Biol. Cybern. 57, 341–347 (1987)

Biological Cybernetics

Bootstrap Variance Estimators for the Parameters of Small-Sample Sensory-Performance Functions

D. H. Foster[1] and W. F. Bischof[2]

[1] Department of Communication and Neuroscience, University of Keele, Keele, Staffordshire ST5 5BG, England
[2] Alberta Centre for Machine Intelligence and Robotics, University of Alberta, Edmonton T6G 2E9, Canada

Abstract. The bootstrap method, due to Bradley Efron, is a powerful, general method for estimating a variance or standard deviation by repeatedly resampling the given set of experimental data. The method is applied here to the problem of estimating the standard deviation of the estimated midpoint and spread of a sensory-performance function based on data sets comprising 15–25 trials. The performance of the bootstrap estimator was assessed in Monte Carlo studies against another general estimator obtained by the classical "combination-of-observations" or incremental method. The bootstrap method proved clearly superior to the incremental method, yielding much smaller percentage biases and much greater efficiencies. Its use in the analysis of sensory-performance data may be particularly appropriate when traditional asymptotic procedures, including the probit-transformation approach, become unreliable.

1 Introduction

In the majority of sensory-performance measurements the typical finding is that the level of performance varies monotonically and nonlinearly with the level of the stimulus. In practice, the set of data relating stimulus level to performance level may be summarized by a single number, the *critical level* of the stimulus that yields a criterion level of performance. Thus when the response is discrete, referring say to the frequency with which a particular stimulus is detected, the critical level of the stimulus, in this case the *threshold*, may be defined as the level which corresponds to a detection frequency of 50%. The performance function is often called the *psychometric function* (see Fig. 1a).

In some situations, it may be possible to replicate the experiment and obtain several estimates of a parameter such as the threshold so that a mean value may be calculated. The reliability of such a point estimate is then typically provided by the variance or standard deviation calculated from the set of individual estimates. But, in other situations, replication of the experiment may be impossible. Given a single set of performance data, an estimate of the standard deviation of a parameter estimate derived from the data set may then be essential in assessing the significance of the parameter estimate, either absolutely or in relation to parameter estimates derived from other distributions. Additionally, the estimate of standard deviation may have an importance in its own right, particularly when there may be sensory pathology (compare Patterson et al. 1980). When replication of the experiment is possible, estimates of the standard deviations of individual parameter estimates may still be useful in forming the best (minimum-variance) estimate of the mean or in assessing the contribution of potential outliers to the mean.

Depending on the method used to fit a model sensory-performance curve to a single set of data, estimates of the variances of the parameters of the model may be derived by classical asymptotic theory. In particular, if $\hat{T}$ is the estimate of the parameter of interest, obtained as the solution of a maximum-likelihood equation, its estimated standard deviation $\widehat{\mathrm{SD}}$ is given by

$$\widehat{\mathrm{SD}} = [-1/(\partial^2 L/\partial T^2)]^{1/2},$$

where L is the likelihood and the partial derivative is evaluated at $\hat{T}$. It is this relationship that is used to estimate the variance of the midpoint ("ED50") and of the slope in the classical *probit-transformation* approach to analysing psychometric functions (Finney 1964). Some computer software packages routinely produce estimates of the standard deviations of parameter estimates derived from the Hessian matrix of 2nd-order partial derivatives of the function describing the goodness of fit.

Little is known, however, about the trustworthiness of these asymptotic formulae when the sample size is small. Substantial errors are certainly possible. For example, biases in the estimated standard deviation of the slope of a fitted performance curve derived in a standard probit analysis may be 20–30% for a full-range (0–100%) curve, such as that in cases 1 or 3, Table 1, based on 25 trials or less, and may become many times larger for a half-range (50–100%) performance curve, such as that in case 5, Table 1, based on 25 trials (see also McKee et al. 1985).

