ESTIMATION BY LEAST SQUARES AND BY MAXIMUM LIKELIHOOD

JOSEPH BERKSON
MAYO CLINIC AND MAYO FOUNDATION*

We are concerned with a functional relation:

$$P_i = F(x_i, \alpha, \beta) = F(Y_i) \tag{1}$$

$$Y_i = \alpha + \beta x_i \tag{2}$$

where P_i represents a true value corresponding to x_i, α, β represent parameters, and Y_i is the linear transform of P_i. At each of $r \geqq 2$ values of x, we have an observation of p_i which at x_i is distributed as a random variable around P_i with variance σ_i^2. We are to estimate the parameters as $\hat{\alpha}$, $\hat{\beta}$ for the predicting equation

$$\hat{p}_i = F(x_i, \hat{\alpha}, \hat{\beta}) . \tag{3}$$

By a least squares estimate of α, β is generally understood one obtained by minimizing

$$\sum \frac{1}{\sigma_i^2}(p_i - \hat{p}_i)^2 . \tag{4}$$

Although statements to the contrary are often made, application of the principle of least squares is not limited to situations in which p is normally distributed. The Gauss-Markov theorem is to the effect that, among unbiased estimates which are linear functions of the observations, those yielded by least squares have minimum variance, and the independence of this property from any assumption regarding the form of distribution is just one of the striking characteristics of the principle of least squares.

The principle of maximum likelihood, on the other hand, requires for its application a knowledge of the probability distribution of p. Under this principle one estimates the parameters α, β so that, were the estimates the true values, the probability of the total set of observations of p would be maximum. This principle has great intuitive appeal, is probably the oldest existing rule of estimate, and has been widely used in practical applications under the name of "the most probable value." If the p_i's are normally distributed about P_i with σ_i^2 independent of P_i, the principle of maximum likelihood yields the same estimate as does least squares, and Gauss is said to have derived least squares from this application.

In recent years, the principle of maximum likelihood has been strongly advanced under the influence of the teachings of Sir Ronald Fisher, who in a renowned paper of 1922 and in later writings [1] outlined a comprehensive and unified system of mathematical statistics as well as a philosophy of statistical inference that has had profound and wide development. Neyman [2] in a fundamental paper in 1949 defined a family of estimates,

* The Mayo Foundation, Rochester, Minnesota, is a part of the Graduate School of the University of Minnesota.

the R.B.A.N. estimates, based on the principle of minimizing a quantity asymptotically distributed as χ^2, which have the same asymptotic properties as those of maximum likelihood.

F. Y. Edgeworth [3] in an article published in 1908 presented in translation excerpts from a letter of Gauss to Bessel, in which Gauss specifically repudiated the principle of maximum likelihood in favor of minimizing some function of the difference between estimate and observation, the square, the cube or perhaps some other power of the difference. Edgeworth scolded Gauss for considering the cube or any other power than the square, and advocated the square on the basis of considerations that he advanced himself as well as on the basis of Gauss's own developments in the theory of least squares. Fisher's revival of maximum likelihood in 1922 is thus seen to be historically a retrogression. Whether scientifically it was also a retrogression or an advance awaits future developments of statistical theory for answer, for I do not think the question is settled by what is now known.

When one looks at what actually has been proved respecting the variance properties of maximum likelihood estimates, we find that it comes to little or nothing, except in some special cases in which maximum likelihood and least squares estimates coincide, as in the case of the normal distribution or the estimate of the binomial parameter. What has been mathematically proved in regard to the variance of maximum likelihood estimates almost entirely concerns asymptotic properties, and no one has been more unequivocal than Sir Ronald Fisher himself in emphasizing that this does not apply to real statistical samples. I hasten to note that, from what has been proved, there is a great deal that reasonably can be inferred as respects approximate minimum variance of the maximum likelihood estimate in large samples. But these are reasonable guesses, not mathematical proof; and sometimes the application in any degree, and always the measure of approximation, is in question. Of greatest importance is this: the maximum likelihood estimate is not unique in possession of the property of asymptotic efficiency. The members of Neyman's class of minimum χ^2 estimates have these properties and he introduced a new estimate in this class, the estimate of minimum reduced χ^2. Taylor's [4] proof that the minimum logit χ^2 estimate for the logistic function and the minimum normit χ^2 estimate for the normal function advanced by me [5], [6] fall in this class directs attention to the possibility of its extension.

