RELATIVE PRECISION OF MINIMUM CHI-SQUARE AND MAXIMUM LIKELIHOOD ESTIMATES OF REGRESSION COEFFICIENTS

JOSEPH BERKSON
MAYO CLINIC

Professor Doob [5], in the charming introduction to his paper delivered at the last Berkeley Symposium, remarked that physicists write like physicists, while mathematicians write like mathematicians and only for posterity. I have been given to wonder at times, for whom it is that mathematical statisticians write, and specifically whether it is for statisticians. The mathematician, we know from a witty and authoritative essay by Bertrand Russell [10], never knows what he is talking about—that is, he deals with generalities, not specificities—and he never knows whether what he is saying is true—that is, he is concerned with logical consistency, not physical facts. This ineffable disinterestedness in the tawdry realities of the physical world, mathematics shares with the sonata, abstract nonrepresentational sculpture, and the classic ballet. It does not share it with statistics. Statistics, however you define it, is very much earthbound and deals with real observable data; what is statistically true must be literally verifiably true for such data.

These remarks are prompted by an experience, which I shall presently describe, with a problem that comes to the laboratory quite frequently, and for which I presented a particular solution [1]. Specifically, in respect to any relevant consideration, so far as I could find, this solution was easier and more satisfactory than the standard one advanced by most mathematical statisticians, but it met a stone wall of opposition in the form of general theorems quoted by these mathematicians. Specifically, it appeared, the method might be fine, but generally speaking, it was no good at all. These theorems, I found when I examined the matter, stated that certain things had been proved for "large samples." Further examination disclosed, however, that if these theorems were valid for large samples, they must refer to *infinitely* large samples, which is to say, samples so large that no statistician ever gets them, at least not on this unpleasant earth.

The problem I refer to is that of bioassay, in which the potency of a drug is estimated from an experiment in which a number of animals are exposed to a series of increasing concentrations of the drug, the fraction of animals succumbing at each dose is noted, and to these fractions a regression function is fitted. The fraction affected, plotted against the dosage, the latter measured logarithmically, in certain instances follows a sigmoidal shaped curve and, as is well known, a widely used method is to fit the integrated normal curve to the data by maximum likelihood,

using probits. I, myself, had for more than twenty years, for reasons which I shall not now attempt to review, used for situations of this sort, not the integrated normal curve, but the logistic function. This function is also sigmoidal, has the pictorial appearance of the integrated normal curve, and is so close to it in this sense, as to be almost superposable. The equation is as follows, and, as is seen, it has a very simple linear transform.

$$P = \frac{1}{1 + e^{-(\alpha+\beta x)}}$$

$$\hat{p} = \frac{1}{1 + e^{-(a+bx)}}$$

$$\text{"logit"} = \ln \frac{p}{1-p} = a + \beta x$$

where P is the "true" value, $\hat{p}$ is the estimate of P, p is the observed value of the fraction affected and x is the logarithm of the dosage.

If p is the fraction affected, the transform quantity, which I have called the "logit," is simply $\ln p/(1 - p)$. So if the logistic holds, the logit of the fraction affected, plotted against x, will be a straight line. Wishing to present the use of this function, the question arose with me, How should it be fitted? Everyone was using maximum likelihood, but instead I used least squares, or more specifically for this situation, minimum χ^2. Why? I suppose partly it was natural cussedness, but also I think it was due to parsimoniousness. I had gone to statistical classes years before, paid tuition fees, and learned about least squares. Why should I waste all that money?

To apply minimum χ^2 directly is somewhat difficult, though essentially no more so than is the standard probit maximum likelihood method for the integrated normal curve. So I devised an easier method which can be considered an approximate minimum χ^2 method of estimate, though I may call it hereafter a minimum logit χ^2 estimate. This may be briefly outlined as follows:

A close approximation to χ^2 is given in terms of logits by

$$\chi^2 = \sum \frac{n}{\hat{p}\hat{q}} (p - \hat{p})^2 \cong \sum npq\,(l - \hat{l})^2$$

where $p = (1 - q)$ is the observed fraction, $l = \ln p/q$ is the logit of p, and $\hat{l}$ is the logit corresponding to the estimated value $\hat{p}$. To be noted is the fact that the coefficient in the logit approximation is in terms of the observations. Now it is a rather remarkable fact that although in the expression for χ^2 itself, the coefficient is in terms of the estimates $\hat{p}$ and $\hat{q}$, the approximation is better if, in the logit expression, the observed values p and q rather than the estimated ones are used. It was, indeed, for this reason, that this particular approximation was chosen. Since $\hat{l}$ is a linear function of the parameters to be estimated, with the coefficients in terms of the known observations, a definitive solution for the estimates can be obtained by a simple weighted least squares solution of a straight line, without iteration. This is in contrast with the usual maximum likelihood solution of the probit equation, which requires an iterative procedure that approaches but never achieves the maximum likelihood estimate. So if we are interested in a solution that

minimizes the classic χ^2, we have a simple definitive method that accomplishes this approximately with great efficiency.

