New Developments and Applications in Experimental Design
IMS Lecture Notes – Monograph Series (1998) Volume 34

EXACT DISTRIBUTIONS OF SEQUENTIAL THRESHOLD ESTIMATORS

BY NANCY FLOURNOY AND DOUGLAS G. KELLY

American University and University of North Carolina

We observe that, for any sequential procedure that is designed to estimate a parameter such as the threshold in a binary response experiment, the distribution of any estimator after n steps is discrete, with at most 2^n possible values. Furthermore, this distribution can be computed exactly, rendering simulations unnecessary. We compute and analyze the distributions of some estimators based on a certain up-down procedure, with a view to their dependence on the initial level and stepsize.

1. Introduction. Binary response experiments and psychometric functions. The ideas and methods considered here apply to any binary response experiment; that is, to any experiment having two possible outcomes, denoted 0 and 1, in which the value of a control variable x (set by the experimenter) determines the probability $\Psi(x)$ of response 1. The general design problem is to choose values of x so that the responses will allow efficient estimation of various features of the function $\Psi(x)$.

Such situations arise in a great many scientific areas. For concreteness we consider the context of psychophysics. In this setting, a stimulus can be delivered to a subject at different levels x, and responses 1 and 0 correspond to correct and incorrect identification, respectively, of some feature of the stimulus. For example, a brush may be stroked lightly on the subject's skin, at a fixed, constant velocity and pressure, and the length of skin traversed, as well as the direction of the brush stroke, can be varied. The subject is to attempt to identify correctly the direction of motion, and if the subject's probability of doing so is $\Psi(x)$ when the traverse length (level) is x, then $\Psi(x)$ is the subject's psychometric function. Another example is in hearing tests, where tones of a fixed frequency but different amplitudes x are delivered to a subject's ear, and response 1 corresponds to the tone's being audible to the subject.

In such studies it is assumed that $\Psi(x)$ is a monotone increasing function of x. It is also of course between 0 and 1, and it is usually assumed to be continuous; but it

---

Received October 1997; revised March 1998.
AMS 1991 subject classifications. Primary 62L12; secondary 62E15.
Key words and phrases. binary responses, sequential methods, up-and-down designs, psychometric functions.

need not be the cumulative distribution function (CDF) of a probability distribution, as its limits at $\pm\infty$ need not be 0 and 1. (It is worth noting that in some studies $\Psi(x)$ is decidedly not the CDF of any distribution. For example, in what is called a 2-alternative forced choice version of the brushstroke experiment described above, two stimuli are delivered with the same traverse length but opposite directions, and the subject is to identify which one was in a specified direction. For such an experiment $\Psi(x)$ has no values lower than 0.5.)

In psychophysics, a typical problem is to estimate the subject's threshold, which is usually defined as the value of x for which $\Psi(x) = 0.5$ [cf. Harvey (1986) and Wetherill (1963)]. We also refer to this as the quantile $q_{0.5}$ of Ψ. In toxicology, $q_{0.5}$ corresponds to the LD50, but one may be more interested in other quantiles such as $q_{0.05}$. In some psychophysical contexts the threshold is defined as the value of x maximizing the derivative $\Psi'(x)$; this may coincide with $q_{0.5}$, but for some commonly-assumed forms of Ψ it corresponds to some quantile other than $q_{0.5}$.

Another common psychophysical problem is the estimation of both the threshold and the maximum slope of $\Psi(x)$, a quantity that reflects the subject's sensitivity or "tuning." It is common to assume a parametrized psychometric function, $\Psi(x) = F(\beta(x - \mu))$, where F is some known function, often a symmetric CDF such as the standard normal or logistic. In such cases μ is the threshold (by any definition), and β is proportional to the maximum value of the slope. Sometimes the extreme-value CDF is used for F (when x is on a logarithmic scale, corresponding to a Weibull psychometric function for positive x in "real" units). In this case μ is taken to be the point of steepest slope, which is not equal to $q_{0.5}$.

We remark again that the subject of this paper does not depend on the context or language of psychophysics. We consider here the evaluation of arbitrary sequential procedures for estimating any scalar or vector parameter of the function $\Psi(x)$ that governs any binary response experiment.