In this paper is presented such an extension applying to a particular situation in which $P_i = 1 - Q_i$ is the conditional probability given x_i of some event such as death, and where $Y_i = \alpha + \beta x_i$ is the linear transform of P_i. This is the situation of bio-assay as it has been widely discussed.

We define a class of least squares estimates either by the minimization of

$$(5) \qquad (A) \qquad \sum w_i (p_i - \hat{p}_i)^2$$

where p_i is an observed relative frequency at x_i, distributed binomially about P_i, $\hat{p}_i$ is the estimate of P_i and $1/w_i$ is any consistent estimate of the variance of p_i; or by the minimization of

$$(6) \qquad (B) \qquad \sum W_i (y_i - \hat{y}_i)^2$$

where y_i is the value of the linear transform Y_i corresponding to p_i, $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$ is the estimated value of the linear transform in terms of $\hat{\alpha}$, $\hat{\beta}$, the estimates of α, β, respec-

tively, and $1/W_i$ is any consistent estimate of the variance of y_i. The quantities (5) and (6) which are minimized are asymptotically distributed as χ^2.

The minimum logit χ^2 estimate and the minimum normit χ^2 estimate fall in the defined class of least squares estimates (B), and, as I mentioned previously, Taylor proved that these are R.B.A.N. estimates. Recently Le Cam kindly examined the class of estimates given by the extended definition and in a personal communication informed me that, on the basis of what is demonstrated in the paper of Neyman previously referred to and Taylor's paper, this whole class of least squares estimates can be shown to have the properties of R.B.A.N. estimates. They are therefore asymptotically efficient.

The defined class contains an infinity of different specific estimates, of which a particular few suggest themselves for immediate consideration.

Suppose we minimize

$$\sum \frac{n_i}{\hat{p}_i \hat{q}_i} (p_i - \hat{p}_i)^2 \tag{7}$$

where n_i is the number in the sample on which the observed p_i is based and the $\hat{p}_i$ of the weight w_i is constrained to be the same value as the estimate $\hat{p}_i$. If $\hat{p}_i$ is a consistent estimate of P_i, then $\hat{p}_i\hat{q}_i/n_i$ is a consistent estimate of the variance of p_i and this estimate falls in the defined class. Now the expression (7) is identically equal to the classic χ^2 of Pearson, so that this particular least squares estimate is identical with the minimum χ^2 estimate.

Suppose we have some other consistent estimate of P_i which we shall symbolize as $\hat{p}_o = 1 - \hat{q}_o$ (omitting the subscripts i) and we minimize

$$\sum \frac{n_i}{\hat{p}_o \hat{q}_o} (p_i - \hat{p}_i)^2 ; \tag{8}$$

then this is a least squares estimate as defined. The weights $w_i = n_i/\hat{p}_o\hat{q}_o$ are now known constants, and to minimize (8) we set the first derivatives equal to zero and obtain the equations of estimate

$$\sum \frac{n_i}{\hat{p}_o \hat{q}_o} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \alpha} = 0 \tag{9}$$

$$\sum \frac{n_i}{\hat{p}_o \hat{q}_o} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \beta} = 0 . \tag{10}$$

If now we specify that *in the conditional equations* (9), (10), $\hat{p}_o = \hat{p}_i$, that is, that the values yielded as the estimates shall be the same as those used in the coefficients, then the equations of estimate become

$$\sum \frac{n_i}{\hat{p}_i \hat{q}_i} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \alpha} = 0 \tag{11}$$

$$\sum \frac{n_i}{\hat{p}_i \hat{q}_i} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \beta} = 0 . \tag{12}$$

The equations (11) and (12) are just the equations of estimate of the M.L.E. Therefore the M.L.E. is also a particular member of the defined class of least squares estimates.

