ON THE RELATIVE EFFICIENCIES OF BAN ESTIMATES[1]

BY LEO KATZ

Michigan State College

1. Introduction. J. Neyman [3] defined BAN (best asymptotically normal) estimates as those functions of observed relative frequencies which i) are consistent, ii) are asymptotically normally distributed, iii) are asymptotically efficient and iv) possess continuous partial derivatives with respect to each relative frequency. He suggested the following two problems; first, to determine the class of estimates which possess the above four properties and second, to investigate this class of estimates to see whether, and under what conditions, the use of some of them is preferable to the use of others. Neyman's paper dealt with the first problem directly and with the second obliquely. With respect to the first problem, he showed that two types of χ^2-minimum estimates belong to the class of BAN estimates as do, obviously, maximum likelihood (ML) estimates. On the second problem, the χ^2-minimum estimates may be more easily computed than the corresponding ML estimates in many cases, the ease of computation being especially pronounced for the modified χ^2 with observed, rather than expected, relative frequencies in the denominators. The present paper contains some additional information regarding the relative merits of these estimates.

For simplicity, we shall consider a random variable taking on values

$$x = 0, 1, 2, 3, \cdots$$

with probabilities $p(x \mid \theta_1, \theta_2, \cdots, \theta_r)$ depending on r parameters. In working with χ^2-minimum estimates, it is almost always necessary to truncate the probability law, taking

$$(1.1)\qquad \begin{aligned} f(x) &= p(x \mid \theta_1, \theta_2, \cdots, \theta_r), \qquad x = 0, 1, \cdots, k-1, \qquad \text{and} \\ f(k) &= \sum_{k}^{\infty} p(x \mid \theta_1, \theta_2, \cdots \theta_r). \end{aligned}$$

The ML estimates are asymptotically efficient, i.e., have minimum variance, with respect to the probability law, $p(x \mid \theta)$, and the χ^2 estimates have the same property with respect to the truncated p. l., $f(x \mid \theta)$. This suggests that the optimum variances of the estimates of the parameters of the two in samples of N may differ and, further, that the minimum variance of the χ^2 estimates may depend essentially upon the choice of k. In the course of some unpublished work by Evelyn Fix and others in the Statistical Laboratory at the University of California on χ^2 estimation of the parameters of several different p. l.'s the same anomalous situation occurred repeatedly. When the observed data were fitted

[1] This paper was presented to a joint meeting of the American Mathematical Society and the Institute of Mathematical Statistics at Boulder, Colorado on September 1, 1949.

by the truncated p. l. with the estimated parameters, the fit appeared to be *improved* when k was chosen smaller. This suggested that perhaps, contrary to intuition, it might be possible to improve the precision of estimation by choosing k smaller, within certain limits. This paper proves that this notion is false and that some other explanation of this phenomenon is needed.

2. Relative efficiency. Cramér [1] has shown, simultaneously with Rao [6], that under mild conditions of regularity, the variance of an unbiased estimate, $\theta^* = \theta^*(x_1, x_2, \cdots, x_N)$, of a single parameter, θ, where $x_1, x_2, \cdots, x_N$ are the observed sample, satisfies the following inequality for fixed N:

$$D^2(\theta^*) \geq \frac{1}{NE\left[\frac{\partial \log p(x)}{\partial \theta}\right]^2}, \tag{2.1}$$

the lower bound being attained only by "efficient" statistics. We may take as a measure of the relative precision attainable in the estimation of the parameter of the truncated p. l. (1.1) the ratio of the lower bounds (2.1) of variances of the estimates of the parameters of the original p. l., $p(x \mid \theta)$, and of the truncated p. l., $f(x|\theta)$. We define

$$\text{Rel. Eff.} = \frac{E\left[\frac{\partial \log f(x)}{\partial \theta}\right]^2}{E\left[\frac{\partial \log p(x)}{\partial \theta}\right]^2}. \tag{2.2}$$

