ESTIMATION OF LINEAR FUNCTIONS OF CELL PROPORTIONS

By John H. Smith

Bureau of Labor Statistics

Summary. In this article certain contributions are made to the theory of estimating linear functions of cell proportions in connection with the methods of (1) least squares, (2) minimum chi-square, and (3) maximum likelihood. Distinctions among these three methods made by previous writers arise out of (1) confusion concerning theoretical vs. practical weights, (2) neglect of effects of correlation between sampling errors, and (3) disagreement concerning methods of minimization. Throughout the paper the equivalence of these three methods from a practical point of view has been emphasized in order to facilitate the integration and adaptation of existing statistical techniques. To this end:

1. The method of least squares as derived by Gauss in 1821–23 [6, pp. 224–228] in which weights in theory are chosen so as to minimize sampling variances is herein called the ideal method of least squares and the theoretical estimates are called ideal linear estimates. This approach avoids confusion between practical approximations and theoretical exact weights.

2. The ideal method of least squares is applied to uncorrelated linear functions of correlated sample frequencies to determine the appropriate quantity to minimize in order to derive ideal linear estimates in sample-frequency problems. This approach leads to a sum of squares of standardized uncorrelated linear functions of sampling errors in which statistics are to be substituted in numerators.

3. A new elementary method is used to reduce the sum of squares in (2)—before substitution of statistics—to Pearson's expression for chi-square. In this result, obtained without approximation, appropriate substitution of statistics shows that the denominators of chi-square should be treated as constant parameters in the differentiation process in order to minimize chi-square in conformity with the ideal method of least squares.

4. The ideal method of minimum chi-square, derived in (3) as the sample-frequency form of the ideal method of least squares, yields ideal linear estimates in terms of the unknown parameters in the denominators of chi-square. When these parameters are estimated by successive approximations in such a way as to be consistent with statistics based on them, it is shown that the method of minimum chi-square leads to maximum likelihood statistics.

5. An iterative method which converges to maximum likelihood estimates is developed for the case in which observations are cross-classified and first order totals are known. In comparison with Deming's asymptotically efficient statistics, it is shown that, in a certain sense, maximum likelihood statistics are superior for any given value of n—especially in small samples.

6. The method of proportional distribution of marginal adjustments is de-

veloped. This method yields estimates of expected cell frequencies whose efficiency is 100 per cent when universe cell frequencies are proportional—a condition closely approximated in most practical surveys for which first order totals are available from complete censuses. Whether this favorable condition is satisfied or not, the method yields results which are easy to interpret and it has many computational advantages from the point of view of economy of time and effort.

Throughout the article discussion is confined to the estimation of parameters whose relationships to cell proportions are linear. However, most of the results can be extended to the case of non-linear relationships, the necessary qualifications being similar to those in curve-fitting problems when the function to be fitted is not linear in its parameters. In this case, of course, least squares estimates are not linear estimates. In particular, obvious extensions of the general proofs in sections 5 and 6 make them applicable to the non-linear case. Thus even when relationships are non-linear, it can be shown that the method of minimum chi-square is the sample-frequency form of the method of least squares which leads (by means of appropriate successive approximations) to maximum likelihood statistics in sample-frequency problems. This principle which establishes the equivalence of the methods of least squares, minimum chi-square, and maximum likelihood greatly facilitates the integration and adaptation of existing techniques developed in connection with these important methods of estimation.

1. Introduction. This article deals with problems of statistical estimation in which the parameters to be estimated are cell proportions or linear functions of them. A simple illustration of this type of problem is that of estimating p, the proportion of white men in a population classified by race and sex. Fom a sample of n persons selected at random from such a population, the desired proportion can be estimated by simply taking the sample proportion of white men as an estimate of the corresponding cell proportion in the population or universe. This estimate is unbiased for all possible values of p and its sampling variance is $p(1 - p)/n$—assuming, for simplicity, that sampling is done with replacements. Whether a more accurate unbiased estimate of p can be derived depends on whether or not any other relevant information concerning the cell proportions in the universe is available. For example, it may be known that all of the white portion of the population is composed of married couples so that in the universe the number of white men is exactly equal to the number of white women. This knowledge implies that half the proportion of whites provides an unbiased estimate of p which is far more accurate than the sample proportion of white men. In fact, the sampling variance of half the proportion of whites is equal to $(2p)(1 - 2p)/4n$—less than half the sampling variance of the proportion of white men.

The term *ideal linear estimate* will be used to refer to any statistic which satisfies the criteria of estimation implied by the foregoing discussion—that is, an ideal linear esimate is any estimate which (1) is a linear function of the sample observations; (2) is recognizable as unbiased by the research worker; and (3) has minimum sampling variance among estimates which have properties (1) and (2). These important criteria of estimation will now be stated in more technical language.

Let n_1, n_2, and n_3 represent the number of (1) white men, (2) white women and (3) non-white persons, respectively, in samples of n persons. Since any linear function with a constant term can be reduced to the homogeneous form by adding an appropriate multiple of the identity

$$n_1 + n_2 + n_3 - n \equiv 0, \tag{1.1}$$

it is possible, without loss of generality, to confine attention to linear estimates of the form

$$T = a_1 n_1 + a_2 n_2 + a_3 n_3, \tag{1.2}$$

which are recognizable as unbiased. In this example, the research worker is assumed to know that the cell proportions in the universe are

$$p_1, p_2, p_3 = p, p, 1 - 2p. \tag{1.3}$$

Hence, absence of bias implies that the expected value of T

$$\begin{aligned} E(T) &= a_1 n p_1 + a_2 n p_2 + a_3 n p_3 \\ &= (a_1 + a_2 - 2a_3) n p + n a_3 \end{aligned} \tag{1.4}$$

is identically equal to p; in other words, that

$$n(a_1 + a_2 - 2a_3) - 1 = 0, \tag{1.5}$$

and

$$n a_3 = 0.$$

The ideal linear esimate is derived by finding values of a_1, a_2, and a_3 which minimize the sampling variance of T subject to equations (1.5) as side conditions.[1] In this way it can be shown that half the sample proportion of whites is actually the ideal linear estimate of p. For more general problems, the process of minimization of sampling variances with the aid of Lagrange multipliers involves expressions which are complicated algebraically. For this reason it is usually easier to derive ideal linear estimates of parameters which are linear functions of cell proportions by the ideal method of least squares which is presented in section 4.