This may be presented more directly in a way that emphasizes an interesting point. Suppose the independently determined consistent estimate $\hat{p}_o$ to be used in the constant weights w_i for minimizing (8) is in fact the one obtained as the solution of (11) and (12). Then $\hat{p}_i$, the estimate obtained, will be the same as was used in the weights and this is the M.L.E. This is clear if we observe that we should obtain these least squares estimates as the solution of (9), (10), and we already have noted that these are

satisfied with $\hat{p}_o = \hat{p}_i$ if $\hat{p}_i$ is the M.L.E. The estimate obtained by minimizing (8) is consistent with the estimate used in the weights, only if the estimate appearing in the weights in equation (8) is the M.L.E. For instance, if we use for $\hat{p}_o$ in the weights w_i not the M.L.E. but the minimum χ^2 estimate, the estimate which will be obtained is not the minimum χ^2 estimate, nor is it the M.L.E., but another estimate which is neither, although it too is asymptotically efficient. This is seen at once if we note that the conditional equations of estimate for the minimum χ^2 estimate [7] are not (11), (12) but

$$\sum n_i \frac{\hat{p}_i q_i + p_i \hat{q}_i}{(\hat{p}_i \hat{q}_i)^2} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \alpha} = 0 \tag{13}$$

$$\sum n_i \frac{\hat{p}_i q_i + p_i \hat{q}_i}{(\hat{p}_i \hat{q}_i)^2} (p_i - \hat{p}_i) \frac{\partial \hat{p}_i}{\partial \beta} = 0 . \tag{14}$$

I should like to make quite clear and convincing that the M.L.E. is derivable as a least squares estimate and that this is not an artificial contrivance used to lure the M.L.E. into the family of defined least squares estimates. In fact there is a reasonable way of proceeding by which the M.L.E. is derived as the most natural or least arbitrary of the least squares estimates of the family (A). Suppose one had never heard of the M.L.E. but only of a least squares estimate in the sense of minimizing (5). To obtain such an estimate, it would be natural to wish to use for $1/w_i$ the value P_iQ_i/n_i. It would be realized that we did not have P_iQ_i/n_i because we did not have the true value $P_i = 1 - Q_i$, and it would be natural to suggest that, not knowing the true value, we should use an estimate. But what estimate? It would be reasonable to say that we should use the same estimate as the one which we were going to use for the final estimate. If this is what we wished to accomplish how would we go about getting the desired result? We might proceed by taking first some provisional estimate, $\hat{p}_o = 1 - \hat{q}_o$. We would then obtain the least squares estimates by differentiating (5), yielding the estimating equations (9), (10). The least squares estimates would thus be the solution of these equations (9) and (10). In general the estimates obtained would not be the same as those used provisionally, that is $\hat{p}_i \neq \hat{p}_o$. We would then take the estimates just obtained and use them as new provisional values to obtain new estimates. We would notice that the new estimates are closer to those used originally and repeating the procedure we would notice that the two estimates, the one used in the weights and the one obtained using these weights, became closer and closer to one another. At some point we would be satisfied that we had fulfilled closely enough the objective of obtaining a least squares estimate minimizing (5) with the weights w_i in terms of the estimates, that is $w_i = n_i/\hat{p}_i\hat{q}_i$. Now the procedure that I have described is just the mechanics of obtaining a M.L.E. in the standard way. For what we would be doing is obtaining by iteration a solution of equations (11), (12), which are the estimating equations of the M.L.E.

