Perception & Psychophysics
1990, 47, 127-134

Confidence intervals for the parameters of psychometric functions

LAURENCE T. MALONEY
New York University, New York, New York

A Monte Carlo method for computing the bias and standard deviation of estimates of the parameters of a psychometric function such as the Weibull/Quick is described. The method, based on Efron's parametric bootstrap, can also be used to estimate confidence intervals for these parameters. The method's ability to predict bias, standard deviation, and confidence intervals is evaluated in two ways. First, its predictions are compared to the outcomes of Monte Carlo simulations of psychophysical experiments. Second, its predicted confidence intervals were compared with the actual variability of human observers in a psychophysical task. Computer programs implementing the method are available from the author.

The performance of an observer in a detection or discrimination task is typically summarized by fitting a psychometric function to the data. Examples of fitting methods include probit analysis (Finney, 1971) and maximum-likelihood fits using the Weibull/Quick psychometric function (Quick, 1974; Watson, 1979; Weibull, 1951). These methods retain an estimate of threshold and a measure of variability (slope).

Whatever fitting method is used, some measure of the variability of the estimated psychometric parameters is needed in order to compare the performances of an observer in two experimental situations, or to compare the performance of an observer to a theoretically motivated value. This article presents a Monte Carlo method for computing the bias, standard deviation, and confidence intervals for maximum-likelihood estimates of the location parameter α and slope parameter β for the Weibull/Quick psychometric function (Quick, 1974; Watson, 1979; Weibull, 1951). The method is derived from Efron's parametric bootstrap and related work (Efron, 1979a, 1981, 1982, 1985). Section 2 contains a description and explanation of the method. Sections 3 and 4 report two evaluations of the method. The outcomes of both evaluations suggest that the parametric bootstrap provides useful estimates of bias, standard deviation, and confidence intervals for the location and slope parameters of the Weibull/Quick psychometric function.

There are many different psychophysical procedures (staircase methods, method of constant stimuli, etc.) to select among in designing an experiment (see Levine & Shefner, 1981), and either forced-choice or yes-no tasks may be used. Furthermore, there are several fitting procedures, among which two, probit analysis and the maximum-likelihood fit to the Weibull/Quick psychometric function, are most frequently employed. The parametric bootstrap method can be adapted to each of the combinations of experimental design and data analysis that could arise. To ease the presentation, the method is first applied to maximum-likelihood fits for a two-parameter Weibull/Quick psychometric function using the method of constant stimuli and assuming a forced-choice task. The changes needed to use the method in other circumstances are then described. The programs psifit, MOCSsim, and anlyz, implementing the method, are described in the Appendix and are available from the author (see Appendix).

For some fitting methods, such as probit analysis (Finney, 1971), approximate confidence intervals can be assigned to a single measured threshold. These intervals are based on the asymptotic normality of maximum-likelihood estimates (discussed in the next section), and they are valid only if "enough" data is collected for each psychometric function fitted. Analyses presented in the next section of this article for the maximum-likelihood fit to the Weibull/Quick psychometric function suggest that typical patterns and quantities of data collected in psychophysical studies do not always satisfy the assumptions of the asymptotic theory for the slope parameter. McKee, Klein, and Teller (1985) reached similar conclusions when using small numbers of trials to estimate threshold via probit analysis. Worst of all, the estimated confidence intervals tend to be smaller than valid confidence intervals, leading the experimenter to find differences where none exist.

THE WEIBULL/QUICK OBSERVER

This section summarizes notation for the Weibull/Quick psychometric function and the maximum-likelihood fitting procedure suggested by Watson (1979). Figure 1 illustrates the experimental parameters for a psychophysi-

I thank Stanley A. Klein, John Krauskopf, Michael S. Landy, and Neil Macmillan for valuable suggestions concerning the presentation. I thank Brian A. Wandell for suggesting this problem to me, and providing encouragement at each stage of the research. Requests for reprints should be sent to Laurence T. Maloney, Department of Psychology, New York University, 6 Washington Place, 8th Floor, New York, NY 10003.

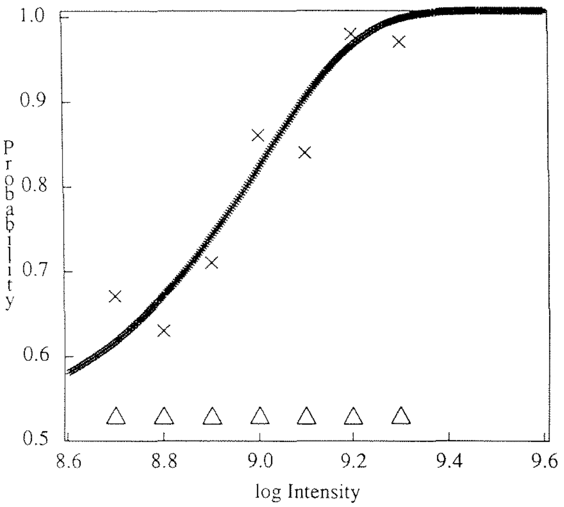

Figure 1. The Weibull/Quick Observer: The psychometric function (curve), intensities at which data is collected (△), and possible data (×).

cal measurement of threshold using the method of constant stimuli and a forced-choice procedure. The experimenter presents stimuli whose intensities are chosen from among a prespecified set of intensity levels, marked by small triangles on the log intensity axis. The observer judges the stimulus on each trial. The observer's performance is summarized by listing the intensities I_1, I_2, ... I_N, the number of trials at each intensity n_1, n_2, ... n_N, and the number of trials of the n_i where the observer succeeded, denoted c_i. If the trials are assumed to be independent, and the true probability of success p_i at level I_i is assumed to be constant over time, then the $3N$ numbers I_i, n_i, and c_i completely capture the observer's performance. Figure 1 records this performance graphically: the proportion of successes (×) is plotted versus intensity.