An actual trial for a series of cases showed that this method yielded a better fit, by the criterion of the χ^2 test for goodness of fit, than did the standard probit procedure for the same data, in all cases tried. This statement may strike you as queer, for you may say, "Since you used a minimum χ^2 method, what else did you expect?" But you must remember (1) that we are comparing two different functions, the logistic and the integrated normal, as well as two different methods of fitting the functions, and (2) the method used was not minimum χ^2, but an approximate minimum χ^2 method, and *so is maximum likelihood an approximate minimum χ^2 method!* It is, indeed, so referred to by Cramér [4, p. 426]. It is very interesting to note that Cramér develops the minimum χ^2 estimator, by taking the derivative of the χ^2 which is to be set equal to zero, observing that some of the terms are negligible in large samples, and equating the remaining terms to zero, to yield what he calls the "modified χ^2 minimum method." This procedure, it turns out, as Cramér says clearly, yields what is otherwise known as the maximum likelihood estimate. So we may say that I was comparing two approximate minimum χ^2 methods; Cramér's or maximum likelihood, and the one I happened to use, which in the present situation is easier. The examination showed that the one I utilized gives a closer approximation to the exact minimization of χ^2 than does the maximum likelihood method. This, of course, is strictly limited to the present case.

Having provided a method that yields a definitive estimate without iteration that on the face of it is asymptotically efficient, and which therefore has all the *proved* desirable attributes of the maximum likelihood estimate, it is natural that the mathematical statisticians who have worked on, or specialized in, the field of bioassay should have welcomed it with great alacrity, as a useful facilitation of procedures, as well as an amusing departure from the rote into which statistical bioassay seems to have got, ever since the invention of the word "probit." I do not mean that exactly this has happened as yet. Doubtless the acclaim of my achievement will come in due time, but so far we have had only the prelude. The prelude is naturally in a different key and it has taken the form of (1) an explicit rejection and (2) a lively, frank and uncompromising attack. Said Finney [7] in an article in the Journal of the Royal Statistical Society, "Berkson has developed a computational procedure for the logistic transformation, which fails to take account of the asymmetry of binomial frequency distributions by the introduction of 'working responses' or otherwise, though he draws attention to this possibility, and thus does not lead to the maximum likelihood solution. He says, 'I believe that the work of fitting the logistic as given here is considerably simpler than that of fitting the normal curve by probits and maximum likelihood as advocated by Bliss and Fisher.' This greater simplicity is entirely a result of the omission of the calculation of working responses." Admittedly easier, and giving a lower χ^2, my method is rejected because it is not maximum likelihood! This appeal to maximum likelihood as the ultimate criterion of perfection is not limited to Finney. One finds in the writings of many mathematical authorities panegyrics to maximum likelihood as being a principle of estimate hardly conceivable of improvement, so near to the ultimate attainable in smallness of sampling error as to make any attempt to better

it a waste of time, and in any case, it is taken as the basis on which all other methods are to be evaluated.

Such confidence in a procedure at hand is rare, and perhaps it is an exhibition of cantankerousness, not to say meanness, for me to disturb it by suggesting that really it is founded on false premises. On what mathematical demonstrations can these opinions be based? The proved optimum properties of maximum likelihood estimates are *asymptotic* properties. To conclude from them that for statistical, that is, finite, samples, the maximum likelihood estimates are necessarily even good enough estimates, is already going beyond what has been proved. Strictly speaking, to put it without equivocation in the words that Lehmann [8, pp. 451–452] used in his paper of the last Symposium, "Actually it seems doubtful that any definition based only on asymptotic properties . . . can be satisfactory, since in practice the sample size is always limited, and since obviously an asymptotic property implies nothing about the behavior of any finite segment. . . ." "Nothing," then, is what the rigidly correct mathematician will accept as having been demonstrated for real statistical samples. As the practical uncritical statistician that I am, I should not go so far. On the basis of what has been proved, taken together with what we know about the statistical properties of many specific maximum likelihood estimates, I should certainly be willing to grant that maximum likelihood gives good acceptable estimates in general.