Objectively an estimate is defined by the estimating equations of which it is the solution, and not by the motivation by which these equations are obtained. The estimating equations (11), (12) can be obtained from a requirement to meet the condition of maximizing the probability of the sample, but they are also derived if a least squares criterion is set up, as just described. It is therefore as legitimate to consider the estimate a least squares estimate as it is to consider it a M.L.E. It suggests itself as a possibility that the minimum variance characteristics of the M.L.E., such as they are, are obtained by this estimate because it minimizes a squared deviation of the observation from the estimate rather than because it maximizes the probability of the sample.

The most familiar consistent estimate of the variance of p_i is given by p_iq_i/n_i where $p_i = 1 - q_i$ is the observed relative frequency and n_i is the number in the sample on which it is based. If we use this estimate to define w_i we shall minimize

$$\sum \frac{n_i}{p_iq_i}(p_i - \hat{p}_i)^2. \tag{15}$$

The expression (15) is equal to the reduced χ^2 of Neyman, so that another familiar estimate in the defined class of least squares estimates is the minimum reduced χ^2 estimate of Neyman.

Now I turn for a moment to the least squares estimates (B) defined in terms of the linear transform y. A consistent estimate of the variance of y is given by

$$\sigma_y^2(\text{asymptotic}) = \frac{pq}{nz^2} \tag{16}$$

where z is the value of $Z = \partial P/\partial Y$ corresponding to the observed p. The corresponding least squares estimate is obtained by minimizing

$$\sum \frac{n_iz_i^2}{p_iq_i}(y_i - \hat{y}_i)^2. \tag{17}$$

This is the estimate which when P is the logistic function I have called the "minimum logit χ^2 estimate" and when P is the integrated normal distribution function I have called the "minimum normit χ^2 estimate." In general we may perhaps call it the "minimum transform χ^2 estimate."

Of the estimates mentioned, the minimum transform χ^2 estimate is the easiest to compute, since it is obtained directly as a weighted least squares estimate of a straight line with known constant weights. The other estimates generally require iterative procedures which are usually time consuming and cumbersome to calculate.

Another pair of estimates in the defined class that have a special interest are those in which for the weight w_i in (5) or the weight W_i in (6) we use the reciprocal of the true variance of p_i or y_i, respectively. These will of course be of no practical application because in practice we shall not know the true variances, since they are functions of the parameters which it is the purpose of the procedure to estimate. Such estimates because of their impracticability were called "ideal" least squares estimates by Smith [8]. But in a number of developments which have been discussed, one can discern a desire to use weights which, as closely as possible, are proportional to the inverse of the true variances [9], [10], [11], [12] and one is curious to know what sort of estimate is obtained if one actually uses the true variances.

The six estimates mentioned will be considered, that is, the M.L.E., minimum χ^2, minimum reduced χ^2, minimum transform χ^2 and the two ideal least squares estimates, the one in terms of the function P, the other in terms of the transform Y. All are asymptotically efficient, but we are interested in their variance properties for finite samples. To get some idea about this I resorted to experimental sampling. The general method of sampling and calculations employed are described in reference [13]. The number of samples on which each estimate is based is given in table I. The situation dealt with simulates that of a bio-assay: in which (i) the probability of death for varying doses is given by (1) the logistic function, and (2) the integrated normal distribution function; (ii) with three equally spaced doses, 10 at each dose; (iii) for dosage arrangement (a)

symmetrically about $P = 0.5$, with $P = 0.3, 0.5, 0.7$, respectively, and (b) asymmetrically with the same spacing of dosages with central $P = 0.8$.

The results are exhibited in tables II through VII. In all the tables the results are shown for both the logistic function and the normal function. Since the relationship of the estimates one to the other was found throughout to be the same for both functions, I shall refer to this or that characteristic of the estimate without specifying the function to which it applies, because it will always apply to both.