2. Sequential procedures. Perhaps the most commonly-used sequential procedures for threshold estimation are loosely termed "up-down" or "staircase" procedures. The use of these for estimating $q_{0.5}$ was proposed by von Békésy (1947) and by Dixon and Mood (1948). The first stimulus is delivered at some prespecified level, and the next stimulus is delivered at a higher level if the response is 0 and a lower level if it is 1. The procedure is repeated, with some rule for determining the successive amounts of increase and decrease (the stepsizes). For the up-down procedure with a fixed stepsize h, Durham and Flournoy (1994) proposed a randomized version of the simple up-down procedure, for which they showed that, given a monotonic psychometric function, the delivered stimulus levels converge to a unimodal distribution whose mode is less than h from the target quantile. Giovagnoli and Pintacuda (1998) embed this procedure into a general class of procedures that do likewise, and show the Durham-Flournoy rule to be optimal in the sense that it is more peaked than others in the class. Giovagnoli and Pintacuda (1996) discuss these and other procedures in the light of optimal design theory.

Here, for the sake of concreteness, we examine an up-down procedure which is

frequently used by psychophysicists in at least one laboratory at the University of North Carolina. It is one of a class of procedures loosely termed PEST, for Parameter Estimation by Sequential Testing, by Taylor and Creelman (1967) and Taylor, Forbes and Creelman (1983). The particular procedure we consider here is actually called "PEST" by Gelfand (1990), but because of the vagueness of this term we call it the Adaptive Stepsize Procedure, or ASP.

There are two parameters in ASP: an initial stimulus level x_1 and an initial stepsize s_0. The stimulus level is decreased by the current stepsize after every correct (1) response and increased by the current stepsize after every incorrect (0) response. The stepsizes are adjusted as follows: the stepsize is divided by 2 whenever a 1 has been followed by a 0 or a 0 by a 1 (this is called a "reversal"), and it is multiplied by 2 whenever there have been three consecutive 1's or three consecutive 0's. In algorithmic form, ASP can be formulated as follows.

Procedure ASP(x_1, s_0) [x_1 =starting stimulus level, s_0 =starting stepsize]
    Set $n = 1$.
    Repeat until stopping rule is met:
        Deliver stimulus at level x_n.
        If latest three responses were 111 or 000, increase the stepsize:
            $s_n = 2s_{n-1}$;
        else if latest two responses were 10 or 01, decrease the stepsize:
            $s_n = s_{n-1}/2$;
        else
            $s_n = s_{n-1}$.
        If latest response was 1,
            $x_{n+1} = x_n - s_n$;
        else
            $x_{n+1} = x_n + s_n$.
    End.

The usual stopping rule (which will play no part in what we do here) is to stop when s_n first becomes smaller than some predefined size s_{end}.

Notice that because the stepsize is adjusted before a step is taken, the first stepsize that is actually used as an increment is s_1. However, s_1 always equals s_0, because the algorithm will not change the stepsize until at least two stimuli have been delivered. Hence we will refer later to the initial stimulus level and stepsize as x_1 and s_1, to avoid confusion.

Figure 1 shows an example of an ASP session that was arbitrarily stopped after 14 trials. The symbol + indicates a correct (1) response, and ∘ indicates an incorrect (0) response. The height of the symbol is the level of the stimulus. The figure shows that the first stimulus was delivered at $x_1 = 0$ and the response was 0. No change in stepsize can occur after the first step, so $s_1 = s_0$ and the stimulus level was increased by $s_1 = 0.5$. The second stimulus, at level $x_2 = 0.5$, elicited a correct response. Since the latest two responses were 01, the stepsize was halved to $s_2 = 0.25$; since the latest response was 1, the level was decreased by s_2. The process continued. Notice

that when there were more than three consecutive 1's, the stepsize was doubled after each succeeding 1, since in each case the latest three responses were 111. Similarly, when there were more than three consecutive 0's, the stepsize was doubled after each succeeding 0.

