Spatial Vision, Vol. 11, No. 1, pp. 135–139 (1997)

Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions*,†

DAVID H. FOSTER[1,‡] and WALTER F. BISCHOF[2,§]

[1] *Department of Vision Sciences, Aston University, Birmingham B4 7ET, UK*
[2] *Department of Psychology, University of Alberta, Edmonton, Alberta, T6G 2E9, Canada*

Received 20 May 1996; accepted 25 September 1996

Abstract—Estimates of the accuracy of a threshold obtained from a psychometric function are often based on asymptotic theory. When the number of trials is small, however, these estimates may be untrustworthy. A computer program is described that uses a more reliable bootstrap approach to obtaining estimates of the standard deviation and confidence limits of a threshold and of the slope and spread of the psychometric function, for any criterion level of performance.

1. STATISTICAL ACCURACY OF THRESHOLDS

It is common practice in psychophysical experiments to obtain a set of data in the form of proportions of responses of a particular type, typically successes in some task, as a function of stimulus level. By fitting a model psychometric function to this empirical data set, one can estimate a *threshold* level of the stimulus corresponding to a particular standard or criterion level of performance, for example, 50% or 75% rate of success.[1] In some circumstances the *slope* or *spread* of the psychometric function at that criterion level may be of more interest.

Although the problems of choosing a model psychometric function and optimally sampling stimulus levels have received some attention in the literature, there has been less concern with the problem of determining the statistical accuracy of the threshold estimate or of other properties of the psychometric function. It may, of course, be possible to repeat the experiment and obtain several estimates of the threshold, in which case an estimate of statistical accuracy such as the standard deviation can be obtained in the usual way.

* http://www.op.umist.ac.uk/dhf.html
† http://www.cs.ualberta.ca/~wfb/
‡ d.h.foster@umist.ac.uk
§ wfb@ualberta.ca

Yet repetition is not always possible, or practicable. Whether it is or not, an estimate of the standard deviation of a single threshold estimate or some other measure of its statistical accuracy can be important for several reasons: (1) The standard-deviation estimate can be used to determine whether the threshold estimate is significantly different from that obtained from a different observer or under different experimental conditions. (2) In the course of a sequential testing procedure, where a threshold estimate is continuously improved as the number of trials increases, the current standard-deviation estimate can be used to define the point at which testing should stop. (3) More generally, when there are several possible psychometric functions available to fit the data, each yielding a threshold estimate, the corresponding standard-deviation estimates can be used to decide which psychometric function is the best, in the sense of yielding the threshold with the greatest statistical accuracy. (4) Even when repetition of the experiment is possible, standard-deviation estimates of individual threshold estimates can still be useful in forming the best estimate of the threshold, namely, the one obtained by weighting the individual estimates by the reciprocals of the squares of the estimated standard deviations.

One of the most popular ways for obtaining an estimate of a threshold and its standard deviation from an empirical data set is by *probit analysis* (Finney, 1952). In this method, the proportion of successes at each stimulus level is transformed by the inverse of a normal cumulative distribution function; a straight line is fitted by weighted linear regression; and estimates of the threshold and of its standard deviation are calculated from the probit-analysis formulae. (The principle of the probit method may itself be traced back to Fechner, 1860.) The probit-analysis formula for the standard deviation, however, is derived from classical asymptotic theory, and its trustworthiness is uncertain when the number of trials is not large (Finney, 1952, pp. 250–251, 1971, p. 57); thus, examples of substantial errors have been reported (McKee *et al.*, 1985; Foster and Bischof, 1987, 1991) with numbers of trials of the order of those often used in practice.

2. THE BOOTSTRAP

An alternative to classical asymptotic theory is offered by the *bootstrap* (Efron, 1982; Efron and Tibshirani, 1993). The bootstrap is a Monte-Carlo resampling technique, which, as Efron (1982) has emphasized, depends on replacing traditional theoretical analysis by computational effort. One of the advantages of the bootstrap in the present context is its potential accuracy with small numbers of trials (Hinkley, 1988).

In brief, the bootstrap approach proceeds as follows. Suppose the empirical data set consists of l proportions $y_1, y_2, \ldots, y_l$, at stimulus levels $x_1, x_2, \ldots, x_l$. Each proportion y_i is given by the number r_i of successes in n_i trials; that is, $y_i = r_i/n_i$, for $i = 1, 2, \ldots, l$. A 'bootstrap' data set $y_1^*, y_2^*, \ldots, y_l^*$ is generated by taking a random sample of size l, drawn with replacement, from the empirical data set $y_1, y_2, \ldots, y_l$ (the star notation indicates that $y_1^*, y_2^*, \ldots, y_l^*$ is not the actual data set $y_1, y_2, \ldots, y_l$, but a randomized, or *resampled*, version of it). How this sampling is done is explained later. Next fit the model psychometric function to the bootstrap data set. Then calculate a bootstrap estimate $\hat{t}^*$ of the unknown threshold t. Repeat this procedure a large number of times, say B times, to obtain estimates $\hat{t}_1^*, \hat{t}_2^*, \ldots, \hat{t}_B^*$ of the threshold. The bootstrap estimate $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$ of the standard deviation is then simply given (Efron, 1982) by the sample standard deviation of the B replications

$$\widehat{\mathrm{SD}}_{\mathrm{BOOT}} = \left[\sum_{b=1}^{B} (\hat{t}_b^* - \hat{t}_\cdot^*)^2/(B-1) \right]^{1/2},$$

where $\hat{t}_\cdot^*$ is the mean, $\hat{t}_\cdot^* = \sum_{b=1}^{B} \hat{t}_b^*/B$.