TABLE I

NUMBER OF SAMPLES USED IN CALCULATION OF STATISTICS

Estimate	Logistic		Normal	
	α Estimated	α and β Estimated	α Estimated	α and β Estimated
1. Maximum likelihood	Total population	Total population	Stratified sample 600	Stratified sample 600
2. Minimum χ^2	Total population	Stratified sample 1,000	Stratified sample 100	Stratified sample 100
3. Minimum reduced χ^2	Stratified sample 100	Stratified sample 100	Stratified sample 100	Stratified sample 100
4. Minimum transform χ^2	Total population	Total population	Stratified sample 600	Stratified sample 600
5. Ideal, least squares, function	Stratified sample 100	Stratified sample 100	Stratified sample 100	Stratified sample 100
6. Ideal, least squares, transform	Total population	Total population	Total population	Total population

TABLE II

β KNOWN, α TO BE ESTIMATED, CENTRAL $P = 0.5$

Estimate	Logistic, $\alpha = 0$ $1/I = .149$			Normal, $\alpha = 0$ $1/I = .056$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	0	.158	.158	0	.058	.058
2. Minimum χ^2	0	.139	.139	.001	.054	.054
3. Minimum reduced χ^2	.002	.229	.229	.001	.086	.086
4. Minimum transform χ^2	0	.137	.137	−.001	.054	.054
5. Ideal, least squares, function	0	.165	.165	0	.061	.061
6. Ideal, least squares, transform	0	.191	.191	0	.066	.066

Table II shows the results for the estimate of α, when β is considered known, for the true P's corresponding to the x's symmetrically disposed about central $P = 0.5$. One finding may occasion surprise. I refer to what is contained in line 6, that is, the ideal least squares estimates in terms of the linear transform. These estimates are obtained from a simple least squares fit of a straight line using as constant weights the reciprocal of the true variances. The estimate is unbiased, but its variance is not the smallest among the estimates; in fact, except for number 3, the minimum reduced χ^2 estimate, it has the largest variance. The estimate is a linear function of the observations, it is unbiased, but it does not have minimum variance. On its face this finding appears to violate the Gauss-

Markov theorem. The explanation of this riddle can be best presented with findings shown in succeeding tables and will be deferred till these are discussed. The ideal least squares estimate in terms of the linear transform, then, does not have minimum variance; and neither, as may be noted, does number 5, which is the ideal least squares estimate in terms of the untransformed variate p. We may now examine the practical least squares estimates, that is, those which are computed in terms of the observations, and which could be applied in practice; these are the ones listed number 1 through 4. Among the practical least squares estimates, it is not the maximum likelihood estimate but the

TABLE III

α, β, BOTH TO BE ESTIMATED; ESTIMATE OF α, CENTRAL $P = 0.5$

ESTIMATE	LOGISTIC, $\alpha=0$			NORMAL, $\alpha=0$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	0	.187	.187	0	.068	.068
2. Minimum χ^2	.002	.179	.179	0	.061	.061
3. Minimum reduced χ^2	.003	.312	.312	.011	.116	.116
4. Minimum transform χ^2	0	.154	.154	0	.058	.058
5. Ideal, least squares, function	−.005	.212	.212	.003	.077	.077
6. Ideal, least squares, transform	0	.191	.191	0	.066	.066

TABLE IV

α, β, BOTH TO BE ESTIMATED; ESTIMATE OF β, CENTRAL $P = 0.5$

ESTIMATE	LOGISTIC, $\beta=.84730$			NORMAL, $\beta=.52440$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	.095	.313	.322	.045	.106	.108
2. Minimum χ^2	.062	.276	.280	.023	.095	.095
3. Minimum reduced χ^2	.213	.464	.509	.115	.171	.184
4. Minimum transform χ^2	.048	.268	.271	.027	.093	.094
5. Ideal, least squares, function	.114	.336	.349	.062	.126	.129
6. Ideal, least squares, transform	.108	.303	.315	.049	.102	.104