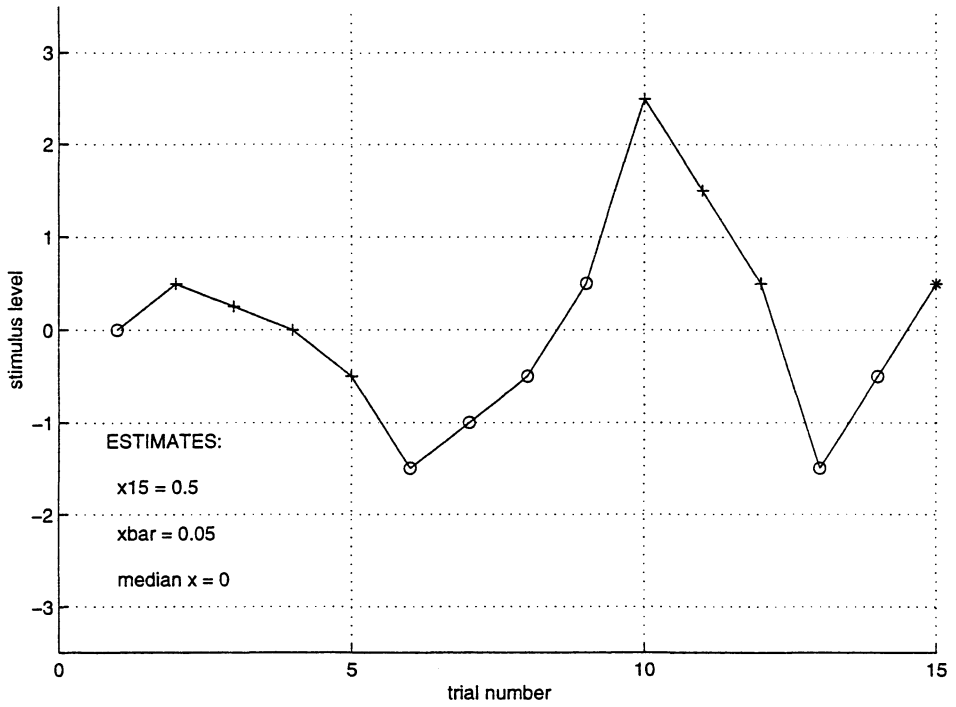

FIG. 1. *An ASP session with* $x_1 = 0$ *and* $s_0 = 0.5$.

Several threshold estimates are possible, based on the results of an ASP session. A common estimate is x_{n+1} (x_{15} in the example of Figure 1), the level of the next stimulus that would have been delivered had the session continued. Others are $\overline{x}$, the mean of all stimuli delivered, x_{med}, the median, and $\overline{x}_k$, the mean of all stimuli delivered in the last k runs, where a run is defined as a sequence of consecutive 1's (resp. 0's) preceded and followed by 0's (resp. 1's). (For example, the last three runs in the session pictured in Figure 1 comprise 10 stimuli, including the undelivered 15^{th}.) Still another threshold estimate that is often used is the mean of the last k reversal points. For the example of Figure 1, with $k = 4$, this would equal $(x_2 + x + x_{10} + x_{13})/4$. (For such an estimate, the stopping criterion would be that a certain even number of reversals have occurred.)

3. Exact distributions of estimates. To evaluate the performance of one of the threshold estimates described above for ASP—or any estimator of any parameter based on any sequential procedure—one thinks first of Monte Carlo simulation. One would simulate sessions in which the procedure is used on a subject with a known

psychometric function Ψ; the simulated subject would respond 1 with probability $\Psi(x)$ and 0 with probability $1 - \Psi(x)$ to any stimulus delivered at level x, independently of other stimuli in the session. (Two assumptions are being made here that are certainly open to question in real psychophysical experiments, but that are widely used in analysis of procedures. One is that the successive trials are independent, and the other is that the psychometric function does not vary with time.)

However, upon thinking about the results of such a simulation, one is led to the following simple observations. First, the set of possible values of the estimate after n steps is finite for any fixed n. For example, if n is only 2, there are just four possible values of the estimator $x_{n+1} = x_3$. If, as in the previous example, $x_0 = 0$ and $s_0 = 0.5$, then these four values are 2, 0.5, -0.5, and -2. In general, there are at most 2^n different values of the estimate, because each value is determined completely by the sequence of subject's responses, and there are 2^n such sequences. Finally, each response sequence also uniquely determines (along with the initial conditions, which in the case of ASP are x_1 and s_0) the sequence of stimulus levels $x_1, x_2, \ldots x_{n+1}$. In turn, these levels determine the probabilities of the responses themselves. The value of the final estimate (whether it be x_{n+1} or some other estimate), and its probability, are also determined uniquely by the response sequence.

The consequence of these observations is that *given a subject's psychometric function, any n-step sequential procedure has a finite probability space of 2^n outcomes whose probabilities can be computed exactly. Any parameter estimator based on the procedure is a random variable on this finite space, whose distribution is therefore known exactly.* The elements of the probability space are pairs (x, r) , where $r = (r_1, r_2, \ldots, r_n)$ is one of the sequences of n 0's and 1's and $x = (x_1, x_2, \ldots, x_n, x_{n+1})$ is the resulting sequence of stimulus levels. The probability of the pair (x, r) is then

$$P(x, r) = \prod_{1 \le j \le n, r_j = 1} \Psi(x_j) \cdot \prod_{1 \le j \le n, r_j = 0} (1 - \Psi(x_j)) \ .$$

The remainder of this paper presents some features of the exact distributions of threshold estimates based on the ASP procedure described above, stopped after 14 steps.

