WHAT IS OPTIMALITY IN SCIENTIFIC INFERENCE?

DAVID A. SPROTT

1. INTRODUCTION

The following are only two examples of an increasing tendency to tie scientific inferences to goals or decisions by imposing diverse extraneous optimality criteria:

(1) *"Thirteen methods for computing binomial confidence intervals are compared based on their coverage properties, widths, and errors relative to exact limits." Statistics in Medicine 12 (1993), 809-823.*

(2) *"Six different statistical methods for comparing limiting dilution assays were evaluated, using both real data and a power analysis of simulated data." In Vitro Cellular & Development Biology 25 (1989), 69-75.* These methods depended on specific statistical goals e.g. minimizing type 1 errors vs. maximizing the ability to discriminate between treatments.

It is reasonable to have six different methods for designing an experiment depending on specific scientific goals. But once the experiment has been performed, yielding an observed sample S_o, it seems contradictory to produce six different quantitative statements of uncertainty about values of θ. The resulting complexity is noteworthy.

It is assumed that science is the study of repeatable phenomena. Its purpose is to predict nature. This requires reproducible experiments. This leads to statistical models of experiments in the form of probability functions $P(S_o;\theta)$ of the observations S_o, usually in terms of unknown parameters θ. The primary problem of inference to which this leads is that of inferential estimation. This consists of specific quantitative statements of the plausibility or support by the observed data S_o of the various possible values of θ. A typical example is $\theta = \bar{y} \pm st_{(n-1)}$ appropriate for normal observations.

KEY WORDS: Bayesian intervals; bias; likelihood intervals; likelihood-confidence intervals; maximum likelihood; normal likelihoods; shortest intervals.

2. Inferential Optimality

The purpose of statistical models P is to generate the probabilities of the observations S_o. For their relevance, inferences about the model should use all the information in S_o. They should therefore be based on the probabilities $P(S_o;\theta)$ that constitute the model. It is suggested here that the optimality of the procedures of scientific inference should likewise be based on these probabilities, which are intrinsic to the model. One possibility is inferential optimality:

> explanations that raise the probability $P(S_o;\theta)$ of the observed sample S_o are more plausible, or more strongly supported by S_o, than explanations of equal complexity that lower the probability of S_o.

This is simply an extension into a graded scale of the single point of zero plausibility whereby any explanation is untenable if it makes the observed data impossible. The reason for limiting the complexity is that the probability $P(S_o;\theta)$ can be raised solely by making P more complex, for example by adding more parameters.

Such a criterion results in considerable simplicity. Imposing extrinsic optimality criteria such as best, shortest, linear, unbiased, coherent, minimum mean square error or variance, is more complicated and, as will be exemplified, can be injurious to quantitative inferential statements. Yet there seems to be an increasing tendency to do this, as the two examples cited at the beginning of Section 1 suggest.

3. Likelihood Intervals

For simplicity what follows is restricted to the estimation of a single scalar parameter θ in a specified model $P(S_o;\theta)$.

Application of this inferential optimality criterion to $P(S_o;\theta)$ implies that a value θ_1 is more plausible or strongly supported by the data S_o than a value θ_2 if and only if $P(S_o;\theta_1) > P(S_o;\theta_2)$.

Extending this to all values of θ produces a ranking given by the likelihood function of θ:

$$L(\theta;S_o) \stackrel{\text{def}}{=} C(S_o)P(S_o;\theta) \propto P(S_o;\theta).$$

For example, $L(\theta_1;S_o)/L(\theta_2;S_o) = P(S_o;\theta_1)/P(S_o;\theta_2) = 4$ means that θ_1 is four times more plausible than θ_2 in the operational sense that the observed S_o will occur four times more frequently in samples from the population defined by θ_1 than in samples

from the population defined by θ_2. This is an objective frequency that can be verified by simulations on a computer.

An optimal inferential set for θ based on S_o is then defined as the set in which the values of θ inside the set make $P(S_o;\theta)$ greater than values of θ outside the set. For a scalar parameter this set will usually be an interval, which is the case in what follows. This is then a likelihood interval.

The value of θ that maximizes the probability of the observed sample is the maximum likelihood estimate $\hat{\theta} = \hat{\theta}(S_o)$. It is the best supported value of θ by S_o in that it makes the observed S_o most probable. It is contained in all of the likelihood intervals. It can therefore serve as a reference point to give the location or the position of the likelihood function on the θ axis. More importantly it can serve as a yardstick by which the plausibility of other values of θ can be judged. This is done by the relative likelihood function

$$R(\theta; S_o) \stackrel{\text{def}}{=} L(\theta; S_o) \Big/ L(\hat{\theta}; S_o) = P(S_o;\theta) \Big/ P(S_o;\hat{\theta}). \tag{3.1}$$

From the above definition, $0 \leq R(\theta; S_o) \leq 1$. Since $P(S_o)$ is a probability function it is necessarily bounded, and so the denominator of (3.1) is finite. Optimal inferential intervals then have the form $R(\theta; S_o) \geq c$, $0 \leq c \leq 1$. A fixed c then gives a level c likelihood interval. As c varies the complete set of nested likelihood intervals is formed reproducing the relative likelihood function. It is essential to emphasize that one such interval does not suffice. Nor does the maximum likelihood estimate alone suffice. As a minimum both are required. But the full information about θ contained in S_o is conveyed by a graph of the relative likelihood function. This is merely the graph of the probability of the observed sample plotted as a function of θ. Thus the inferences are fully conditioned on the observed sample. Likelihood intervals have the frequency interpretation that in repetitions (simulations) any value of θ within a likelihood interval will produce S_o more frequently than will any value of θ outside this interval.

No optimality criterion can be justified or proved solely on mathematical or logical grounds. It must be judged by its practical consequences. Do the resulting inferences seem reasonable? This is clearly a matter of opinion, and so must be discussed in terms of practical experimental examples.

Example 1. A capture-recapture problem.