The observer's performance is then reduced to two parameters by fitting a psychometric function to the data (the solid line in Figure 1). In this paper, the two-parameter Weibull cumulative distribution function is used. Its equation is

$$p_D = 1 - e^{-(I/\alpha)^\beta},\ I \in [0,\infty), \qquad (1)$$

where p_D is the probability of correct detection at intensity I, and α and β are the parameters adjusted in fitting the data. The parameter α is a location parameter semilog coordinates; changing α shifts the curve to the left or right without changing its shape. The parameter β is a scale parameter in semilog coordinates that determines the slope of the function.

An observer succeeds in a two-alternative forced-choice (2AFC) trial with probability one half by guessing alone. Consequently, the probability of success is related to the probability of detection by $p_C = ½ + ½p_D$.

The fitting method used is maximum-likelihood estimation of the two parameters α and β from the data as described in Watson (1979). This procedure maximizes the combined likelihood of the N independent binomial outcomes c_i given n_i, I_i by choice of estimates $\hat{\alpha}, \hat{\beta}$. The program psifit described in the Appendix implements this fitting method.

The *bias* of the estimator $\hat{\beta}$ is the expected value of the discrepancy between $\hat{\beta}$ and β. Estimates of the slope of the psychometric function, for example, are biased in typical experimental conditions: Estimates of a slope whose true value is 2 could average around 2.1 for particular experimental conditions. Knowing the bias of an estimate is important when we compare an observer's performance against a theoretical prediction of that performance, derived, for example, from an ideal observer model.

The variance is $\mathbf{E}[(\hat{\beta} - \mathbf{E}(\hat{\beta}))^2]$, and the standard deviation (*SD*) is the square root of the variance. A 95% nonparametric confidence interval (NCI) is gotten by estimating the 2.5th percentile and the 97.5th percentile of the distribution of $\hat{\beta}$ (Efron, 1981). Bias, standard deviation, and confidence intervals are defined in an analogous fashion for $\hat{\alpha}$.

The bias, standard deviation, and confidence intervals for the maximum-likelihood estimates $\hat{\alpha}, \hat{\beta}$ depend in a complex way on the true values of α and β, the N intensity levels I_i, and the number of trials n_i taken at each intensity level. An alternative method for computing the standard deviation and confidence intervals (mentioned in the previous section) would make use of the remarkable fact that maximum-likelihood estimators are asymptotically Gaussian with mean value the true value of the parameter, and standard deviation computable in theory from the distribution (Cox & Hinkley, 1974, pp. 283-304; Mood, Graybill, & Boes, 1974, pp. 358-362; and, for the multiparameter case, Kendall & Stuart, 1979, pp. 59-64). If sufficient trials are taken, then these asymptotic results could in principle be used to estimate standard deviation and establish confidence intervals, but, in practice, the equations are too complex to solve analytically. In any case, asymptotic results cannot be used to estimate bias in the estimators, for maximum-likelihood estimators are asymptotically unbiased.

In Figure 2, we plot the distributions of $\hat{\alpha}$ and $\hat{\beta}$ for an observer whose performance is described by a Weibull/Quick psychometric function with $\log_{10}\alpha = 9.0$ and $\beta = 2.5$. The values of α and β chosen are realistic for an observer in many vision experiments (Wandell, 1985).[1] The number of different intensities is as in Figure 1: N is 7, and the intensities $\log_{10}(I_i)$ are taken to be 8.7, 8.8, 8.9, 9.0, 9.1, 9.2, and 9.3. The number of trials at each intensity is taken to be 30 (210 trials total).

Figure 2A is a histogram of fitted values of $\hat{\alpha}$ and 2B a histogram of $\hat{\beta}$, for 1,000 simulated replications of the experiment (computed using MOCSsim; see below and Appendix). The values of $\log_{10}\hat{\alpha}$ are approximately unbiased and normally distributed. Estimates of β are markedly skewed and non-normal. Asymptotic methods (inappropriate for estimating bias in any case) are not applicable to slope estimates in this experimental situation.