We remark here that the computation of exact distributions is not restricted to sequential procedures that are truncated after a fixed number of steps. The method can be used for a procedure with any stopping rule, in the sense that for any $\epsilon > 0$ one can compute the exact probabilities of a finite set of outcomes whose probability exceeds $1 - \epsilon$. One simply computes, for each n, the probabilities of all outcomes that have terminated in n or fewer steps, increasing n until the probabilities of all outcomes computed so far add to more than $1 - \epsilon$. In such cases, means and standard deviations of the distributions cannot be computed reliably because of the possible presence of very large values whose probabilities are not included. But more robust measures of location and scale will not suffer.

4. Exact distributions of ASP-based threshold estimates. Here we evaluate and compare estimates based on a 14-step ASP procedure, for a subject whose

psychometric function $\Psi(x)$ is the standard normal CDF $\Phi(x)$. For this psychometric function, $\mu = 0$ and $\beta = 1$. We consider four sets of starting conditions, determined by two values each of the starting stimulus level x_1 (0 and 3) and the starting stepsize s_0 (0.5 and 2).

We begin with the estimate x_{15}. Figure 2a is a histogram of the exact distribution of x_{15} for procedure parameters $x_1 = 0$ and $s_0 = 0.5$, and Figure 2b is the same for $x_1 = 0$ and $s_0 = 2$. Note that not every possible value of the distribution is within the range plotted (the total probability of all values depicted is given on the figure); nevertheless, the reported means and standard deviations of the distributions are correct for the complete exact distributions.

Comparing the two histograms in Figures 2a and 2b, one sees that in this case ($x_1 = 0$) where the initial stimulus level equals the subject's actual threshold, a smaller stepsize appears better than a larger one. The procedure is unbiased in either case, of course, but the larger stepsize produces an estimator with slightly higher variance. The estimate from the larger stepsize also has a greater tendency to assume values in the discrete set $\{0, \pm\frac{1}{4}, \pm\frac{1}{2}, \ldots\}$.

Figures 3a and 3b are similar to Figures 2a and 2b, except that now the starting stimulus level is $x_1 = 3$ while the subject's threshold remains 0. In this case, one sees that the larger stepsize is better, producing an estimate with smaller bias and smaller variance. Thus, based on these four combinations of x_1 and s_1, we see that deciding on an optimum stepsize will be difficult: smaller stepsizes appear to be better if the initial stimulus level is close to the true threshold, but larger stepsizes seem better otherwise.

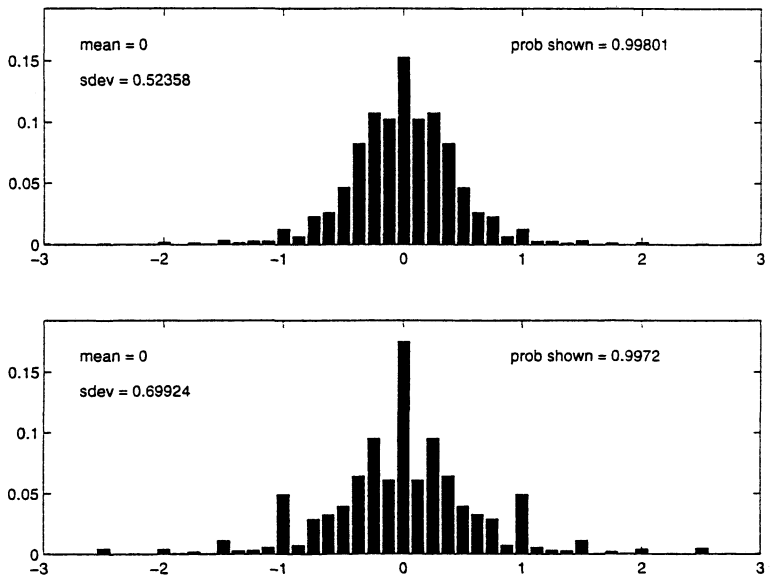

FIG. 2. *Distribution of x_{15} for ASP, 14 steps. Subject's $\mu = 0$, $\beta = 1$. Top: $x_0 = 0$, $s_0 = 0.5$. Bottom $x_0 = 0$, $s_0 = 2$.*