This is very different, however, from using the superlative "smallest variance," taken from asymptotic theorems, as applicable in the realm of finite samples. If it can be readily granted as a practical generality that the maximum likelihood estimate is good enough, it cannot be granted at all that it is better, or as good, or nearly as good, as another estimate—any other estimate—until this is shown specifically to be so. Statements of a contrary implication are particularly inappropriate, coming from the school of statistics that has argued with such force that inefficient estimates must not be used, because of the smaller variance of efficient estimates. Am I incorrect in saying that there is not any reason to believe that in the finite realm, the difference between the variance of an inefficient estimate and an efficient estimate is necessarily of a larger order of magnitude than the difference between one efficient estimate and another efficient estimate? If estimator B must replace estimator A because B has a smaller variance than A, then estimator C should replace B if it has an equally smaller variance than B. The inefficient estimate is represented by A, the maximum likelihood estimate by B, and it is entirely possible that the minimum χ^2 estimate could be represented by C.

We shall return to a specific inquiry as to whether this is so, but first I must mention briefly the direct attack that has been made on the minimum χ^2 estimate, to which I made previous reference. This came from Mr. Jerome Cornfield of the United States Public Health Service in a paper delivered on behalf of himself and Mantel [3] at a Biometrics Section meeting, distributed in mimeographed form, and presumably to be published in the Journal of the American Statistical Association. In what Cornfield designated as a "devastating" analysis, he referred not only to deficiencies of the minimum χ^2 estimate, but to its absurdity. The paper was mathematical and I did not follow it entirely, so we shall have to wait for its publication to have it get proper consideration. It appeared clear, however, that

the burden of his argument was that the minimum χ^2 estimate is severely biassed as compared with the maximum likelihood estimate. Although I never was able to substantiate this, let me take a moment to pay my disrespects to the consideration of bias. An estimator which is asymptotically biassed, I will admit freely, is repugnant intuitively. It seems to imply, though I am not sure whether it does, that with such an estimator, even if you sample the entire urn, you do not estimate the right number of black balls, and this seems repugnant, not to say intolerable. Bias in finite samples, however, is another matter, since it means only that the *average* of the estimates does not happen to be the true value, not that the estimates themselves do not frequently evaluate the correct quantity or come close to it or both. I may recall the story you have all heard, of the sportsman who returned curiously early from the hunt, and, asked for an explanation, said, "My first shot was a foot over the bird, my second a foot below, the third a foot to the right and the fourth a foot to the left. Being a statistician I figured my average, and finding it exactly on the mark, I came home." In the present situation, we are dealing with the estimate of the potency of a drug. It is hard to see, if one makes an error understating the potency, resulting in death of several patients from overdose, how this is in any way mitigated by making another error overstating the potency, so that in an equivalent number of instances the patients are not cured, even if they are not killed. Bias of an estimate is of secondary importance, if only because one can eliminate it by applying an unbiassing factor, actually if its amount is known, or symbolically when it is not, if for analytical purposes one wishes to do so.

The question of whether the minimum χ^2 estimator is as good or possibly a materially better one than the maximum likelihood estimator, is then, a real and meaningful one, not prejudiced by what has been proved, favorable as regards maximum likelihood or unfavorable as respects minimum χ^2. What shall we take to mean by "better"? The number of conceivable criteria is indefinitely large. The criterion of size of sampling variance is the one most frequently referred to, though there are others that I have used, all suggested directly by practical problems.

In the present paper—which I am at last about to begin—I shall confine myself to the relative variance and report the results of some sampling experiments to determine this for the minimum χ^2 estimate, as compared with the maximum likelihood estimate, in some regression situations simulating the bioassay experiment. For variance I use the mean square error of the estimate, which I call the "error variance," rather than the variance about the mean, since the estimates may be biassed. By minimum χ^2 estimate I mean now the minimization of the classic Pearson χ^2, not the approximate logit χ^2.