Like other least squares estimates, an ideal linear estimate of a linear function of cell proportions depends on ideal least squares weights. Since these weights

[1] In this example, it is possible to solve equations (1.5) for a_2 in terms of a_1, drop subscripts, and substitute in the formula for the sampling variance of T to obtain a quadratic in a to be minimized.

are, in general, functions of variances and covariances of sample frequencies, the theoretical connotation of the term "ideal" makes it preferable to other terms such as "optimum" and "best." In this connection it should be emphasized that (1) the sampling variance of linear estimates is insensitive to small errors in estimating ideal weights, and (2) the process of deriving practical approximations to ideal linear estimates automatically provides maximum likelihood estimates of the ideal weights. Thus the estimation of weights is perfectly objective and the best practical approximations to ideal linear estimates are expressed in terms of sample observations. This degree of objectivity is rare in statistical estimation as a brief consideration of regression problems will illustrate.

In ordinary regression problems, the ideal weights are inversely proportional to error variances. It is usually necessary to draw upon past experience to estimate relative weights because satisfactory estimates of error variances are rarely available in terms of sample observations. From the present point of view, the widespread use of equal weights implies the *subjective* "assumption" that all error variances are equal. (Maximum likelihood estimates of regression coefficients require, in addition, the even more subjective assumption of normality.) In spite of these (usually implicit) subjective assumptions, discussions of optimum properties of least squares regression coefficients based on *ideal* weights in terms of *unknown parameters* are highly commendable because (1) sampling variance is not very sensitive to small errors in weights and (2) properties of theoretical ideal linear estimates furnish a simple basis for discussion of the properties of practical statistics based on any reasonably good approximations to the exact ideal weights. In any case, it is important to know what the ideal weights are in terms of unknown parameters because research workers can make better estimates if they know what quantities should be estimated than they could otherwise.

2. Estimation of a single parameter. In sample-frequency problems, least squares weights are rarely given explicitly or even implied by information available to the research worker. Since the hypothetical example used in Section 1 is a trivial special case from this point of view, a more realistic example is presented in this section. Since the biological interpretation of this problem is presented in detail in all but the first of the many editions of Fisher's well-known book [3] it is sufficient here to consider only the statistical problem. The four cell proportions are

$$p_1, p_2, p_3, p_4 = (2 + \theta)/4, (1 - \theta)/4, (1 - \theta)/4, \theta/4, \tag{2.1}$$

and the parameter θ is to be estimated from the set of sample frequencies

$$n_1, n_2, n_3, n_4 = 1997, 906, 904, 32, \tag{2.2}$$

obtained in a sample of $n = 3839$ selected at random from an infinite universe. Fisher considers five different statistics—T_1, T_2, T_3, T_4, and T_5—so it will be convenient to use the symbol T_6 for the ideal linear estimate. Consider the class of linear unbiased estimates of the form

$$T = a_1 n_1 + a_2 n_2 + a_3 n_3 + a_4 n_4 , \tag{2.3}$$

where absence of bias implies that

$$2a_1 + a_2 + a_3 = 0 \tag{2.4}$$

and

$$a_1 - a_2 - a_3 + a_4 - 4/n = 0.$$

Minimizing the sampling variance of T in equation (2.3) subject to side conditions based on equations (2.4) yields the ideal linear estimate T_6 defined by the equation

$$n(1 + 2\theta)T_6 = 3\theta n_1 - 3\theta n_2 - 3\theta n_3 + (4 - \theta)n_4 . \tag{2.5}$$

The exact sampling variance of T_6,

$$\sigma_6^2 = \frac{2\theta(1 - \theta)(2 + \theta)}{n(1 + 2\theta)} , \tag{2.6}$$

is used by Fisher as the asymptotic sampling variance of any efficient estimate of θ. The exact sampling variance of the ideal linear estimate is especially appropriate as the asymptotic sampling variance of the maximum likelihood estimate T_4 because T_4 is the limit of an iterative process designed to estimate T_6 as closely as possible from sample data by using successive approximations to T_6 for θ in equation (2.5). The limit of this process (which is, of course, only an approximation to T_6) can be obtained by substituting the symbol T_4 for both T_6 and θ in equation (2.5) and solving the resulting quadratic equation which can be reduced to

$$nT_4^2 - (n_1 - 2n_2 - 2n_3 - n_4)T_4 - 2n_4 = 0, \tag{2.7}$$

an equation which is identical, except for notation, with Fisher's equation of maximum likelihood of which T_4 is the positive solution.

The foregoing result is a comparatively simple illustration of the general principle that the maximum likelihood estimate of any linear function of cell proportions is the limit of an iterative process designed to approximate the corresponding linear estimate as closely as possible by means of sample frequencies. Since the accuracy of estimates of least squares relative weights increases with size of sample, maximum likelihood statistics have, in an asymptotic sense for large samples, the same optimum properties which are possessed in an exact sense (even for small samples) by the corresponding ideal linear estimates. Thus the results obtained by means of the theory of large samples are supported by the approach to estimation problems by means of ideal linear estimates. In addition, the later approach facilitates the integration of available techniques as explained in later sections.

It is true that the optimum properties of maximum likelihood statistics can be presented in terms of the theory of large samples, but the fact that a given method of estimation yields a statistic whose asymptotic sampling variance is a minimum does not imply that the same technique will yield a minimum variance statistic for any given small value of n. For example, it is well known that the median is a maximum likelihood estimate of the midpoint of a double exponential universe. Nevertheless, in samples of three observations from such a universe, another statistic—4/9 of the mean plus 5/9 of the median—has greater relative advantage over the median than the median has over the mean.

Fisher's discussion of the relative efficiencies of his five alternative consistent statistics suggests that it is impossible to formulate objective criteria for making choices among alternative statistics such that each statistic will be used whenever its sampling variance is smallest. Consider the sequence of universes generated by letting θ vary from zero to unity. In general, each value of θ would determine which of Fisher's five statistics would have smallest sampling variance for that particular universe for any given value of n. In comparison with any other single statistic, the statistic T_4 would usually have smaller sampling variance, but there are notable exceptions. For example, in the absence of linkage when θ is equal to one-fourth, the statistic T_2 is the ideal linear estimate and its sampling variance is smaller than that of T_4—at least for certain small values of n. For this reason, Fisher used T_2 in preference to T_4 as the basis for testing the significance of linkage. The statistic T_5—derived by Fisher's method of minimum chi-square—is also of special interest. Fisher's method of minimum chi-square yields statistics which differ from the corresponding maximum likelihood statistics because Fisher considers the denominators as variables in the process of differentiation instead of considering them as unknown parameters to be estimated by identifying them with the corresponding statistics in the numerators *after* differentiation. Arguments of later sections tend to show that the latter method is more appropriate. In this example, it can be shown that if T_5 were substituted for the corresponding parameter in the denominators of chi-square (*and treated as a parameter*) the minimization of chi-square with respect to statistics in its numerators would be exactly equivalent to substituting 0.035785, the numerical value of T_5 for θ in equation (2.5) and solving for T_6 to obtain 0.035717, a value which is much closer to 0.035712, the numerical value of the maximum likelihood estimate T_4 than to Fisher's T_5. In problems of estimation chi-square should be minimized in order to obtain efficient statistics—not to obtain a small criterion for testing goodness of fit—and it should be minimized in a manner consistent with this purpose. Whether or not it is possible to derive an even smaller value for a quantity called chi-square should be considered to be irrelevant in either estimation problems or tests of significance. It is difficult to present these ideas in more technical language because it is possible to construct trivial hypothetical universes for which Fisher's method of minimum chi-square provides statistics which are superior in certain respects to the corresponding maximum likelihood statistics. Nevertheless, it seems clear that the ideal linear estimate usually has smaller sampling variance than the maximum likelihood statistic which, in turn, usually has smaller sampling variance than any other given practical statistic. Evidence presented in later sections tends to show that these advantages are more important in small samples than in cases in which the theory of large samples is applicable.

