each stage, from the latest approximation. Or we could introduce further terms into (I.8).

5. *Normal distribution.*—As an example, take the case of a normal distribution. Let the categories be determined by values $Y_{-\frac{1}{2}} (= -\infty)$, $Y_{\frac{1}{2}}$, $Y_{\frac{3}{2}} \ldots Y_{m+\frac{1}{2}} (= +\infty)$ of a variate Y. Then

$$P_f = A_{f+\frac{1}{2}} - A_{-\frac{1}{2}}, \quad \text{(I.9)}$$

where

$$A_{f+\frac{1}{2}} = \int_{-\infty}^{X_{f+\frac{1}{2}}} Z\,dX, \qquad Z = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}X^2}, \qquad X_{f+\frac{1}{2}} = \frac{Y_{f+\frac{1}{2}} - M}{D}. \quad . \quad . \quad \text{(I.10)}$$

Using the notation

$$[A]_f \equiv A_{f+\frac{1}{2}} - A_{f-\frac{1}{2}}, \text{ etc.,}$$

and writing

$$q_{0f} \equiv \frac{[Z]_f}{[A]_f},\ q_{1f} \equiv \frac{[XZ]_f}{[A]_f},\ q_{2f} \equiv \frac{[X^2Z]_f}{[A]_f},\ q_{3f} \equiv \frac{[X^3Z]_f}{[A]_f}, \quad . \quad . \quad . \quad . \quad \text{(I.11)}$$

it will be found that the equations (I.8) become

$$\left.\begin{aligned} \sum_f (q_{1f} + q^2_{0f})\, n_f \,.\, \psi + \sum_f (q_{2f} - q_{0f} + q_{0f}\, q_{1f})\, n_f \,.\, \omega &= - \sum_f q_{0f}\, n_f \,.\, D \\ \sum_f (q_{2f} - q_{0f} + q_{0f} q_{1f})\, n_f \,.\, \psi + \sum_f (q_{3f} - 2q_{1f} + q^2_{1f})\, n_f \,.\, \omega &= - \sum_f q_{1f}\, n_f \,.\, D \end{aligned}\right\} \quad \text{(I.12)}$$

6. *Numerical example.*—For a numerical example, take the case constructed in Appendix II. The data are

f.	$Y_{f-\frac{1}{2}}$.	$Y_{f+\frac{1}{2}}$.	n_f.
0	$-\infty$	-6	47494
1	-6	-2	55965
2	-2	$+1$	52182
3	$+1$	$+4$	51458
4	$+4$	$+8$	52141
5	$+8$	$+\infty$	40760
		Total	300000

For a 1st approximation we obviously take

$$M = +2/3, \quad D = 20/3.$$

Using these values for finding q_{0f}, etc., we get

$$\psi = +0{\cdot}00991\ \ 51046, \quad \omega = +0{\cdot}00281\ \ 86887,$$

so that our 2nd approximation is

$$M = +0{\cdot}67658\ \ 17713, \quad D = 6{\cdot}66948\ \ 53554.$$

A 3rd approximation gives

$$M = +0{\cdot}67658\ \ 99206, \quad D = 6{\cdot}66950\ \ 00707,$$

which is good enough to work with.

In this example the n's are all large. But this is not necessary. The method is applicable even if the n's are small; and the values of the constants can be worked out as accurately as we please, or, rather, as is permitted by the mathematical tables we use and the capacity of our calculating machine.

Appendix II.

Construction of illustrative examples.

1. *Principle of construction.*—In the theory of statistical frequency, and in the calculus of observations generally, there is usually a difficulty in providing suitable material for illustration of theory or of formula. Two methods are in use. The first is to take a set of actual observations, and deal with them according to some assumption as to the law connecting the observed quantities. The second is to construct a theoretical set of observations, following exactly (or to the nearest multiple of the unit of measurement) a hypothetical law, and deal with the figures as if we were dealing with a mathematical table. Neither of these methods is quite satisfactory. The defect of the first is that we do not know the constants involved, and indeed can rarely be quite certain even as to the underlying law. The defect of the second is that the quantities in our constructed scheme, being calculated exactly, do not contain errors such as those of measurement or of random sampling, and therefore are not the sort of data that we actually have to deal with in practice. What we need is something that shall combine the virtues of the two methods. In the theory of statistical frequency, for instance, we should assume a certain distribution of values in a population, and should then construct a table showing the sort of result that might be obtained by random sampling from this population. The table so constructed could be used for illustrating or testing any relevant method of treatment.

2. *Errors independent.*—In the simplest type of case, the errors which we have to introduce are independent of one another, and each error is one of a set of errors distributed about a mean value zero, according to the Gaussian law, with a known s.d. We begin with a series of quantities $U_1, U_2 \ldots$ calculated on some hypothesis; they might, for instance, be the values of some simple function of Y corresponding to definite values $Y_1, Y_2 \ldots$ of Y. Each U, say U_f, is to be affected with an error ε_f, the mean value of which is 0, and its s.d. D_f. From a distribution of values of ξ with frequencies given by $\zeta \propto e^{-\frac{1}{2}\xi^2}$ we take values $\xi_1, \xi_2 \ldots$ at random; and we replace $U_1, U_2 \ldots$ by $U_1 + \xi_1 D_1$, $U_2 + \xi_2 D_2 \ldots$. The set of quantities thus obtained is an illustrative set.