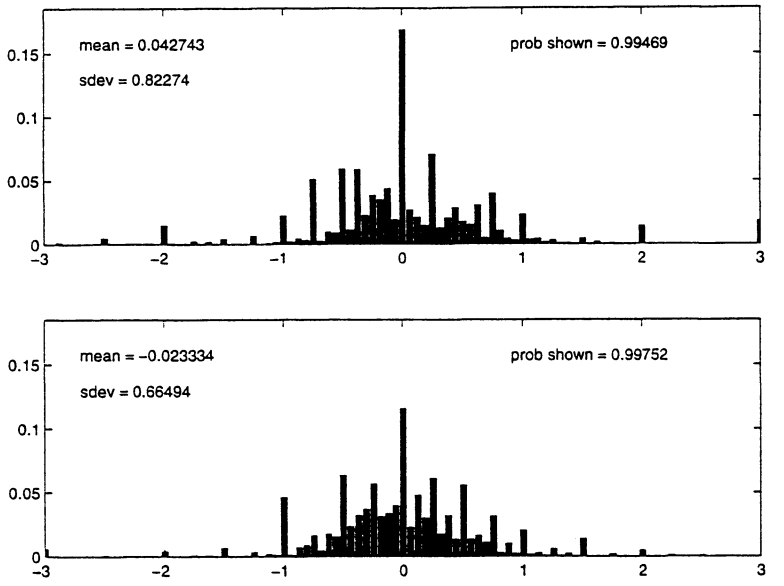

FIG. 3. *Distribution of x_{15} for ASP, 14 steps. Subject's $\mu = 0$, $\beta = 1$. Top: $x_0 = 3$, $s_0 = 0.5$. Bottom $x_0 = 3$, $s_0 = 2$.*

This is confirmed by the data in Table 1, showing the expected value, standard deviation, and root mean squared error of x_{15} for two starting values, $x_1 = 0$ and 3, and initial stepsizes in the range $s_1 = 0.25 : 0.25 : 2.00$.

These numbers appear to confirm that, if x_1 is at the subject's unknown threshold μ, then x_{15} is unbiased for all stepsizes, but that smaller stepsizes produce smaller mean squared errors than larger ones. On the other hand, if x_1 is not equal to μ, then the bias of x_{15} appears smallest when s_1 is a little less than half of $|x_1 - \mu|$, while larger stepsizes produce smaller mean squared errors.

In Figures 4 through 7 we study three different ASP-based estimates, for procedures truncated after n steps, where n ranges from four to 14. The three estimates are:

x_{n+1}, the level of the first undelivered stimulus (this is the estimate whose exact distribution is shown in Figures 2 and 3);

$\overline{x}$, the mean of the stimulus levels $x_1, x_2, \ldots, x_{n+1}$;

x_m, the median of the stimulus levels $x_1, x_2, \ldots, x_{n+1}$.

As before, the subject's psychometric function $\Psi(x)$ is the standard normal CDF $\Phi(x)$, so that the subject's threshold is $\mu = 0$ and the slope parameter is $\beta = 1$. Also as before, we studied the ASP procedure with four combinations of initial conditions, corresponding to initial level $x_1 = 0$ or 3 and initial stepsize $s_1 = 0.5$ or 2. Each of Figures 4 through 7 corresponds to one of these four sets of initial conditions for the procedure, and it plots, for $n = 4, 5, \ldots, 14$, the median and the 2.5 and 97.5 percentiles of the exact distribution of each of these three estimates.

TABLE 1

Expected values, standard deviations, and root mean square errors of the estimate x_{15} for the 14-step ASP, for starting stimulus levels $x_1 = 0$ and 3 and various initial stepsizes.

x_1	s_1	$\mathrm{E}(x_{15})$	$\mathrm{SD}(x_{15})$	$\mathrm{RMSE}(x_{15})$
0	0.25	0.0000	0.4293	0.4293
0	0.50	0.0000	0.5236	0.5236
0	0.75	0.0000	0.5650	0.5650
0	1.00	0.0000	0.5966	0.5966
0	1.25	0.0000	0.6285	0.6285
0	1.50	0.0000	0.6560	0.6560
0	1.75	0.0000	0.6789	0.6789
0	2.00	0.0000	0.6992	0.6992
3	0.25	0.0564	0.7497	0.7519
3	0.50	0.0427	0.8238	0.8250
3	0.75	0.0248	0.7738	0.7742
3	1.00	−0.0147	0.6977	0.6979
3	1.25	−0.0102	0.6820	0.6820
3	1.50	−0.0154	0.6873	0.6875
3	1.75	−0.0198	0.6815	0.6818
3	2.00	−0.0233	0.6653	0.6658

Notice that the data in these figures do not enable the establishment of confidence intervals for μ. They provide statements of the form

$$1 - \alpha = P\left[\text{est} - c_1(d) \leq \mu \leq \text{est} + c_2(d)\right],$$

where d is the unknown difference between the starting stimulus level x_1 and the subject's threshold, and the numbers $c_1(d)$ and $-c_2(d)$ are the 97.5 and 2.5 percentiles, respectively. Our calculations give the values of $c_1(d)$ and $c_2(d)$ for $d = 0$ and $d = 3$. Thus, if one is able to assume an upper bound on d, one can obtain conservative confidence bounds on μ using calculations like the ones shown here.