3. The "ideal" method of least squares. When sample observations are uncorrelated in successive samples and parameters to be estimated are linear functions of the expected values of the sample observations, the method of least squares yields ideal linear estimates of the parametes provided that the weight of each observation is inversely proportional to its variance in successive samples. Although the minimum sampling variance property among linear unbiased estimates is seldom stressed, this principle of weighting has been presented in connection with the method of least squares for more than a hundred years. In order to emphasize the theoretical nature of weights which depend on variances which are usually unknown in practice and to distinguish the method based on such weights from the more familiar method of least squares with equal weights, the method which yields ideal linear estimates will be called the *ideal method of least squares.*

Discussion of the general problem of estimating linear functions of cell proportions can be facilitated by making use of results obtained by other writers—notably Gauss (as reported by Whittaker and Robinson [6]) and Pearson [4]. According to Whittaker and Robinson, "the first writer to connect the method [of ideal least squares] with the theory of probability was Gauss" [6, p. 224]. In his Theoria Motus proof of 1809, Gauss derived the "most probable value" [6, p. 223] of a parameter (i.e., the statistic which satisfies the criterion now called maximum likelihood) for the case in which sample observations are statistically independent and normally distributed. In his Theoria Combinationis proof of 1821–23, Gauss "abandoned the 'metaphysical' basis" [6, p. 220] of his earlier work and derived the method herein called the ideal method of least squares (without approximation) from the criteria of (1) minimum variance and (2) absence of bias for the case in which "the mean value of [the covariance of a pair of errors] is zero" [6, p. 224]. Since the covariances of *uncorrelated* linear functions are zero whether they are *statistically independent* or not, it follows from the work of Gauss that the ideal method of least squares applied to uncorrelated linear functions of sample frequencies yields ideal linear estimates. In other words, the ideal method of least squares implies the following six steps:

1. From the set of $k + 1$ sample frequencies construct k linear functions which are uncorrelated in successive samples.
2. From each function subtract its expected value in terms of the unknown parameters to find its sampling error.

3. Write the ratio of each sampling error to its own standard error in the form of a fraction.
4. Sum the squares of these standardized uncorrelated sampling errors to obtain a quantity called chi-square.
5. Substitute statistics[2] for the parameters in the *numerators* of chi-square.
6. Minimize the sum of squares of residuals with respect to each statistic in turn (subject to appropriate side conditions in case linear functions not implied in preceding steps are known).

This series of six steps can be summarized by the single statement that the function to minimize is the sum of squares of standardized uncorrelated residuals. Actually this statement is oversimplified because even though sampling errors are both uncorrelated and standardized, the corresponding residuals are, in general, neither standardized nor uncorrelated.

4. Pearson's expression for chi-square. As defined by Pearson [4], chi-square is the sum of squares of a set of k standardized uncorrelated linear functions of sampling errors in a set of $k + 1$ correlated sample frequencies. A set of k standardized uncorrelated linear functions can be constructed in an infinite number of ways, but each set can be obtained from any of the others by means of an orthogonal transformation. Thus the sum of squares is the same no matter what set is originally chosen. As his set of standardized uncorrelated linear functions, Pearson chose those determined by the axes of the correlation ellipse for which he gave the required sum of squares in terms of "minors" or cofactors of the correlation determinant of the first k sample frequencies. Pearson reduced this complicated expression to the now familiar form

$$\chi^2 = \sum_{i=1}^{k+1} (n_i - np_i)^2/np_i, \tag{4.1}$$

where p_i is the proportion in the ith cell in the universe and n_i is the frequency in the ith cell of a sample of n observations selected at random from an infinite universe (or with replacements from a finite universe).

The widespread misunderstanding of the nature of chi-square seems to be based primarily on the facts that

1. Pearson's rule for degrees of freedom is inadequate (see section 5), and
2. Pearson's expression for chi-square can be derived by approximate methods as well as by exact methods.

Pearson's derivation of the expression for chi-square by exact methods is sufficient to show that its derivation by approximate methods involves a paradox in which different sets of approximations offset each other; however, Pearson's article is relatively inaccessible and, in addition, his algabraic reductions involve

[2] It is convenient to call these variable symbols "statistics"; the quantities whose squares are summed, "residuals"; and the whole expression "chi-square," even though, from a certain point of view, these terms are strictly applicable only after the minimization process. This usage should always be clear from its context.

the minors of a general determinant of the kth order. For these reasons, the following exact derivation is presented in terms of elementary algebra.

Since the sum of squares is the same for any set of k standardized uncorrelated linear functions of the sampling errors in $k + 1$ correlated frequencies, a set should be chosen for which the algebraic reductions are as easy as possible. From this point of view a satisfactory set, which can be written in any of three forms, is given by

$$\begin{aligned} y_i &= p_i n_{i+} - p_{i+} n_i \\ &= p_i e_{i+} - p_{i+} e_i \\ &= -p_i e_{i-} - (p_i + p_{i+}) e_i \end{aligned} \tag{4.2}$$

where $e_i = n_i - np_i$ and $i+$ and $i-$ refer to classes formed by combining all classes above the ith class and below the ith class, respectively.

By means of the known variances and covariances of the sample frequencies in expected value form,

$$E(e_i^2) = np_i(1 - p_i), \tag{4.3}$$

and

$$E(e_i e_j) = -np_i p_j, \tag{4.4}$$

it can be shown that the variance of y_i is

$$E(y_i) = np_i p_{i+}(p_i + p_{i+}), \tag{4.5}$$

and, by using the third expression in equation (4.2) for y_i and the second for y_j, it can be shown that any pair of y's are uncorrelated because

$$E(y_i y_j) = 0, \qquad (i < j). \tag{4.6}$$

Let z_i represent the variable y_i expressed in standard-deviation units. The square of this standardized uncorrelated linear function of correlated sampling errors can be written

$$z_i^2 = \frac{(p_i e_{i+} - p_{i+} e_i)^2}{np_i p_{i+}(p_i + p_{i+})}. \tag{4.7}$$

It remains to show that Pearson's expression for chi-square can be obtained by adding the k values of z_i^2 in succession. For this purpose it is convenient to define

$$\chi_r^2 = \sum_{i=1}^{r} \frac{e_i^2}{np_i} + \frac{e_{r+}^2}{np_{r+}}, \tag{4.8}$$

obtained by combining all classes above the rth class.