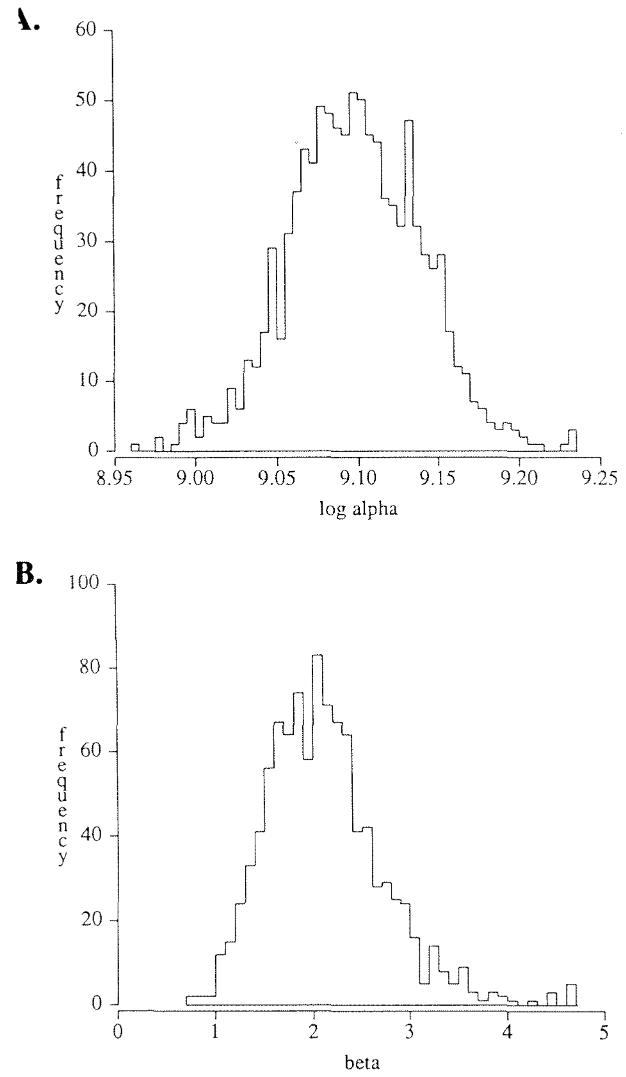

Figure 2. Repeated estimates of $\log_{10}\hat{\alpha}$ and $\hat{\beta}$, based on simulation of the observer and experiment specified in Figure 1 with 30 trials at each intensity.

About 500 trials total (70 per intensity level) are needed before the distribution of $\hat{\beta}$ is not obviously skewed. In conclusion, asymptotic methods are analytically intractable and not always appropriate for realistic experimental conditions.

In the next section, a method based on Efron's parametric bootstrap is described (see Efron, 1979a, 1981, 1982, 1985), which uses resampling and Monte Carlo simulation to estimate the bias, standard deviation, and confidence intervals for $\hat{\alpha}$ and $\hat{\beta}$ given I_i, n_i, c_i, $i = 1, 2, \ldots N$, the observer's measured performance.

ESTIMATING BIASES AND STANDARD DEVIATIONS OF PSYCHOMETRIC PARAMETERS

The experimenter's task is to estimate the values of threshold α and slope β that characterize a particular observer's performance in a particular experimental situation. Figure 3 represents the experimental situation. Values of α are plotted on the horizontal axis (logarithmic scale) and values of β on the vertical. The observer's psychometric function corresponds to some point on this plane, as yet unknown to the experimenter. The experimenter places forced-choice trials at N values of intensity, I_i, represented by solid triangles along the horizontal axis, and, on the basis of the outcome of these trials, estimates the location of the point corresponding to the observer's psychometric function. Suppose now that, unknown to us, the observer's psychometric function corresponds to the point marked by a ● in Figure 3. Assume that 120 2AFC trials are taken at each intensity marked by a triangle (840 trials total). The maximum-likelihood fitting procedure provides us with estimates of the two parameters, $\hat{\alpha}$ and $\hat{\beta}$, that can also be plotted as a single point on the plane in Figure 3.

Repeated measurements under these conditions would result in multiple estimates $\hat{\alpha}, \hat{\beta}$, suggested on the figure by the points marked with ×s. These values were obtained by Monte Carlo simulation of the experiment described: The program MOCSsim described in the Appendix takes as input a specification of an ideal Weibull/Quick observer (that is, α and β), as well as a specification of an experimental situation (intensity levels and number of trials). It then repeatedly simulates an experimental session, fits the data (in the same way as psifit does), and outputs the estimates $\hat{\alpha}, \hat{\beta}$.

The discrepancies between these estimates and the true values of the parameters have two sources: the variability of the estimation procedure (the spread of the cloud

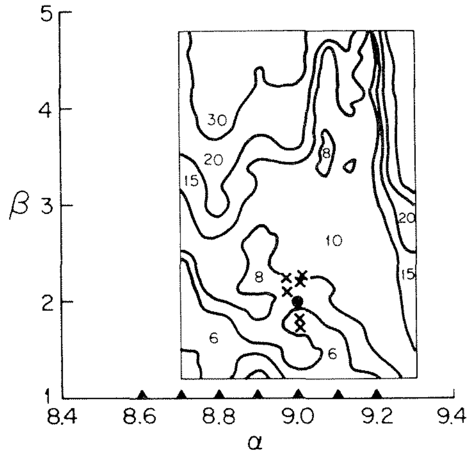

Figure 3. Log threshold α for an observer is plotted on the horizontal axis and slope β on the vertical. Each point corresponds to a possible psychometric function. The solid triangles (▲) on the horizontal axis represent intensities where the experimenter places 2AFC trials. The bullet (●) represents an observer's true psychometric function, and the ×'s represent possible estimates of the true values in the experiment where 120 2AFC trials are taken at each location marked by a triangle. The standard deviation of estimated β for each possible psychometric function is plotted as a contour plot. The numbers in each region represent the standard deviation of β times 100. See text for details.

of ×s) and the bias of the estimation procedure (the extent to which the cloud is not centered at the point α,β). As noted above, the bias and variability of these estimates depend in a complicated way on the number of trials taken and the position of the trials relative to the true location parameter α. But, if we knew the true values α and β for a given observer (which, of course, we never do), we could compute the bias and variability of the estimates obtained for that observer in a measurement by direct Monte Carlo simulation of that measurement. We could compute biases, standard deviations, or confidence intervals for $\hat{\alpha}$ or $\hat{\beta}$ (or confidence regions for the joint parameters), simply by simulating the experiment many times and computing the distribution of the estimated parameters based on the simulated outcomes. The estimated distribution of the parameters converges uniformly to the true distribution, since they are maximum-likelihood estimates (Cox & Hinkley, 1974, pp. 283-304; Kendall & Stuart, 1979, pp. 59-64; Mood et al., 1974, pp. 358-362). Consequently, with weak assumptions, measures derived from the simulated distribution, such as its variance, converge to the corresponding values for the true distribution. We can therefore compute values of interest, as, for example, the standard deviation of $\hat{\beta}$, for any point α,β in the plane.