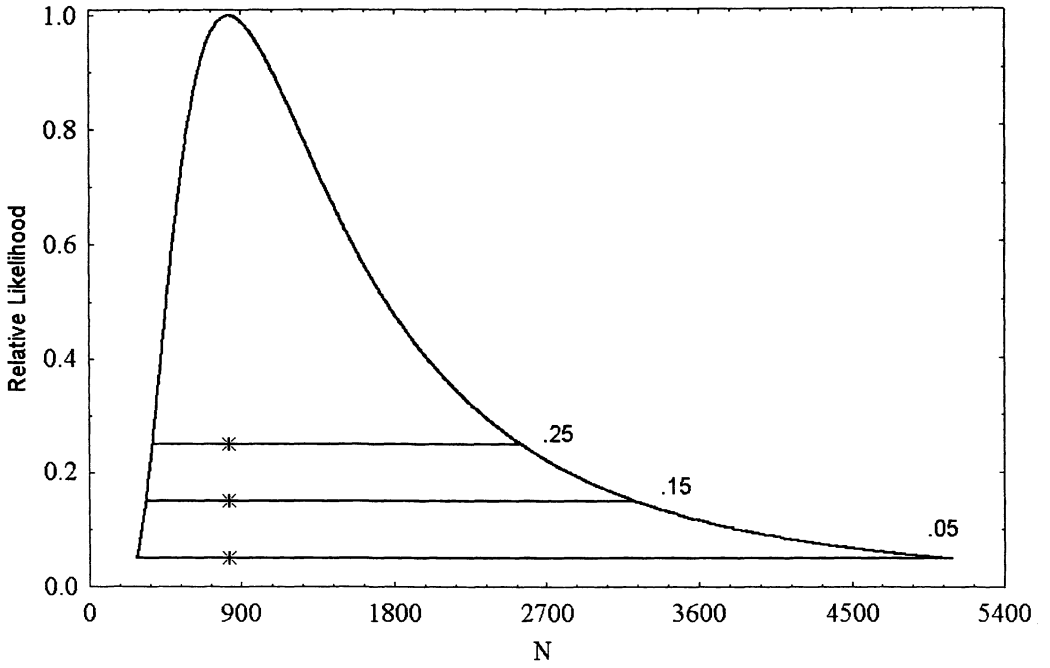

FIGURE 1. Relative likelihood function, Example 1

Animal population sizes N are often estimated using mark-capture-recapture techniques. The observations are $f_1, f_2, \ldots$, where f_i animals are caught i times. Then f_0 is unobserved and $N = \sum_0^\infty f_i$.

Let $s = \sum_{i=1}^\infty i f_i$, the number of animals caught, and $r = \sum_{i=1}^\infty f_i$, the number of distinct animals caught, so that $f_0 = N - r$. Assume the probability of any specified animal being caught is $1/N$, irrespective of how often it has been caught previously. This is the classical occupancy model, leading to the probability model

$$\begin{aligned} P(f_1, \ldots; N|\, s) &= N^{-s} N! s! / (N-r)! \prod_{i=1}^{\infty} f_i!(i!)^{f_i} \\ &\propto N^{-s} \binom{N}{r} \propto L(N; r, s), \quad N \geq r. \end{aligned}$$

An example using butterflies gave

$$S_o = \{f_1 = 66,\ f_2 = 3\ , f_i = 0,\ i \geq 3\}, \quad r = 69,\ s = 72,$$

Craig (1953), Darroch and Ratcliff (1980). Figure 1 shows the relative likelihood function along with $c = .05, .15, .25$ likelihood intervals, giving the position of the maximum likelihood estimate in each.

The main feature of Figure 1 is the extreme skewness. Values $N > \hat{N} = 828$ are much more plausible than values $N < \hat{N}$. Failure to convey this would result in seriously understating the population size N.

Quoting $\hat{N} = 828$ alone conveys practically no information. Exhibiting a single likelihood interval alone fails to convey much information. For example the .05 likelihood interval by itself simply shows that values of N between 280 and 5,089 have relative likelihoods $c \geq .05$. This gives no indication of the behavior of the plausibility of points within the interval, in particular the skewness. But supplementing the likelihood interval by the position of the maximum likelihood estimate $\hat{N} = 828$ conveys considerably more information. Its deviation from the geometrical center of the interval, 2,684, indicates the extreme skewness. From (3.1) it is apparent that while N varies from 280 to 828, the probability of the observed sample, $P(S_o; N)$, rises from .05 of its maximum to its maximum. But for this probability to decrease back to .05 of its maximum requires N to increase from 828 all the way to 5,000. It is preferable to present some additional likelihood intervals such as the .15 and .25 in Figure 1. The endpoints of all of the likelihood intervals (the complete set) reproduce the likelihood function.

Ignoring the extreme asymmetry underlined by Figure 1 would result in seriously understating the magnitude of N. This is reinforced numerically by the corresponding likelihood intervals below, in which $\hat{N} = 828$ has been included in each interval to emphasize the skewness:

$$\begin{array}{lll} c = .25 & 375, \quad 828, & 2{,}548, \\ c = .15 & 336, \quad 828, & 3{,}225, \\ c = .05 & 280, \quad 828, & 5{,}089, \end{array}$$

Example 2. Poisson dilution series. This is the second of the two examples cited at the beginning of Section 1.

The density of organisms in a given medium is θ per unit volume. The original medium is successively diluted by a factor a to obtain a geometric series of $k + 1$ solutions with densities