When $r = k$, the expression in equation (4.8) is the expression to be derived. It remains to show that χ_k^2 is the sum of squares of k standardized uncorrelated linear functions of sampling errors; i.e.,

$$\chi_k^2 = \sum_{i=1}^{k} z_i^2. \tag{4.9}$$

For the first cell $e_{1+} = -e_1$ and $p_{1+} = 1 - p_1$. Hence y_1 reduces to the negative of the error in the first frequency and

$$\begin{aligned} \chi_1^2 &= e_1^2/np_1(1 - p_1) \\ &= e_1^2/np_1 + e_{1+}^2/np_{1+} \qquad (p_{1+} = 1 - p_1), \end{aligned} \tag{4.10}$$

a special case expressed in the required form. The general case is established by showing that

$$\chi_{r-1}^2 + z_r^2 = \chi_r^2, \tag{4.11}$$

or, alternatively, that

$$\begin{aligned} z_r^2 &= \chi_r^2 - \chi_{r-1}^2 \\ &= e_r^2/np_r + e_{r+}^2/np_{r+} - (e_r + e_{r+})^2/n(p_r + p_{r+}) \\ &= \frac{(p_{r+} e_r^2 + p_r e_{r+}^2)(p_r + p_{r+}) - p_r p_{r+}(e_r^2 + 2e_r e_{r+} + e_{r+}^2)}{np_r p_{r+}(p_r + p_{r+})} \\ &= \frac{p_r^2 e_{r+}^2 - 2p_r p_{r+} e_r e_{r+} + p_{r+}^2 e_r^2 \equiv (p_r e_{r+} - p_{r+} e_r)^2}{np_r p_{r+}(p_r + p_{r+})}, \end{aligned} \tag{4.12}$$

thus establishing the derivation of Pearson's expression for chi-square.

When sampling is done without replacement each variance and covariance is multiplied by $(N - n)/(N - 1)$ where N is the number of observations in the universe. Hence, chi-square for this case can be written

$$\chi^2 = \frac{N - 1}{N - n} \sum_{i=1}^{k+1} \frac{e_i^2}{np_i}. \tag{4.13}$$

This expression shows that the factor involving sampling errors is the same whether sampling is done with replacement or without replacement. Hence, the derivation of least squares statistics is the same for either method of sampling, but sampling variances for the simpler case are multiplied by the factor $(N - n)/(N - 1)$ when sampling is done without replacement.

5. The method of minimum chi-square. The derivation of Pearson's expression for chi-square completes first four steps of the ideal method of least squares outlined in section 3. Hence, the method of minimum chi-square is the sample-frequency form of the ideal method of least squares in which only two of the six steps remain to be taken.

In his original article [4] Pearson pointed out that the use of statistics instead of parameters would affect the value of chi-square but that such effects would usually be so small that no allowance need be made for them in connection with tests of significance. It is now well known that the average value of chi-square is reduced approximately one unit for each parameter estimated from the sample, and that the main portion of this effect is on the numerators; i.e., in large samples the effect of substituting statistics for parameters in the denominators usually has a negligible effect on the value of chi-square. By confining the discussion to the case in which parameters are used in the denominators, it is possible to make simple exact statements concerning the main effects in terms of the number of squares of standardized uncorrelated linear functions—also known as the number of degrees of freedom and the mean value of chi-square.

When the expected values in the numerators of chi-square can be expressed as linear functions of r algebraically independent parameters, ideal linear estimates of the r parameters are determined by substituting statistics for the r parameters and minimizing the resulting expression wth respect to each statistic. In general, such a substitution of statistics for parameters in the numerators of chi-square reduces the number of degrees of freedom by one unit for every parameter estimated; that is, the appropriately minimized chi-square can be analyzed into $k - r$ squares of standardized uncorrelated linear functions of sampling errors.

The r ideal linear estimates are linear functions of the sample frequencies. Let $(v_1, v_2, \cdots, v_r)$ be a set of standardized uncorrelated linear functions of the correlated sampling errors in these statistics and let $(v_1, v_2, \cdots, v_k)$ be a set of linear functions obtained from the z_i's of section 3 by an orthogonal transformation. Since the sum of squares is not changed by such a transformation, chi-square is the sum of the k values of v_i^2. The process of substituting statistics for the r parameters in the numerators of chi-square reduces the values of the first rv_i^2's to zero without affecting the values of the other $(k - r)v_i^2$'s.

Thus the appropriately minimized chi-square can be analyzed into $k - r$ squares of standardized uncorrelated linear functions of sampling errors and is therefore said to have $k - r$ degrees of freedom. The mean value of each square is the variance of a standardized linear function of sampling errors and is therefore unity by definition. Hence the mean value of the appropriately minimized chi-square (with parameters in the denominators) is exactly $k - r$ when r statistics are estimated from a set of $k + 1$ sample frequencies.

The expression to be minimized is

$$\chi^2 = \sum \frac{(n_i - m_i')^2}{np_i} \tag{5.1}$$

where m_i' is the ideal linear estimate of np_i. The set of statistics described by the equation

$$m_i' = n_i, \tag{5.2}$$

reduces the value of chi-square to zero—its minimum value. This shows that the sample cell proportion is the ideal linear estimate of the corresponding parameter.

Whenever a linear function independent of the sum of the cell proportions is known, it is possible to take advantage of additional information provided by the known function by minimizing chi-square subject to an appropriate side condition. When side conditions are used in this way, the number of degrees of freedom for the minimized chi-square is equal to the number of side conditions which are algebraically independent of each other (and of the sum of the cell proportions). Let the known linear function be written

$$\Sigma a_i n p_i - m = 0. \tag{5.3}$$

In order to facilitate comparison of the typical equation of maximization with the corresponding equation of the method of maximum likelihood, it is convenient to minimize chi-square by maximizing $-\chi^2/2$ subject to a side condition based on (5.3). The function to be maximized can be written

$$-\chi^2/2 = \Sigma(n_i - m_i')^2/(-2np_i) + h(\Sigma a_i m_i' - m), \tag{5.4}$$

where h is a Lagrange multiplier. Setting the partial derivative of $-\chi^2/2$ with respect to m_i' equal to zero, the typical equation for minimizing chi-square can be written

$$(n_i - m_i')/np_i + ha_i = 0, \tag{5.5}$$

a form which shows that, in general, ideal linear estimates are defined in terms of unknown parameters. Fortunately, these parameters can usually be approximated closely by an iterative process. Substituting m_i for both np_i and m_i' in equations (5.5) the typical equation in the limiting values of such a process can be reduced to

$$n_i/m_i - 1 + ha_i = 0, \tag{5.6}$$

a form which is identical with the typical equation (6.6) of maximum likelihood derived in section 6. This equality of typical equations implies that whenever the denominators of chi-square are estimated in such a way as to be consistent with least squares statistics based on them, the method of minimum chi-square always leads (by means of approximations necessary in practice) to maximum likelihood estimates of parameters which are linear functions of cell proportions.