Figure 3 is actually a three-dimensional contour plot. The third dimension, vertical to the page, is the standard deviation of $\hat{\beta}$ as a function of true α and true β when $\hat{\beta}$ is estimated in the 840-trial method-of-constant-stimuli experiment described above (Figure 2B). Details of the computation of the contour plot are described in the Appendix under MOCSsim. The numbers between the contours are this standard deviation times 100. For example, the point marked with a bullet (●) corresponds to a Weibull observer with $\alpha = 9$ and $\beta = 2$. Inspecting the contour plot, we note that the expected standard deviation of estimates of β returned by this observer over the course of a large number of repetitions of the 840-trial measurement is about 0.08. In contrast, consider the Weibull observer with $\alpha = 8.8$ and $\beta = 4$, whose standard deviation is about 0.30. The placement and spacing of trials in the 840-trial measurement does not permit as reliable a measure of β for this observer as for observer ●.

As mentioned above, the contours in this plot are determined by the number and spacing of trials. Note that the variability of the estimate is lowest at the bottom center of the figure. This outcome is expected for the following reasons: If the true value of α is much higher or lower than the bulk of the trial intensities, the estimate of $\hat{\beta}$ suffers. As β increases, the extreme trial intensities fall on regions of the psychometric function where they are either never seen or always seen and variability also increases. The two effects combine to produce the sharp rise in the "northeast" quadrant of the figure.

If we knew the true value of α,β for the observer in this experiment, we could read off the standard deviation of $\hat{\beta}$ from Figure 3. Instead, we know estimates of α and β obtained experimentally, $\hat{\alpha}$ and $\hat{\beta}$. The parametric bootstrap is used to estimate the value of the standard deviation $\hat{\beta}$ in the following way: Assume that $\hat{\alpha},\hat{\beta}$ are the true values for an observer, the bootstrap observer. Compute the standard deviation of estimates for this bootstrap observer by simulation. Use the computed estimates of bias and standard deviation of the bootstrap observer as estimates of the corresponding parameters for the true observer at α,β.

The contour plot serves the purpose of explaining the intuition behind the parametric bootstrap method. In Figure 3, the height of the contours above each of the ×s provides a good estimate of the height of the contours above the ●, which is the desired standard deviation of $\hat{\beta}$. If our estimates $\hat{\alpha},\hat{\beta}$ are "close enough" to the true values α,β, the estimated standard deviation based on $\hat{\beta}$ will be close to the desired value based on β. "Closeness" here depends critically on how flat the contour plot is in the vicinity of the true point. The flatter it is, the greater the error in the estimate $\hat{\alpha},\hat{\beta}$ that can be tolerated and still produce good approximations to the desired estimate of the standard deviation of β.

The justification of the procedure is the theorem mentioned above: For sufficiently many data points, $\hat{\alpha},\hat{\beta}$ converge to the true values α,β, since the estimates are joint maximum-likelihood estimates (Kendall & Stuart, 1979, pp. 59-64). Eventually, then, the point $\hat{\alpha},\hat{\beta}$ will very likely fall in a small neighborhood of the true point over which the contour surface of the standard deviation of $\hat{\beta}$ changes little.

Contour plots similar to those in Figure 3 can be computed for the standard deviation of $\hat{\alpha}$, for the bias of $\hat{\alpha}$, and for the bias of $\hat{\beta}$ (see Appendix). The four contour plots (including Figure 3) form an excellent summary of the statistical characteristics of a particular experimental design (number of trials, intensities), but, in actual use, no plots are involved in computing bootstrap estimates of the standard deviations and biases of the estimates of the parameters α,β. The parametric bootstrap uses the program MOCSsim to estimate the bias and standard deviations of the estimates of α and β directly by simulation. The actual procedure reduces to the following:

1. Do an experiment. Fit the data (psifit) to get estimates $\hat{\alpha},\hat{\beta}$ as usual.

2. Use MOCSsim to perform n Monte Carlo simulations (n in the range 500-1,000) of the original experiment (termed *bootstrap replications*) using $\hat{\alpha}$ and $\hat{\beta}$ in place of the (unknown) parameters α and β, but using the same selection of intensities and the same number of trials at each intensity as in the original experiment. After each replication, the simulated responses are fitted just as the original observer's data were fitted, obtaining bootstrap estimates α_i^* and β_i^* ($i = 1, 2, \ldots n$), one pair of estimates for each replication.

3. Compute the bias and standard deviations of the bootstrap estimates. The bootstrap estimate of the bias and the standard deviation of $\hat{\beta}$ are the bias and the standard deviation of the bootstrap estimates:

$$\text{Bias}^*(\hat{\beta}) = \bar{\beta}^* - \hat{\beta} \tag{3}$$

and

$$SD^*(\hat{\beta}) = \sqrt{\sum_{i=1}^{n} \frac{(\beta_i{}^* - \bar{\beta}^*)^2}{n-1}} \qquad (4)$$

here $\bar{\beta}^*$ is the mean of the bootstrap estimates. The cor-esponding quantities for α are defined analogously. The rogram anlyz described in the Appendix computes these alues and others from the output of MOCSsim.