$$\theta/a^0 = \theta, \ \theta/a, \ \theta/a^2, \ \ldots, \ \theta/a^k.$$

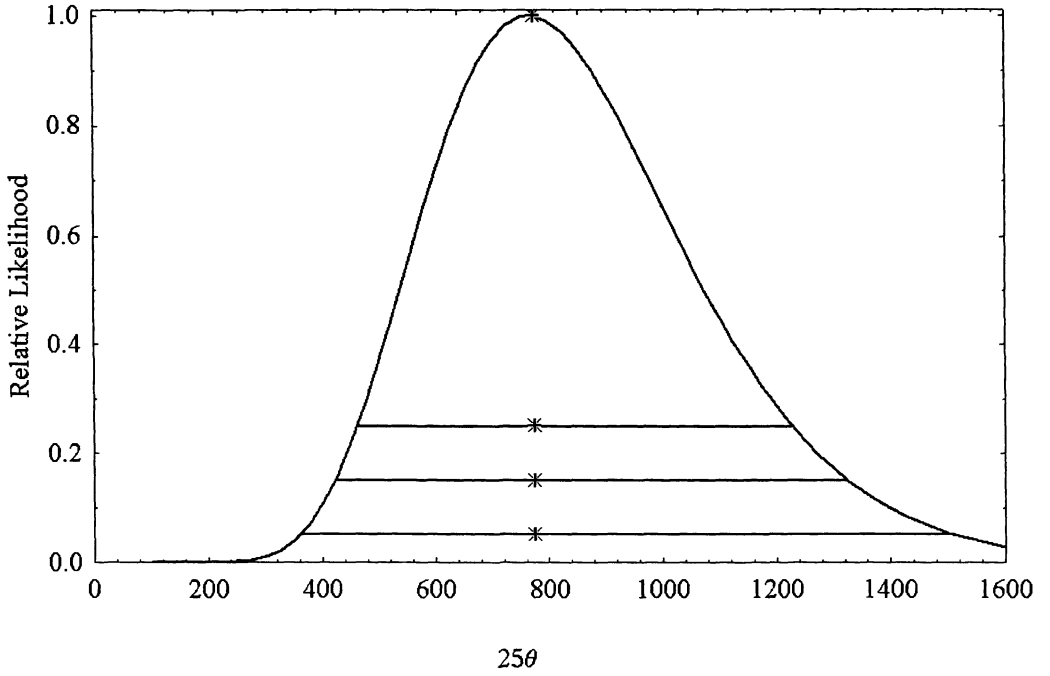

FIGURE 2. Relative likelihood function, Example 2

A unit volume of the solution with density θ/a^i is injected into each of n_i plates. Only the presence or absence of organisms can be detected.

Thus the observations are then $y = y_0, y_1, \ldots, y_k$, where y_i is the number of sterile plates out of the n_i at dilution level i. Assuming a Poisson distribution of the organisms, and that a sterile plate is one that had zero organisms to begin with, the probability of a sterile plate at level i is

$$p_i = \exp(-\theta/a^i),\ i = 0, 1, \ldots, k.$$

The probability of a fertile plate at level i is $1 - p_i$. Assuming independence, the likelihood function is proportional to the product of $(k+1)$ binomial likelihoods

$$L(\theta; y) \propto \prod_{i=0}^{k} p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Fisher and Yates (1963 p. 9) give the following data: $a = 2$, $k + 1 = 10$, $\{n_i\} = 5$, and $S_o = \{y_i\} = \{0, 0, 0, 0, 1, 2, 3, 3, 5, 5\}$. The unit volume was 1 cc, which contained

.04 gm of the material (potato flour) containing the organisms. Thus if θ is the number of organisms per cc, the number of organisms per gm of potato flour is 25θ.

The maximum likelihood estimate is $\hat{\theta} = 30.65$ organisms/cc, so that the maximum likelihood estimate of the number of organisms per gm of potato flour is $25\hat{\theta} = 766$. Again the graph of the likelihood function is asymmetric, Figure 2. Thus the comments of the previous Example also apply here. The information in the sample may be summarized, as before, by the .05, .15, .25 likelihood intervals along with the maximum likelihood estimate:

$$\begin{array}{rcl|lll} c & = & .25 & 461, & 766, & 1{,}228, \\ c & = & .15 & 422, & 766, & 1{,}325, \\ c & = & .05 & 358, & 766, & 1{,}512, \end{array}$$

organisms per gm.

Example 3a. Exponential failure times t_i,

$$f(t_i;\theta) = \theta \exp(\theta t_i).$$

The relative likelihood function of θ based on n independent observations $S_o = t_1, \ldots, t_n$ is

$$\begin{aligned} R(\theta;t) &= \left(\theta/\hat{\theta}\right)^n \exp\left(n - n\theta/\hat{\theta}\right), \\ \hat{\theta} &= n/t,\ t = \sum t_i, \end{aligned}$$

This is shown in Figure 3 for $n = 2$ along with the .05, .15, .25 likelihood intervals.

For future reference t has the gamma (n, θ) distribution

$$f(t;\theta) = \frac{1}{(n-1)!} t^{n-1}\theta^n \exp(-\theta t).$$

Example 3b. Censored exponential failure times.

Suppose the failure times above were censored. The n items were observed for fixed periods of time $T_1, \ldots, T_n$ and r of the items were observed to fail at times $t_1, \ldots, t_r$; the remaining $(n - r)$ items were observed to survive their periods of observation, $t_i > T_i$. The observed sample is thus $S_o = t_1, \ldots, t_r, r, T_{r+1}, \ldots, T_n$. Since $P(t_i > T_i) =$