6. The method of maximum likelihood. Maximum likelihood estimates of linear functions of cell proportions can be obtained by (1) expressing the probability function (general term of the multinomial expansion) in terms of the r parameters to be estimated; (2) substituting r statistics for the r parameters; and (3) maximizing with respect to the r statistics. In practice, this is usually accomplished by maximizing the logarithm of the variable factor in step (3) which can be written,

$$L = \Sigma n_i \log m_i, \tag{6.1}$$

where m_i is the maximum likelihood estimate of np_i, the expected value of the ith frequency n_i in a sample of n observations classified into $(k + 1)$ classes or cells. It is evident that L as written has no maximum with respect to any m_i since it increases without bound as m_i increases, but it sometimes has a uniquely determined maximum when each of the m_i's is written explicitly in terms of less than $k + 1$ algebraically independent statistics. In the general case it is easier to maximize L subject to an appropriate set of side conditions, one of which must be equivalent to

$$m_1 + m_2 + \cdots + m_{k+1} - n = 0. \tag{6.2}$$

When no linear function except the sum is known, the likelihood function can be written

$$L = \Sigma n_i \log m_i - (\Sigma m_i - n), \tag{6.3}$$

a function which, subject to equation (6.2), is always equal to that in equation (6.1) but which has a uniquely determined maximum. The typical equation of maximum likelihood, obtained by setting the partial derivative of L with respect to m_i equal to zero, is

$$n_i/m_i - 1 = 0, \tag{6.4}$$

an equation which shows that each sample frequency is a maximum likelihood estimate of its own expected value.

When a linear function such as that in equation (5.3) is known, an improved set of maximum likelihood statistics can be found by maximizing

$$L = \Sigma n_i \log m_i - (\Sigma m_i - n) + h(\Sigma a_i m_i - m). \tag{6.5}$$

The typical equation of maximization is found to be

$$n_i/m_i - 1 + h a_i = 0, \tag{6.6}$$

an equation which, as stated above, is identical with equation (5.5). Since equation (5.5) was obtained as the limit of an iterative process from the typical equation (5.4) for minimizing chi-square subject to the same side condition and since each additional side condition affects the typical equation of each method in exactly the same way, the method of minimum chi-square and the method of maximum likelihood are equivalent for the general case in the sense that the method of minimum chi-square always leads to maximum likelihood statistics as limits of an iterative process.

7. Second-order tables with known expected marginal totals. As stated in section 2, the integration of available techniques is facilitated by regarding maximum likelihood statistics as the best practical approximations to the corresponding ideal linear estimates. Since this important principle may not be immediately obvious, it will be illustrated for the important special case of second-order tables for which the expected marginal totals are known.

Consider a sample of n observations arranged on two bases of classification and presented in a table containing r rows and s columns. The universe of N

observations has been completely enumerated and classified on each basis separately but not cross-classified; i.e., universe totals of first order classes are known.

For the cell in the ith row and the jth column, let p_{ij} represent the universe cell proportion; n_{ij}, the sample frequency; np_{ij}, the expected value of n_{ij}; and m_{ij}, the maximum likelihood estimate of np_{ij}. Indicating summation by substituting a dot for the letter over which summation is to be performed, the known marginal totals satisfy the equations

$$Np_{i.} - N_{i.} = 0, \tag{7.1}$$
$$Np_{.j} - N_{.j} = 0,$$

where $p_{i.}$ and $p_{.j}$ are the universe proportions and $N_{i.}$ and $N_{.j}$ are the known universe totals in the ith row and the jth column, respectively.

When n observations of a random sample are arranged according to two bases of classification in a table with r rows and s columns for which the $r + s$ marginal totals are known, the typical equation of maximum likelihood can be obtained by maximizing, subject to side conditions based on equations (7.1), the likelihood function

$$L = \Sigma\Sigma n_{ij} \log m_{ij} - \Sigma a_i(m_{i.} - n_{i.}) - \Sigma b_j(m_{.j} - n_{.j}), \tag{7.2}$$

with respect to the maximum likelihood estimates m_{ij}, where a_i and b_j are typical Lagrange multipliers. Setting the partial derivative with respect to m_{ij} equal to zero and transposing, the typical equation of maximum likelihood can be written

$$n_{ij}/m_{ij} = a_i + b_j. \tag{7.3}$$

Since equations (7.3) are not linear in their unknowns, the reader's first reaction might well be to agree with a certain anonymous critic that "their solution is difficult." This impression of great difficulty is probably the chief reason that previous writers have not used the method of maximum likelihood for this type of problem even after they had developed a set of techniques adequate for the solution of the equations of maximum likelihood. In other words, all that was needed was the integration of available techniques as will now be shown.

In 1940, Deming and Stephan [2] derived a set of normal equations for the adjustment of a set of second-order cell frequencies to known expected marginal totals by the method of least squares in which each sample frequency is weighted by its own reciprocal. This method yields statistics which are efficient according to the theory of large samples, but they do not satisfy the criterion of maximum likelihood exactly. In the same article was presented an easier method of *iterative proportions*, which, unfortunately, does not yield least squares statistics. In 1942, Stephan [5] developed an improved iterative process which yields statistics which satisfy the criterion of least squares with arbitrarily chosen weights. The foregoing developments are presented in greater detail in Deming's book [1] in which Deming adapts Stephan's iterative method to the particular case in which each sample frequency is weighted by its own reciprocal so as to yield solutions for the normal equations derived in the joint article [2].

In Deming's notation, equation 8 of Stephan's article [5, p. 169] can be written

$$m_{ij} = c_{ij}(p_i + q_j - 1) + n_{ij}, \tag{7.4}$$

an expression obtained by substituting c_{ij} for np_{ij} in the denominators of chi-square and minimizing with respect to the statistics in the numerators. Hence, if exact values of the np_{ij} were used for the c_{ij}, the Stephan iterative method would yield ideal linear estimates. Unless these parameters are implied by some hypothesis to be tested, it is necessary, in practice, to estimate the np_{ij} from sample data. In order to secure maximum likelihood estimates of expected cell frequencies by means of the Stephan iterative method, the adjusted frequencies based on first approximations to the c_{ij} should be used as second approximations to the c_{ij}, etc. In this way, maximum likelihood statistics can be derived to any desired degree of approximation. At this point it should be emphasized that the preceding statement applies not only to the class of problems considered in this section but also to the wider class of problems for which the Stephan iterative method provides solutions.