Figure 4 illustrates the parametric bootstrap computa-ion. The true (unknown) values characterizing the psy-hometric function are plotted as a ●. A single ex-erimental measurement of $\hat{\alpha}, \hat{\beta}$ is plotted as an ×. This oint corresponds to the bootstrap observer. A few boot-trap replications are plotted as asterisks. The similarity etween the distribution of the bootstrap replication round the point × in Figure 4 to the distribution of the ×s around the point ● in the previous figure is the key dea behind the parametric bootstrap estimate. To the ex-ent that the estimates $\hat{\alpha}$ and $\hat{\beta}$ are close to the true values α, β, and to the extent that the "landscape" of the con-our plot near α, β is flat, the obtained bootstrap estimates re accurate. The parametric bootstrap is computation-lly intensive, like the maximum-likelihood Weibull es-imation procedure that it employs. Estimation of bias and onfidence intervals for psychometric functions using a UN 3/160 computer consumed approximately 5 min of PU time per function, considerably less time than was equired to collect the data initially. The use of such com-utationally intensive methods is now common in statis-ics (Efron, 1979b).

COMPUTATIONAL EVALUATION OF THE METHOD

When does the method work, and how acurate are the stimates?

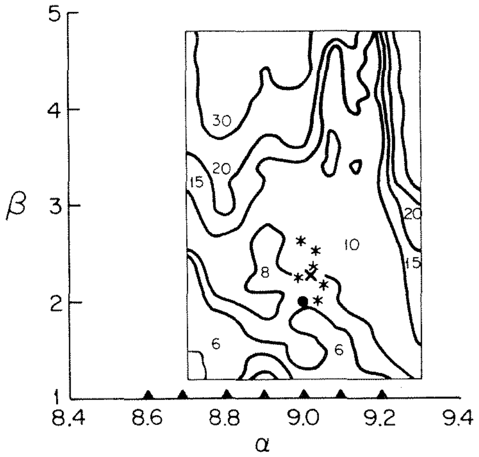

Figure 4. The axes are as in Figure 3. The point marked with a bullet (●) represents the true (unknown) psychometric function. The point marked with an × represents the outcome of a single experiment. The asterisks (*) represent bootstrap replications based on the experimental measurement.

Table 1
Estimates of Bias and Standard Deviation (*SD*), and Nonparametric Estimates of Variability, for the Location Parameter $\log_{10}\alpha$ in Simulated Experiments

Bias	*SD*	*IQR*	*NCI*/4
	Condition 1 (210 Trials)		
−0.0015	0.0410	0.0425	0.0408
−0.0026	0.0365	0.0365	0.0364
−0.0018	0.0468	0.0448	0.0474
−0.0014	0.0446	0.0441	0.0458
−0.0015	0.0429	0.0422	0.0417
	Condition 2 (420 Trials)		
−0.0002	0.0287	0.0293	0.0273
0.0011	0.0383	0.0373	0.0371
−0.0011	0.0281	0.0282	0.0272
−0.0007	0.0291	0.0282	0.0283
−0.0003	0.0305	0.0296	0.0293
	Condition 3 (700 Trials)		
−0.0002	0.0212	0.0210	0.0203
0.0003	0.0264	0.0261	0.0261
−0.0005	0.0182	0.0182	0.0175
−0.0001	0.0201	0.0197	0.0193
−0.0003	0.0196	0.0189	0.0192

Note—*IQR* = difference between upper quartile and lower quartile, divided by 1.35. *NCI* = difference between 97.5th percentile and 2.5th percentile. The three simulated experiments (Conditions 1-3) were done according to the method of constant stimuli, with seven equally spaced intensities as shown in Figure 1, and with 30, 60, and 100 trials per intensity level (210, 420, and 700 trials total) respectively. The correct values for each simulated experiment are given in the first row, followed by the results of four simulated applications of the parametric bootstrap to that experiment.

Direct Monte Carlo simulation as in Figure 3 allows us to evaluate how well the bootstrap estimate will do in a given experimental context. The program MOCSsim is the main tool used. The observer's true performance is assumed to correspond to the psychometric function plotted in Figure 1 ($\log_{10}\alpha = 9.0$, $\beta = 2.0$). MOCSim is then used to simulate an experiment using the method of constant stimuli and the intensities shown in Figure 1. The data from the simulated experiment is then fitted to obtain estimates $\hat{\alpha}, \hat{\beta}$, and the parametric bootstrap is applied to estimate bias and three measures of variability denoted *SD*, *IQR*, and *NCI*/4. *SD* is standard deviation as above. *IQR* is the difference between the upper quartile and the lower quartile divided by 1.35. With that scaling factor (1.35), the *IQR* will have the same average value as *SD* when the sampling distribution is Gaussian. When the sampling distribution is non-Gaussian, it is less sensitive to outliers. *NCI* is the difference between the 97.5th percentile and the 2.5th percentile, a nonparametric 95% confidence interval. *NCI*/4 will approximate *SD* for a Gaussian distribution. The bootstrap estimates, based on a single experiment, can be compared to the true values of *SD*, *IQR*, and *NCI*/4.

Tables 1 and 2 report the results of four simulated experiments for each of three experimental conditions. Condition 1 had 30 trials at each intensity level (210 total), Condition 2 had 60 (420), and Condition 3 had 100 (700).