In the case of functions depending on several parameters, $p(x \mid \theta_1, \theta_2, \cdots, \theta_r)$, and unbiased estimates, θ_i^*, which are functions of the observed relative frequencies, with non-singular covariance matrix $\| L_{ij} \|$, Cramér [1] showed that the fixed ellipsoid,

$$N \sum_{i=1}^{r} \sum_{j=1}^{r} \delta_{ij} t_i t_j = r + 2, \tag{2.3}$$

where

$$\delta_{ij} = E\left[\frac{\partial \log p(x)}{\partial \theta_i} \frac{\partial \log p(x)}{\partial \theta_j}\right],$$

lies wholly within the concentration ellipsoid,

$$\sum_{i=1}^{r} \sum_{j=1}^{r} L^{ij} t_i t_j' = r + 2, \tag{2.4}$$

where $\| L^{ij} \| = \| L_{ij} \|^{-1}$. The two ellipsoids coincide if and only if the θ_i^* are joint efficient estimates of the θ_i. Thus, the covariance matrix of a set of joint efficient estimates is $\| N\delta_{ij} \|^{-1}$. In this case, we may define separately the relative efficiency with respect to each of the parameters as in (2.2) or we may consider the set of estimates for one function to possess greater concentration than the set for the other function if the fixed ellipsoid (2.3) for the first lies wholly within the similar ellipsoid for the second. The latter will be the procedure we adopt in section 5.

3. Estimation of a single parameter. With $p(x \mid \theta)$ and $f(x \mid \theta)$ defined as in (1.1), form the difference

$$\phi(k) = E\left[\frac{\partial \log p(x)}{\partial \theta}\right]^2 - E\left[\frac{\partial \log f(x)}{\partial \theta}\right]^2. \tag{3.1}$$

The regularity conditions under which the Cramér-Rao inequality (2.1) holds involve existence of $\partial p(x)/\partial\theta$ for all x and absolute convergence of

$$\sum_x \frac{\partial p(x)}{\partial \theta}.$$

Assuming we have a regular case of estimation in Cramér's sense so that these conditions hold, we may write

$$\phi(k) = \sum_k^{\infty} \frac{1}{p(x)}\left[\frac{\partial p(x)}{\partial \theta}\right]^2 - \frac{1}{f(k)}\left[\frac{\partial f(k)}{\partial \theta}\right]^2, \tag{3.2}$$

and, since $\partial f(k)/\partial\theta = \sum_k^{\infty} (\partial p(x)/\partial\theta)$ by the second of the regularity conditions above and $f(k) = \sum_k^{\infty} p(x)$ by (1.1),

$$\phi(k)f(k) = \sum_k^{\infty} p(x) \sum_k^{\infty}\left[\frac{1}{\sqrt{p(x)}}\frac{\partial p(x)}{\partial \theta}\right]^2 - \left[\sum_k^{\infty}\frac{\partial p(x)}{\partial\theta}\right]^2. \tag{3.3}$$

By the Cauchy inequality, the right member of (3.3) is non-negative and, since $f(k) > 0$, it follows that $\phi(k) \geqq 0$, with the sign of equality holding only when $\partial p(x)/\partial\theta$ is proportional to $p(x)$ for all $x \geqq k$. In this event, $p(x) = K_\theta e^{g(x)}$, where K_θ is a constant depending on θ. Now, if $g(x)$ is constant, $p(x)$ is a rectangular p. l. On the other hand, if $g(x)$ is not constant, there are two cases which must be considered, namely:

$$\begin{aligned} &\text{a) } p(x) = K_\theta e^{g(x)}, && x \geqq 0, \quad \text{and} \\ &\text{b) } p(x) = p_1(x \mid \theta), && 0 \leqq x < a \leqq k, \\ &\qquad\quad\;\; = K_\theta e^{g(x)}, && x \geqq a. \end{aligned}$$

In the first case, $K_\theta = (\sum_{x=0}^{\infty} e^{g(x)})^{-1}$ and is independent of θ, so that we do not have a case of estimation at all. In the second case, each $p(x)$ for $x \geqq a$ is known *a priori* to within a multiplicative constant depending on θ and, hence, no essential information is lost in truncation. Thus, except in these trivial cases, the relative efficiency is less than unity.

It then appears that, in every case of regular estimation, the variance of an efficient estimate of the parameter of the p. l. $p(x \mid \theta)$ is less than the corresponding variance for the truncated p. l. $f(x \mid \theta)$ and that, as an immediate consequence, the ML estimate in general is capable of greater precision than the χ^2-minimum estimate for fixed N. This is the result mentioned in the first paragraph of section 1. It should be pointed out that the regularity conditions for the Cramér-Rao inequality are stringent enough to give this result. To complete the argument for estimation of a single parameter, form the function

$$\psi(k) = p(k) \sum_{k+1}^{\infty} p(x) \sum_{k}^{\infty} p(x)[\phi(k) - \phi(k+1)], \tag{3.4}$$

where $\phi(k)$ is defined by (3.1). Using (3.1) and (1.1), we may write

$$\phi(k) - \phi(k+1) = \frac{1}{p(k)}\left[\frac{\partial p(k)}{\partial \theta}\right]^2 + \frac{1}{\sum_{k+1}^{\infty} p(x)}\left[\sum_{k+1}^{\infty}\frac{\partial p(x)}{\partial \theta}\right]^2 - \frac{1}{\sum_{k}^{\infty} p(x)}\left[\sum_{k}^{\infty}\frac{\partial p(x)}{\partial \theta}\right]^2. \tag{3.5}$$

Making use of (3.5), straightforward algebraic reduction of (3.4) gives

$$\psi(k) = \left[\frac{\partial p(k)}{\partial \theta}\sum_{k+1}^{\infty} p(x) - p(k)\sum_{k+1}^{\infty}\frac{\partial p(x)}{\partial \theta}\right]^2 \geqq 0, \tag{3.6}$$

the sign of equality holding again only for the p. l.'s discussed after (3.3). Since the first three factors in the right member of (3.4) are positive, it follows that $\phi(k)$ is a strictly decreasing function of k. Thus, the variance of an efficient estimate of the parameter of a truncated p. l., $f(x)$, depends upon the choice of k and decreases in strictly monotone fashion to the variance of the original p. l., $p(x)$, as limit. As a result, the anomalous situation mentioned in the second paragraph of section 1 does *not* arise through irregularity in the behavior of this variance.

4. Poisson and binomial probability laws. The Poisson p. l., $p(x \mid \lambda) = e^{-\lambda}\lambda^x/x!$ gives immediately

$$E\left[\frac{\partial \log p(x)}{\partial \lambda}\right]^2 = \frac{1}{\lambda}, \tag{4.1}$$

whence, from (2.1), we obtain the usual result that the variance of the best unbiased estimate of λ is λ/N. The truncated p. l. has $\partial \log f(x)/\partial\lambda = (x/\lambda) - 1$ for $x \leqq (k-1)$, and $(\partial \log f(k))/\partial\lambda = p(k-1)/\sum_k^\infty p(x)$.
Thus,

$$E\left[\frac{\partial \log f(x)}{\partial \lambda}\right]^2 = \frac{1}{\lambda}\left[\sum_{0}^{k-1} p(x) + (\lambda - k)p(k-1)\right] + \frac{[p(k-1)]^2}{\sum_{k}^{\infty} p(x)}. \tag{4.2}$$

Writing $P(k-1)$ for $\sum_0^{k-1} p(x)$, we obtain finally,

$$\text{Rel. Eff.}_{\text{Poisson}}(k) = P(k-1) + (\lambda - k)p(k-1) + \frac{\lambda[p(k-1)]^2}{1 - P(k-1)}. \tag{4.3}$$

Values of $p(k)$ and $1 - P(k - 1)$ are given directly in Molina's Tables [2] for integer values of k and $\lambda = .001\ (.001)\ .01\ (.01)\ .3(.1)\ 15(1)\ 100$, or may be obtained indirectly from Pearson's Tables [4] of the incomplete Γ-function. In the classical example of a Poisson p. l. quoted by von Bortkiewicz, relating to numbers of deaths due to kicks by horses in Prussian Army Corps, $N = 200$ and the average number of deaths per corps-year is .61. Either χ^2 procedure would take $k = 2$ and $\lambda = .6$, approximately. Using these values, we find that Rel. Eff. $(k = 2 \mid \lambda = .6) = .9508$, i.e., the loss in efficiency incurred by using a χ^2 estimate rather than a ML estimate is of the order of five per cent.

The binomial p. l. is given by $p(x \mid n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$, $x = 0, 1, \cdots, n$, where n is a known parameter and θ is the parameter to be estimated from a sample of N observations. We obtain directly $E[(\partial \log p(x))/\partial\theta]^2 = n/(\theta(1 - \theta))$. Computing a similar quantity for the truncated p. l. and making use of the notations $p(x; n) \equiv \binom{n}{x} \theta^x (1 - \theta)^{n-x}$ and $P(a; n) \equiv \sum_0^a p(x; n)$, we obtain, after some reduction,

$$\begin{aligned} \text{Rel. Eff.}_{\text{binomial}}(k) &= \frac{\theta}{1-\theta}\Bigg[(n - 1)P(k - 3; n - 2) \\ &\quad + \frac{1 - 2n\theta}{\theta} P(k - 2; n - 1) + nP(k - 1; n) \\ &\quad + \frac{n\{P(k - 1; n) - P(k - 2; n - 1)\}^2}{1 - P(k - 1; n)}\Bigg]. \end{aligned} \tag{4.4}$$

The form (4.4) is suitable for computation if tables, such as Pearson's Tables [5], of the incomplete B-function are available covering a range up to the parameter n. If such tables are not available (4.4) is inconvenient since it involves probabilities associated with three different binomial laws. In this case we may use the relations

$$\begin{aligned} P(a; n) - P(a - 1; n - 1) &= (1 - \theta)p(a; n - 1), \\ p(a; n) &= \frac{n\theta}{a} p(a - 1; n - 1) \quad \text{and} \\ p(a; n - 1) &= \frac{(n - a)\theta}{a(1 - \theta)} p(a - 1; n - 1) \end{aligned} \tag{4.5}$$

to obtain the alternative form

$$\begin{aligned} \text{Rel. Eff.}_{\text{binomial}}(k) &= P(k - 1; n - 1) + (n\theta - k)p(k - 1; n - 1) \\ &\quad + \frac{n\theta(1 - \theta)[p(k - 1; n - 1)]^2}{1 - P(k - 1; n - 1) + \theta p(k - 1; n - 1)}, \end{aligned} \tag{4.6}$$

which involves only the one binomial p. l., $p(x \mid n - 1, \theta)$.

As an example, consider the probability situation in which ten independent trials are made, each with the same probability of success, θ. The number of successes in each set of ten trials is one observation. On the basis of N observations, we are to estimate θ. We shall investigate the relative efficiencies when $\theta = .10$. Taking $n = 10$ and $\theta = .10$ in (4.6) we compute the following table of relative efficiencies for different choices of k:

Relative efficiencies of χ^2 estimates in the case of the binomial p. l., $n = 10$, $\theta = .10$

k	Rel. Eff.
2	.8993
3	.9828
4	.9979
5	.9998

It is obvious from the table that the loss in efficiency is not great when $k \geqq 3$ and, hence, the variances of the χ^2 estimates are practically equal to the variance of the ML estimate. But, in ordinary practice, N, the number of sets of ten trials each, would have to be over 140 before k could be safely chosen as large as $k = 3$, and even $k = 2$ requires $N \geqq 38$. Cases in which we seek to estimate parameters on the basis of about 100 observations are not rare; in the present instance, use of a χ^2 estimate would produce about 11% greater variance than the use of a ML estimate.

The two elementary examples considered in this section provide only very fragmentary evidence of the need for caution in employing χ^2-minimum estimates; much numerical work would have to be done to provide any reliable guide to the relative efficiency of such estimates.

5. Estimation of two or more parameters. Consider the p. l. $p(x \mid \theta_1, \theta_2, \cdots, \theta_r)$, $x = 0, 1, 2, \cdots$, with ellipsoid of concentration for a set of joint efficient estimates given by (2.3). The truncated p. l. given by (1.1) has a corresponding ellipsoid of concentration

$$N \sum_{i=1}^{r} \sum_{j=1}^{r} \delta'_{ij} t_i t_j = r + 2, \tag{2.3$'$}$$

with $\delta'_{ij} = E\left[\frac{\partial \log f(x)}{\partial \theta_i} \frac{\partial \log f(x)}{\partial \theta_j}\right]$. We shall show, in this section, that the ellipsoid (2.3) lies wholly within (2.3′); this is so if the left member of (2.3) is uniformly greater than the left member of (2.3′), for every choice of the t_i, $i = 1, 2, \cdots, r$. Accordingly, we form the difference,

$$Q(k) = \sum_{i=1}^{r} \sum_{j=1}^{r} (\delta_{ij} - \delta'_{ij}) t_i t_j. \tag{5.1}$$

Adopting the notations,

$$p_i(x) \equiv \frac{\partial p(x)}{\partial \theta_i} \quad \text{and} \quad f_i(x) \equiv \frac{\partial f(x)}{\partial \theta_i},$$

we obtain by direct subtraction,

$$Q(k) = \sum_{i=1}^{r}\sum_{j=1}^{r}\left[\sum_{x=k}^{\infty}\frac{p_i(x)p_j(x)}{p(x)} - \frac{f_i(k)f_j(k)}{f(k)}\right]t_i t_j. \tag{5.2}$$

Equation (5.2) is unchanged if the right member is written in the form

$$Q(k) = \sum_{i=1}^{r}\sum_{j=1}^{r}\left[\sum_{x=k}^{\infty}\left\{\frac{p_i(x)p_j(x)}{p(x)} - \frac{f_i(k)}{f(k)}p_j(x) - \frac{f_j(k)}{f(k)}p_i(x) + \frac{f_i(k)f_j(k)}{f(k)}\frac{p(x)}{f(k)}\right\}\right]t_i t_j. \tag{5.3}$$