Unfortunately, theoretical discussions of previous writers contain confusing compensating errors which (1) present their own methods in an unnecessarily unfavorable light and (2) increase the difficulties involved in the introduction of the improvements in techniques suggested in section 9 which involve some degree of adaptation of techniques already available. For these reasons, it seems necessary to follow the arguments of previous writers in order to show the points at which improvements are needed. This can be done most effectively in connection with Deming's book [1] where the method of least squares is presented in great detail.

For the special case in which the sampling errors in the observations are uncorrelated, the ideal criterion of least squares implies that the weight of each observation should be inversely proportional to its sampling variance. This criterion is accepted as well known by Deming who says that "the principle of least squares requires the minimizing of the sum of the weighted squares of the residuals" [1, p. 14] where "the weights of two functions are inversely proportional to their variances" [1, p. 22]. Deming assumes that "there is no correlation between the errors in the observations" with the qualification that "this assumption covers a wide class of problems, but does fail to cover some." [1, p. 49]. This assumption of uncorrelated errors is not applicable to sample-frequency problems, of course, because the sample frequencies are correlated with each other in such a way that the reciprocals of the ideal least squares weights are not proportional to the sampling variances $np_{ij}q_{ij}$ but rather to the expected frequencies np_{ij} which appear in the denominators of chi-square.

In this connection it is interesting to note that Deming himself insists that "there is only one principle of least squares, namely, the minimizing of χ^2." [1, p. 51]. However, the method currently in use for the minimizing of chi-square was that given by Fisher [3] which leads to equations which are difficult to solve even for such a simple example as the one presented in section 2 above.

Deming and Stephan are to be commended for seeking an easier method but there is no justification (even as a device for saving effort) for their modification of the "principle of least squares" so as to imply erroneously that

(1) weights of correlated sample frequencies are inversely proportional to their variances, and
(2) sample frequencies are, in general, approximately proportional to their own sampling variances.

Strangely enough, these two errors were applied in combination by Deming and Stephan to obtain good practical approximations to the ideal least squares weights. It might be argued that the second misleading implication is really not an error because it is offered as a simplifying approximation, but it is an integral part of both the normal equations approach in the joint article [2] and Deming's adaptation [1] of the Stephan iterative method; that is, in each case the method would have to be revised if better approximations to the ideal least squares weights were used. More explicitly, Deming (1) uses n_{ij} for Stephan's c_{ij} in equation (7.4); (2) identifies it with the other n_{ij} in the same equation; and (3) reduces the equation to a different form thus effectively preventing the use of successive approximations to the c_{ij} without returning to Stephan's iterative method in the general form given by equation (7.4) above which Deming does not present at all. Results of the joint article [2] are quoted by Stephan [5] without any explanation of the nature of the errors, but none of these results are used in the development of his iterative method which as noted above, is applicable to any arbitrarily chosen set of weights. The fact that Stephan corrected the second error without correcting the first implies that the weights he actually used are unsatisfactory. In Deming's adaptation of the Stephan iterative method, a much better set of weights is obtained, not by correcting the first offsetting error overlooked by Stephan, but by resurrecting the second offsetting error which Stephan had corrected. Since this error is an integral part of Deming's adaptation, Deming's theoretical discussion implies that his own efficient statistics are only rough approximations which are definitely inferior to the inefficient statistics obtained by means of the weights chosen by Stephan. These inconsistencies are most clearly brought out by Deming when he says:

> "Strictly, in random sampling, the reciprocal of the weight of n_{ij} is $np_{ij}q_{ij}$, which is nearly equal to $n_{ij}q_{ij}$ where p and q have their usual connotations. But since factors proportional to the weights may be substituted for them, it is sufficient to use n_{ij} as the reciprocal of the weight in cell ij, since the values of q_{ij} do not usually vary much over the table." [1, p. 102.]

In any given problem the seriousness of the error in the first statement in the foregoing quotation depends on the variation among the q_{ij}'s. In the particular example used by Deming the error is of considerable importance because the largest q_{ij} is more than 40 per cent larger than the smallest q_{ij}. The weights actually used by Deming agree with weights implied by the ideal method of least squares except for sampling errors in the n_{ij}; hence, the error in any relative weight converges stochastically to zero so that Deming's statistics are efficient according to the theory of large samples. The efficiency of Deming's statistics is inconsistent with the theory presented by Deming which implies erroneously that efficiency of estimation depends on approximate equality of cell proportions. If this argument were true it would apply also to the method of maximum likelihood and all other methods which yield efficient practical statistics in sample-frequency problems. The foregoing discussion, together with the results of section 8 show that the theory as presented by Deming has the following seriously misleading features:

(1) it is based on a paradox in which a good final result is obtained by means of compensating errors;
(2) it presents his efficient statistics in an unnecessarily unfavorable light;
(3) it emphasizes the irrelevant condition of approximate equality of universe cell proportions;
(4) it fails to mention the important condition of proportionality by rows and columns; and
(5) it makes least squares, minimum chi-square, and maximum likelihood seem to be competing alternative methods.

Of these undesirable characteristics, the last two are probably the most serious because they make the effective integration and adaptation of statistical techniques more difficult. As has been shown in sections 4, 5, and 6, the sample-frequency form of the ideal method of least squares is the method of minimum chi-square which always leads (by means of appropriate practical approximations to unknown weights) to maximum likelihood statistics; in other words, the methods are equivalent from a practical point of view.

Since the ideal method of least squares based on the unknown np_{ij} determines fully efficient, but theoretical, ideal linear estimates, the efficiency of practical approximations to ideal linear estimates depends on the accuracy with which the denominators of chi-square are estimated. For the unknown denominators np_{ij}, Deming uses the sample frequencies n_{ij} while the method of maximum likelihood implies the use of the corresponding maximum likelihood estimates—statistics which, in general, have smaller sampling variances. The foregoing argument suggests that maximum likelihood statistics are slightly superior to Deming's statistics for any given finite value of n and that their relative advantage increases as the sample size decreases. In large samples both methods yield efficient statistics because the relative errors in the weights implied by either method converge stochastically to zero as n increases. Although the advantage of maximum likelihood statistics over Deming's statistics is unimportant except in small samples, it can be shown that Deming's choice of weights leads to imperfectly compensated negative errors of estimation even in his large sample of 33,837 observations.