Two regressions were dealt with, the straight line and the logistic function, the straight line because of its fundamental interest in the field of regression, the logistic because of its immediate interest for bioassay. The logistic was used rather than the integrated normal curve, primarily because it is less difficult to make the necessary calculations for the logistic. Three dosages were used, and the true P's at these dosages were initially taken at 0.3, 0.5 and 0.7, which determined the values of the parameters of the regressions for all subsequent experiments. In the first instance the experiments were centered about $P = 0.5$, because a well designed bioassay experiment does this as nearly as possible, the theoretical sampling errors

of the estimates being minimal for this arrangement. Estimates were obtained by iterative methods [2], the iterations being continued until constancy in the third decimal figure was attained. This required usually about three or four iterations, but sometimes many more than this. At each dose there were, for one series of experiments $n = 2$ exposed, for another $n = 10$. In the case of the experiments with $n = 2$, the possible results at each dose are 0, 0.5 and 1 and there being three doses, there are only 27 possible combinations. I obtained the solutions for each of the possible samples and weighted the results in proportion to the total probabilities, thus obtaining the distribution for the entire population, without sampling error. The results for $n = 2$ had a surprising character for which the extensive literature of statistical bioassay did not prepare me. I found that in a large fraction of the samples, about 25 per cent for the straight line and about 40 per cent for the logistic, it was impossible to obtain finite estimates for a and b by either method, the insoluble samples being the same for both methods. Referring to the population of samples for which estimates were obtainable, I found the error variance in the case of the straight line to be smaller for the minimum χ^2 estimate than for the maximum likelihood estimate, for both a and b. The same was true for the logistic. The detailed findings will be given in a later publication reporting these results together with those of experiments still in progress, but there is not time to discuss them now. Actually the most important result for samples with $n = 2$ relates to the finding that so few samples yield any finite estimate at all; therefore the conclusion from these experiments with $n = 2$ for three dose assay is—don't use samples of 2!

Now for $n = 10$. Here 100 samples were obtained for each experiment, stratified and randomized in an appropriate manner. For the linear case and for the logistic, and for both a and b the error variance was smaller for the minimum χ^2 estimate than for the maximum likelihood estimate.

It should be understood that all the experiments summarized thus far were for dosages centered at true $P = 0.5$. These results were distributed preliminarily last year, and several of my mathematical friends more than hinted that my findings, favorable to χ^2, might be related to the circumstance that I had arranged the experiments with the median dosage placed at the L.D. 50. The only reply I could give at that time was that, with only a limited amount of energy available for these experiments, I had done them so as to simulate the best arranged bioassay.

At this point I had an experience that suggested a modification of the experiment, which would make the calculating job less arduous. I found that a large pharmaceutical firm made estimates of dosage on the basis of an already established evaluation of β and estimated the L.D. 50, which is given by the ratio of a to b, by evaluating a alone. An experiment to determine a alone with β known, I could do as for dosages with central P other than 0.5, as well as for dosages with this central P. The various experiments were done as for three evenly spaced doses, centered successively at P's 0.5, 0.6, 0.7, 0.8, 0.85, 0.9. I had intended to go to 0.99, but found that already at $P = 0.9$, now with samples of $n = 10$, I encountered the difficulty of a considerable fraction of the samples being insoluble for the estimates, there being 3 per cent of such samples for the central $P = 0.9$. Figure 1 shows in the plotted points the results as determined from the experimen-

tal sample; the curves are drawn freehand through them as an estimate of the true values. As is seen, the variance is lower for the χ^2 estimator than for the maximum likelihood estimator at all points, not only at $P = 0.5$. Indeed the difference seems to be larger, perhaps, for the outlying P's than for $P = 0.5$. The ratio of the