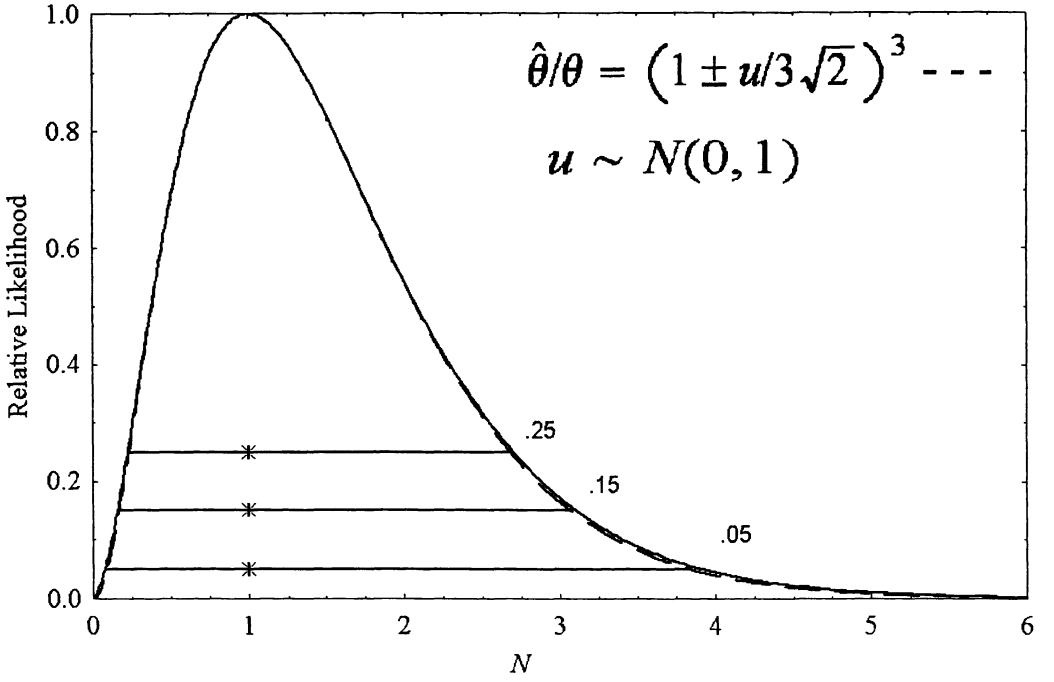

FIGURE 3. Relative likelihood, Example 3, and its normal approximation (2), Examples 4 and 5

$\exp(-\theta T_i)$, the resulting likelihood function of θ is

$$L(\theta; t, r) \propto \left[\prod_1^r \theta \exp(-\theta t_i)\right]\left[\prod_{r+1}^n \exp(-\theta T_i)\right].$$

The relative likelihood function is algebraically the same as the uncensored case with n replaced by r and t by

$$t = \sum_{i=1}^{r} t_i + \sum_{i=r+1}^{n} T_i.$$

Thus the likelihood inferences are unaffected by the censoring and so retain their simplicity. In contrast, the sample space and probability functions are much more complicated. In particular t no longer has the gamma (n, θ) distribution of Example 3a.

The foregoing should not be confused with the traditional Method of Maximum Likelihood. The Method of Maximum Likelihood focuses on the asymptotic behavior of the maximum likelihood estimate. The present approach focuses on the inferential

use of the *whole* likelihood function as furnishing measures of relative plausibility. Here the behavior of the maximum likelihood estimate is not relevant.

4. Likelihood-Probability (Confidence) Intervals

In Example 3a, $t \sim$ gamma (n, θ), giving

$$t\theta \sim \text{ gamma } (n) \Longleftrightarrow z = 2t\theta \sim \chi^2_{(2n)},$$

so that z is a $\chi^2_{(2n)}$ pivotal quantity. That is, for any $a < b$,

$$a \leq z \leq b \Longleftrightarrow a/2t_o \leq \theta \leq b/2t_o$$

the probability of which can be calculated using the $\chi^2_{(2n)}$ distribution of z. In this way a pivotal quantity z can be used to establish probability relationships between θ and the observed t_o. This allows the assignment of probabilities to parametric intervals. The result is a probability or a confidence interval, depending on its interpretation.

For $n = 2$, $z \sim \chi^2_{(4)}$. Then, for example,

$$P(0 \leq z \leq 9.488) = .95, \text{ and } 0 \leq z \leq 9.488 \Longleftrightarrow 0 \leq \theta \leq 4.744/t_o$$

gives a .95 confidence interval for θ.

Similarly

$$.7108 \leq z \leq \infty \Longleftrightarrow .3554/t_o \leq \theta < \infty$$

is a .95 confidence interval, and

$$\begin{aligned} (0 \leq z \leq 3.199) \ \cup \ (3.519 \leq z \leq \infty) \Longleftrightarrow \\ (0 \leq \theta \leq 1.599/t_o) \ \cup \ (1.7595/t_o \leq \theta \leq \infty) \end{aligned}$$

is a .95 union of such intervals.

Thus confidence intervals at any specified confidence level are not unique. This raises the question of optimal confidence intervals. Constancy of the coverage frequency is a relatively mild restriction and does not guarantee sensible intervals. In fact it is easy to produce exact confidence intervals that do not even depend on the data.

Such deficiencies can be avoided if confidence intervals are likelihood intervals. Likelihood-confidence intervals are optimal confidence intervals in the sense that they include the most plausible values of the parameter. They are also unique.

For example, from the $\chi^2_{(4)}$ distribution of z, the $c = .15$ likelihood interval

$$.17\hat{\theta}_o \leq \theta \leq 3.06\hat{\theta}_o \Longleftrightarrow .68 \leq 4\theta/\hat{\theta} = z \leq 12.16$$

is the unique .937 likelihood-confidence interval.

It should be emphasized that likelihood measures the relative plausibility of *specific* values of the parameter within an interval. Likelihood cannot measure the plausibility of the intervals themselves.

Probability measures the plausibility of the intervals. Probability cannot measure the relative plausibility of specific values of a continuous parameter, since the probability of any specific value of a continuous variable is zero.

The domains of application of likelihood and of probability are therefore complementary.

5. Normal Likelihoods

Normal likelihoods are useful for simplifying the inferential structure and for obtaining approximate likelihood-confidence intervals.

A normal likelihood has the form

$$R_N(\delta; y) = \exp\left(-\tfrac{1}{2}u_\delta^2\right), \quad u_\delta = (\hat{\delta} - \delta)\sqrt{I(\hat{\delta}; y)},$$

$$\text{where} \quad I(\hat{\delta}; y) = -\frac{\partial^2}{\partial \hat{\delta}^2} \log R(\delta; y)$$

is the observed information.

If this holds in repeated samples it suggests that u_δ can be treated as a $N(0,1)$ pivotal quantity. The inferences then take the simple classical form of the estimation statements

$$\delta = \hat{\delta} \pm su, \quad s = 1\Big/\sqrt{I(\hat{\delta}; y)}, \; u \sim N(0,1).$$

This is a complete set of nested likelihood-confidence intervals, thus reproducing the likelihood function. Functional invariance gives corresponding likelihood-confidence intervals for any 1-1 function of δ by ordinary algebraic substitution.

These results seldom hold exactly, but often hold approximately, as the following examples show.

Example 4. Exponential failure times, Example 3a.