Table 2
Estimates of Bias and Standard Deviation (*SD*), and Nonparametric Estimates of Variability, for the Slope Parameter β in Simulated Experiments

Bias	*SD*	*IQR*	*NCI*/4
	Condition 1 (210 Trials)		
0.1204	0.6058	0.5480	0.5958
0.1750	0.6482	0.5530	0.6652
0.0891	0.4940	0.4620	0.4882
0.1090	0.5590	0.5155	0.5567
0.1028	0.5553	0.4873	0.5357
	Condition 2 (420 Trials)		
0.0613	0.4013	0.3897	0.3876
0.0298	0.3336	0.3282	0.3296
0.0682	0.4022	0.3959	0.3901
0.0578	0.3920	0.3715	0.3831
0.0510	0.3793	0.3791	0.3760
	Condition 3 (700 Trials)		
0.0493	0.3021	0.2808	0.3015
0.0345	0.2493	0.2409	0.2362
0.0752	0.3751	0.3651	0.3646
0.0518	0.3205	0.2930	0.3213
0.0637	0.3720	0.3503	0.3830

Note—*IQR* = difference between upper quartile and lower quartile, divided by 1.35. *NCI* = difference between 97.5th percentile and 2.5th percentile. The three simulated experiments (Conditions 1-3) were done according to the method of constant stimuli, with seven equally spaced intensities as shown in Figure 1, and with 30, 60, and 100 trials per intensity level (210, 420, and 700 trials total) respectively. The correct values for each simulated experiment are given in the first row, followed by the results of four simulated applications of the parametric bootstrap to that experiment.

The entry labeled, for example, "Cond 1" shows the true *SD*, *IQR*, *NCI*/4 for that condition, and the *SD*, *IQR*, and *NCI*/4 for the true $\alpha = 9.0$, $\beta = 2.0$ observer. The four experimental simulations evaluating the bootstrap method follow the entry for each condition. Table 1 reports results for $\log_{10}\alpha$; Table 2 reports results for β. For example, in Table 2, Condition 1, the "true" value of *SD* for $\hat{\beta}$ is 0.6058, and the four bootstrap estimates are 0.6482, 0.4940, 0.5590, and 0.5553. Each of these represents an estimate of *SD* that could have been obtained in an experimental situation by using the parametric bootstrap. The bootstrap values of Bias, *SD*, *IQR*, and *NCI*/4 are usable estimates of the corresponding true values in both tables.

One measure of the usefulness of the bootstrap method is to ask how much effort on the part of the experimenter (and observer) is saved by using it. Suppose that, instead of using the bootstrap method, the experimenter decided to estimate the *SD* of $\hat{\beta}$ by replicating the experiment n times and computing the *SD* of the n estimates. How big would n have to be before the ratio of the empirically estimated *SD* was typically so close to 1 as the estimates obtained by the bootstrap method in Tables 1 and 2?

In the analysis that follows, the distribution of $\hat{\beta}$ is assumed to be approximately Gaussian. The ratio of an estimate of the standard deviation of a Gaussian random variable with samples of size n to the true *SD* is distributed as the square root of an F distribution with $(n-1, \infty)$ degrees of freedom. Assume that the smaller of the true *SD* and the estimated *SD* is placed in the denominator so that the ratio is greater than or equal to 1 (see Hays, 1988, chap. 9). In Tables 1 and 2, we have reported four replications of each simulated experiment. The largest ratio between the simulated parametric bootstrap *SD* and the true *SD* in Condition 1 is 0.0468/0.0410 = 1.14. The largest ratios for *SD* in the three conditions for Conditions 1, 2, and 3 for $\hat{\alpha}$ are 1.14, 1.33, and 1.24, respectively. For $\hat{\beta}$ in Table 2, the corresponding values are 1.33, 1.19, and 1.24.

The distribution of the maximum ratio of four samples from an F distribution with $F(\mathrm{n}-1,\infty)$ degrees of freedom is computable from the F distribution (it is not an F distribution).[2] The 75th percentile of the distribution of the maximum ratio is 1.33 for $n = 10$, 1.24 for $n = 20$. These results suggest that the estimate of *SD* obtained using the parametric bootstrap method replaces about 10 replications.

AN EMPIRICAL TEST OF THE METHOD

In this section, we test the applicability of the bootstrap method by comparing it directly to empirical data. A basic assumption of the bootstrap method is that we can treat a bootstrap replication as if it were a replication of the original measurement, at least for the purposes of computing means and standard deviations of α and β. Consequently, if we actually replicate the empirical experiment, requiring the observer to estimate multiple psychometric functions under identical conditions, the observed variability of the repeated empirical estimates $\hat{\alpha}, \hat{\beta}$, should match the estimated variability computed by means of the parametric bootstrap[3].

This section summarizes results of experiments in which observers were asked to repeatedly measure threshold for 2AFC detection of spectrally narrowband, 4°, 50-msec test lights against a 10° bright 510-nm field. Observers used the method of constant stimuli. Each observer repeated the measurements several times. The experiment involved only three test wavelengths: 440, 560, and 670 nm. Subjects made 30 trials at each of 7 intensity levels spaced at 0.1 log unit intervals around threshold (threshold was previously estimated by a staircase procedure). Details of the empirical procedure and motivation for the measurements are to be found in Maloney (in press).

