Perception & Psychophysics
1986, 39 (2), 151-153

Observations on maximum-likelihood and Bayesian methods of forced-choice sequential threshold estimation

PHILLIP L. EMERSON
Cleveland State University, Cleveland, Ohio

Quite a lot of research associated with signal detection theory (Green & Swets, 1966) has shown simple yes-no measures of sensitivity to be seriously confounded with variations of the observer's criterion. For that reason, forced-choice methods have been adopted widely for measuring thresholds. The forced-choice methods seem to solve the criterion problem, but to introduce new problems of their own. Kershaw (1985) and McKee, Klein, & Teller (1985) found two-alternative forced-choice (2AFC) thresholds to have undesirable sampling characteristics. For cases in which the number of trials is small (30, 45, and 60), the findings showed remarkably large standard errors and negative bias associated with a long downwardly skewed tail of the sampling distribution. The standard practice of specifying symmetrical confidence intervals therefore can be quite misleading. Users of 2AFC methods would be well advised to consult those findings, especially if maximum-likelihood (ML) estimation is used.

I became somewhat aware of ML-2AFC problems when considering the extension of an approximate ML sequential method from a yes-no procedure (Emerson, 1984) to a 2AFC procedure. With the 2AFC ML method, the threshold estimate seemed to become trapped far below the true threshold, and to remain there for many trials before eventually recovering. Although these occurrences were infrequent in some cases, they contributed greatly to the standard error and negative bias. A first hypothesis was of some artifact of numerical ill-conditioning of the computations. However, the same phenomenon was observed when different computational methods and various checks against numerical underflow and overflow in the computer programs were used. The reports of Kershaw (1985) and McKee et al. (1985) then made it seem even more plausible that the fault lay with the 2AFC procedure itself, or with ML estimation combined with the 2AFC procedure.

The undesirable characteristics of a particular ML-2AFC sequential method are confirmed by data presented here, and the main problems are more definitely identified to be due to the combination of ML estimation with the 2AFC procedure. It is fortunate that a Bayesian method utilizes all the same computations as does the ML. Thus, the Bayesian method provides a control against numerical ill-conditioning as the cause of the problems. The worst of the problems therefore are identified as ensuing from ML estimation. The Bayesian method seems to offer a less troublesome alternative when the necessary computing facilities are available.

The author's mailing address is Department of Psychology, Cleveland State University, Cleveland OH 44115.

The ML Method

The ML method is essentially the same as one used by Shelton, Picardi, and Green (1982), Shelton (1983), and Shelton and Scarrow (1984). The assumed psychometric function is

$$P(x,m) = .5 + .5\{1 + \exp[k(m-x)]\}^{-1}.$$

This is the modified logistic ogive in which $P(x,m)$ is the probability of correct response when the true 75th percentile is m and the stimulus value is x. $P(x,m)$ is also a function of the slope parameter, k, but this method requires the use of an a priori assumed value of k. Many researchers have made use of this assumed psychometric function for the 2AFC procedure, for several reasons of convenience. It has an inflection at the point $x=m$, so it is nearly linear in that region. Note, though, that, to map the a priori 50%-correct stimulus level to $-\infty$, it requires that x be a logarithmic transformation of stimulus energy. The single parameter, m, is to be estimated. The ML criterion is to choose m so as to maximize

$$L = \Pi P(x_1,m)\Pi[1-P(x_0,m)],$$

where the two products are over stimuli yielding correct responses (x_1) and incorrect responses (x_0), respectively.

The computations are done on the logarithm of this compound product by a method described by Shelton (1983). The method represents $\log L$ as a discrete array of numbers in computer memory, typically with about 50 equally spaced elements. These span a finite stimulus range that is thought almost surely to contain m. When a new stimulus is presented in a sequence of trials, the response is recorded as correct or incorrect. Then the $\log L$ array is updated by adding in the elements of either of two arrays representing $\log P(x_1,m)$ or $\log[1-P(x_0,m)]$, as the case may be. For speed of execution, these two arrays are precomputed for twice the stimulus range, before the start of a run of trials. Thus, the trialwise updating reduces to a series of about 50 offset-lookups and additions. The ML solution for m is then obtained after the updating on each trial, by scanning the $\log L$ array for its maximal element. With this method, the next value of x to be presented is always taken equal to the most recent estimate of m.