Recently, Bradley Efron has developed a "bootstrap" method for estimating the standard deviation of a point estimate of a parameter, or any other aspect of its distribution (Efron 1982; Efron and Tibshirani 1986). In essence, the method entails approximating the theoretical distribution of the empirical observations by what Efron refers to as the bootstrap distribution. This distribution is obtained by taking the empirical distribution of the data in the definition of the parameter and then repeatedly sampling the data, with replacement, to produce a Monte Carlo distribution of the resulting random variable. (A more formal definition is given in Sect. 2.)

The bootstrap method has been found to be very powerful. It requires few modelling assumptions and little analysis, and can be applied automatically to almost any situation (Efron 1982), the success of the method depending on replacing traditional theoretical analysis by computing effort. The purpose of this report is to demonstrate the use of the bootstrap approach to the problem of estimating standard deviations of the estimated midpoints and spreads of sensory-performance functions based on small data sets. As will be shown, the bootstrap method appears well suited to dealing with small samples, and is clearly superior to general asymptotic methods.

2 Estimation of Standard Deviation by the Bootstrap Method

Without loss in generality suppose that the performance measure is discrete, for example per-cent correct in a simple detection task, as illustrated in Fig. 1a. Let $(Y_1, Y_2, \ldots, Y_l)$ be the observed set of l scores measured at levels $x_1, x_2, \ldots, x_l$ of the stimulus. The scores Y_i each represent the proportion of r_i successes out of n_i trials, $i=1, 2, \ldots, l$, resulting from an unknown theoretical distribution F. Let $\hat{T}$ be the estimate of the parameter of interest (the midpoint α, say) derived from the observed data by some procedure g; thus $\hat{T}=g(Y_1, Y_2, \ldots, Y_l)$. The standard deviation $\sigma(F, \hat{T})$ of the statistic $\hat{T}$ cannot be written explicitly. Following Efron (1982), we use a Monte Carlo algorithm.

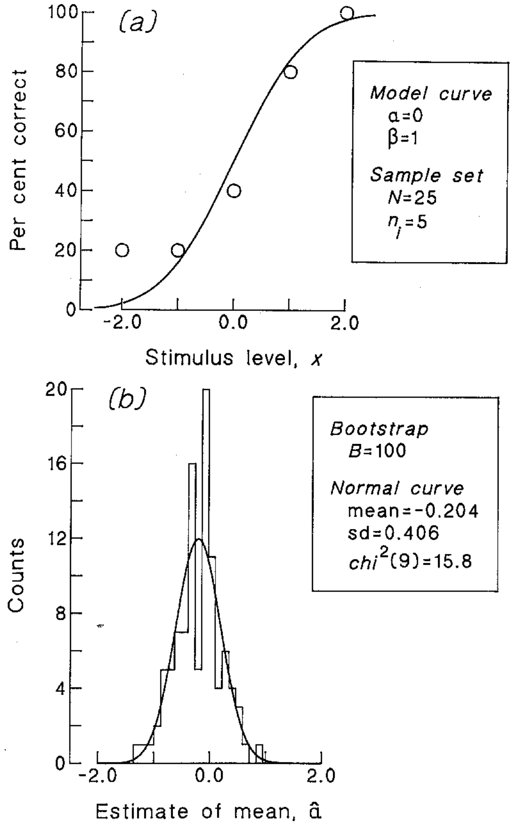

Fig. 1. a A performance curve based on a cumulative normal function [Eq. (3a) with $\gamma \to \infty$, $\alpha=0$, $\beta=1$] and a sample data set based on 5 trials at each stimulus level $x=-2, -1, \ldots, 2$. **b** Histogram of values of the estimated midpoint $\hat{\alpha}$ based on 100 bootstrap replications generated from the sample data set in **a**. The smooth curve is a normal curve with the same mean and standard deviation as the histogram

1. Construct $\hat{F}$, the empirical distribution of $(Y_1, Y_2, \ldots, Y_l)$, i.e., the distribution obtained by placing the rescaled binomial $\mathrm{Bi}(n_i, r_i/n_i)/n_i$, with n_i draws and probability r_i/n_i, at each level x_i, $i=1, 2, \ldots, l$.