3. *Errors correlated.*—In the more common case, the quantities $U_1, U_2 \ldots$ are liable to errors which are not independent but correlated. What we have to do, then, after having formed our hypothesis and determined what the U's would be if there were no errors, is to calculate a new set of quantities $V_1, V_2 \ldots$ which are definite functions of the U's and are so related to them that their errors will be independent. We proceed as in the previous case, constructing values $V_1 + \xi_1 D_1$, $V_2 + \xi_2 D_2 \ldots$, where $D_1, D_2 \ldots$ are the s.dd. of the errors of the V's, and then find the corresponding U's, which we may call $u_1, u_2 \ldots$. These u's are the set of quantities we require.

For a frequency-distribution of N individuals, a simple method is to use the property mentioned in § 15. Taking the $m+1$ frequencies in any order, let $T_0\ (=0)$, T_1, T_2 ... T_m, $T_{m+1}\ (=N)$ be their successive sums, and let $R_1 \equiv T_1/T_2$, $R_2 \equiv T_2/T_3$ ... $R_m \equiv T_m/N$ be the ratios of consecutive T's. Then the errors to which the R's are liable are independent, and the m.s.e. for R_f is $R_f^2\left(\frac{1}{T_f} - \frac{1}{T_{f+1}}\right)$. Adding errors to the R's accordingly, we obtain new values $r_1, r_2 \ldots r_m$; and the corresponding t's are formed by successive multiplications, beginning with $t_m = r_m N$. The differences of the t's are the entries for the required table.

It has been assumed, in the above, that the frequencies are all fairly large, so that the error of any quantity is small in comparison with the quantity. Where this is not the case, special methods must be employed, which need not be considered here.

4. *Numerical example.*—The following is an example of a Gaussian distribution constructed by the above method.

We start with a representative distribution for which

$$\text{mean value of } Y = \tfrac{2}{3}\,;\ \text{s.d. of } Y = \tfrac{20}{3}\,;\ N = 300{,}000.$$

The equation to the frequency-figure is then

$$\text{ordinate} = 300{,}000 \times \frac{1}{\sqrt{2\pi}\,.\,20/3}\, e^{-\frac{1}{2}\left(\frac{3Y-2}{20}\right)^2}.$$

As determining values of the classification ($m = 5$) we take

$$Y = -6, \quad -2, \quad +1, \quad +4, \quad +8.$$

The following extracts from the numerical work will show the method: some of the entries have been abbreviated. The ξ's were taken from a table of 100 ξ's, constructed for use in cases of this kind.

f	Y_f.	T_f.	$R_f = T_f/T_{f+1}$.	ξ_f.	$r_f = R_f + \xi_f D_f$.	t_f.	$t_{f+1} - t_f$.
5	+ 8	259300·182	0·8643339	−0·3196812	0·8641341	259240	40760
4	+ 4	207438·738	0·7999946	−1·4329574	0·7988688	207099	52141
3	+ 1	155981·642	0·7519408	−0·4336674	0·7515292	155641	51458
2	− 2	103373·477	0·6627285	+1·6702496	0·6647301	103459	52182
1	− 6	47596·576	0·4604332	−0·8866451	0·4590592	47494	55965
0	− ∞	—	—	—	—	0	47494
						Total	300000

REFERENCES.

(1) A. L. Bowley and L. R. Connor (1923). "Tests of Correspondence between Statistical Grouping and Formulæ." 'Economica,' No. 7, p. 1.

(2) R. A. Fisher (1921). "On the Mathematical Foundations of Theoretical Statistics." 'Phil. Trans.,' A., vol. 222, p. 309.

(3) R. A. Fisher (1922). "On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P." 'J. R. Stat. Soc.,' vol. 85, p. 87.

(4) R. A. Fisher (1923). "Statistical Tests of Agreement between Observation and Hypothesis." 'Economica,' No. 8, p. 1.

(5) R. A. Fisher (1924). "The Conditions under which χ^2 Measures the Discrepancy between Observation and Hypothesis." 'J. R. Stat. Soc.,' vol. 87, p. 442.

(6) K. Pearson (1900). "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.,' vol. 50, p. 157.

(7) W. F. Sheppard (1898). "On the Application of the Theory of Error to Cases of Normal Distribution and Normal Correlation." 'Phil. Trans.,' A, vol. 192, p. 101.

(8) G. Udny Yule (1922). "On the Application of the χ^2 Method to Association and Contingency Tables, with Experimental Illustrations." 'J. R. Stat. Soc.,' vol. 85, p. 95.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES