Maximum Likelihood; An Introduction*

L. Le Cam

Department of Statistics

University of California

Berkeley, California 94720

1 Introduction

One of the most widely used methods of statistical estimation is that of maximum likelihood. Opinions on who was the first to propose the method differ. However Fisher is usually credited with the invention of the name ‘‘maximum likelihood’’, with a major effort intended to spread its use and with the derivation of the optimality properties of the resulting estimates.

Qualms about the general validity of the optimality properties have been expressed occasionally. However as late as 1970 L.J. Savage could imply in his ‘‘Fisher lecture’’ that the difficulties arising in some examples would have rightly been considered ‘‘mathematical caviling’’ by R.A. Fisher.

Of course nobody has been able to prove that maximum likelihood estimates are ‘‘best’’ under all circumstances. The lack of any such proof is not sufficient by itself to invalidate Fisher's claims. It might simply mean that we have not yet translated into mathematics the basic principles which underlied Fisher's intuition.

The present author has, unwittingly, contributed to the confusion by writing two papers which have been interpreted by some as attempts to substantiate Fisher's claims.

* Prepared with the partial support of AROD grant DA AG 29 76 G 0167 and NSF Grant DMS87-01426.
Key Words. Estimation, maximum likelihood, one-step approximations.

To clarify the situation we present a few known facts which should be kept in mind as one proceeds along through the various proofs of consistency, asymptotic normality or asymptotic optimality of maximum likelihood estimates.

The examples given here deal mostly with the case of independent identically distributed observations. They are intended to show that maximum likelihood does possess disquieting features which rule out the possibility of existence of undiscovered underlying principles which could be used to justify it. One of the very gross form of misbehavior can be stated as follows:

Maximum likelihood estimates computed with all the information available may turn out to be inconsistent. Throwing away a substantial part of the information may render them consistent.

The examples show that, in spite of all its presumed virtues, the maximum likelihood procedure cannot be universally recommended. This does not mean that we advocate some other principle instead, although we give a few guidelines in Section 6. For other views see the discussion of the paper by J. Berkson (1980).

This paper is adapted from lectures given at the University of Maryland, College Park, in the Fall of 1975. We are greatly indebted to Professor Grace L. Yang for the invitation to give the lectures and for the permission to reproduce them.

2 A Few Old Examples

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed observations with values in some space $\{\mathcal{X}, \mathcal{A}\}$. Suppose that there is a σ-finite measure λ on $\mathcal{A}$ and that the distribution P_θ of X_j has a density $f(x,\theta)$ with respect to μ. The parameter θ takes its values in some set Θ.

For n observations $x_1, x_2, \ldots, x_n$ the maximum likelihood estimate is any value $\hat{\theta}$ such that

$$\prod_{j=1}^{n} f(x_j, \hat{\theta}) = \sup_{\theta \in \Theta} \prod_{j=1}^{n} f(x_j, \theta).$$

Note that such a $\hat{\theta}$ need not exist, and that, when it does, it usually depends on what version of the densities $f(x,\theta)$ was selected. A function $(x_1, \ldots, x_n) \rightsquigarrow \hat{\theta}(x_1, \ldots, x_n)$ selecting a value $\hat{\theta}$ for each n-tuple $(x_1, \ldots, x_n)$ may or may not be measurable. However all of this is not too depressing. Let us consider some examples.

Example 1. (This may be due to Kiefer and Wolfowitz or to whoever first looked at mixtures of Normal distributions.). Let α be the number $\alpha = 10^{-10^{137}}$. Let $\theta = (\mu,\sigma)$, $\mu \in (-\infty,+\infty)$, $\sigma > 0$. Let $f_1(x,\theta)$ be the density defined with respect to Lebesgue measure λ on the line by

$$f_1(x,\theta) = \frac{1-\alpha}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x-\mu)^2\right\} + \frac{\alpha}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}.$$

Then, for $(x_1, \ldots, x_n)$ one can take $\mu = x_1$ and note that $\sup_{\sigma} \prod_{j=1}^{n} f_1(x_j; \mu,\sigma) = \infty$. If $\sigma = 0$ was allowed one could claim that $\hat{\theta} = (x_1,0)$ is maximum likelihood.

Example 2. The above Example 1 is obviously contaminated and not fit to drink. Now a variable X is called log normal if there are numbers (a,b,c) such that

$$X = c + e^{aY+b}$$

with a Y which is $\mathbf{N}(0,1)$. Let $\theta = (a,b,c)$ in R^3. The density of X can be taken zero for $x \leq c$ and, for $x > c$, equal to

$$f_2(x,\theta) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2a^2}[\log(x-c) - b]^2\right\} \frac{1}{a}\,\frac{1}{x-c}.$$

A sample $(x_1, \ldots, x_n)$ from this density will almost surely have no ties and a unique minimum $z = \min_j x_j$. The only values to consider are those for which $c < z$. Fix a value of b, say $b = 0$. Take a $c \in (z - \frac{1}{2}, z)$ so close to z that $|\log(z-c)| = \max_j |\log(x_j - c)|$. Then the sum of squares in the exponent of the joint density does not exceed $\frac{1}{a^2} n |\log(z-c)|^2$. One can make sure that this does not get too large by taking $a = \sqrt{n}\,|\log(z - c)|$. The extra factor in the density has then a term of the type

$$[\sqrt{n}\,|\log(z - c)|]^{-n} \frac{1}{z - c}$$

which can still be made as large as you please.

If you don't believe my algebra, look at the paper by Bruce Hill, (1963).

Example 3. The preceding example shows that the log normal distribution misbehaves. Everybody knows that taking logarithms is unfair. The following shows that three dimensional parameters are often unfair as well. (The example can be refined to apply to $\theta \in R^2$.)

Let $X = R^3 = \Theta$. Let $\|x\|$ be the usual Euclidean length of x. Take a density

$$f_3(x,\theta) = C \frac{e^{-\|x - \theta\|^2}}{\|x - \theta\|^\beta}$$

with $\beta \in (0,1)$ fixed, say $\beta = \frac{1}{2}$. Here again $\prod_{j=1}^{n} f_3(x_j,\theta)$ will have a supremum equal to $+\infty$. This time it is even attained by taking $\theta = x_1$, or x_2.

One can make the situation a bit worse selecting a dense countable subset $\{a_k\}$, $k = 1,2,\ldots$ in R^3 and taking

$$f_4(x,\theta) = \sum_k C(k) \frac{\exp\{-\|x - \theta - a_k\|^2\}}{\|x - \theta - a_k\|^{1/2}}$$

with suitable coefficients C(k) which decrease rapidly to zero.

Now take again $\alpha = 10^{-10^{137}}$ and take

$$f_5(x,\theta) = \frac{1 - \alpha}{(2\pi)^{3/2}} e^{-\frac{1}{2}\|x - \theta\|^2} + \alpha f_3(x,\theta).$$

If we do take into account the contamination $\alpha f_3(x,\theta)$ the supremum is infinite and attained at each x_i. If we ignore it everything seems fine. but then the maximum likelihood estimate is the mean $\bar{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ which, says C. Stein, is not admissible.

Example 4. The following example shows that, as in examples 2 and 3, one should not shift things. Take independent identically distributed observations $X_1, \ldots, X_n$ from the gamma density shifted to start at ξ so that it is $f(x,\theta) = \beta^{\alpha}\Gamma^{-1}(\alpha)\, e^{-(x-\xi)}(x-\xi)^{\alpha-1}$ for $x \geq \xi$ and zero otherwise. Let β and α take positive values and let ξ be arbitrary real. Here, for arbitrary $0 < \alpha < 1$, and arbitrary $\beta > 0$, one will have $\sup_{\xi} \prod_{j=1}^{n} f(x_j, \theta) = \infty$. One can achieve $+\infty$ by taking $\hat{\xi} = \min X_j$, $\hat{\alpha} \in (0,1)$ and $\hat{\beta}$ arbitrary. The shape of your observed histogram may be trying to tell you that it comes from an $\alpha \geq 10$, but that must be ignored.

Example 5. The previous examples have infinite contaminated inadmissible difficulties. Let us be more practical. Suppose that $X_1, X_2, \ldots, X_n$ are independent uniformly distributed on $[0, \theta]$, $\theta > 0$. Let $Z = \max X_j$. Then $\hat{\theta}_n = Z$ is the m.l.e. It is obviously pretty good. For instance

$$E_{\theta}(\hat{\theta}_n - \theta)^2 = \theta^2 \frac{2}{(n+1)(n+2)}.$$

Except for mathematical caviling, as L.S. Savage says, it is also obviously best for all purposes. So, let us not cavil, but try $\theta_n^* = \frac{n+2}{n+1} Z$. Then

$$E_{\theta}(\theta_n^* - \theta)^2 = \theta^2 \frac{1}{(n+1)^2}.$$

The ratio of the two is

$$\frac{E_{\theta}(\hat{\theta}_n - \theta)^2}{E_{\theta}(\theta_n^* - \theta)^2} = 2\frac{n+1}{n+2}.$$

This must be less than unity. Therefore one must have $2(n+1) \leq n+2$ or equivalently $n \leq 0$.

It is hard to design experiments where the number of observations is strictly negative. Thus our best bet is to design them with $n = 0$ and uphold the faith.

3 A More Disturbing Example

This one is due to Neyman and Scott. Suppose that (X_j, Y_j), $j = 1,2,...$ are all independent random variables with X_j and Y_j both Normal $\mathbf{N}(\xi_j, \sigma^2)$. We wish to estimate σ^2. A natural way to proceed would be to eliminate the nuisances ξ_i and use the differences $Z_j = X_j - Y_j$ which are now $\mathbf{N}(0, 2\sigma^2)$. One could then estimate σ^2 by

$$s_n^2 = \frac{1}{2n} \sum_{j=1}^{n} Z_j^2 .$$

That looks possible, but we may have forgotten about some of the information which is contained in the pairs (X_j, Y_j) but not in their differences Z_j. Certainly a direct application of maximum likelihood principles would be better and much less likely to lose information. So we compute $\hat{\sigma}^2$ by taking suprema over all ξ_j and over σ.

This gives

$$\hat{\sigma}_n^2 = \frac{1}{2} s_n^2 .$$

Now, we did not take logarithms, nothing was contaminated, there was no infinity involved. In fact nothing seems amiss.

So the best estimate must be not the intuitive s_n^2 but $\hat{\sigma}^2 = \frac{1}{2} s_n^2$.

The usual explanation for this discrepancy is that Neyman and Scott had too many parameters. This may be, but how many is too many? When there are too many should one correct the m.l.e. by a factor of two or $\frac{n+2}{n+1}$ as in Example 5, or by taking a square root as in the m.l.e. for star-like distribution? For this latter case, see Barlow *et. al* (1972).

The number of parameters, by itself, does not seem to be that relevant. Take, for instance, i.i.d. observations $X_1 X_2, \ldots, X_n$ on the line with a totally unknown distribution function F. The m.l.e. of F is the empirical cumulative F_n. It is not that bad. Yet, a crude evaluation shows that F depends on very many parameters indeed, perhaps even more than Barlow *et. al* had for their star-like distributions.

Note that in the above examples we did not let n tend to infinity. It would not have helped, but now let us consider some examples where the misbehavior will be described as $n \to \infty$.

4 An Example of Bahadur

The following is a slight modification of an example given by Bahadur in 1958. The modification does not have the purity of the original but it is more transparent and the purity can be recovered.

Take a function, say h, defined on $(0,1]$. Assume that h is decreasing, that $h(x) \geq 1$ for all $x \in (0,1]$ and that $\int_0^1 h(x)dx = \infty$. Select a number c, $c \in (0,1)$ and proceed as follows. One probability measure, say p_0, on $[0,1]$ is the Lebesgue measure λ itself. Define a number a_1 by the property

$$\int_{a_1}^{1} [h(x) - c]dx = 1 - c.$$

Take for p_1 the measure whose density with respect to λ is c for $0 \leq x \leq a_1$ and $h(x)$ for $a_1 < x \leq 1$.

If $a_1, a_2, \ldots, a_{k-1}$ have been determined define a_k by the relation

$$\int_{a_k}^{a_{k-1}} [h(x) - c]dx = 1 - c$$

and take for p_k the measure whose density with respect to λ on $[0,1]$ is c for $x \notin (a_k, a_{k-1}]$ and $h(x)$ for $x \in (a_k, a_{k-1}]$.

Since $\int_0^1 h(x)dx = \infty$ the process can be continued indefinitely, giving a countable family of measures p_k, $k = 1,2,\ldots$ Note that any two of them, say p_j and p_k with $j < k$, are mutually absolutely continuous.

If $x_1, x_2, \ldots, x_n$ are n observations taken on $[0,1]$ the corresponding logarithm of likelihood ratio is given by the expression:

$$\Lambda_j^k = \log \prod_{i=1}^{n} \frac{dp_k(x_i)}{dp_j(x_i)}$$

$$= \sum_i^{(k)} \log \frac{h(x_i)}{c} - \sum_i^{(j)} \log \frac{h(x_i)}{c}$$

where the first sum $\Sigma^{(k)}$ is for $x_i \in (a_k, a_{k-1}]$ and the second is for $x_i \in (a_j, a_{j-1}]$.

Now assume that the $X_1, \ldots, X_n$ are actually i.i.d. from some distribution p_{j_0}. They have a minimum $Z_n = \min_i X_i$. With probability unity this will fall in some interval $(a_{k_n}, a_{k_n-1}]$ with $k_n = k_n(Z_n)$. Fix a value j and consider $\frac{1}{n}\Lambda_j^{k_n}$. This is at least equal to

$$\frac{1}{n} \log \frac{h(Z_n)}{c} - \frac{1}{n} \nu_{j,n} \log \frac{h(a_j)}{c}$$

where $\nu_{j,n}$ is the number of X_j's which fall in $(a_j, a_{j-1}]$.

According to the strong law of large numbers $\frac{1}{n}\nu_{j,n}$ converges to some constant $p_{j_0,j} \le 1$. Also, j_0 being fixed, Z_n tends almost surely to zero. In fact if $y < a_{j_0}$ one can write

$$p_{j_0}\{Z_n > y\} = (1 - cy)^n \le e^{-ncy}.$$

Thus, as long as $\sum_n e^{-ncy_n} < \infty$ it will be almost certain that eventually $Z_n < y_n$. In particular Z_n may have a limiting distribution but nZ_n^2 almost certainly tends to zero.

This being the case take $c = 9/10$ and $h = \exp\{\frac{1}{x^2}\}$. Then $\frac{1}{n} \log h(Z_n) = (nZ_n^2)^{-1}$ tends to infinity almost surely.

Thus if we take any finite set $J = (1, 2, \ldots, j_1)$, for any fixed j_0 there will almost surely be an integer N such that $\hat{\theta}_N$ cease to be in J from N on.

It might be thought that such a disgraceful behavior is due to the vagaries of measure theory. Indeed the variables X_j used here are continuous variables and everybody knows that such things do not really exist.

However, replace the measures p_k used above by measures q_k whose density on (a_k, a_{k-1}) is constant and equal to $[a_{k-1} - a_k]^{-1} \int_{a_k}^{a_{k-1}} h(x)dx$. Then, there is no need to record the exact values of the observations X_j. It is quite enough to record in which interval $(a_k, a_{k-1}]$ they fall. The parameter θ is itself integer valued. However the same misbehavior of m.l.e. will still occur. (This is essentially equivalent to Bahadur's first construction.)

In the present construction the parameter θ is integer valued. It is easy to modify the example to obtain one in which θ takes values, say, in $(1, \infty)$ and in which the observable variables have densities $f(x, \theta)$ which are infinitely differentiable functions of θ. For this purpose define p_k as above. Let u be a function defined on $(-\infty, +\infty)$ constructed so that $u(x) = 0$ for $x \leq 0$, and $u(x) = 1$ for $x \geq 1$. One can find functions u of that kind which are strictly increasing on $(0, 1)$ and are infinitely differentiable on $(-\infty, +\infty)$.

Now let $p_\theta = p_k$ if θ is equal to the integer k. If $\theta \in (k, k+1)$ let $p_\theta = [1 - u(\theta - k)]p_k + u(\theta - k)p_{k+1}$.

Taking for each p_k the densities f_k used previously, we obtain similarly densities

$$f(x, \theta) = [1 - u(\theta - k)]f_k(x) + u(\theta - k) f_{k+1}(x).$$

The function u can be constructed, for instance, by taking a multiple of the indefinite integral of the function

$$\exp\left\{-[\frac{1}{t} + \frac{1}{1-t}]\right\}$$

for $t \in [0, 1)$ and zero otherwise. If so $f(x, \theta)$ is certainly infinitely differentiable in θ. Also the integral $\int f(x, \theta) dx$ can be differentiated infinitely under the integral sign. There is a slight annoyance that at all integer values of θ all the derivatives vanish. To cure this take $\alpha = 10^{-10^{137}}$ and let

$$g(x, \theta) = \frac{1}{2}[f(x, \theta) + f(x, \theta + \alpha e^{-\theta^4})].$$

Then, certainly, everything is under control and the famous conditions in Cramér's text are all duly satisfied. Furthermore, $\theta \neq \theta'$ implies $\int | g(x,\theta) \ - \ g(x,\theta') | dx > 0$.

In spite of all this, whatever may be the true value θ_0, the maximum likelihood estimate still tends almost surely to infinity.

Let us return to the initial example with measures p_k, $k = 1,2,...$, and let us waste some information. Having observed $X_1, \ldots, X_n$, according to one of the p_k take independent identically distributed $\mathbf{N}(0,10^6)$ variables $Y_1, \ldots, Y_n$ and consider $V_j = X_j + Y_j$ for $j = 1,2, \ldots, n$.

Certainly one who observes V_j, $j = 1, \ldots, n$ instead of X_j, $i = 1, \ldots, n$ must be at a gross disadvantage!

Maximum likelihood estimates do not really think so.

The densities of the new variables V_j are functions, say ψ_k, defined, positive analytic, etc. on the whole line $\mathbb{R} = (-\infty, +\infty)$. They still are all different. In other words $\int | \psi_k(x) - \psi_j(x) | dx > 0$ if $k \neq j$.

Compute the maximum likelihood estimate $\hat{\theta}_n = \hat{\theta}_n(v_1, \ldots, v_n)$ for these new observations. We claim that

$$p_j[\hat{\theta}_n(V_1, \ldots, V_n) \ = \ j] \ \rightarrow \ 1$$

as $n \rightarrow \infty$.

To prove this let $\sigma = 10^3$ and note that $\psi_j(v)$ is a moderately small distortion of the function

$$\Psi_j(v) \ = \ c\int_0^1 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v-\xi)^2}{2\sigma^2}} d\xi \ + \ (1-c)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v-a_j)^2}{2\sigma^2}} .$$

Furthermore, as $m \rightarrow \infty$ the function $\psi_m(v)$ converges pointwise to

$$\psi_\infty(v) \ = \ c\int_0^1 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v-\xi)^2}{2\sigma^2}} d\xi \ + \ (1-c)\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{v^2}{2\sigma^2}} .$$

Thus, we can compactify the set $\theta = \{1,2,...\}$ by addition of a point at infinity with $\psi_\infty(v)$ as described above.

We now have a family $\{\psi_\theta ; \theta \in \overline{\Theta}\}$ such that $\psi_\theta(v)$ is continuous in θ for each v. Also $\sup_{k \geq m} [\log \frac{\psi_k(v)}{\psi_j(v)}]^+$ does not exceed $\frac{10^6}{2} |(v-1)^2 - v^2|$. Since this is certainly integrable, the theorem due to Wald (1949) is applicable and $\hat{\theta}$ is consistent.

So throwing away quite a bit of information made the m.l.e. consistent. Here we wasted information by fudging the observations. Another way would to enlarge the parameter space and introduce irrelevant other measures p_θ.

For this purpose consider our original variables X_j, but record only in which interval $(a_k, a_{k-1}]$ the variable X_j falls. We obtain then discrete variables, say Y_j such that $P_i[Y_j = k]$ is the integral $q_i(k)$ of $p_i(x)$ on $(a_k, a_{k-1}]$. Now, the set Θ of all possible discrete measures on the integers k = 1,2,... can be metrized, for instance by the metric $\| Q_s - Q_r \| = \sum_k | q_s(k) - q_r(k) |$. For this metric the space is a complete separable space.

Given discrete observations Y_j, j = 1, . . . , n we can compute a maximum likelihood estimate, say θ_n^* in this whole space Θ. The value of θ_n^* is that element θ_n^* of Θ which assigns to the integer k a probability $\theta_n^*(k)$ equal to the frequency of k in the sample. Now, if θ is any element whatsoever of Θ, for every $\varepsilon > 0$, $P_\theta\{\| \theta - \theta_n^* \| > \varepsilon\}$ tends to zero as $n \rightarrow \infty$. More precisely $\theta_n^* \rightarrow \theta$ almost surely.

The family we are interested in, the q_i, i = 1,2,..., constructed above form a certain subset, say Θ_0, of Θ. It is a nice closed (even discrete) subset of Θ.

Suppose that we do know that $\theta \in \Theta_0$. Then, certainly, one should not waste that information. However if we insist on taking a $\hat{\theta}_n \in \Theta_0$ that maximizes the likelihood there, then $\hat{\theta}_n$ will almost never tend to θ. If on the contrary we maximize the likelihood over the entire space of all probability measures on the integers, we get an estimate θ_n^* that is consistent.

It is true that this is not the answer to the problem of estimating a θ that lies in Θ_0. May be that is too hard a problem? Let us try to select a point $\overline{\theta}_n \in \Theta_0$ closest to θ_n^*. If there is no such closest point just take $\overline{\theta}_n$ such that $\| \theta_n^* - \overline{\theta}_n \| \leq$

$2^{-n} + \inf\{\|\theta_n^* - \theta\|; \theta \in \Theta_0\}$. Then $P_\theta\{\overline{\theta}_n = \theta$ for all sufficiently large $n\} = 1$. So the problem cannot be too terribly hard. In addition Doob (1948) says that if we place on Θ_0 a prior measure that charges every point, the corresponding Bayes estimate will behave in the same manner as our $\overline{\theta}_n$.

As explained this example is imitated from one given by Bahadur (1958). Another example imitated from Bahadur and from the mixture of example 1 has been given by T.S. Ferguson (1982). Ferguson takes $\Theta = [0,1]$ and considers i.i.d. variables taking values in $[-1,+1]$. The densities, with respect to Lebesgue measure on $[-1,-1]$, are of the form

$$f(x,\theta) = \frac{\theta}{2} + \frac{1-\theta}{\delta(\theta)}\left[\frac{1-|x-\theta|}{\delta(\theta)}\right]^+$$

where δ is a continuous function that decreases from 1 to 0 on $[0,1]$. If it tends to zero rapidly enough as $\theta \to 1$, the peaks of the triangles will distract the m.l.e. from its appointed rounds. In example 1, section 2, the m.l.e. led a precarious existence. Here everything is compact and continuous and all of Wald's conditions, except one, are satisfied. To convert the example into one that satisfies Cramér's conditions, for $\theta \in (0,1)$, Ferguson replaces the triangles by Beta densities.

The above example relies heavily on the fact that ratios of the type $f(x,\theta)/f(x,\theta_0)$ are unbounded functions of θ. One can also make up examples where the ratios stay bounded and m.l.e. still misbehaves.

A possible example is as follows. For each integer $m > 1$ divide the interval $(0,1]$ by binary division, getting 2^m intervals of the form $(j2^{-m},(j+1)2^{-m}]$, $j = 0,1,\ldots,2^m - 1$.

For each such division there are $\begin{bmatrix} 2^m \\ 2^{m-1} \end{bmatrix}$ ways of selecting 2^{m-1} of the intervals. Make a selection s. On the selected ones, let $\phi_{s,m}$ be equal to 1. On the remaining ones let $\phi_{s,m}$ be equal to (-1).

This gives a certain countable family of functions.

Now for given m and for the selection s let $p_{s,m}$ be the measure whose density with respect to Lebesgue measure on $(0,1]$ is

$$1 + \frac{1}{2}(1 - e^{-m})\phi_{s,m}.$$

In this case the ratio of densities is always between $\frac{1}{2}$ and 2. The measures are all distinct from one another.

Application of a maximum likelihood technique would lead us to estimate m by $+\infty$.

(This is essentially equivalent to another example of Bahadur.)

5 An Example from Biostatistics

The following is intended to show that even for ''straight'' exponential families one can sometimes do better than the m.l.e.

The example has a long history, which we shall not recount. It occurs from the evaluation of dose responses in biostatistics.

Suppose that a chemical can be injected to rats at various doses $y_1, y_2, \ldots, y_i > 0$. For a particular dose, one just observes whether or not there is a response. There is then for each y a certain probability of response. Biostatisticians, being complicated people, prefer to work out not with the dose y but with its logarithm $x = \log y$.

We shall then let $p(x)$ be the probability of response if the animal is given the log dose x.

Some people, including Sir Ronald, felt that the relation $x \rightarrow p(x)$ would be well described by a cumulative normal distribution, in standard form $p(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$. I do not know why. Some other people felt that the probability p has a derivative p' about proportional to p except that for p close to unity (large dose) the poor animal is saturated so that the curve has a ceiling at 1.

Thus, somebody, perhaps Raymond Pearl, following Verhulst, proposed the ‘‘logistic’’ $p'(x) = p(x)[1 - p(x)]$ whose solution is

$$p(x) = \frac{1}{1 + e^{-x}}$$

give or take a few constants.

Therefore, we shall assume that $p(x)$ has the form

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

with two constants α and β, $\beta > 0$.

Since we are not particularly interested in the actual animals, we shall consider only the case where β is known, say $\beta = 1$ so that α is the only parameter and

$$p(x) = \frac{1}{1 + e^{-(\alpha + x)}}.$$

Now we select a few log doses $x_1, x_2, \ldots, x_m$. At x_j we inject n_j animals and count the number of responses r_j. We want to estimate α.

For reasons which are not biostatistical but historical (or more precisely routine of thought) it is decided that the estimate $\hat{\alpha}$ should be such that

$$R(\hat{\alpha}, \alpha) = E_\alpha(\hat{\alpha} - \alpha)^2$$

be as small as possible.

A while back, Cramér and Rao said that, for unbiased estimates, $R(\hat{\alpha}, \alpha)$ cannot be smaller than $1/I(\alpha)$ where $I(\alpha)$ is the Fisher information

$$I(\alpha) = \Sigma n_j p(x_j)[1 - p(x_j)].$$

So, to take into account the fact that some positions of α are better than some others we shall use instead of R the ratio

$$F = I(\alpha) E_\alpha(\hat{\alpha} - \alpha)^2.$$

The joint density of the observations is easy to write. It is just

$$\prod_j \binom{n_j}{r_j} [p(x_j)]^{r_j} [1 - p(x_j)]^{n_j - r_j} = \prod_j \binom{n_j}{r_j} [\frac{1 - p(x_j)}{p(x_j)}]^{s_j} [p(x_j)]^{n_j}$$

where s_j is the number of non respondents at log dose x_j.

Now

$$\frac{1 - p(x_j)}{p(x_j)} = e^{-(\alpha + x_j)}$$

so that the business term of the above is just

$$\prod_j e^{-(\alpha + x_j)s_j}$$

in which one recognizes immediately a standard one-dimensional, linearly indexed exponential family, with sufficient statistic Σs_j.

The first thing to try is of course the best estimate of all, namely m.l.e. That leads to a nice equation

$$\Sigma\, s_j = \Sigma\, n_j [\, 1 - p(x_j)\,]$$

which unfortunately is hard to solve.

So, somebody decided, let us try the next best, namely the minimum χ^2. This leads to minimize

$$\Sigma \frac{n_j [\, f_j - p(x_j)\,]^2}{p_j(1 - p_j)}$$

with $f_j = \frac{r_j}{n_j}$. It is even a worse affair. However Joe Berkson had the bright idea of taking the log of $\frac{1-p}{p}$ and noticing that it is *linear in* α which leads to the question why not just apply least squares and minimize $\Sigma\,[\log \frac{1 - f_j}{f_j} - (\alpha + x_j)\,]^2$?

Well, as Gauss said, that will not do. One should divide the square terms by their variances to get a good estimate. The variance of $\log \frac{1 - f_j}{f_j}$? Oops! It is infinite. Too bad, let us approximate. After all if ϕ is differentiable, then $\phi(f) - \phi(p)$ is about $(f - p)\phi'(p)$, so its variance is almost $(\phi'(p))^2 \mathrm{Var}(f - p)$ give or take a mile and a couple of horses tails. If $\phi(f)$ is $\log \frac{1-f}{f}$, that gives $\phi'(p) = -[\,p(1 - p)\,]^{-1}$. Finally, we would want to minimize

$$\Sigma n_j p(x_j)[1 - p(x_j)] \left\{ \log [\frac{1 - f_j}{f_j}] - (\alpha + x_j) \right\}^2 .$$

Not pleasant! All right, let us replace the coefficients $p(x_j)[1 - p(x_j)]$ by estimates $f_j(1 - f_j)$.

Now we have to minimize

$$\Sigma n_j f_j (1 - f_j) [\log \left[\frac{1 - f_j}{f_j} \right] - (\alpha + x_j)]^2 ,$$

a very easy matter.

After all these approximations nobody but a true believer would expect that the estimate so obtained would be any good, but there are people ornery enough to try it anyway. Berkson was one of them. he found to his dismay, that the estimate had, at last at one place, a risk ratio $F = I(\alpha) E_\alpha (\hat{\alpha} - \alpha)^2$ strictly less than unity. Furthermore, that was a point where the estimate was in fact unbiased! So Joe was ready to yell ‘‘down with Cramér-Rao!’’ when Neyman pointed out that the derivative of the bias was not zero, and that Fréchet before Cramér and Rao had written an inequality which involves the derivative of the bias.

To make a long story short, they looked at the estimate. Then Joe Berkson and Joe Hodges Jr. noticed that one could Rao-Blackwellize it. Also these two authors tried to find a minimax estimator. They found one which for most purposes is very close to being minimax.

Their work is reported in 4th Berkeley Symposium, volume IV. Numerical computations show that for certain log doses ($x_1 = -\log \frac{7}{3}$, $x_2 = 0$, $x_3 = \log \frac{7}{3}$) with 10 rats at each dose, the minimum logit estimate is definitely better than m.l.e.

Some numbers are as follows; the label P_2 means $p(x_2) = (1 + e^{-\alpha})^{-1}$. The entry is $F(\alpha)$. The minimum logit estimate Rao-Blackwellized is B and H is the Berkson-Hodges near minimax estimate.

(12

	0.5	0.7	0.85
m.l.e.	1.0575	1.1034	1.2740
min logit	.9152	.9589	.9109
B	.8778	.8935	.8393
H	.8497	.8502	.8465

Now conventional wisdom has it that, for any reasonable estimate, the ratio called $F(\alpha)$ above should have the form

$$F^*(\alpha) = 1 + \frac{1}{n}(A + B + C) + \varepsilon$$

where ε is negligible (give or take a couple of horses hairs).

The quantities A, B and C are all positive numbers.

One of them, say A, is a Bhattacharya curvature term which depends on the parametrization but not the estimate. The second, say B, is the Efron curvature term, which depends neither on the estimate, nor on the parametrization.

Finally C depends on many things but it is zero for the m.l.e. From this one concludes that m.l.e. is best, or if one wants in the preceding table that $127 \le 85$.

Something must be amiss somewhere. One possibility is that there are too many horses hairs in ε.

This might be so here. It might be so also in some other cases, such as the ''dilution series'' studied by T. Ferguson and perhaps also in the case that you may want to study.

However, there is also another reason. In the derivation of expansions such as the F^* given above, most authors first *correct the estimates* to make them almost unbiased or to make them have about the same bias as the local m.l.e., depending on circumstances.

Why would one want to do that? Well, for one thing, the bias would introduce in F^* terms which are not of order $1/n$ but of order unity and can be positive or *negative*.

They would overwhelm the (C/n). We cannot allow any such thing, of course. This would send us back to the beginning and point out that we do not have a standard procedure for controlling the first order term.

So, never mind what the first order terms do, m.l.e. will control the second order terms for you, if it happens to be in the neighborhood of the true value.

This author has heard rumors to the effect that measuring the risk by expected square deviations is foolish. Of course it is. In fact in the present case I forgot to say that the m.l.e. can take infinite values with positive probability.

Berkson and Hodges had replaced those infinite values by 3.37 in all the computations.

To end on a more cheerful note, let us remark that the minimum logit is also pretty badly behaved at times.

Bill Taylor points out that if one takes a fixed number m of log doses $x_1, x_2, \ldots, x_m$ and let the number n_j at each grow, like say $c_j N$, $N = \sum_j n_j$, c_j constant, then both m.l.e. and min logit are consistent.

However, if one scatters the rats between log doses x_j, $a < x_j < b$, with numbers of rats at each dose bounded, the m.l.e. stays consistent while min logit does not. So there!

6 ‘‘What Should I Do?’’

If the hallowed maximum likelihood principle leads us to difficulties, maybe some other principle will save us.

There is indeed such a principle. It is as follows:

Basic Principle 0. Don't trust any principle.

This applies in particular to the principles and recommendations listed below and should be kept in mind any time one encounters a problem worth studying. Anyway, here are some other principles.

Principle 1. Have clear in your mind what it is that you want to estimate.

Principle 2. Try to ascertain in some way what precision you need (or can get) and what you are going to do with the estimate when you get it.

Principle 3. Before venturing an estimate, check that the rationale which led you to it is compatible with the data you have.

Principle 4. If satisfied that everything is in order, try first a crude but reliable procedure to locate the general area in which your parameters lie.

Principle 5. Having localized yourself by (4), refine the estimate using some of your theoretical assumptions, being careful all the while not to undo what you did in (4).

Principle 6. Never trust an estimate which is thrown out of whack if you suppress a single observation.

Principle 7. If you need to use asymptotic arguments, don't forget to let your number of observations tend to infinity.

Principle 8. J. Bertrand said it this way: ''Give me four parameters and I shall describe an elephant; with five, it will wave its trunk.''

Counting the Basic Principle, this makes a total of nine. Confucius had many more, but I tire easily.

Incidentally Principle 6 sounds like something Hampel would say. However I learned it in 1946 from E. Halphen.

Now I will try to illustrate the principles by telling a story.

It is a fictitious story, but it could have happened to me (and almost did). However to protect everybody, I shall pretend that it happened to my friend Paul. I have also changed some of the formulas and taken some other liberties.

Paul was going to take n independent observations X_i from a certain distribution, reported to be a gamma distribution with density

$$f(x, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha - 1}, \quad x > 0.$$

As we shall see later, the reasons for the choice of this formula were almost as good as any one encounters in practical settings.

It is true that in certain cases one does have better reasons. For instance physicists are fairly certain that waiting times to disintegration for deuterium are exponentially distributed. But even there, the precision of the measurements, or some unknown feature of the measurement process may make the actual observations deviate from exponential. See the article by J. Berkson on the subject.

Paul's situation was somewhat different and more typical. In his case, as in many other ones, α and β are there because something has to be there, but they are not the quantities of direct interest. One might be contemplating instead the possibility of estimating a median, an expectation, an interquartile range, the point t such that $P[X > t] = 10^{-3}$ or the point s such that $P[X < s] = 10^{-2}$ or even the average of another independent sample of 21 observations, discounting the largest and smallest. Principle 1 says that if Paul wants to estimate t, he should not try his best to estimate s instead.

Paul wanted to estimate t because his boss told him to do so. In fact the boss is an engineer who has to decide on the size of the spillway needed for a certain dam. Paul will live some 50 years after the dam is completed. If the spillway is too small and a mighty flood occurs, the dam will overturn and Paul will lose his pension and his fishing rights. However if Paul's estimate is too large and out of line with usual practices, his boss will tell him to go soak his head and stop wasting taxpayers' money.

Then the boss will take Paul's estimate and multiply it by a safety factor of two no matter what.

With all this information I shall let you apply Principle 2. Remember that if Paul knew t exactly and used that for the design, his probability of losing his pension would be about 5%.

The observations X_i that Paul can use are the yearly peak floods for the past 26 years. They were measured very accurately as things go. For the large ones the precision is reported to be 20% or so. However the year 2000 was very very dry. Even the peak flood ran mostly under the gravel of the river bed. It is admitted that the measurements that year could be off by a factor of 5 or maybe 7 either way.

Furthermore the peak flood of 2001 occurred early and most of the rain replenished the ground water and did not reach the river. Paul applies Principle 3 and decides that (i) the X_i are independent, (ii) there is not much trend in them even though silt deposited in the river bed and that was not taken into account in the measurements, and (iii) the X_i are indeed gamma distributed.

Paul is absolutely certain about (iii) even though he carried out a Kolmogorov-Smirnov test which gave him a distance

$$\sqrt{n} \sup_x | F_n(x) - F(x, \hat{\alpha}, \hat{\beta}) | = 1.6.$$

He has to be certain because his boss spent the best part of his life proving that floods are indeed gamma distributed and independent from year to year. He has published numerous papers and a book on the subject.

Paul is satisfied and applies Principle 4. He uses auxiliary estimates, getting a confidence interval for the median from suitable order statistics. He does the same for the interquartile range and ventures a guess for t.

The boss says ‘‘That won’t do. Fisher has shown that maximum likelihood is best. Where have you learned your statistics?’’

Paul agrees reluctantly. The m.l.e. depend on $\sum_i \log X_i$ and $\sum_i X_i$. They are easily computable. However they do not even fall in the range indicated by the previous computations on medians and interquartile ranges. This is partly due to that very very dry year of 2000 which may be off by a factor of seven.

Paul tries to explain Principles 5 and 6 to his boss. The boss refuses to listen. Paul is miffed and decides to see what would happen if he used the formulas proposed by some of his boss' competitors and enemies.

One of the competitors, a Dr. G., has a perfectly rigorous demonstration that the formula to use should not be gamma but such that

$$P[X \leq x] = \exp\{-\exp(\frac{x-a}{b})\}$$

for some numbers a and b. Another competitor, Dr. S., has proved long ago that the X_i are log normal or at least that there are always constants a, b, c and σ such that

$$X_i = [e^{\sigma Z} + a][be^{\sigma Z_i} + c]^{-1}$$

with Z_i Gaussian, $\mathbf{N}(0,1)$. The log normal case is in the particular case where $b = 0$

Still another competitor Dr. F. says that the procedure to use is to plot the data cumulatively on log-probit paper, fit a polynomial and extrapolate it.

Dr. F's rationale is that you cannot assume a particular analytic form for the distribution, but you can mimic anything you please by polynomials of high enough order. See the recent paper by L. Breiman and C.J. Stone (1985).

To check what would happen, Paul tries a Monte Carlo experiment using Dr. F's method with polynomials of increasingly higher degree. He finds that linear or quadratic is not too bad, but that for higher degrees the prediction so obtained is totally unreliable. That reminds him of Bertrand's elephant and makes him suspicious of Dr. S' general formula also.

Indeed he quickly realizes that he can fix the median of X_i and a point, say x_0 such that $P[X \geq x_0] = 10^{-1}$ and still vary the point t such that $P[X \geq t] = 10^{-3}$ by factors of 2 or 3.

In fact many of the curves he can obtain readily fit his observations, if one can judge by χ^2 or Kolmogorov-Smirnov. However they do lead to vastly different estimates of t.

Even if he takes the simplest log normal $X_i = e^{\mu+\sigma Z_i}$, he finds that this leads to estimates which are about 1.2 times those obtained from gamma and about 1.5 times those obtained from Dr. G's formula. What Paul finally did, I will not tell you, except that he asked to have a few years to study the problem.

However to obtain good estimates in Dr. S' formulas was not too easy. Paul had to have recourse to asymptotic arguments. We have already seen in Section 2 that the maximum likelihood technique does not work well in the case where b = 0. It is even worse if b can be positive.

All is not lost however. One can often obtain fairly easily estimates which are asymptotically best in some sense. I will now describe one possible procedure for the case of the three parameter log normal, that is Dr. S' formula with b = 0.

It is a technique that relies on the fact that near the true value of the parameter the log likelihood $\sum_i \log f(x_1 ; \theta)$ are often approximable by expression that are linear-quadratic in θ. This is true for instance in examples 1, 2, 3 of section 2. It is also true for example 4 if it happens that $\alpha > 2$. It is of course not true for example 5 and suggests terrible things for the Neyman Scott example. The idea is to localize the problem (Principle 4) and then fit a linear-quadratic expression to the log likelihood. One then treat that fitted expression as if it was a Gaussian log likelihood. Sometimes one needs to take a few extra precautions as in example 1 or example 3. There, one needs to avoid working around values where the likelihood function misbehaves. Officially and asymptotically this can be achieved by suitable discretization of auxiliary estimates. For practical purposes Principle 0 must be duly honored and one must check that everything is under control.

To return to the log normal, let m be the median of X and let q_1 and q_2 be the quartiles such that $P[X < q_1] = P[X > q_2] = \frac{1}{4}$.

If Z_i is still $\mathbf{N}(0,1)$ one can rewrite the log normal expression in the form.

$$X_i = m + (q_2 - m)\left(\frac{V^{cZ_i} - 1}{V - 1}\right)$$

where $V = \dfrac{q_2 - m}{m - q_1}$ and where c is a constant $c = (.6745)^{-1}$.

The parameters m, q_1 and q_2 can be estimated by their empirical counterparts, say $\overline{m}$, $\overline{q}_1$, $\overline{q}_2$. The point t that Paul wanted to estimate could be estimated by substituting these estimates in $t = m + (q_2 - m)\left(\dfrac{V^k - 1}{V - 1}\right)$ with k = 4.5815.

These empirical estimates $\overline{m}$, $\overline{q}_q$, $\overline{q}_2$ may not be the best obtainable here, at least if one assumes that the X_i are indeed log normal. A possible improvement procedure (Principle 5) is as follows. Let $\theta = (m, q_1, q_2)$ for short. For two values θ_0 and θ_1 let $\Lambda_n(\theta_1, \theta_0)$ be the logarithm of likelihood ratio

$$\Lambda_n(\theta_1, \theta_0) \ = \ \Sigma \log \frac{f(X_i, \theta_1)}{f(X_i, \theta_0)}.$$

Let $\overline{\theta} = (\overline{m}, \overline{q}_1, \overline{q}_2)$ and compute $\Lambda_n(\overline{\theta}_1, \overline{\theta})$ for values $\overline{\theta}_1$ of the form $\overline{\theta}_1 = \overline{\theta} + (u_j + u_k)/\sqrt{n}$, j = 0,1,2,3, k = 0,1,2,3, where $u_0 = 0$ and the u_j, j = 1,2,3 form a basis in three dimensional space. For instance one may take $u_{1`=}$ (1,0,0), u_2 = (0,1,0) and u_3 = (0,0,1).

The next step is to fit a quadratic to the values of Λ_n so obtained. One then takes for improved estimate the point T_n which maximizes the quadratic.

This sounds very much like trying to get the m.l.e. by approximation. However it is not the same thing as looking for the m.l.e. We have already seen that in the present case this does not work.

On the contrary one can show that for the log normal case the estimate T_n is asymptotically normal, asymptotically sufficient, etc.

Here we have used a particular auxiliary estimate $\overline{\theta}$, but the choice of preliminary estimate matters relatively little asymptotically, as long as $\{\mathbf{L}[\sqrt{n}(\overline{\theta}_n - \theta) | \theta]$ is a relatively compact sequence.

If the influence of the choice of $\overline{\theta}$ is deemed pernicious, one can repeat the above improvement procedure using T_n instead of $\overline{\theta}_n$. This will give a second estimate T_n'. The influence of $\overline{\theta}_n$ is now (relatively little)^2.

However one should be careful. Iterating the procedure might conceivably lead toward the m.l.e. and that must be avoided in all but special cases. What one might want to do is try the procedure several times with different bases to see whether it makes any substantial difference and to check on the quality of the quadratic approximation to the log likelihood. One might also trace the contours of the likelihood function to see how they look. See for instance J. Hodges (1987).

In any event I recommend the general procedure enthusiastically and without any reservations for all those cases where it does work.

In spite of the fact that the procedure does work, at least at times, some people are never satisfied and they have complained. The procedure is too arbitrary. It entails the choice of an auxiliary estimate, of bases and even something more dangerous: The choice of a parametrization. Indeed we have used a local approximation of the log likelihood by a quadratic expression. Quadratic in what variables? The problem goes away ''as n tends to infinity'' but it does not for the particular n you have. For instance, in a binomial situation should one parametrize by the binomial p? By $\arcsin\sqrt{p}$? By $\log\frac{p}{1-p}$? Who says ''quadratic'' says Euclidean or Hilbert space. There are indications that the ''best'' parametrization is the one where the usual Euclidean distance between parameter values u and v is close to $\{-\log\int [dP_u dP_v]^{1/2}\}^{1/2}$. See for instance E. Mammen (1988). However that work deals only with the case where the matrix in the quadratic is nearly non-random. Since the estimation method works even in cases where the matrix in the quadratic is genuinely random, something remains to be studied. I have not done all my homework. In spite of this, I will still recommend the procedure whenever it works.

I shall not insist on all of this. It is more important to return to Principle 5. Presumably the passage from $\bar{\theta}$ to T_n was supposed to be an improvement. However in the computation of the Λ_n used above one will encounter terms which look like $\Sigma \log(X_i - c)$ or $\Sigma[\log(X_i - c)]^2$. It is true that they find themselves multiplied by factors of the order $(1/\sqrt{n})$. However this is not enough to prevent them from doing some mischief.

Principle 5 says that T_n is supposed to improve on $\overline{\theta}$, not undo it.

Now one can readily find confidence intervals for (m, q_1, q_2) using order statistics. For instance for m one can use order statistics with ranks about $(\frac{n}{2} \pm k\sqrt{n})$, with a suitable k.

What is one supposed to do if the improved estimate $T_n = (\hat{m}, \hat{q}_1, \hat{q}_2)$ is such that $\hat{m}$ is not in the confidence interval so obtained? Or if plotting the cumulative at the estimated values T_n one finds it outside of the Kolmogorov-Smirnov band?

There are at present no standard procedures to handle this. One *ad hoc* procedure is to eliminate from the computation all the $|\log(X_i - c)|$ which are too big, or restrain them in some way. Another procedure would be to decide in advance that all values of X_i outside a certain interval $[\gamma_1, \gamma_2]$ will be thrown away.

If, however, the purpose is to estimate Paul's t, I would not want to eliminate the large values of the X_i.

When T_n differs too much from $\overline{\theta}$, there are usually some reasons. One of them is that the improvement procedure is only justified if the function $\theta \rightsquigarrow \Lambda_n(\theta, \theta_0)$ is indeed close to a quadratic. (If it was *known* to be close to some other smooth function, one might want to use that function instead.) Also, the quadratics are often so flat that their maximum is badly determined.

The same kind of thing can happen in other similar methods, such as ''scoring'' or ''Newton-Raphson''.

There is not too much one can do about it except let n tend to infinity (Principle 7), but Paul who got only one observation per year could not wait that long, even though his boss allowed him a few years to study the situation.

However, at least in some cases, the fact that the ''improved'' estimate T_n falls out of all reasonable bounds is imply due to the fact that the specific model used to compute the Λ_n is too far from reality. One can check that, but only to a very limited extent, by standard tests. The fact that T_n differs too much from $\overline{\theta}$ provides another such test.

Here we have worked within an assumed model. R. Beran (1981) has proposed a clever procedure to work around an assumed model instead of within it. The procedure can be misrepresented as follows. Suppose that the model specifies a certain family of densities $\mathbf{F} = \{f(x,\theta)\,;\,\theta \in \Theta\}$ for which you would not hesitate to use a Newton-Raphson type of procedure to produce an m.l.e. Suppose for simplicity that the observations are real variables and construct their empirical cumulative F_n. Let $\rho_n = \inf_\theta \sup_x \sqrt{n}\,|F_n(x) - F(x,\theta)|$ and assume that $\bar{\theta}_n$ achieves that infimum. Take the score function $\phi(x,\theta) = \frac{\partial}{\partial\theta} \log f(x,\theta)$. Instead of using Newton-Raphson starting with $\frac{1}{\sqrt{n}} \Sigma\, \phi(x_i,\bar{\theta}_n)$, tone down ϕ replacing $\phi(x_i,\bar{\theta}_n)$ by something like $\min[\,a_n, \max(-a_n), \phi(x_i\,;\,\bar{\theta}_n)\,]$ where a_n is taken large if ρ_n is small and closer and closer to zero as ρ_n increases. (Beran's procedure is more complex, but this does not pretend to be a description of it, only a misrepresentation!) Beran shows that the procedure has definite merits. It works efficiently if the model is correct. If the model does not fit it too well the estimate one gets is close to $\bar{\theta}_n$. It still makes some sense and is robust.

There is a slight difficulty with this approach. What if ρ_n appears too large, for instance larger than three or four? This is a strong indication that the postulated model is too far from reality. One might then want to review the situation and replace it by a different model. That is all right. However fiddling with the models until they fit is not always a commendable thing to do unless the result can be checked on a new independent set of data. Nowadays people do not hesitate to jackknife, bootstrap or otherwise resample with great abandon and thereby convince themselves that they have successfully bypassed all essential difficulties. However I feel more secure with a new independent sample.

Some of the recipes to be found in the literature amount to deciding in advance that the ''true'' $\Lambda_n(\theta_1,\theta_0)$ is likely to misbehave and replace in Λ_n the functions $\log \frac{f(x,\theta_2)}{f(x,\theta_1)}$ by some other functions, say $\psi(x,\theta_2,\theta_1)$ chosen in a fairly arbitrary manner. One is led then to the M-estimates.

This being said, we should return to Principle 7. It has several aspects. I will touch only on one of them here, because it tends to be forgotten.

Suppose that you do have a large number of observations, say 10^9. This is still very far from infinity, but suppose that you hesitate between two estimates, say T_n' and T_n'', and that you are told that T_n'' has an expected square deviation smaller than that of T_n' by an amount of the order $n^{-2} = 10^{-18}$.

If deviations of that size do matter to you (and they might at times), you had better check *everything* very carefully. There are probably a lot of little things which could have happened to your observations and could make vastly larger differences. Also the particular formula you use for $f(x, \theta)$ needs to be very seriously justified.

In other words the criteria you might use to choose between T_n' and T_n'' may well look at tiny differences which get magnified as $n \to \infty$ but are of little interest to you for the particular n you have. I don't mean to imply that one should not look at tiny things. What I am trying to say is that, if possible, the method or procedure, or optimality principle used to select the estimation procedure should preferably have some sort of stability, so that its dictates would not be grossly affected by deviations from the assumptions which are invisible on the data.

If the method is not stable in this sense it may be reasonable to check afterwards that the estimate it suggests behaves in a reasonable manner.

Finally suppose that in spite of all of this you have decided to use the T_n obtained by improving $\overline{\theta}$, partly because you can show that it is asymptotically sufficient, etc. and partly because you did not think of anything else.

That does not necessarily mean that you should use T_n *itself* as your estimate.

There may be functions or modifications of T_n which are better suited to your purposes.

That is where we left Berkson and Hodges trying to find functions of the sufficient statistics which had a desirable bias and is perhaps as good a place as any to stop.

References

Bahadur, R.R. (1958). Examples of inconsistency of maximum likelihood estimates. *Sankhyā* **20**, 207-210.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference Under Order Restrictions*. John Wiley, N.Y.

Beran, R. (1981). Efficient robust estimates in parametric models. *Z. Wahrscheinlichkeitstheorie u.v. G.* **55**, 91-108.

Berkson, J. (1955). Maximum likelihood and minimum χ^2 estimates of the logistic function. *J. Amer. Statist. Assoc.* **50**, 130-162.

Berkson, J. (1975). Do radioactive decay events follow a random Poisson-exponential? *Int. J. of Applied Radiation and Isotopes* **26**, 543-549.

Berkson, J. (1980). Minimum chi-square, not maximum likelihood with discussion by B. Efron, J.K. Ghosh, L. Le Cam, J. Pfanzagl and C.R. Rao. *Ann. Statist.* . **8**, 457-487.

Berkson, J. and Hodges, J.L. Jr. (1961). A minimax estimator for the logistic function. *Proc. 4th Berkeley Symp. Math. Statist. Prob.* **4**, 77-86.

Breiman, L. and Stone, C.J. (1985). Broad spectrum estimates and confidence intervals for tail quantiles. Tech Report No. 46. U.C. Berkeley.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Doob, J.L. (1948). Application of the theory of martingales. *Le calcul des probabilités et ses applications*. Paris C.N.R.S.

Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency). *Ann. Statist.* **3**, 1189-1242.

Ferguson, T.S. (1958). A method of generating best asymptotically normal estimates with application to the estimation of bacterial densities. *Ann. Math. Stat.* **29**, 1046-1062.

Ferguson, T.S. (1982). An inconsistent maximum likelihood estimate. *J. Amer. Statist. Assoc.* **77**, 831-834.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambridge Phil. Soc.* **22**, 700-725.

Fréchet, M. (1943). Sur l'extension de certaines évaluations statistiques au cas des petits èchantillons. *Rev. Inst. Int. Statist.* **11**, 183-205.

Hampel, F. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.* **42**, 1887-1896.

Hill, B. (1963). The three parameter log normal distribution and Bayesian analysis of a point source epidemic. *J. Amer. Statist. Assoc.* **58**, 72-84.

Hodges, J.S. (1987). Assessing the accuracy of the normal approximation. *J. Amer. Statist. Assoc.* **82**, 149-154.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many parameters. *Ann. Math. Statist.* **27**, 887-906.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. Calif. Publ. Statist.* **3**, 37-98.

Mammen, E. (1988). Local optimal Gaussian approximation of an exponential family. *Probab. Theory Related Fields* . **76**, #1, 103-109.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1-32.

Pfanzagl, J. and Wefelmeyer, W. (1978). A third order optimum property of the maximum likelihood estimator. *J. Multivariate Analysis* **8**, 1-29.

Savage, L.J. (1976). On rereading R.A. Fisher. *Ann. Statist.* **4**, 441-500.

Taylor, W.F. (1950). On tests of hypotheses and best asymptotically normal estimates related to certain biological tests. Ph.D. Thesis, U.C. Berkeley. Unpublished.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann.*

Math. Statist. **20**, 595-601.