For the parameter $\delta = \theta^{1/3}$,

$$R(\delta; t) \approx R_N(\delta; t), \text{ with } I(\hat{\delta}; t) = 9n/\hat{\delta}^2.$$

The inferences take the revealingly simple form of estimation statements

$$\delta = \hat{\delta} \pm u\hat{\delta}/3\sqrt{n} \iff \theta = \hat{\theta}(1 \pm u/3\sqrt{n})^3, \quad u \sim N(0,1). \tag{5.1}$$

As an example of functional invariance, the likelihood function and corresponding likelihood-confidence intervals of the survivor function $\pi = \exp(-\tau\theta) = P(t > \tau)$ can immediately be obtained from (5.1),

$$\tau = -\hat{\theta}^{-1} \log \pi (1 \pm u/3\sqrt{n})^3.$$

This gives inferences about τ given π, π given τ, or π, τ jointly.

These are complete sets of reasonably accurate nested likelihood-confidence intervals. They are shown for θ when $n = 2$ in Figure 3. The approximate probabilities and likelihoods are obtained from $u \sim N(0,1)$ and $\exp(-\frac{1}{2}u^2)$, respectively. For example setting $u = 1.96$ gives the approximate .146 likelihood .95 confidence interval $\theta = \hat{\theta}(1 \pm 0.4620u)^3$. In this case the exact probabilities can be calculated for comparison from the $\chi^2_{(4)}$ distribution of z. Some results are given in the following Table

u_o	$1-\alpha_N$	$1-\alpha$	c_L, c_U	c_N
2.576	.99	.991	.024, .032	.036
1.960	.95	.946	.131, .139	.146
1.645	.90	.892	.246, .252	.258
0.674	.50	.483	.796, .796	.797

Here, $1-\alpha_N$, $1-\alpha$ are the approximate normal and the corresponding exact probabilities; c_L, c_U are the exact left and right hand relative likelihoods and c_N is the corresponding approximate normal relative likelihood of the resulting intervals. For exact likelihood intervals $c_L = c_U$.

These results apply to any gamma likelihood. They therefore apply to the censored Example 3b with n replaced by r provided $r \neq 0$. If $r = 0$ the resulting likelihood cannot be approximated by normality. Thus the accuracy of the coverage frequencies in repeated samples will decrease for values of θ that make $P(r = 0)$ non-negligible.

The corresponding result for $\delta = \log\theta$ is

$$\delta = \hat{\delta} \pm \frac{u}{\hat{\theta}\sqrt{I(\hat{\theta};y)}} \Longleftrightarrow \theta = \hat{\theta}\exp\left(\pm\frac{u}{\hat{\theta}\sqrt{I(\hat{\theta};y)}}\right). \tag{5.2}$$

6. Bias Corrections

In contrast with the above approach, inferences are usually based on a statistic $t(y)$. A statistic is usually presented as a method of data reduction to produce an estimate $\tilde{\theta} = \tilde{\theta}(y)$ of θ. This leads to a study of optimal properties of estimates. Unbiased estimates with uniformly minimum variance (UMVs) are particularly favored; also best linear unbiased estimates (BLUEs). Since the maximum likelihood estimate $\hat{\theta}$ is usually biased, this leads to attempts to correct $\hat{\theta}$ for its bias and to calculate the variance $\sigma^2_{\tilde{\theta}}$ of the resulting approximately unbiased estimate $\tilde{\theta}$. The following examples illustrate the undesirable scientific consequences this approach can produce.

Example 5. Exponential failure times of components connected in series.

Miyramura (1982) applied the above procedure to systems of components connected in series, assuming individual exponential failure times. The results for a single component were:

$$\begin{aligned}
\tilde{\theta} &= [1-(2/\tilde{\nu})]\hat{\theta}, \\
{\tilde{\sigma}_{\tilde{\theta}}}^2 &= \tilde{\theta}[1-(2/\tilde{\nu})+(4\tilde{m}\tilde{\theta}/\tilde{\nu}^2)]/[\tilde{m}(1-4/\tilde{\nu}], \\
\text{where } \hat{\nu} &= 2\left(\sum_{i=1}^{n} r_i z_i/\tilde{\beta}_i\right)^2 \Bigg/ \left(\sum_{i=1}^{n} r_i z_i^2/\tilde{\beta}_i^2\right), \\
\tilde{m} &= \sum_{i=1}^{n} r_i z_i/\tilde{\beta}_i, \quad \tilde{\beta}_i = (r_i-1)/t_i,
\end{aligned}$$

where t_i are failure times, r_i are determined by the censoring mechanism, and the z_i determine the structure of the system in series. The properties of the above estimates $\tilde{\theta}$, ${\tilde{\sigma}_{\tilde{\theta}}}^2$, were assessed by simulations.

One of the numerical examples given yielded $\hat{\theta} = .035$, $\tilde{\theta} = .028$, $\tilde{\sigma}_{\tilde{\theta}} = 0.024$ in a sample of $n = 2$. Viveros (1991) noted that the use of this to produce intervals gives the .95 confidence interval $-.019 \leq \theta \leq .075$. Since values $\theta < 0$ are impossible, such intervals may be called incredible.

Although the probability of the entire system in series is complicated, because of the underlying exponential distributions the resulting likelihood function is a simple gamma $\theta^k \exp(t\theta)$. This suggests using the normal approximation to the likelihood of $\delta = \theta^{1/3}$ of Example 4. Using (5.1) gives the estimation statements

$$\theta = \hat{\theta}(1 \pm u/3\sqrt{n})^3, \quad u \sim N(0,1).$$

The simplicity of this may be compared with the unbiased results above. Also for $\hat{\theta} = .035$, $n = 2$, the resulting .95 likelihood-confidence interval is $.005 \leq \theta \leq .109$, shown in Figure 3, which is at least credible.

Simulations show that the coverage frequency of intervals produced this way are very close to those obtained by assuming $u \sim N(0,1)$, Viveros (1991). Thus, these are a highly efficient set of nested approximate likelihood-confidence intervals.