Execution speed is fast enough to be practical on a common microcomputer, but a compiled version of the program is recommended, rather than an interpretive one as is common with BASIC. One further operation was included in the implementation used in collecting the data reported here. It is the subtraction of the mean of the elements of $\log L$ from each of the elements, after each updating step. This keeps the range of the elements centered

approximately at zero, which improves the numerical conditioning for some of the computations.

The Bayesian Method

With the Bayesian method, the fundamental calculations are the same as those described above, using a discrete representation of $\log L$ with the same updating procedure. However, the estimate of m is determined not from the maximum of the elements of $\log L$, but as the normalized expected value of m, with the exponentials of the elements of $\log L$ being treated as unnormalized probabilities. This involves quite a lot more computations, about 50 exponential operations per trial, and it slows down execution considerably. For 50 elements, the computations on each trial take about ⅓ sec when one of the faster microprocessers available (an 8-mHz 80186) and single-precision arithmetic (Software Toolworks C compiler and Mathpak) are used. Watson and Pelli (1983) described a similar "Bayesian" method that is faster because it uses the mode of the likelihood function rather than the mean. For that reason, however, it is essentially the same as the ML method.

General Procedure

Comparison data were obtained from Monte-Carlo runs. An advantage of this technique is that the true values of k and m can be controlled exactly for the investigation of various cases. A possible danger is that not all the foibles of human and animal subjects are built into the model that generates the Monte-Carlo data. Thus, the absolute performances of the methods may appear better than can be expected with live subjects. However, comparisons of relative performances should be less affected by such discrepancies. In all of the simulations reported here, the initial likelihood function was taken to be uniform over a finite stimulus range and zero elsewhere. Also, the initial starting estimate of the threshold was always the midrange point. The model of the subject was simply a uniform pseudorandom number generator, together with the assumed psychometric function. However, the true value of m was varied relative to the starting midrange value. The assumed value of k was taken to be equal to the true value in all cases ($k=1$, for convenience).

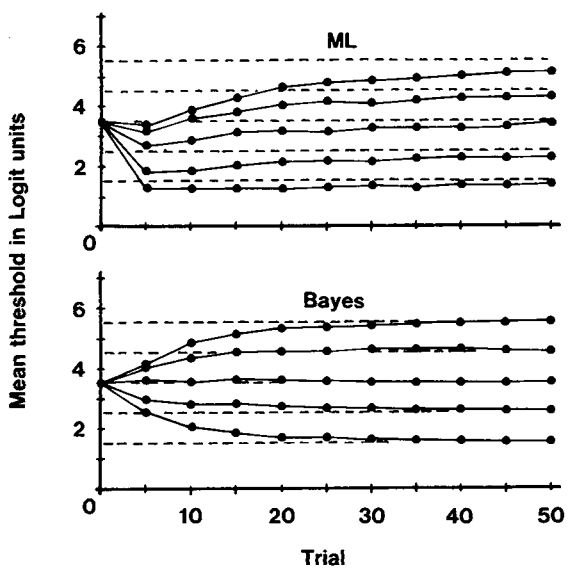

Figure 1. Mean threshold estimates for the two methods. Each array of connected dots is from 200 Monte-Carlo runs with the same true value of m. The five true levels are indicated by the horizontal dashed lines. The stimulus range was 7 logit units, and $\log L$ was represented over this range by a discrete array of 50 equally spaced elements. The random number generator was seeded differently for each set of 200 runs.