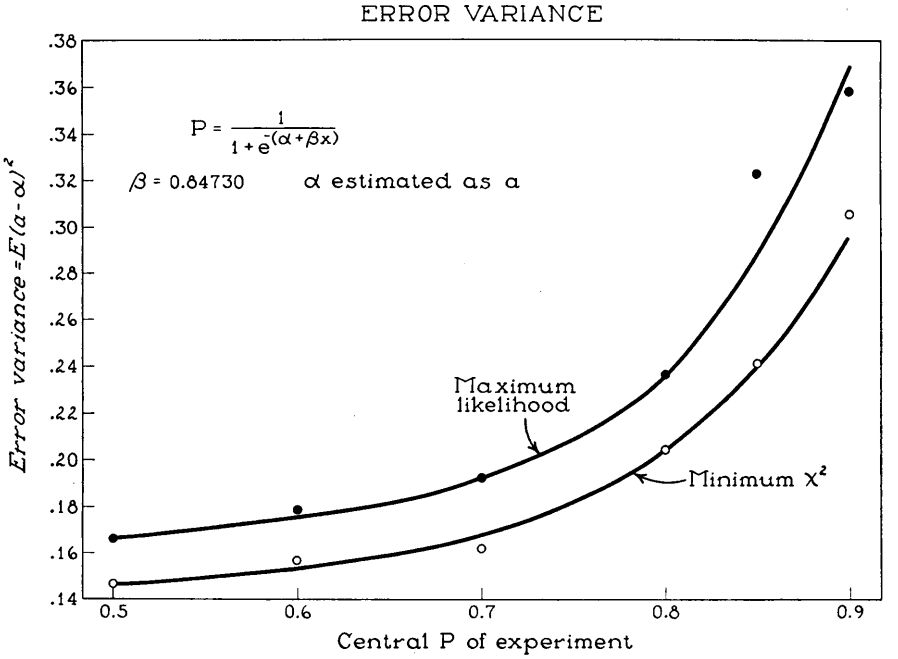

FIGURE 1

Error variance of minimum χ^2 and maximum likelihood estimate of α with β known. Experiment as for three unit spaced doses, with central one of the three at true P indicated; 10 exposed at each dose. One hundred samples for each experiment. The circles correspond to the values calculated from the sample; the smooth curves through these are drawn by eye.

variances is about 1.12 and this is of the same order, say, as the ratio which characterizes the relative efficiency of the estimate of the standard deviation of a normal distribution as computed from the mean deviation, which is an inefficient estimate. Here we see illustrated concretely that the difference of sampling error between two efficient estimates may be of the same size as the difference between an inefficient estimate and the maximum likelihood estimate.

A point of considerable passing interest is the fact that the values for the minimum χ^2 estimate are not only lower than for the maximum likelihood estimate, but lower than the lower bound for the variance given by the Fisher [6, chap. 9]—Cramér [4, pp. 486–487]—Rao [9, pp. 81–91] inequality! All the experimental values except that for the one corresponding to central $P = 0.6$ fell below the minimum variance for an unbiassed estimator, by an average amount of about 2 per cent. As an explanation of this rather disturbing finding, it is to be recalled that the present inquiry disclosed that in the bioassay experiment, there are always a finite number of samples for which neither maximum likelihood nor minimum χ^2 yields

finite estimates. The findings of Cramér [4, pp. 486–487] and Rao [9] for finite samples presumably refer to situations in which all samples of the random population yield estimates. Possibly in this consideration is contained the resolution of the seeming contradiction found in the present investigation. Perhaps account can be taken of the insoluble samples to evaluate a lower bound for the variance, appropriate to the bioassay situation.

I then went back to the case of the estimation of both parameters, α and β. I performed the experiment with central $P = 0.8$, unable to take a larger value, for

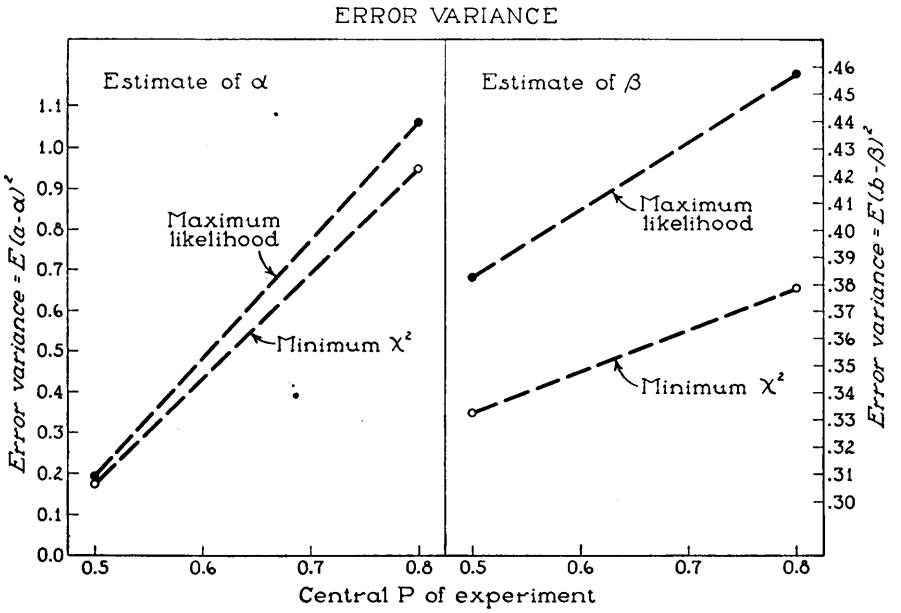

FIGURE 2

Error variance of minimum χ^2 and maximum likelihood estimates of both α and β. Experiments performed as described in the legend of figure 1. The circles correspond to the values calculated from the samples; the interrupted lines connect these points.