2. Draw a bootstrap sample $(Y_1^*, Y_2^*, \ldots, Y_l^*)$ from $\hat{F}$ and calculate $\hat{T}^*=g(Y_1^*, Y_2^*, \ldots, Y_l^*)$.

3. Independently repeat step 2 a large number B of times, obtaining bootstrap replications $\hat{T}^{*1}$, $\hat{T}^{*2}, \ldots, \hat{T}^{*B}$, and calculate

$$\widehat{\mathrm{SD}}=\left[\frac{1}{B-1}\sum_{b=1}^{B}(\hat{T}^{*b}-\hat{T}^{*\cdot})^2\right]^{1/2}, \tag{1}$$

where $\widehat{\mathrm{SD}}=\sigma(\hat{F}, \hat{T})$, the bootstrap estimate of the standard deviation of $\hat{T}$, and $\hat{T}^{*\cdot}=\sum_{b=1}^{B}\hat{T}^{*b}/B$.

In the present Monte Carlo studies (Sect. 4), the number B of bootstrap replications was set at 100 (see

Efron 1982; Efron and Tibshirani 1986). Figure 1b shows a histogram of values of the estimated midpoint $\hat{\alpha}$ generated from the sample data set shown in Fig. 1a.

3 Estimation of Standard Deviation by the Incremental Method

The performance of the bootstrap estimator of the standard deviation was assessed against an alternative, general estimator obtained by an approximate method belonging to the classical study of the "combination of observations". Like the bootstrap method it does not depend upon a particular theoretical model and it involves little analytic effort. Its application in the present context is described more fully in Foster (1986).

With notation as in Sect. 2, consider the estimate $\hat{T}=g(Y_1, Y_2, \ldots, Y_l)$, and suppose that the estimated variances $\hat{\sigma}_i^2$ of the Y_i, $i=1,2,\ldots,l$, are not too large. Then, provided that some additional conditions are satisfied (including the smoothness of g and the independence of the Y_i), the estimated standard deviation $\widehat{\mathrm{SD}}$ is given approximately by

$$\widehat{\mathrm{SD}}=\left[\sum_{i=1}^{l} (\partial g/\partial Y_i)^2 \hat{\sigma}_i^2\right]^{1/2}, \qquad (2)$$

where the partial derivatives $\partial g/\partial Y_i$ are evaluated at $(Y_1, Y_2, \ldots, Y_l)$ (Foster 1986). The $\hat{\sigma}_i^2$ are given by the usual binomial formula $Y_i(1-Y_i)/n_i$. It should be noted that if $Y_i=0$ or 1 the ith term contributes nothing to $\widehat{\mathrm{SD}}$. The method is also known as the *incremental method.*

4 Monte Carlo Studies

4.1 Data Simulation

The bootstrap and incremental methods for obtaining the estimated standard deviation $\widehat{\mathrm{SD}}$ were tested by applying them to simulated sensory-performance data generated in the following way. The underlying performance curve was assumed to be of the form of the traditional cumulative normal function

$$y_i=1/\gamma+(1-1/\gamma)(2\pi)^{-1/2} \int_{-\infty}^{z} \exp\left(-\frac{1}{2}u^2\right) du,$$
$$z=(x_i-\alpha)/\beta, \quad i=1,2,\ldots,l, \qquad (3a)$$

where the constants α and β define the midpoint of the curve and its spread [reciprocal of the slope at the midpoint except for the factor $(1-1/\gamma)(2\pi)^{-1/2}$], and the constant γ, when finite, corresponds to the number of alternatives in the MAFC task giving rise to the data. Thus, in the extreme, in a 2AFC task, the stimulus