Looking at Figures 4 and 5, corresponding to $d = 0$ (initial stimulus level equals subject's threshold), we confirm that in this case all three estimates are unbiased, and we see that the smaller stepsize produces narrower confidence intervals. We further see that the confidence-interval widths based on $\overline{x}$ and x_m are nearly equal, and are smaller than those based on the "usual" estimate x_{n+1}.

Turning to Figures 6 and 7, for the case $d = 3$ (initial level differs from true threshold), we see again that a larger stepsize is better; the estimates tend to have less bias and smaller variance. It is less clear which of the three is preferred, however, as confidence intervals based on x_{n+1} will be slightly wider than those based on $\overline{x}$ or x_m, but the bias of x_{n+1} is much less than that of the other estimates.

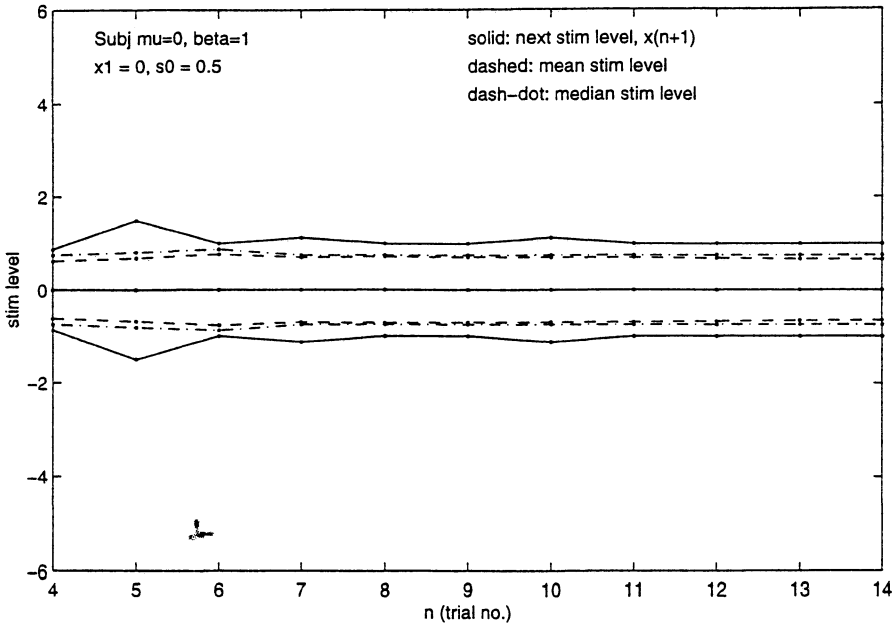

FIG. 4. *Median and 2.5% tails of ASP estimates.*

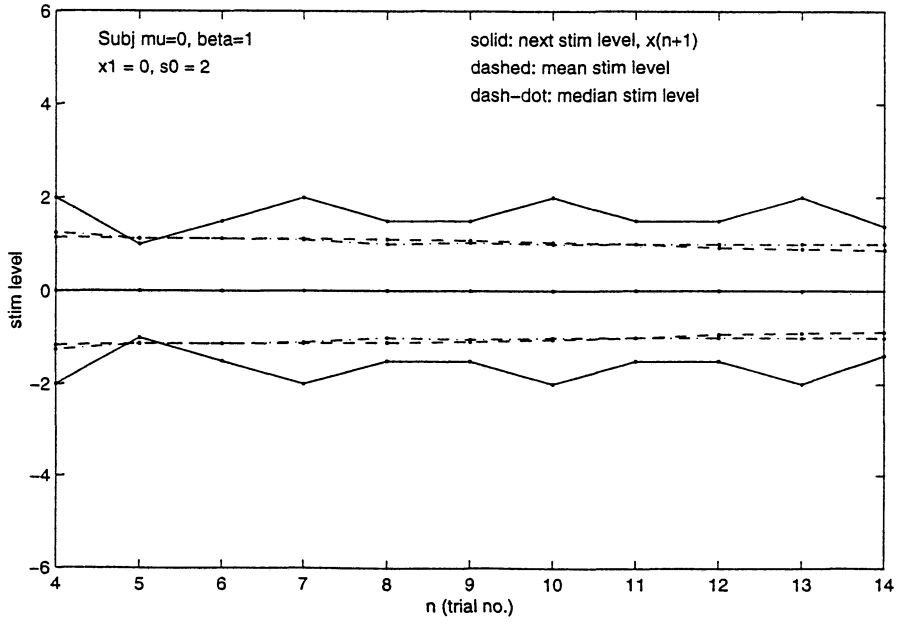

FIG. 5. *Median and 2.5% tails of ASP estimates.*

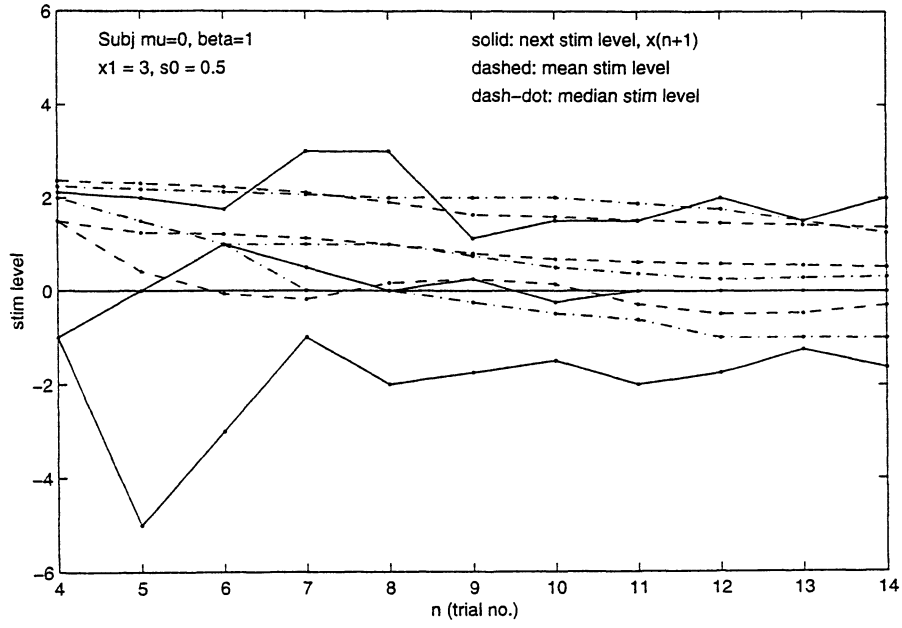

FIG. 6. *Median and 2.5% tails of ASP estimates.*

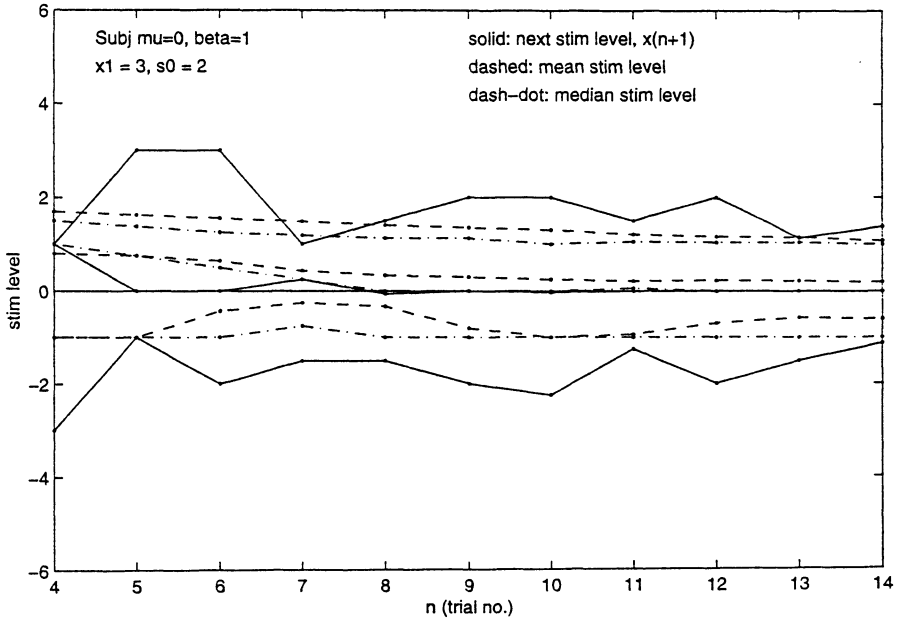

FIG. 7. *Median and 2.5% tails of ASP estimates.*

Notice that because the computed distributions are exact, one actually knows the bias of any estimate, except that again one knows it only as a function of the unknown distance $d = x_1 - \mu$. A full study of any estimate, based on any procedure, will require calculating its exact distribution for a range of values of d.