Deming weights each sample frequency by its own reciprocal. Positive errors of sampling decrease the value of the reciprocal and thus increase the absolute size of the required negative adjustments. Negative errors of sampling increase the value of the reciprocal and thus decrease the size of the positive adjustment. Thus every error of sampling (either positive or negative) leads to a negative error of estimation due to inappropriate weighting. Because the sum of all adjustments must be zero, these negative errors of estimation are compensated on the average but more or less imperfectly. The net effect of this imperfect compensation of negative errors of estimation is that Deming's statistics are too small in those cells in which the relative adjustments (either positive or negative) are large, and vice versa. In a preliminary draft of this article, this type of error of estimation was studied by comparing Deming's statistics with the corresponding maximum likelihood statistics in conection with Deming's example involving 33,837 observations. Although errors of estimation of the type under discussion are apparent, they are, of course, extremely small in such a large sample. For this reason the large-sample comparson has been deleted in favor of simple hypothetical examples designed to throw light on similar errors of estimation in statistics derived by Fisher's method of minimum chi-square as well as in those derived by Deming's adaptation of Stephan's iterative method.

Consider a set of sample frequencies in a two-by-two table for which all expected marginal totals are equal. For this special case, the cell proportions on each diagonal are equal and the ideal linear estimate (which is also the maximum likelihood estimate) of any cell proportion is the mean of the two sample cell proportions on its diagonal. For the same case, Deming's adaptation of the Stephan iterative method yields an estimate for each cell which is proportional to the harmonic mean of sample proportions on its diagonal while Fisher's method of minimum chi-square yields estimates proportional to the corresponding quadratic means.

As a numerical example of the foregoing problem consider the set of frequencies

$$n_{11}, n_{12}, n_{21}, n_{22} = 1, 4, 3, 2, \tag{7.5}$$

obtained in a sample of 10 observations selected at random from a universe in which the cell poportions are known to be

$$p_{11}, p_{12}, p_{21}, p_{22} = p, 0.5 - p, 0.5 - p, p. \tag{7.6}$$

As estimates of the parameter p, the ideal linear estimate is .15, Deming's adaptation of the Stephan iterative method yields .14, and Fisher's method of minimum chi-square yields .1545 to four decimal places, the other two estimates being exact. The results illustrate the imperfectly compensated errors of estimation explained previously. The two sample frequencies on the principal diagonal (n_{11} and n_{22}) have greater relative dispersion than the frequencies on the other diagonal. For this reason, the relative adjustments made by Deming's method are greater and according to the principle of imperfectly compensated negative errors of estimation, the estimate of p obtained by Deming's method is smaller than the ideal linear estimate of p. Fisher's method of minimum chi-square yields an estimate of p which is greater than the ideal linear estimate. In fact, one should usually expect imperfectly compensated errors of estimation in statistics derived by Fisher's method of minimum chi-square to be opposite in sign and about half as large as those in the corresponding statistics derived by means of Deming's adaptation of the Stephan iterative method.

At this point, it should be emphasized that Fisher does not recommend his own method of minimum chi-square in preference to the method of maximum likelihood. In fact, he presents the theory of estimation in such a way as to imply correctly that the method of maximum likelihood is superior, especially in small samples. Other writers have noted the small differences between equations of maximum likelihood and those for minimizing chi-square by Fisher's method and some have even derived one set of equations from the other by neglecting higher order terms in a Taylor series expansion. These derivations are of no interest here because they seem to justify the method of maximum likelihood as a simple approximation to some more complicated method. This type of justification is both unnecessary and undesirable. It is more useful to regard the method of maximum likelihood as an approximation to a method—least squares—for which the theory is simpler.

Skeptical readers who find the foregoing argument unconvincing may be able to profit from the following example. Consider the problem of estimating the parameter p where $2p$ is the proportion of white balls in an urn. A sample of 10 balls is selected and classified by the following process. Each white ball is placed in one of the cells on the principal diagonal of a two-by-two table, the particular cell being decided by the toss of a coin. A similar method is used for non-white balls placed in cells on the other diagonal. Assuming that the results of this process are given by equation (7.5), which of the three alternative estimates of p given above should be preferred? Belief in the general superiority of Fisher's method of minimum chi-square seems to imply that the device of coin-tossing described in this example can be used in practical problems involving the estimation of the proportion of "successes" to secure estimates which are superior to the sample proportion—the ideal linear estimate in such cases. Even if it is possible to construct trivial special case examples supporting some complicated method for such problems the general use in practical problems of the coin-tossing device in connection with either Fisher's method of minimum chi-square or Deming's adaptation of the Stephan iterative method would be absurd as this example is intended to emphasize.

8. The method of proportional distribution of marginal adjustments. The method of proportional distribution of marginal adjustments is a general method of adjusting sample frequencies so that their row and column totals agree with known expected marginal totals. In other words, the adjusted frequency for the cell in the ith row and the jth column is given by the equation

$$m_{ij}^{*} = n_{ij} - p_{i.}d_{.j} - p_{.j}\,d_{i.}\,, \tag{8.1}$$

where

$$d_{i.} = m_{i.} - n_{i.}\,,$$

and

$$d_{.j} = m_{.j} - n_{.j}\,,$$

are the net adjustments in the sample cell frequencies of the ith row and the jth column, respectively. The asterisk is used to distinguish maximum likelihood estimates m_{ij} and the ideal linear estimates m'_{ij} from the set of statistics based on equation (8.1).

The method of proportional distribution of marginal adjustments yields ideal linear estimates when the universe cell proportions are proportional by rows and by columns; i.e., when

$$p_{ij} = p_{i.}p_{.j}\,. \tag{8.2}$$

This important principle can be established by substituting in equation (7.4) of section 7 the quantities

$$c_{ij} = np_{i.}p_{.j}\,, \tag{8.3}$$

$$p_i = 0.5 + d_{i.}/np_{i.}\,,$$

and

$$q_j = 0.5 + d_{.j}/np_{.j}\,,$$

and reducing the typical equation of the ideal method of minimum chi-square to the form of equation (8.1) which defines the method of proportional distribution of marginal adjustments.

Even in the absence of exact proportionality, under which it yields fully efficient statistics, the method of proportional distribution of marginal adjustments has the following relative advantages over other available methods:

(1) ease of extension to tables of higher order;
(2) *exact* agreement with known (expected) marginal totals;
(3) simplicity of interpretation;
(4) independence of computational errors;
(5) rapidity of processing;
(6) economy of effort; and
(7) fully efficient criteria for testing the significance of departures from proportionality of rows and columns.

Ease of extension to tables of higher order is a desirable property of the method of proportional distribution of marginal adjustments. Equation (8.1) applies to the special case in which there are only two bases of classification. In the more general case sample observations are cross-classified according to r bases of classification, each cell frequency in an rth order table being the number of observations in the corresponding rth order class whose expected value is to be estimated. The required adjustment for each first order class (obtained by subtracting the sample total from its known expected value) is distributed among the various cells in proportion to the universe totals of the corresponding $(r - 1)$th order classes to which the cells belong. The general process is illustrated by

$$m^*_{ijk} = n_{ijk} + p_{i..}d_{.jk} + p_{.j.}d_{i.k} + p_{..k}d_{ij.}, \tag{8.4}$$

the formula for estimating the expected frequency in the general cell of a third order table.