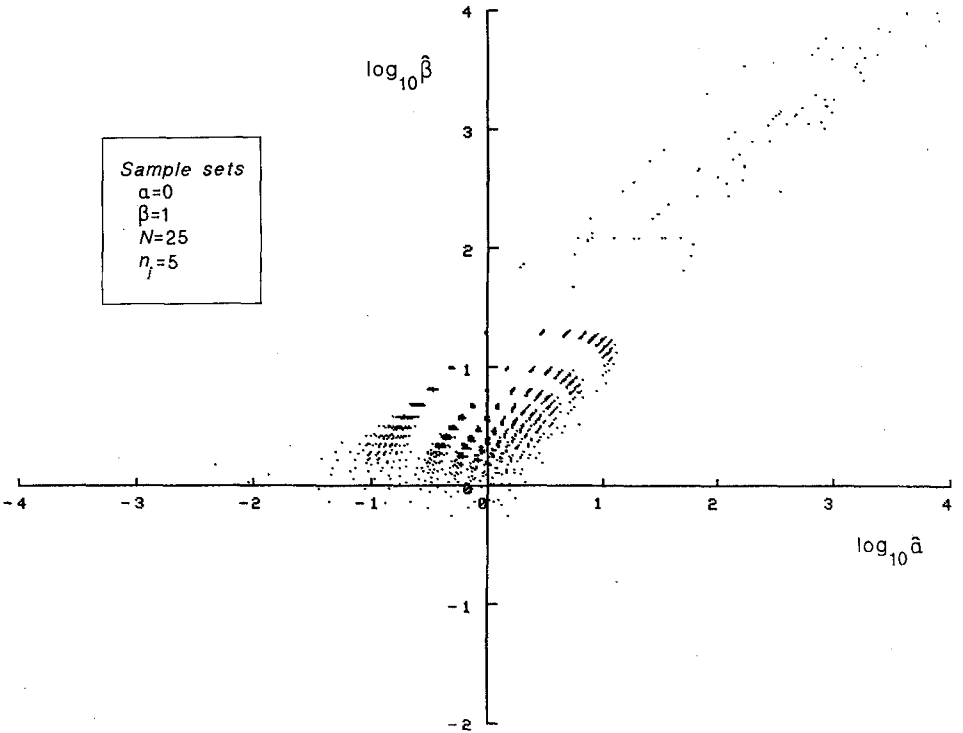

Fig. 2. Plot of $\log_{10}\hat{\beta}$ against $\log_{10}\hat{\alpha}$. Values of the estimated midpoint $\hat{\alpha}$ and spread $\hat{\beta}$, constrained to be positive finite, were obtained from the set of all data sets $(Y_1, Y_2, \ldots, Y_5)$ generated by Eqs. (3) with $\alpha=0$, $\beta=1$, $\gamma\to\infty$, $x_i=-2, -1, \ldots, 2$, and $n_i=n=5$ (as in Fig. 1a)

level x at the midpoint α corresponds to a performance level y of 75%, and, in a "yes-no" task where $\gamma \to \infty$, it corresponds to a performance level y of 50%. Given n_i trials at level x_i, sequences of scores Y_i were drawn from the corresponding rescaled binomial distributions

$$Y_i \sim \mathrm{Bi}(n_i, y_i)/n_i, \quad i = 1, 2, \ldots, l. \tag{3b}$$

In Fig. 1a, the performance curve had constants $\alpha = 0$, $\beta = 1$, and $\gamma \to \infty$, and the data set was derived with $x_i = -2, -1, \ldots, 2$, and $n_i = n = 5$. No special significance should, however, be attached to the present use of the cumulative normal curve. Differences between it and the logistic function, also often used to model sensory-performance curves, are small over most of the range.

For each set of simulated data $(Y_1, Y_2, \ldots, Y_l)$ generated by (3) for some $\alpha, \beta, \gamma, x_i, n_i, i = 1, 2, \ldots, l$, a curve of the form (3a) was fitted by maximizing the likelihood over α and β to obtain "new" estimates $\hat{\alpha}$ and $\hat{\beta}$. Details of the estimation method are given in Foster (1986).