The authors plan such studies of the estimates x_{n+1} and x_m in the future, along with studies of other estimates, using procedures with larger numbers of steps. In particular, we plan studies of *sequential maximum-likelihood procedures* [see Hall (1968), Wu (1985), or Harvey (1986)]. In such a procedure one assumes the subject's psychometric function is $\Psi(x) = F(\beta(x - \mu))$ for some known function F. After a small initial set of stimuli, at levels chosen according to one of several rules, one computes the maximum-likelihood estimate (MLE) $\widehat{\mu}$ of μ, and then delivers the next stimulus at that level. Alternatively, one can find both $\widehat{\mu}$ and $\widehat{\beta}$ and deliver the next two stimuli at $\widehat{\mu} \pm c\widehat{\beta}$ where c can be chosen for D-optimality or c-optimality [see Kalish (1985)]. The process is iterated, computing new MLEs after each step, until some stopping condition is met.

5. Conclusion. In summary, we have pointed out that, rather than using Monte Carlo methods to investigate the distributions of estimates of parameters of psychometric functions by sequential procedures, one can find the distributions exactly. This is based on the simple observation that if the procedure is stopped after n steps, then there are 2^n possible outcomes and the distribution of the estimate can be computed exactly for any given psychometric function. If the procedure is not arbitrarily stopped after n steps, but has some other stopping rule, one can still find the exact distribution except on a set of probability ϵ. Planned future studies will investigate and compare the behaviors of several commonly-used threshold estimation procedures.

Acknowledgments. The authors thank the editor, William F. Rosenberger, and the anonymous referees for helpful suggestions which have improved this paper.

REFERENCES

VON BÉKÉSY, G. (1947). A new audiometer. *Oto-laryngology* **35** 411–422.

DIXON, W. J. AND MOOD, A. M. (1948). A method for obtaining and analyzing sensitivity data. *J. Am. Statist. Assoc.* **43** 109–126.

DURHAM, S. D. AND FLOURNOY, N. (1994). Random walks for quantile estimation. In *Statistical Decision Theory and Related Topics V* (S. S. Gupta and J. O. Berger, eds.), 467–476. Springer-Verlag, New York.

GELFAND, S. A. (1990). *Hearing An Introduction to Psychological and Physiological Acoustics.* Marcel Dekker, New York.

GIOVAGNOLI, A. AND PINTACUDA, N. (1996). Discussion of A.C. Atkinson: The usefulness of optimum experimental designs. *J. Roy. Statist. Soc. B.* **58** 98.

GIOVAGNOLI, A. AND PINTACUDA, N. (1998). Properties of frequency distributions induced by general "up-and-down" methods for estimating quantiles. *J. Statist. Plann. Inference*, in press.

HALL, J. L. (1968). Maximum-likelihood sequential procedure for estimation of psychometric functions. *J. Acoust. Soc. Amer.* **44** 370.

HARVEY, L. O., JR. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, and Computers* **6** 623–632.

KALISH, L. A. (1990). Efficient design for estimation of median lethal dose and quantal dose-response curves. *Biometrics* **46** 737-748.

TAYLOR, M. M. AND CREELMAN, C. D. (1967). PEST: Efficient estimates on probability functions. *J. Acoust. Soc. Amer.* **41** 782-787.

TAYLOR, M. M., FORBES, S. M. AND CREELMAN, C. D. (1983). PEST reduces bias in forced choice psychophysics. *J. Acoust. Soc. Amer.* **74** 1367-1374.

WETHERILL, G. B. (1963). Sequential estimation of quantal response curves. *J. Roy. Statist. Soc. B* **25** 1-48.

WU, C. F. J. (1985). Efficient sequential designs with binary data. *J. Amer. Statist. Assoc.* **85** 156-162.

NANCY FLOURNOY
DEPARTMENT OF MATHEMATICS AND STATISTICS
AMERICAN UNIVERSITY
WASHINGTON, DC 20016-8050, U.S.A.

DOUGLAS G. KELLY
DEPARTMENT OF STATISTICS AND DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NC 27599-3260, U.S.A.