Exact agreement with marginal totals follows easily from the method of proportional distribution and can be established algebraically by summing the estimation equation by first order classes; e.g., summing equation (8.1) by rows and columns. In practice, discrepancies are always either errors of rounding or mistakes in computation; they are never due to lack of convergence of iterative processes as is often true in alternative methods of estimation.

Although simplicity of interpretation is desirable in general, it is especially important when random sampling is an unrealistic abstraction. For example, the method of proportional distribution of marginal adjustments has been used to estimate the cell proportions in a two-way classification of incomes from known marginal proportions and a detailed cross classification at an earlier date. In this problem known shifts in income distributions made it evident that certain cells previously vacant should not have the zero proportions which would be estimated for them by other available methods of estimation. The ease with which the effects of the method of adjustment can be traced is important also in the analysis of the results of sample surveys in which various types of bias are important.

The method of proportional distribution of marginal adjustments yields the estimated expected frequency for any cell by a single sequence of computations which is independent of the corresponding process for any other cell. Errors made in computing the estimate for any cell appear in marginal totals of estimates for all first order classes which include that cell. If only a few errors are made in a table they can be localized immediately and can be corrected without recomputing any estimates which are correct.

In certain types of social surveys, rapidity of processing is so important that, as Deming puts it, "the delay of only the brief time required for adjustment may not be advisable." [1, p. 102]. Under these conditions, it is important to have a simple formula like equation (8.1) in which substitutions can be made rapidly. Even when the time element is relatively unimportant, the economy of effort and the ease of explaining the method to clerical assistants are often of practical importance.

Finally, departures from proportionality among rows and columns often provide the chief element of interest in research studies—not only in social surveys of the type illustrated in Deming and Stephan's example but also in biological sciences. The most effective tests of significance for the purpose of presenting statistical evidence of lack of proportionality are those based on statistics like those derived by the method of proportional distribution of marginal adjustments whose efficiency is 100 per cent when proportionality is exact.

Even when proportionality is not exact, the efficiency of statistics derived by proportional distribution may be close to 100 per cent under fairly typical problem conditions such as those in the example by Deming and Stephan wherein the other more complicated methods require several times as much computational effort, but have little advantage over the easier method with respect to efficiency of estimation in this particular problem.

9. Suggested improvements in techniques. In section 7, a method was outlined by which it is possible to derive sets of maximum likelihood statistics by merely integrating available techniques without changing any of them. In this section a number of improvements are suggested. At this point it should be emphasized that a given change is not an improvement merely because it yields slightly more accurate estimates or makes possible a slight saving of time and effort. In each case the research worker should consider saving of time and effort and accuracy of estimation simultaneously. In particular, it seems likely that most social surveys of the type considered by Deming and Stephan are characterized by approximate proportionality by rows and by columns—conditions relatively favorable to the simple method of proportional distribution of marginal adjustments. It should be clearly understood that suggestions in this section are intended for those research workers whose problems justify a great deal more effort than is required to adjust sample frequencies by this simple method.

Assuming that the problem at hand warrants the effort required to derive maximum likelihood estimates, the first consideration is the derivation of a set of $m_{ij}(1)$, first approximations to the m_{ij}, and a set of values of $p_i(1)$, first approximations to the p_i. Even if proportionality by rows and by columns is not closely approximated use of values of the $p_i(1)$ provided by equation (8.3) are especially to be recommended. In the example used by Deming these values for the $p_i(1)$ are so much better than the values recommended by Deming that they save a large proportion of the effort required by the iterative process. If rows and columns are approximately proportional, equation (8.1) should be used to provide values of the $m_{ij}(1)$, in which case it is possible to use an iterative process similar to the one used by Deming but based on the typical equation of maximum likelihood (7.3) to achieve a given degree of accuracy in the maximum likelihood estimates with even less effort. Under favorable conditions such as those in Deming's example the suggested iterative process yields excellent approximations to maximum likelihood estimates by means of the following steps:

1. Construct a set of first approximations to the r row components of the rs maximum likelihood divisors $(a_i + b_j)$ by means of the equation

$$a_i(1) = n_{i.}/np_{i.} - 1/2. \tag{9.1}$$

2. Compute successive approximations to the a_i and b_j by means of the equations

$$b_j(g) = [n_{.j} - \Sigma m_{ij}(1)a_i(g)]/np_{.j}, \tag{9.2}$$

$$a_i(g + 1) = [n_{i.} - \Sigma m_{ij}(1)b_j(g)]/np_{i.}, \tag{9.3}$$

where $m_{ij}(1)$, the first approximation to m_{ij}, is derived by means of equation (8.1). Just as in Deming's iterative process, the expression in brackets is a series of products which can be subtracted in a single sequence of machine operations and the final division can be performed without having to record any of the intermediate results.

3. Divide the sample frequencies by the maximum likelihood divisors to obtain the maximum likelihood estimates

$$m_{ij} = n_{ij}/(a_i + b_j), \tag{9.4}$$

where limiting values of a_i and b_j are approximated as closely as desired by successive approximations in the preceding equations.

Under unfavorable conditions, the iterative process of this section is not always the easiest way to obtain satisfactory estimates. For example, when samples are small and/or rows and columns are not approximately proportional, it is better to use the iterative method as originally presented by Stephan where sample frequencies can be used for first approximations to the c_{ij} and these may be replaced by successively better approximations.

The point made in the final paragraph of Fisher's well-known book [3] that "in practice one need seldom do more than solve, at least to a good approximation, the equation of maximum likelihood," is strongly supported by the developments of this article. In addition, the proof that the method of least squares and the method of minimum chi-square always lead (by means of approximations to ideal weights) to maximum likelihood statistics greatly facilitates the adaptation of techniques developed in connection with these hitherto competing methods.

REFERENCES

[1] W. Edwards Deming, *Statistical Adjustment of Data*, John Wiley & Sons, 1943.

[2] W. Edwards Deming and Frederick F. Stephan, "On a least squares adjustment of a sample frequency table when the expected marginal totals are known," *Annals of Math. Stat.*, Vol. 11 (1940), pp. 427–444.

[3] R. A. Fisher, *Statistical Methods for Research Workers*, 6th ed., Oliver and Boyd, 1936, Ch. 9.

[4] KARL PEARSON, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Phil. Mag.*, Vol. 50 (1900), pp. 157–175.

[5] FREDERICK F. STEPHAN, "An iterative method of adjusting sample frequency tables when expected marginal totals are known," *Annals of Math. Stat.*, Vol. 13 (1942), pp. 166–178.

[6] E. T. WHITTAKER, AND G. ROBINSON, *The Calculus of Observations*, D. Van Nostrand Company, 1924, Ch. 9.