With small data sets, there was an increased risk that values of $\hat{\alpha}$ and $\hat{\beta}$ for the fitted curve would take extreme values; in particular, $\hat{\alpha}$ could become positive infinite or negative infinite, and $\hat{\beta}$ negative, zero, or positive infinite. It was not sufficient, however, to exclude just these values. Figure 2 shows, for positive finite values of $\hat{\alpha}$ and $\hat{\beta}$, values of $\log_{10}(\hat{\beta})$ plotted against values of $\log_{10}\hat{\alpha}$ for the set of all data sets $(Y_1, Y_2, \ldots, Y_l)$ generated by (3) with $\alpha = 0$, $\beta = 1$, $\gamma \to \infty$, $x_i = -2, -1, \ldots, 2$, and $n_i = n = 5$. The total number of points plotted is 1878. Despite the restriction of $\hat{\alpha}$ and $\hat{\beta}$ to positive finite values, there are a number of extreme values that would have a destabilizing effect on the computation of the standard deviation of the midpoint or of the spread of the distribution. To avoid this problem, sample data sets yielding values of $\hat{\alpha}$ or $\hat{\beta}$ greater than 20 times the range of the stimulus were excluded. For the case illustrated in Fig. 2, the range $x_5 - x_1$ was 4.0, and data sets were therefore excluded if either $\log_{10}\hat{\alpha}$ or $\log_{10}\hat{\beta}$ exceeded 1.9. The proportion of points thereby eliminated was 4.2%.

For a given set of constants $\alpha, \beta, \gamma, x_i, n_i, i = 1, 2, \ldots, l$, characterizing the model, either 5000 or 10000 such sets of admissible data were generated, in turn yielding either 5000 or 10000 estimates of $\hat{\alpha}$ and $\hat{\beta}$. The standard deviations of these distributions were calculated and used as the "true" values of $\mathrm{Sd}(\hat{\alpha})$ and $\mathrm{Sd}(\hat{\beta})$.

4.2 Assessment of Standard-Deviation Estimators

The bootstrap and incremental methods were each tested by applying them to 1000 new "trial" sets $(Y_1, Y_2, \ldots, Y_l)$ of simulated data, a set at a time, generated as above but independently of that exercise. A modification was made, however, to the implementation of the algorithms of Sects. 2 and 3. For a given trial set $(Y_1, Y_2, \ldots, Y_l)$, the model curve was fitted to the data, yielding estimates $\hat{\alpha}$ and $\hat{\beta}$ for the midpoint and spread, and a replacement data set $(Y_1', Y_2', \ldots, Y_l')$ calculated, where $Y_i' = y_i$ defined by (3a) with $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$, the other constants remaining unaltered. Each score in the original trial set was thus replaced by its "smoothed" value on the fitted curve[1]. The algorithms were applied to these smoothed data sets.

As noted earlier (Sect. 2), bootstrap estimates of the standard deviation were each based on 100 bootstrap replications. Because of the sensitivity of the bootstrap method to extreme values, each Monte Carlo distribution (of the estimates of the mean and of the spread) was symmetrically 2-fold Winsorized, that is, the values $\hat{T}^{*b}$ in (1) were re-ordered linearly $\hat{T}^{*(1)}$, $\hat{T}^{*(2)}, \ldots, \hat{T}^{*(B)}$, and $\hat{T}^{*(1)}$, $\hat{T}^{*(2)}$ each replaced by $\hat{T}^{*(3)}$, and $\hat{T}^{*(B)}, \hat{T}^{*(B-1)}$ each replaced by $\hat{T}^{*(B-2)}$. Winsorization was essentially a safety measure, and, if omitted, would not have led to very large increases in estimated standard deviations[2]. Winsorization was preferred to trimming, for in addition to imparting robustness it made some use of the extreme values. More serious difficulties due to extreme values would have arisen, however, if variances rather than standard deviations had been estimated from the empirical distributions. Winsorization was not applied to the distributions used to estimate the true values of $\mathrm{Sd}(\hat{\alpha})$ and $\mathrm{Sd}(\hat{\beta})$ (Sect. 4.1). Although these values were based on 5000–10000 draws, they may therefore have been vulnerable to occasional extreme values.