If this latter is now written as

$$Q(k) = \sum_{i=1}^{r}\sum_{j=1}^{r}f(k)\left[\frac{1}{f(k)}\sum_{x=k}^{\infty}\left\{\left(\frac{p_i(x)}{p(x)} - \frac{f_i(k)}{f(k)}\right)\left(\frac{p_j(x)}{p(x)} - \frac{f_j(k)}{f(k)}\right\}p(x)\right]t_i t_j, \tag{5.4}$$

it is evident that the expression in square brackets in the right hand member is precisely the mean value of the expression in curly brackets taken over the set $x \geqq k$. If we denote by $\underset{x\geqq k}{E}\{g(x)\}$ the expected value of $g(x)$ over the set $x \geqq k$, we have

$$Q(k) = \sum_{i=1}^{r}\sum_{j=1}^{r}f(k)\underset{x\geqq k}{E}\left\{\left(\frac{p_i(x)}{p(x)} - \frac{f_i(k)}{f(k)}\right)t_i\left(\frac{p_j(x)}{p(x)} - \frac{f_j(k)}{f(k)}\right)t_j\right\}. \tag{5.5}$$

Finally, since the (finite) sum of the expected values is equal to the expected value of the sum, we have,

$$Q(k) = f(k)\underset{x\geqq k}{E}\left\{\sum_{i=1}^{r}\left[\frac{p_i(x)}{p(x)} - \frac{f_1(k)}{f(k)}\right]t_i\right\}^2. \tag{5.6}$$

Since $f(k) > 0$, $Q(k) \geqq 0$. We need only note that $Q(k) = 0$ only if the linear form in curly brackets in (5.6) is identically zero, i.e., if each coefficient of t_i vanishes. This can happen only in the trivial cases analogous to those described in Section 3.

It has been shown that the ellipsoid of concentration of a set of joint efficient estimates of the parameters of a p. l. lies wholly within the corresponding ellipsoid of the truncated p. l. Therefore, the best procedure for estimating the parameters of a truncated p. l. cannot attain the precision of an efficient procedure for estimating those of the original p. l.

In order to complete the argument for the general case, we form the difference

$$Q(k) - Q(k+1) = \sum_{i=1}^{r}\sum_{j=1}^{r}\left[\frac{p_i(k)p_j(k)}{p(k)} - \frac{f_i(k)f_j(k)}{f(k)} + \frac{f_i(k+1)f_j(k+1)}{f(k+1)}\right]t_i t_j. \tag{5.7}$$

Making use of the two relationships $f(k) = p(k) + f(k+1)$ and $f_i(k) =$

$p_i(k) + f_i(k + 1)$, we have

$$Q(k) - Q(k+1) = \frac{p(k)f(k+1)}{f(k)} \left\{ \sum_{i=1}^{r} \left[\frac{p_i(k)}{p(k)} - \frac{f_i(k+1)}{f(k+1)} \right] t_i \right\}^2. \tag{5.8}$$

The right member of (5.8) being positive except in the trivial cases, it is clear that $Q(k)$ is a strictly monotone function of k.

6. Conclusions. It has been shown that the efficiency of χ^2-minimum estimates, or any other estimates which involve computation in terms of a truncated p. l., is necessarily less than the efficiency of corresponding ML or other estimates based on the original p. l. and, further, that the efficiency increases with the point of truncation. This was established for estimates of a single parameter and, also, for joint estimates of several parameters. Examples given indicate that, in any case of regular estimation, use of χ^2-minimum estimates rather than ML estimates should be accompanied by an investigation into the loss in efficiency.

The author is indebted to Professor J. Neyman, who suggested the problem.

REFERENCES

[1] H. Cramér, "Contributions to the theory of statistical estimation," *Skandinavisk Aktuarietidskrift*, Vol. 29 (1946), pp. 85–94.
[2] E. C. Molina, *Poisson's Exponential Binomial Limit*, Van Nostrand, 1942.
[3] J. Neyman, "Contribution to the theory of the χ^2 test," *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 1949, pp. 239–273.
[4] K. Pearson, *Tables of the Incomplete* Γ*-function*, Cambridge University Press, 1922.
[5] K. Pearson, *Tables of the Incomplete* B*-function*, Cambridge University Press, 1934.
[6] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta Math. Soc.*, Vol. 37 (1945), pp. 81–91.