TABLE V

β KNOWN, α TO BE ESTIMATED; CENTRAL $P = 0.8$

ESTIMATE	LOGISTIC, $\alpha=0$ $1/I=.208$			NORMAL, $\alpha=0$ $1/I=.071$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	.056	.246	.249	.027	.089	.090
2. Minimum χ^2	−.059	.207	.211	−.035	.075	.076
3. Minimum reduced χ^2	.189	.261	.296	.090	.097	.105
4. Minimum transform χ^2	−.097	.187	.196	−.038	.064	.066
5. Ideal, least squares, function	.051	.217	.220	.040	.116	.118
6. Ideal, least squares, transform	.059	.181	.184	.103	.053	.064

minimum transform χ^2 estimate that shows the smallest variance, while the largest variance is shown by the minimum reduced χ^2 estimate, which is the only one exceeding that of the M.L.E. Before proceeding to the next table, I should like to add to the riddle about line 6 a riddle about lines 4 and 2. At the head of the table you will see given the numerical value of $1/I$, where I is the amount of information. This, it is to be recalled, is the lower bound for the variance of an unbiased estimate. On line 4 and line 2 it may be seen that the minimum transform χ^2 estimates and minimum χ^2 estimates are unbiased, but that their variances are less than this lower bound value.

TABLE VI

α, β, BOTH TO BE ESTIMATED; ESTIMATE OF α, CENTRAL $P = 0.8$

ESTIMATE	LOGISTIC, $\alpha=0$			NORMAL, $\alpha=0$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	−.026	1.102	1.103	−.019	.311	.311
2. Minimum χ^2	.037	.970	.972	.039	.273	.275
3. Minimum reduced χ^2	−.049	1.221	1.223	.032	.397	.398
4. Minimum transform χ^2	.084	.682	.689	.036	.268	.270
5. Ideal, least squares, function	−.029	1.066	1.067	.030	.303	.304
6. Ideal, least squares, transform	.133	.873	.891	.081	.254	.261

TABLE VII

α, β, BOTH TO BE ESTIMATED; ESTIMATE OF β, CENTRAL $P = 0.8$

ESTIMATE	LOGISTIC, $\beta=.84730$			NORMAL, $\beta=.52440$		
	Bias	Variance	M.S.E.	Bias	Variance	M.S.E.
1. Maximum likelihood	.088	.458	.466	.023	.123	.123
2. Minimum χ^2	−.019	.392	.392	0	.111	.111
3. Minimum reduced χ^2	.002	.570	.570	.087	.158	.165
4. Minimum transform χ^2	−.077	.202	.208	−.052	.076	.079
5. Ideal, least squares, function	.088	.411	.419	.042	.110	.112
6. Ideal, least squares, transform	−.044	.258	.260	−.044	.077	.079

Table III shows the results for estimate of α with central $P = 0.5$, when α and β are both considered unknown and are estimated simultaneously. If line 6, which records the results for the ideal linear χ^2 estimate, is examined, it will be noted that again it fails to achieve the smallest variance. Among the practical least squares estimates exhibited in lines 1 through 4, the minimum transform χ^2 estimate shows the smallest variance, the next in order being the minimum χ^2, the M.L.E., and minimum reduced χ^2. Table IV shows the estimate for β in the same conditions as just described for α. As can be seen on line 6, it is again found that the ideal linear χ^2 estimate fails to achieve smallest variance, but now this riddle is wrapped in an enigma for the estimate is not only not minimum variance, but it is not unbiased. One can easily find textbook proofs that a linear regression fitted by least squares is unbiased in the estimates of the parameters. What is the explanation of the present finding? It is, I should say, that while we are dealing with a linear functional relation, we are not dealing with a linear regression. By a regression of y

on x I understand the conditional expectation of a random variable y, given x. Now the primary functional relation P_i is a regression, because while it is written as for the true value P_i, it is a fact that the expectation of the observed values of the binomial variate p_i is the true value P_i. Therefore the function gives the expected value of the dependent variate p_i at the same time that it gives the true value of P_i. In the linear transform equation (2) Y_i, the true value of the transform (that is, the transform of the true P_i), is a straight line function of x_i, but the expectation of y_i, the observed values of the transform (that is, the transform value of the observed p_i), does not have as its expectation the true transform Y_i. We are dealing with an example of the elementary and well known fact that the average of a function is not necessarily equal to the function of the average. In all the discussion that has taken place in recent years about fitting a function by way of the linear transform, it appears that this point has been overlooked.