For the incremental method, the partial derivatives in (2) were each estimated by finite-difference approximations, with averages of forward and backward differences being taken.

The principal measure of performance of the bootstrap and incremental estimators $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ and $\widehat{\mathrm{SD}}_{\mathrm{INC}}$ for the parameters $\hat{T} = \hat{\alpha}, \hat{\beta}$ was *percentage bias*, that is, the difference between the average of the

[1] This parametric version of the bootstrap was employed to avoid obtaining spuriously small standard deviation estimates from data sets in which several of the Y_i were zero or unity. Such data sets were particularly likely to occur when the total number of trials was small and stimulus levels were widely spaced

[2] Without Winsorization, the worst biases in the estimates given in Table 1 were in case 1, where the bias in $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{\beta})$ increased from 1.5% to 16%, and in case 2, where the bias in $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{\alpha})$ increased from 9% to 26%; all other biases were about the same as or less than those with Winsorization

estimate and the true value, expressed as a percentage of the true value, thus

$$[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{T}))-\mathrm{Sd}(\hat{T}))/\mathrm{Sd}(\hat{T})]\times 100,$$

and

$$[(\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{T}))-\mathrm{Sd}(\hat{T}))/\mathrm{Sd}(\hat{T})]\times 100.$$

Additionally, the *relative efficiency* of $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ with respect to $\widehat{\mathrm{SD}}_{\mathrm{INC}}$ was calculated as the inverse ratio of variances of the estimates, thus

$$\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{T}))/\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{T})).$$

Although large data sets were not investigated here, it was confirmed in separate studies that $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ and $\widehat{\mathrm{SD}}_{\mathrm{INC}}$ each behaved as consistent estimators[3].

The analysis was carried out for two different values of γ, two different numbers of levels l, and two different spacings of the levels, $x_1, x_2, \ldots, x_l$, the constants chosen to encompass typical experimental values. The values of the midpoint α, spread β, and number n_i of trials at each level were kept fixed, $\alpha=0, \beta=1, n_i=5, i=1,2,\ldots,l$. Computations were carried out in FORTRAN on two mainframe computers, a Cyber 176 and a CDC 7600, each with floating-point precision of 15 significant decimal digits. The NAG routine G05EYF was used to generate pseudo-random integers (Numerical Algorithms Group 1984).

4.3 Results

Table 1 shows the results of the Monte Carlo studies. Summary data are given for the estimates $\widehat{\mathrm{SD}}(\hat{\alpha})$ and $\widehat{\mathrm{SD}}(\hat{\beta})$ of the standard deviation of the estimated midpoint and spread, respectively, for the five different

Table 1. Comparison of bootstrap and incremental methods of estimating standard deviations of estimated midpoint $\hat{\alpha}$ and spread $\hat{\beta}$ of an underlying performance curve [Eq. (3)]. For each method, five Monte Carlo experiments were performed, each comprising 1000 trials $(Y_1, Y_2, \ldots, Y_l)$ distributed according to Eq. (3), with $\alpha=0, \beta=1, n_i=n=5$, and other constants as indicated. "True" values of the standard deviations were each based on 5000 or 10000 trials. Relative efficiency of the bootstrap estimator was determined with respect to the incremental estimator