This example also illustrates that if simulations are required, the right quantities should be simulated. One should simulate samples from the original model. But to set up confidence intervals it is rarely appropriate to examine the frequency properties of the resulting estimates, including the maximum likelihood estimates, and their variances formed from these simulated samples. The frequency properties of the resulting quantities u above formed from the simulated samples should be examined. These quantities have the form of a Student t pivotal, a function of *two* random variables.

Example 6. Capture-recapture.

For the capture-recapture data of Example 1 Darroch and Ratcliff (1980) estimated the bias of $\hat{N}$ to be 230, which looks large. They obtained the estimate $\tilde{N} = 668$ with a reduced bias, and quoted its standard error as $s = 333$. The corresponding estimation statement would be

$$N = 668 \pm 333u, \quad u \sim N(0,1).$$

Some of these intervals are shown in the Figure 4 in relation to $R(N; r, s)$.

The same comments apply as in the preceding example. The intervals are shifted well to the left of the likelihood function. Their lower bounds include values of N which, if true, would make it impossible to have obtained the observed sample. The lower .99 confidence bound, $u = 2.576$ gives $N = -147$ which is again incredible; the lower .95 bound, $u = 1.96$, gives $N = 15$, which is unrealistic since $r = 69$ distinct animals have already been observed; essentially the same is true of the lower .90 bound.

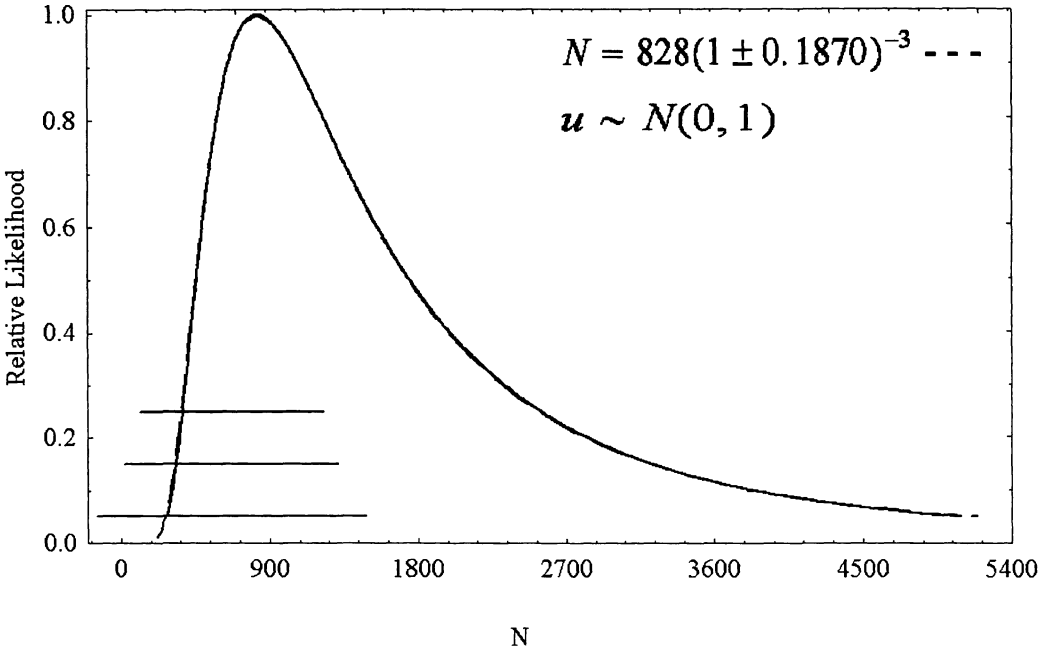

FIGURE 4. Confidence intervals corrected for bias, and normal approximation (2) to likelihood, Example 6

Conversely, the upper bounds of the intervals exclude values of N that would make the observed sample have high probabilities, .644, .75, .81 or more relative to the most probable sample.

Again the bias reduction along with the use of the asymptotic variance, without regard for the shape of the likelihood function, drastically understates the magnitude of N.

The use of $\delta = N^{-1/3}$ with a modification of (5.1) gives the complete set of nested approximate likelihood-confidence intervals as

$$N = 828(1 \pm 0.1870u)^{-3}, \quad u \sim N(0, 1),$$

shown also in Figure 4. The accuracy of the resulting confidence intervals N_L, N_U, as determined by the tail probabilities $P(r \leq 69|N_U)$, $P(r \geq 69|N_L)$, using the exact distribution $P(r; N|s) = N^{-s}\binom{N}{r}\Delta^r 0^s$, has been examined by Viveros (1992). This

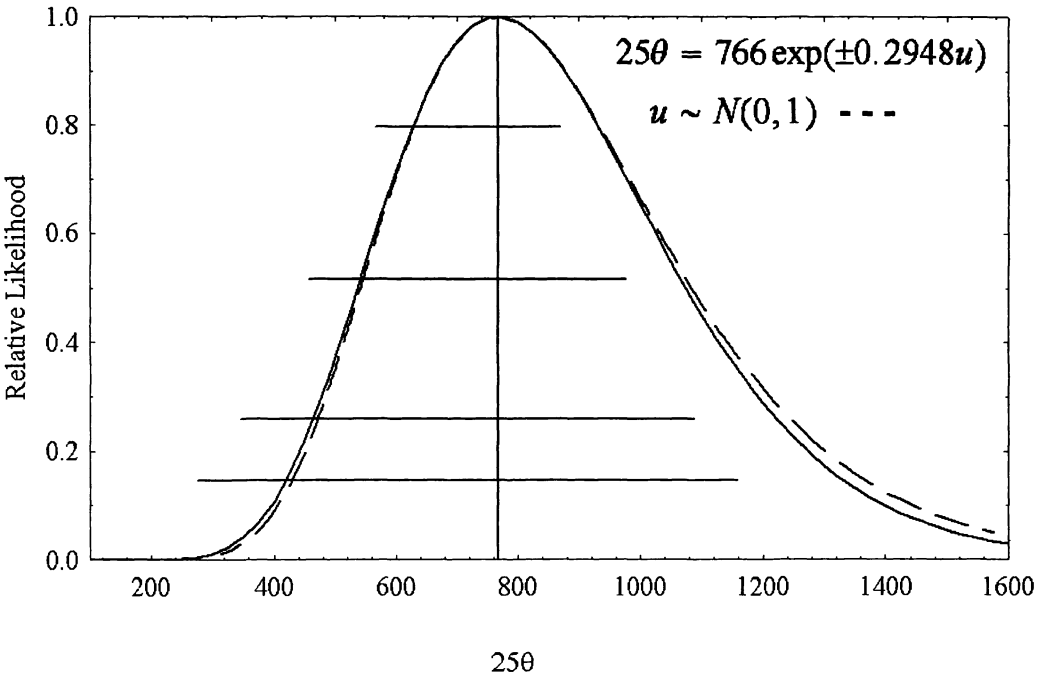

FIGURE 5. Confidence intervals corrected for bias, and normal approximation (3) to likelihood, Example 7