If we consider the enigma explained, we still have to deal with the riddle it contained. In the case of estimating α for central $P = 0.5$, the estimate of α is unbiased, but still the ideal linear χ^2 estimate did not attain minimum variance. The explanation, I believe, is this: when we say that the Gauss-Markov theorem refers to unbiased estimates of the parameter, we should understand that this means unbiased for all values of the parameter. An estimate may be unbiased for some special situation, as it is for the estimate of α, with symmetrical disposition of the P's about central $P = 0.5$, but if it is not unbiased for all other situations, then it may not have minimum variance even where it is unbiased. The same consideration of the estimator's being sometimes but not always unbiased is the explanation of the apparent paradox shown in table II, of the attainment by the minimum transform χ^2 estimate and the minimum χ^2 estimate of a variance which is less than $1/I$, even though they are in that experiment unbiased. The complete formula for the lower bound contains the term $\partial b/\partial \alpha$ where b is the bias, and if the value of the quantity is negative, it is possible for the variance of the estimate to be less than $1/I$ even if the value of $b = 0$.

Tables V, VI and VII show results respectively similar to those shown in tables II, III and IV except that the dosage arrangement now corresponds to true P's asymmetrically disposed around $P = 0.8$. In table V are shown the results for estimate of α, β known. Inspecting line 6 again it is seen that the estimate of α is biased, which no longer surprises us; but the variance and mean square error paradoxically are now smallest among the estimates listed. This finding, however, is just an exhibition of the mischievousness of this estimate, which seems bent on creating confusion; we shall see presently that it is not always of smallest variance when it is biased. Returning to the orderly part of the comparisons, we find that among the practical least squares estimates, the minimum transform χ^2 estimate has smallest variance and mean square error, the minimum Pearson χ^2 next, the maximum likelihood next, and the minimum reduced χ^2 estimate the largest. Also we may note again that the minimum χ^2 estimates attain variances less than $1/I$. Table VI exhibits the results for the estimate of α when both parameters are to be estimated. In line 6 we may now observe that here the minimum linear χ^2 estimate does not have minimum variance and neither is it unbiased. The practical least squares estimates are in the same order in magnitude of the variances and the mean square errors as previously noted. In table VII the results are shown for estimate of β; the same order is shown again in the magnitudes of the variances and mean square errors.

I may summarize as follows: for a situation involving the binomial variate, a class of least squares estimates has been defined which are R.B.A.N. estimates. An estimate of

this class has been derived which is identical with the maximum likelihood estimate. While it is in this class of best estimates, the M.L.E. is not necessarily the best among the best. In some circumstances the M.L.E. may be among the worst of all best possible estimators. For the logistic function and the normal function, the M.L.E. was compared in a specified situation with three other least squares estimates, the minimum χ^2, the minimum reduced χ^2 and the minimum transform χ^2, with respect to variance and mean square error. The order of the magnitude of the errors was found to be smallest for the minimum transform χ^2, next for the minimum χ^2, next for the maximum likelihood, and largest for the minimum reduced χ^2.

◊ ◊ ◊ ◊ ◊

Note added in proof. After the presentation of this paper, an extension of the class of least squares estimates (A) was developed, applicable to a multinomial variable.

Suppose there are $r \geqq 2$ classes, the probability of the events in which are, respectively, $P_1 = F_1(\theta)$, $P_2 = F_2(\theta), \cdots, P_r = F_r(\theta)$, where θ represents one or more parameters. When n trials have been made, a generalized minimum χ^2 estimate is obtained as the value of θ which minimizes

$$\chi^2_0 = \sum_{i=1}^{i=r} \frac{(o_i - e_i)^2}{e_0} \tag{18}$$

where

$e_i = nF_i(\theta)$ is the "expected" number in the ith class,

o_i is the observed number in the ith class,

$e_0 = nF_i(\hat{\theta}_0)$, where $\hat{\theta}_0$ is any consistent estimate of θ.