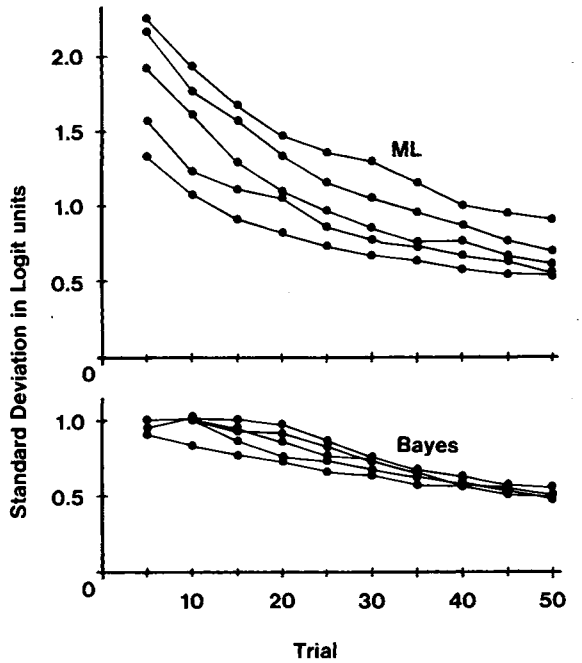

Figure 2. Standard deviations of threshold estimates for the two methods. These are from the same data whose means are plotted in Figure 1. The ML curves are ordered according to the true means, 5.5, 4.5, 3.5, 2.5, and 1.5, from top to bottom. The Bayes curves are ordered in that way at Trials 20 and 25, but they cross frequently, and this ordering is not maintained elsewhere.

Results

Figures 1 and 2 present the means and standard deviations of threshold estimates for five different values of the true threshold. A stimulus range of 7 logit units was used, and this range was broken into 50 intervals for the array of $\log L$. The range of 7 logit units would seem to be somewhat typical of real applications. Shelton and Scarrow (1984) specified a range of about 4.4, in terms of their assumed logit scale factor, but it is fairly clear that they underestimated the value of k for the conditions of their experiments by a factor of about 2.

The top graph of Figure 1 shows the general negative bias of the ML method. All runs were started with the initial estimate of the threshold at midrange, 3.5. It is not likely that other choices of the starting estimate would eliminate the negative bias. The bias is greatest when the true threshold is near the top of the range and is still of considerable magnitude after 50 trials. The bottom graph of Figure 1 shows the means from the Bayesian method. There is no apparent overall bias, negative or positive. When the true threshold is not equal to the initial midrange estimate, there is bias toward the midrange point. However, this bias becomes essentially negligible after

about 30 trials. Even at 20 trials, the biases with the Bayesian method are generally less than those after 50 trials with the ML method.

The standard deviations from the same runs are shown in Figure 2. As with the means, the worst cases for the ML method are those in which the true threshold is toward the top of the range. The sampling error when the true threshold is 5.5 is generally almost twice the sampling error when the true threshold is 1.5. For the Bayesian method, the sampling errors vary much less as a function of the true threshold. Also, the standard deviations are generally smaller than those from the ML method. For a true threshold of 5.5, the ML method yielded a standard deviation of .9 on Trial 50. For the same case, the Bayesian method achieved the same standard deviation by Trial 25.

It might be argued that the Bayesian method benefits more from the artifact of range truncation. That may be true, but it can account for only a minor component of the difference in performances of the two methods. If it were of major significance, one would expect much larger biases toward midrange with the Bayesian method. Moreover, the data of Figure 3 show that range extension does not eliminate the difference in performances. Indeed, the difference tends to increase markedly as the range increases. For the calculations with ranges greater than about 12 logit units, it was found necessary to use double-precision arithmetic (the Ecosoft C88 C compiler was used) to avoid the log0 artifact in calculating $\log P$ and $\log(1-P)$. Double-precision arithmetic further slows down the computations, but such ranges are probably larger than needed for most practical situations. The same log0 problem emerged, even with double-precision arithmetic, for ranges greater than about 36 logit units.