	Bootstrap estimate $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{T})$				Incremental estimate $\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{T})$			True value $\mathrm{Sd}(\hat{T})$
	Ave	Std dev	% Bias	Rel Eff	Ave	Std dev	% Bias	
Curve range 0–100% ($\gamma\to\infty$) Sample size $\sum n_i=N=25$								
Model curve 1: number of levels $l=5$, $x_i=-2, -1, 0, 1, 2$								
$T=\alpha$	0.330	0.089	− 7.4	0.89	0.324	0.084	− 9.1	0.356
$T=\beta$	0.320	0.202	1.5	0.76	0.320	0.176	1.5	0.315
Model curve 2: number of levels $l=5$, $x_i=-1, -0.5, 0, 0.5, 1$								
$T=\alpha$	0.389	0.265	9.0	67	0.497	2.17	39.2	0.357
$T=\beta$	0.799	0.740	− 5.6	137	1.285	8.66	52.0	0.846
Curve range 0–100% ($\gamma\to\infty$) Sample size $\sum n_i=N=15$								
Model curve 3: number of levels $l=3$, $x_i=-1, 0, 1$								
$T=\alpha$	0.474	0.162	− 1.4	614	0.606	4.01	25.9	0.481
$T=\beta$	0.681	0.315	− 9.1	2336	1.32	15.2	76.8	0.749
Curve range 50–100% ($\gamma=2$) Sample size $\sum n_i=N=25$								
Model curve 4: number of levels $l=5$, $x_i=-1, -0.5, 0, 0.5, 1$								
$T=\alpha$	0.552	0.255	−21.1	87000	10.6	75.2	1410	0.700
$T=\beta$	0.824	0.466	−27.7	278000	35.2	246	2990	1.140
Model curve 5: number of levels $l=5$, $x_i=-2, -1, 0, 1, 2$								
$T=\alpha$	0.779	0.385	−13.5	3100	1.99	21.4	121	0.900
$T=\beta$	0.965	0.678	−19.9	12200	5.18	74.8	330	1.200

[3] For sufficiently large samples, the probability of the estimate being different from the "true" value could be made arbitrarily small; that is, for any small $\varepsilon>0$, $\mathrm{prob}\{[\widehat{\mathrm{SD}}(N)-\mathrm{Sd}]>\varepsilon\}\to 0$ as $N\to\infty$, where N is the sample size

model performance curves [4]. Results have been divided according to curve range and sample size. The variations of the true values of the standard deviations (last column in Table 1) may be noted. Decreasing the number l of stimulus levels but keeping constant the spacing between levels and the number n_i of trials performed at each level (cases 1 and 3) led to increases in the values of $\mathrm{Sd}(\hat{\alpha})$ and $\mathrm{Sd}(\hat{\beta})$. Changing the value of γ so that the curve range contracted from 0–100% ("yes-no" task) to 50–100% (2AFC task) also led to increases the values of $\mathrm{Sd}(\hat{\alpha})$ and $\mathrm{Sd}(\hat{\beta})$ (cases 1 and 5, and cases 2 and 4). Reducing the spacing between levels for $\gamma \rightarrow \infty$ and for $\gamma = 2$ (cases 1 and 2, and cases 5 and 4) had small but opposite effects, which presumably resulted from the introduction of respectively larger and smaller binomial variances at the altered stimulus levels.

5 Discussion

The superiority of the bootstrap method is evident in Table 1. For all the full-range model curves (cases 1–3), percentage bias in $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{\alpha})$ and $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}(\hat{\beta})$ ranged from -9.1% to 9.0%, including case 3 where the total number N of trials was 15. Percentage bias for the incremental method ranged from -9.1% to 76.8%, the latter occurring in case 3 for $\widehat{\mathrm{SD}}_{\mathrm{INC}}(\hat{\beta})$, although there were also large values in case 2. For the half-range model curves (cases 4 and 5), the performance of both methods deteriorated, but much more seriously for the incremental method. For the bootstrap method percentage bias ranged from -13.5% to -27.7%, whereas for the incremental method it ranged from 121% to 2990%. The relative efficiency of the bootstrap method with respect to the incremental method was also generally high, particularly for the half-range model (cases 4 and 5) where values ranged from 3100 to 278 000.

For bootstrap estimation with some parameters, for example the off-centre criterion level ED75, the distribution of the estimator may be sufficiently skewed that the standard deviation no longer provides an appropriate summary of its variability. Percentiles of the bootstrap distribution might then be used to estimate confidence limits for the true parameter value, although the number of bootstrap replications may have to be increased (Efron and Tibshirani 1986).