It may be shown (see Le Cam [14]) that any estimate falling in the defined class is R.B.A.N. and therefore asymptotically efficient, and that the χ^2_0 of (18) is distributed asymptotically as χ^2 with $(r - 1 - s)$ D.F., where s is the number of parameters estimated.

By suitably defining $\hat{\theta}_0$, the estimate obtained is the maximum likelihood estimate, the minimum χ^2 (Pearson) estimate or the minimum reduced χ^2 estimate.

A general two-step procedure for obtaining the estimate is as follows:

Obtain a preliminary estimate $\hat{\theta}_0$ as the estimate of θ which minimizes

$$\sum (o_i - e_0)^2 \tag{19}$$

where $e_0 = nF_i(\theta)$. This estimate $\hat{\theta}_0$ will be consistent [2]. Compute the respective values of e_0, and, using these as constants, minimize (18).

Employing the example presented by Sir Ronald Fisher [15], the maximum likelihood estimate, the minimum (Pearson) χ^2 estimate, the minimum reduced χ^2 estimate, and the "two-step" generalized χ^2 estimate were compared with respect to variance and mean-square-error. The results obtained so far are only preliminary, but it is clear that the maximum likelihood estimate is not always the best among those compared.

REFERENCES

[1] R. A. FISHER, *Contributions to Mathematical Statistics*, New York, John Wiley and Sons, Inc., 1950.

[2] JERZY NEYMAN, "Contribution to the theory of the χ^2 test," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1949, pp. 239–273.

[3] F. Y. EDGEWORTH, "On the probable error of frequency-constants," *Jour. Roy. Stat. Soc.*, Vol. 71 (1908), pp. 381–397.

[4] W. F. TAYLOR, "Distance functions and regular best asymptotically normal estimates," *Annals of Math. Stat.*, Vol. 24 (1953), pp. 85–92.

[5] JOSEPH BERKSON, "A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function," *Jour. Amer. Stat. Assoc.*, Vol. 48 (1953), pp. 565–599.

[6] ———, "Estimate of the integrated normal curve by minimum normit chi-square with particular reference to bio-assay," *Jour. Amer. Stat. Assoc.*, Vol. 50 (1955), pp. 529–549.

[7] ———, "Minimum χ^2 and maximum likelihood solution in terms of a linear transform, with particular reference to bio-assay," *Jour. Amer. Stat. Assoc.*, Vol. 44 (1949), pp. 273–278.

[8] J. H. SMITH, "Estimation of linear functions of cell proportions," *Annals of Math. Stat.*, Vol. 18 (1947), pp. 231–254.

[9] R. A. FISHER, "The analysis of variance with various binomial transformations," *Biometrics*, Vol. 10 (1954), pp. 130–139.

[10] M. S. BARTLETT, "A comment on the use of the square-root and angular transformations" (discussion of reference [9]), *Biometrics*, Vol. 10 (1954), pp. 140–141.

[11] F. J. ANSCOMBE, "Comments" (discussion of reference [9]), *Biometrics*, Vol. 10 (1954), pp. 141–144.

[12] W. G. COCHRAN, "The relation between simple transforms and maximum likelihood solutions" (discussion of reference [9]), *Biometrics*, Vol. 10 (1954), pp. 144–147.

[13] JOSEPH BERKSON, "Maximum likelihood and minimum χ^2 estimates of the logistic function," *Jour. Amer. Stat. Assoc.*, Vol. 50 (1955), pp. 130–162.

[14] LUCIEN LE CAM, Personal communication.

[15] R. A. FISHER, *Statistical Methods for Research Workers*, Ed. 11, Edinburgh, Oliver and Boyd, 1950, pp. 299–325.