with both parameters to be estimated, even at central $P = 0.9$, there are an excessive number of samples without finite solution. Figure 2 shows the results. The variance of the χ^2 estimator is smaller than that of the maximum likelihood estimator, at central $P = 0.8$ as well as at $P = 0.5$.

The results then are, summarily, that for both regressions, linear and logistic, as respects each of the parameters estimated, and for all positions of the dosages, the minimum χ^2 estimator was found to have the smaller error variance. How far can we generalize this? Alas, we are statisticians, not mathematicians; so we cannot generalize at all! Is it reckless, however, to suggest that these findings indicate that this may be generally true: whenever a minimum χ^2 estimate can be defined and it is different from the maximum likelihood estimate, the minimum χ^2 estimator has the smaller sampling error. Is there anything I can do, anything unpleasant enough I can say, to goad the mathematicians into investigating this

question *mathematically?* If any of you do this, I shall be very happy, and I promise to retract all the nasty things I have said about mathematicians.

REFERENCES

[1] J. BERKSON, "Application of the logistic function to bio-assay," *Jour. Amer. Stat. Assoc.*, Vol. 39 (1944), pp. 357–365.

[2] ———, "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.

[3] J. CORNFIELD and N. MANTEL, "Simplified methods for computing the maximum likelihood estimate of the dosage-response curve," abstract 77, *Biometrics*, Vol. 5 (1949), p. 85.

[4] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[5] J. L. DOOB, "Time series and harmonic analysis," *Berkeley Symposium on Mathematical Statistics and Probability*, edited by Jerzy Neyman, University of California Press, Berkeley (1949), pp. 303–343.

[6] R. A. FISHER, *Statistical Methods for Research Workers*, 7th ed., Oliver and Boyd, Edinburgh, 1938.

[7] D. J. FINNEY, "The principles of biological assay," *Jour. Roy. Stat. Soc., Suppl.*, Vol. 9 (1947), pp. 46–81.

[8] E. L. LEHMANN, "Some comments on large sample tests," *Berkeley Symposium on Mathematical Statistics and Probability*, edited by Jerzy Neyman, University of California Press, Berkeley (1949), pp. 451–457.

[9] C. R. RAO, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, Vol. 37, No. 3 (1945), pp. 81–91.

[10] B. RUSSELL, "Mathematics and metaphysicians," *Mysticism and Logic and Other Essays*, Chap. 5, Longmans, Green and Co., London, 1921.

Note added in proof, September, 1951: In the statements made respecting the finding that the mean square error and variance about the mean of the minimum χ^2 estimate was below the "lower bound," the bound was calculated as $1/I$, where $I = E(d \ln \phi/d\alpha)^2$, ϕ being the probability of the total sample, α the parameter to be estimated. Since the time that the report given above was made, my attention has been called to the fact that the pertinent inequality as given by Cramér [4] is:

$$E(\alpha^* - \alpha)^2 \geqq \frac{(1 + db/d\alpha)^2}{I}$$

where b is the bias, and it was pointed out that $db/d\alpha$ *could be negative.* Fortunately, the experiments described provided observations from which $db/d\alpha$ could be estimated; it was found to be negative for the minimum χ^2 estimate, positive for the maximum likelihood estimate.

When the estimated values for $db/d\alpha$ were used to evaluate the right hand side of the inequality of Cramér, the values of the χ^2 estimates as well as those of the maximum likelihood estimates were found to be slightly above their respective lower bounds.

Thus for the situation described the findings are:

(1) The root mean square error for the maximum likelihood estimator is larger than that for the minimum χ^2 estimator, and the last is less than $1/I$.

(2) The Cramér inequality for the root mean square error sets a lower bound for the maximum likelihood estimator which is above that which it sets for the minimum χ^2 estimate and the last is less than $1/I$.

(3) The relations given in (1) and (2) apply if the experiment is performed with central dosage at the LD-50, even though, under these conditions, the estimates are unbiassed.