It is worth emphasizing that the application of the bootstrap method to estimating the distributional characteristics of sensory-performance functions is not tied to the use of either the cumulative normal curve (3a) or the binomial distribution (3b). The method should be equally useful for other sigmoidal performance functions, including the logistic function, and, more generally, the power-law increment-threshold function (Foster 1986).

Although not explored here, a characteristic consequence of using the bootstrap method (or any Monte Carlo method) to analyse an individual set of empirically generated data is that replication of the analysis need not yield precisely the same value of the estimated standard deviation. Provided that the model curve is an appropriate representation of the underlying performance and that the relative efficiencies given in Table 1 offer a reliable guide, the differences in these values should be small. Simulation errors can of course be reduced by increasing the number B of bootstrap replications, or by introducing techniques that increase the efficiency of the simulations (e.g., Davison et al. 1986).

The numbers of trials associated with each model curve were here constrained to range from 15 to 25. Most experimental designs (including fixed-levels methods and adaptive procedures such as PEST; Taylor and Creelman 1967, and hybrid PEST; Hall 1981) would usually specify more than 25 trials overall, but there are some circumstances where it may be possible to perform only a few trials, for example, when estimating a threshold or spread for a sensory system whose characteristics are changing fairly rapidly, as in some sensory-adaptation paradigms, or when for other reasons there is little time for measurement, as in some clinical situations. Bootstrap estimation of the standard deviations of these parameters appears to offer an acceptably reliable method of assessing their significance.

A FORTRAN listing of the main programs used in this study is available from the first author on written request.

Acknowledgements. We thank A. C. Davison, A. M. Herzberg, and S. R. Pratt for critical review of the manuscript. WFB was supported by Grant No. 81.166.0.84 from the Swiss National Science Foundation.

References

Davison AC, Hinkley DV, Schechtman E (1986) Efficient bootstrap simulation. Biometrika 73:555–566

Efron B (1982) The Jackknife, the bootstrap and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38; Philadelphia, PA, Society for Industrial and Applied Mathematics

Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statist Sci 1:54–75

Finney DJ (1964) Probit analysis. Cambridge University Press, Cambridge

[4] Average values of $\hat{\alpha}^{*\cdot}$ and $\hat{\beta}^{*\cdot}$ did not differ from their "true" values by more than 0.009 and 0.094 respectively in cases 1–3, and by more than 0.163 and 0.211 respectively in cases 4 and 5

Foster DH (1986) Estimating the variance of a critical stimulus level from sensory performance data. Biol Cybern 53:189–194

Hall JL (1981) Hybrid adaptive procedure for estimation of psychometric functions. J Acoust Soc Am 69:1763–1769

McKee SP, Klein SA, Teller DY (1985) Statistical properties of forced-choice psychometric functions: implications of probit analysis. Percept Psychophys 37:286–298

Numerical Algorithms Group (1984) FORTRAN library manual, Mark 11, vol 5. Numerical Algorithms Group, Oxford

Patterson VH, Foster DH, Heron JR (1980) Variability of visual threshold in multiple sclerosis. Effect of background luminance on frequency of seeing. Brain 103:139–147

Taylor MM, Creelman CD (1967) PEST: Efficient estimates on probability functions. J Acoust Soc Am 41:782–787

Received: May 29, 1987

Dr. D. H. Foster
Department of Communication and Neuroscience
University of Keele
Keele
Staffordshire ST5 5BG
United Kingdom

Verantwortlich für den Textteil: Prof. Dr. W. Reichardt, Max-Planck-Institut für biologische Kybernetik, Spemannstr. 38, D-7400 Tübingen. Verantwortlich für den Anzeigenteil: E. Lückermann, Springer-Verlag, Heidelberger Platz 3, D-1000 Berlin 33, Fernsprecher: (030)8207-0, Telex: 01-85411. © Springer-Verlag Berlin Heidelberg 1987. Druck der Brühlschen Universitätsdruckerei, Gießen. Printed in Germany. – Springer-Verlag GmbH & Co KG, 1000 Berlin 33