accuracy is affected by the extreme discreteness of the observations. This can be improved by using a continuity correction, Sprott (1981).

Example 7. Dilution series.

Mehrabi and Mathews (1995) applied a bias reduction to the maximum likelihood estimate in the Poisson dilution series model of Example 2. For the data of Example 2, the resulting approximately unbiased estimate is $\tilde{\theta} = 28.666$, with estimated standard error $s = 1 \Big/ \sqrt{I(\hat{\theta}; y)} = 9.0351$, giving

$$25\theta = 717 \pm 226u, \quad u \sim N(0,1),$$

as shown in Figure 5. The intervals again understate the magnitude of θ. The lower limits include values of θ that would make the observed sample have probabilities as small as .04, or even .005, of the maximum probability. The upper limits exclude values

that would make the observed sample have probabilities as large as .35, or .47 or more of the maximum possible.

In this case the appropriate procedure is to use $\delta = \log\theta$ (as did Fisher), the likelihood function of which is almost normal, giving from (5.2) the complete set of approximate likelihood-confidence intervals

$$25\theta = 766\exp(\pm.2948u), \quad u \sim N(0,1),$$

shown also in Figure 5.

These examples indicate that reducing statistical bias can introduce a more important and obvious scientific bias. The positive statistical bias in these examples is important in forcing attention to values of the parameter larger than the maximum likelihood estimate, thus reinforcing the message conveyed by the asymmetry of the likelihood. Ignoring these facts results in seriously understating the value of the parameter.

However, if the problem involves the combination of a large number of estimates a bias correction may be necessary. For otherwise the biases may accumulate and result in a misleading likelihood function as in the next Example.

Example 8. The common variance problem.

An example is the common variance problem, the most extreme case of which is pairs of observations $(x_i, y_i) \sim N(\xi_i, \sigma^2)$, independently. Then $d_i = (x_i - y_i) \sim N(0, 2\sigma^2)$. The maximum likelihood estimate is $\hat{\sigma}_i^2 = d_i^2/4$ with expected value $\sigma^2/2$. The maximum likelihood estimate based on k such pairs is $\hat{\sigma}_k^2 = \sum_{i=1}^{k} \hat{\sigma}_i^2/k$, the expected value of which is $\sigma^2/2$. Thus $\hat{\sigma}_k^2 \to \sigma^2/2$ as $k \to \infty$, which is inconsistent. Essentially the profile likelihood assigns $2k$ degrees of freedom to estimating σ^2 instead of k degrees of freedom. This equally implies that care is necessary in eliminating incidental parameters ξ_i.

It is interesting that this difficulty does not arise in the corresponding problem of the common mean, where $(x_i, y_i) \sim N(\delta, \sigma_i^2)$. In the case of the common variance the difficulty is avoided by using the marginal likelihood of σ arising from the marginal distribution of ${\hat{\sigma}_i}^2$. This is based on the $\chi^2_{(k)}$ distribution, and so assigns k degrees of freedom to estimating σ^2.

It might be argued that it was never the intention to interpret $(\tilde{\theta}, s)$ in the form of the estimation statement $\theta = \tilde{\theta} \pm su$, $u \sim N(0,1)$. This would simply raise the question, What quantitative interpretation does $(\tilde{\theta}, s)$ then have?

The scientific inferential interpretation of maximum likelihood estimation requires the existence of an appropriate approximate normal likelihood $R_N(\delta; y)$ which is summarized by the *un*adjusted maximum likelihood estimate $\hat{\delta}$ and the *observed* information $I(\hat{\delta}; y) = 1/s^2$. In the common variance problem these are supplied by the marginal likelihood function.

7. Length of Intervals

Much the same can be said about length as about bias. Any likelihood-confidence interval obtained from a skew likelihood can easily be shortened. Merely replace a large interval in the long tail of the likelihood function by a short interval in the short tail. The decrease in probability produced by deleting a larger part of the long tail, where the change in probability is slow, is restored by the increase in probability produced by adding a smaller part of the short tail, where the change in probability is rapid. This invariably shortens the interval and destroys its likelihood property.

Example 3a, Figure 3 provides a simple example of this. The interval $0.166\hat{\theta} \leq \theta \leq 3.09\hat{\theta}$ is a .146 likelihood-.94 confidence interval. Its length is $2.92\hat{\theta}$. The interval $0 \leq \theta \leq 2.26\hat{\theta}$ is a .94 confidence interval of length $2.26\hat{\theta}$. This decrease in length is obtained at the price of including values of θ near zero that make it impossible to have obtained the observed sample, and excluding large values that make the observed sample relatively probable.

Like unbiasedness, length is not functionally invariant. Shortest intervals in θ will not be a shortest interval in $\delta(\theta)$, such as the survivor function $\pi = \exp(-\theta\tau)$. Thus the justification for such a procedure involves motives that rest outside the field of objective scientific inference.

8. Bayesian Intervals

The purpose of prior distributions appears to be to convert likelihoods into probabilities, thus eliminating likelihood. Additionally, Bayesians calculate (shortest) intervals of highest probability density. This eliminates the advantage of the functional invariance of both probability and of likelihood, introducing the problems discussed in the previous two sections. The same can be said about equal tail intervals. Additionally, equal tail intervals do not seem to be reasonable if the posterior distribution is highly skewed. (This seems similar to difficulties with two-tail tests of significance.)