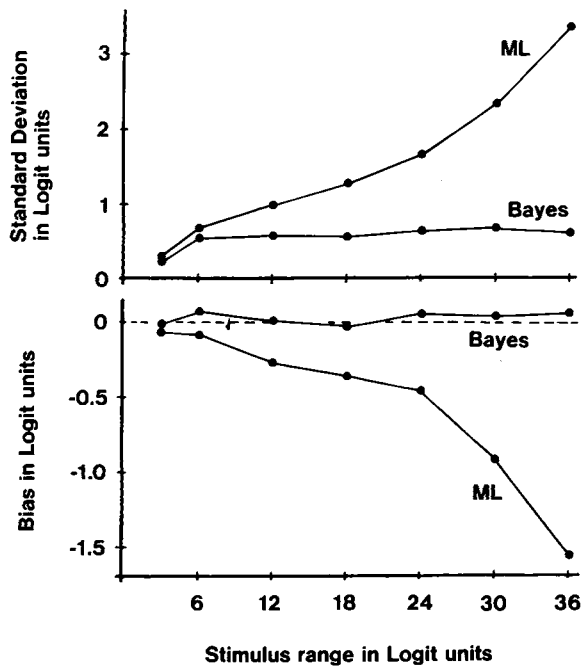

Figure 3. Biases and standard deviations of threshold estimates on the 50th trial, as functions of stimulus range, for the two methods. Each plotted point is from 200 runs. The true threshold was at the range midposition in each case, as was the starting estimate. The log$L$ function was represented by $10R$ equally spaced numerical elements, where R is the stimulus range. Double-precision arithmetic was used, and the random number generator was seeded differently for each set of 200 runs.

In summary, these data confirm and further illustrate the adverse sampling characteristics of ML estimation with 2AFC data. The data suggest, further, that the Bayesian method is almost devoid of these adverse sampling characteristics. The Bayesian method yields less systematic bias and generally less variability of estimates. The data of Figure 3 indicate that the ML method degenerates systematically and rapidly as the stimulus range is increased, whereas the Bayesian method does not.

Remaining Problems

These data may revive some hopes for a well-behaved statistical method of 2AFC sequential threshold estimation. However, further work is needed to achieve a routinely useful Bayesian method. First, one would like some convenient way to obtain usable estimates of k—perhaps a systematic formulation of heuristic techniques used by psychophysicists who routinely employ "assumed-slope" methods. Another aspect of this problem is the evaluation of the consequences of mismatches between assumed and true values of k. Very likely, this question is complicated by dependencies on the stimulus range used, unless very large stimulus ranges can be used.

The second main problem is the time-consuming computations of the Bayesian method outlined here. As is, the method is probably usable in real time only on some of the faster microprocessors now available.

REFERENCES

EMERSON, P. L. (1984). Observations on a maximum likelihood method of sequential threshold estimation and a simplified approximation. *Perception & Psychophysics*, **36**, 199-203.

GREEN, D. M., & SWETS, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

KERSHAW, C. D. (1985). Statistical properties of staircase estimates from two alternative forced choice experiments. *British Journal of Mathematical & Statistical Psychology*, **38**, 35-43.

MCKEE, S. P., KLEIN, S. A., & TELLER, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception & Psychophysics*, **37**, 286-298.

SHELTON, B. R. (1983). Rapid calculation procedures for the maximum likelihood method in adaptive psychophysics. *Behavior Research Methods & Instrumentation*, **15**, 87-88.

SHELTON, B. R., PICARDI, M. C., & GREEN, D. M. (1982). Comparison of three adaptive psychophysical procedures. *Journal of the Acoustical Society of America*, **71**, 1527-1533.

SHELTON, B. R., & SCARROW, I. (1984). Two-alternative versus three-alternative procedures for threshold estimation. *Perception & Psychophysics*, **35**, 385-392.

WATSON, A. B., & PELLI, D. G. (1983). QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, **33**, 113-120.

(Manuscript received January 3, 1986;
accepted for publication February 5, 1986.)