The data were analyzed as follows: For each repetition by each observer at each wavelength in the experiment, the 210 trials were fit using psifit. This procedure provided several estimates of α and β for each observer and wavelength based on 210 trials each. Bootstrap estimates of bias and variability were based on the median value of obtained estimates for α and the median value of estimates of β (one estimate of bias and one estimate of standard deviation for each observer and each wavelength) using MOCSsim. The distribution of $\hat{\beta}$ is very skewed

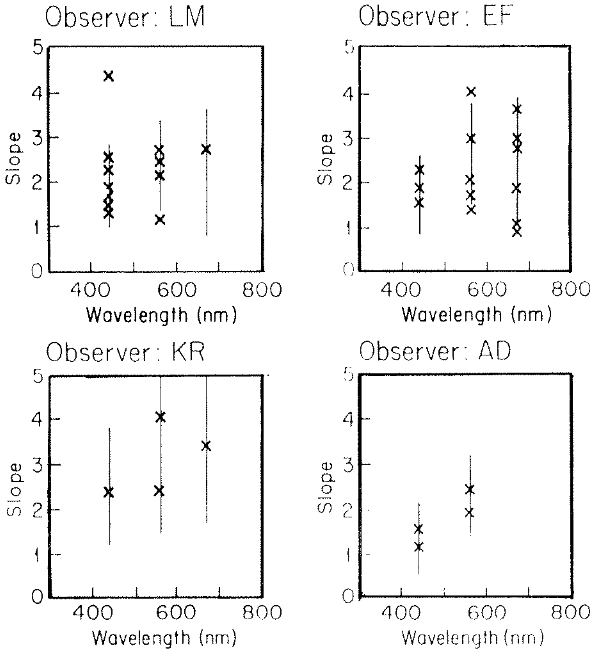

Figure 5. Estimates of β for each of four wavelengths for 4 observers. The vertical bars delimit the 95% confidence intervals computed using the bootstrap method on the median of the estimates for each observer, for each wavelength. The confidence intervals here are based on the 2.5th and 97.5th percentiles of the bootstrap replications, since the actual distribution of $\hat{\beta}$ is markedly skewed, with only 210 trials per psychometric function.

with so few trials, dictating the use of the median. For the same reason, the 2.5th and 97.5th percentiles of the bootstrap distribution were used as a 95% confidence interval (the confidence interval given by anlyz). These are difficult conditions under which to estimate β; consequently, they are good conditions under which to test the bootstrap method.

Figure 5 contains the estimated slopes from the experiment for each of three observers. The error bars are the bootstrap 95% confidence intervals about the unbiased median value of slope for each wavelength and observer.

Estimates of slope under these conditions are biased and very variable. Where a single observer measured multiple estimates, we can compare the variability of the data with the predicted variance gotten from the bootstrap by an *F* test (Hays, 1988, pp. 334-335). The confidence intervals agree with the actual variability of the observers, except for one point at 440 nm for observer L.M. The parametric bootstrap predicts the empirical variability of estimates of β under these experimental conditions. Estimates of α (not shown) were also in good agreement with predicted confidence intervals.

EXTENSIONS

The method is extendable to other psychophysical methods and models. Such extensions are most easily described in terms of changes to the programs psifit and MOCSsim:

nAFC methods. For forced-choice methods with *n* alternatives, $p_C = \frac{1}{2} + \frac{1}{2}p_D$ is replaced by $p_C = \gamma + (1-\gamma)p_D$ where $\gamma = 1/n$. The programs psifit and MOCSsim accept γ as an input parameter. No other changes are needed.

Yes-no rather than forced-choice tasks. Set γ to 0 if detection is assumed to result in a "yes" response. Otherwise, see Nachmias (1981) for a discussion of γ.

Other parameters. Experimenters may prefer to use i_δ, the intensity at which the observer is correct with probability δ, as an index of threshold. If program MOCSsim is modified to compute and print out $\hat{I}_\delta$ in place of $\hat{\alpha}$, anlyz will provide estimates of the bias and variability of $\hat{I}_\delta$ in place of the corresponding estimates for $\hat{\alpha}$.

A different family of psychometric functions. All that is needed is to change the computation of likelihood in psifit and MOCSsim to agree with the new family of functions. The new family may have more or fewer parameters than the two-parameter Weibull. Nachmias (1981), for example, suggests that γ may also be estimated from observers' data in yes-no tasks; it need not be assumed to be 0. Then the estimate $\hat{\gamma}$ can be assigned a confidence interval by the parametric bootstrap method.

Staircase methods. The computation in psifit is not affected by the order in which experimental trials were taken. However, in staircase methods, the choice of intensities does depend on performance during the experiment. The program MOCSsim, useful for method-of-constant-stimuli experiments, must be replaced by a program that simulates staircase experiments. The other programs are unaffected.

SUMMARY

Bootstrap methods and related sampling methods are now frequently used in statistics. They permit computation of statistical estimates that are analytically intractable or not adequately approximated by asymptotic methods. The parametric bootstrap presented here permits computation of estimates of bias, standard deviation, and confidence intervals for the parameters α and β of the Weibull/Quick psychometric function. It is readily adapted to other experimental conditions and choices of psychometric function.

Evaluations of the method for the ideal Weibull/Quick observer, and for human observers, suggests that it provides useful estimates of bias, standard deviation, and confidence intervals for the purpose of testing hypotheses concerning psychophysical performance using MOCSsim with 200-300 or more trials.

REFERENCES

Akima, H. A. (1978). A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software*, **4**, 148-164.

Becker, R. A., & Chambers, J. M. (1984). *S: An interactive environment for data analysis and graphics*. Belmont, CA: Wadsworth.

Chandler, J. P. (1975). *STEPT—Direct search optimization; solution of least squares problems* (Quantum Chemistry Program Exchange, QCEP Program No. 307). Bloomington, IN: Indiana University, Department of Chemistry.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: Chapman & Hall.

Efron, B. (1979a). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, **7**, 1-26.

Efron, B. (1979b). Computers and the theory of statistics: Thinking the unthinkable. *SIAM Review*, **21**, 460-480.