And *neither* probability *nor* probability density is capable of making comparisons of plausibility of points, θ_1 vs. θ_2. For the probabilities of θ_1 and of θ_2 are zero; and the density ratio $f(\theta_1)/f(\theta_2)$ is not well-defined owing to lack of functional invariance. For example, if $\delta = \log\theta$

$$f(\delta_1)/f(\delta_2) = [f(\theta_1)/f(\theta_2)]\,(\theta_1/\theta_2) \neq f(\theta_1)/f(\theta_2).$$

As was the case of shortest intervals, the justification for such a procedure rests outside of scientific inference.

In keeping with the criterion of inferential optimality, the optimal inferential use of Bayes would appear to be

(1) retain the likelihood function for establishing likelihood intervals;

(2) use the posterior density function to assign probabilities to these likelihood intervals.

The Bayesian approach would then consist in supplementing the likelihood with probability, retaining both as complementary measures of uncertainty.

9. Discussion

The essential role of the likelihood function in inductive inference has been stressed repeatedly by Fisher since Fisher (1921, 1925 and every subsequent edition, e. g. 14th ed. pp. 10, 11).

But the first use of the whole likelihood function as an inferential entity on which to base inferences, as in Section 3, was in Fisher (1956, p. 73), arguing "The likelihood supplies a natural order of preference among the possibilities under consideration.". He also exhibited a graph of a relative binomial likelihood function (Fisher 1956 p. 76). Since then there has been an increasing use of the likelihood function to measure inductive uncertainty or plausibility directly in terms of likelihood, see, for example, Edwards (1992), Sprott (2000). In all of the preceding there is no mention of optimality principles. But the natural order mentioned above by Fisher is the ordering that ranks evidence in favor of values of θ by how probable these values make the observed sample. This implicitly implies the criterion of inferential optimality in Section 2.

For the sake of simplicity, except for Example 8 attention has been restricted to estimation statements in models having a single scalar parameter θ. That the result is not entirely academic is shown by the variety of practical examples it covers.

However, multiparametric problems $\theta = \delta, \xi$, are more complex. One of the difficulties is that of separate estimation – obtaining estimation statements about the parameter of interest δ as in Sections 3-5, which are valid in the absence of knowledge of the incidental parameters ξ. This gives rise to marginal, as in Example 8, conditional, pivotal, integrated, profile likelihoods, and perhaps other kinds of likelihoods. Sometimes an appropriate likelihood cannot be found; the parameter δ cannot be adequately separated from ξ. For an example, see Farewell, Viveros, and Sprott (1993). General methods of handling these and associated problems in models $P(S_o; \theta) = f(\{y_i\}; \{\delta_i\}, \{\xi_i\})$, such as the assessment of the assumed f and of the repeatability of the parameter of interest, $\delta_i = \delta$, are discussed and exemplified in Sprott (2000).

Some have criticized this paper as being unbalanced in portraying only the attractive side of likelihood. For instance it is said that the paper ignores the devastating examples of completely misleading maximum likelihood estimates such as those of Ferguson (1982) and of Le Cam (1990, specifically p. 158). But the examples of the mixture of normal distributions and of the three parameter log normal distribution cited by Ferguson and by Le Cam are caused by approximating probability functions by density functions, and can thereby be remedied, e. g. Barnard (1966), Edwards (1992, Section 8.4). However there may be other mathematically contrived examples that cannot be resolved in this way. Interested readers can study the above references for themselves, bearing in mind that this is not a paper on maximum likelihood estimation. Moreover the paper is restricted to *scientific* inference, that is, the information that a given body of data contains about the issues under study.

Acknowledgement

I should like to thank Dr. E. Díaz-Francés and Professor M. D. Vogel-Sprott for reading preliminary versions of this paper and for helpful suggestions. I should also like to thank the referees for helpful suggestions.

References

[1] Barnard G. A. (1966). The use of the likelihood function in statistical practice. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1, 27-40.

[2] Craig C.C. (1953). On the utilization of marked specimens in estimating populations of flying insects. *Biometrika* 40, 170-176.

[3] Darroch, J.N. and Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics* 36, 149-153.

[4] Edwards, A.W.F. (1992). *Likelihood, Expanded Edition.* The Johns Hopkins Press: Baltimore and London.

[5] Farewell, V. T., Viveros, R., and Sprott, D. A. (1993). Statistical consequences of an adaptive treatment allocation in a clinical trial. *Canadian Journal of Statistics* 21, 21-27.

[6] Ferguson, T. S. (1982). An inconsistent maximum likelihood estimate. *Journal of the American Statistical Association* 77, 831-834.

[7] Fisher, R.A. (1921). On the 'Probable Error' of a coefficient of correlation deduced from a small sample. *Metron* 1, 3-32.

[8] Fisher, R.A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd: Edinburgh.

[9] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd: Edinburgh.

[10] Fisher, R.A. and Yates, F. (1963). *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed. Oliver and Boyd: Edinburgh.

[11] Le Cam, L. (1990). Maximum likelihood: An introduction. *International Statistical Review* 58, 153-171.

[12] Mehrabi, Y. and Mathews, J.N.S. (1995). Likelihood-based methods for bias reduction in limiting dilution assays. *Biometrics* 51, 1543-1549.

[13] Miyramura, T. (1982). Estimating component failure rates from combined component and systems data: exponentially distributed component lifetimes. *Technometrics* 24, 313-318.

[14] Sprott, D.A. (1981). Maximum likelihood applied to a capture-recapture model. *Biometrics* 37, 371-375.

[15] Sprott, D.A. (2000). *Statistical Inference in Science.* Springer-Verlag, New York.

[16] Viveros, R. (1991). Combining series system data to estimate component characteristics. *Technometrics* 33, 13-23.

[17] Viveros, R. (1992). Conditional inference under incomplete power-series samples. *Utilitas Mathematica* 41, 65-73.

DAVID A. SPROTT
CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS
GUANAJUATO, MÉXICO.
AND
UNIVERSITY OF WATERLOO
ONTARIO, CANADA.

sprott@cimat.mx