Efron, B. (1981). Nonparametric standard errors and confidene intervals (with discussion). *Canadian Journal of Statistics*, **9**, 139-172.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, **72**, 45-58.

Finney, D. J. (1971). *Probit analysis*. Cambridge: Cambridge University Press.

Hays, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

Kendall, M. K., & Stuart, A. (1979). *The advanced theory of statistics: Vol. 2. Inference and relationship* (4th ed.). New York: Macmillan.

Levine, M. V., & Shefner, J. M. (1981). *Fundamentals of sensation and perception*. Reading, MA: Addison-Wesley.

Maloney, L. T. (in press). The slope of the psychometric function at different wavelengths. *Vision Research*.

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.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**, 215-233.

Quick, R. F. (1974). A vector magnitude model of contrast detection. *Kybernetik*, **16**, 65-67.

Wandell, B. A. (1985). Color measurement and discrimination. *Journal of the Optical Society of America A*, **2**, 62-71.

Watson, A. B. (1979). Probability summation over time. *Vision Research*, **19**, 515-522.

Weibull, W. (1951). Statistical distribution function of wide applicability. *Journal of Applied Mechanics*, **18**, 292-297.

NOTES

1. The many parameters that affect the estimates can be simplified slightly. First, only the position of the intensities $\log_{10}I_i$ relative to the location parameter (in semilog coordinates), $\log_{10}\alpha$, matter. If a common offset is added to $\log_{10}\alpha$ and each of the intensities, the estimates $\hat{\alpha},\hat{\beta}$ are shifted. Second, β is a scale parameter: If the grid of intensities I_i is shrunk or expanded linearly around the point $\log_{10}\alpha$, the resulting change is equivalent to scaling β by the same factor. For convenience, the same values $\log_{10}\alpha = 9$, $\beta = 2$, with a method-of-constant-stimuli grid of test intensities at 8.7, 8.8, 8.9, 9.0, 9.1, 9.2, and 9.3, were used throughout this paper in examples and simulations. The values $\log_{10}\alpha = 10$, $\beta = 2$, and a grid of intensities at 9.7, 9.8, 9.9, 10.0, 10.1, 10.2, and 10.3 would produce identical results.

2. To compute the 75th percentile of the maximum of 4 samples from an F distribution with $(n-1,\infty)$ degrees of freedom, note that

$$0.75 = P[\mathrm{MAX} < x] = F(x,n-1,\infty)^4,$$

whenever $F(x,n-1,\infty) = 0.931$. For various values of n, we can compute x. The statistical language S was used to compute the values in the text (Becker & Chambers, 1984), taking infinity to be 500. $F(x,9,\infty) = 0.931$, for example, when $x = 1.78$. The value reported in the text is the square root of 1.78, 1.33.

3. If the observer's true slope changes from session to session, the effect will be to inflate the *SD* estimated from the empirical replications relative to the true values.

APPENDIX

Send requests for programs to Itm@kepler.psych.nyu.edu. Include a full electronic mail address and a full street address. Programs will be sent to the electronic mail address.

Currently available programs are written in FORTRAN 77 under SUN UNIX 4.3BSD. The FORTRAN 77 version of the programs psifit and MOCSsim use the STEPT 74 package (Chandler, 1975), obtained separately from The Quantum Chemistry Program Exchange, Chemistry Department, Indiana University, Bloomington, IN 47405.

1. *psifit*: Fits the two parameter Weibull/Quick psychometric function as described in Watson (1979). Output is estimates $\hat{\alpha},\hat{\beta}$, various thresholds, and a summary of the fit.

2. *MOCSsim*: Repeatedly simulates the performance of a Weibull/Quick observer with parameters α, β, and γ (Nachmias, 1981), and fits the two-parameter Weibull/Quick psychometric function to the data. Output is as many simulated $\hat{\alpha},\hat{\beta}$ pairs as desired. This program is also used to compute the parametric bootstrap described in the text.

   Random number generators: MOCSsim uses the *random* function available in UNIX 4.3BSD. Other uniform RNGs may be substituted as documented in the program text.

   Contour maps: Computation of the contour maps (Figures 3 and 4) is not part of the bootstrap computation. However, MOCSsim can be easily adapted to compute them if desired. In constructing Figures 3 and 4, the program MOCSsim was used to simulate 200 replications at each location and estimate $SD_{\hat{\beta}}$ at each of 100 locations, α_i, β_j, i = 1, 2, ... 10, j = 1, 2, ... 10. The values α_i, β_j were equally spaced in the rectangle drawn in Figures 3 and 4. The square grid of computed $SD_{\hat{\beta}}$ values was used to estimate the contour plot via Akima's method (Akima, 1978) using the *interp* function of the statistical language S (Becker & Chambers, 1984). Any other contour fitting program could be substituted.

3. *anlyz*: Accepts α,β and multiple $\hat{\alpha},\hat{\beta}$ pairs and computes several measures of variability. For each of $\hat{\alpha}$ and $\hat{\beta}$ these measures include:

   MEAN: mean of the estimates.

   BIAS: bias of the estimates.

   SD: estimated standard deviation about the mean.

   MEDIAN: median of the estimates.

   Q1, Q3: quartiles.

   IQR: interquartile range (Q3-Q1)/1.35. With the scaling factor 1.35, the reported value should approximate the *SD* when the distribution of the estimate is Gaussian. For non-Gaussian distributions, it is less sensitive to outliers.

   NCI: the estimated 2.5th percentile and 97.5th percentile, a nonparametric confidence interval (Efron, 1981).

(Manuscript received April 20, 1989;
revision accepted for publication July 17, 1989.)