This estimate has been compared (Foster and Bischof, 1991) with estimates obtained by two other methods: an incremental method, which depends on forming a Taylor series approximation (Foster, 1986); and the original asymptotic method from probit analysis. The comparison was based on 12 kinds of psychometric function, for each of which 1000 data sets were generated by Monte-Carlo simulation. The quality of the estimates from the three methods was assessed by two measures. The first was the *percentage bias*, that is, the difference between the average of the estimate $\widehat{\mathrm{SD}}$ taken over the 1000 samples and the true standard deviation Sd, expressed as a percentage of the true standard deviation; for example, for the bootstrap estimate $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$, the percentage bias was $\{[\mathrm{Ave}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}}) - \mathrm{Sd}]/\mathrm{Sd}\} \cdot 100$. The second quality measure was the *relative efficiency*, that is, the inverse ratio of the variance of the corresponding estimates (the lower the variance, the higher the efficiency); so, for the bootstrap estimate $\widehat{\mathrm{SD}}_{\mathrm{BOOT}}$, the relative efficiency with respect to the probit estimate $\widehat{\mathrm{SD}}_{\mathrm{PROBIT}}$ was $\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{PROBIT}})/\mathrm{Var}(\widehat{\mathrm{SD}}_{\mathrm{BOOT}})$. The bootstrap was found to be superior to the other two methods, especially with small numbers of trials. For further details, see Foster and Bischof (1987, 1991). For an application of the bootstrap to psychometric functions based on the Weibull cumulative distribution function rather than on the normal cumulative distribution function, see Maloney (1990).

3. A BOOTSTRAP PROBIT PROGRAM

A computer program that uses the bootstrap approach to estimating the accuracy of a threshold obtained by probit analysis is available from the authors. Some details relevant to its implementation are as follows.

Suppose, as before, that the empirical data set consists of l proportions $y_1, y_2, \ldots, y_l$, at stimulus levels $x_1, x_2, \ldots, x_l$. Suppose that the number of trials performed at level x_i is n_i, where $i = 1, 2, \ldots, l$. The program fits by weighted linear regression a normal cumulative distribution function to this data set. Any lack of fit is indicated by the value of a chi-squared statistic. (This statistic, which should not be confused with the popular chi-squared test, is computed after transformation of the data by the empirical logistic transform, which is more appropriate than the 'crude' logistic transform when there are few trials; see Cox, 1970, Chapt. 3.) For a given criterion level of observer performance, the program gives an estimate of the threshold and of the slope and spread of the fitted function, and, for each of these quantities, an estimate of their standard deviations and of their 90% and 95% confidence limits. The bootstrap data sets from which these accuracy measures are calculated are based not on the original proportions y_i, but on their smoothed values $\hat{y}_i$. These smoothed values are simply the values of the fitted function at the stimulus levels x_i. Making this substitution ensures that none of the y_i is 0 or 1. Such extreme values are likely to occur with small numbers of trials and they contribute nothing to the standard-deviation estimate.

To generate each of the bootstrap data sets $y_1^*, y_2^*, \ldots, y_l^*$, rescaled binomial random-number generators $\mathrm{Bi}(n_i, \hat{y}_i)/n_i$ are assigned to the stimulus levels x_i. Thus, on each bootstrap replication, a sequence of proportions is generated: $y_1^* = r_1^*/n_1$, $y_2^* = r_2^*/n_2, \ldots, y_l^* = r_l^*/n_l$, where r_i^* is the number of successes in n_i bootstrap trials at level x_i for which the probability of success in a single trial is $\hat{y}_i$, the smoothed value mentioned earlier. Notice that the n_i (which are the same as in the original data set) may each be as small as 1, and that there is no need for the n_i to be the same or for the x_i to be evenly spaced. The number B of bootstrap replications is chosen to lie typically between 200 and 1000. In the sense that specific assumptions are made about the probability distributions underlying the observed data (a normal integral psychometric function with binomial generators), the bootstrap should here be referred to as the *parametric bootstrap* (Efron and Tibshirani, 1993, Sect. 6.5). A good parametric analysis, when appropriate, can be far more efficient that its non-parametric counterpart (Efron and Gong, 1983).

The source code of the program is written in ANSI C. It is freely available, with some additional documentation and two sets of test data, from the authors' web sites or by E-mail.

Acknowledgements

We are grateful to C. J. Savage and S. A. Randle for critical comments on the manuscript.

NOTE

[1] Some sequential testing procedures, such as staircase methods and PEST (Taylor and Creelman, 1967), provide a threshold without having to fit a psychometric function.

REFERENCES

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. In: *CBMS-NSF Regional Conference Series in Applied Mathematics.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am. Statist.* **37**, 36–48.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Fechner, G. T. (1860). *Elemente der Psychophysik.* Breitkopf und Härtel, Leipzig.

Finney, D. J. (1952). *Probit Analysis*, 2nd edn. Cambridge University Press, Cambridge.

Finney, D. J. (1971). *Probit Analysis*, 3rd edn. Cambridge University Press, Cambridge.

Foster, D. H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biol. Cybernet.* **53**, 189–194.

Foster, D. H. and Bischof, W. F. (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biol. Cybernet.* **57**, 341–347.

Foster, D. H. and Bischof, W. F. (1991). Thresholds from psychometric functions: superiority of bootstrap to incremental and probit variance estimators. *Psychol. Bull.* **109**, 152–159.

Hinkley, D. V. (1988). Bootstrap methods. *J. Roy. Statist. Soc.* **B50**, 321–337.

Maloney, L. T. (1990). Confidence intervals for the parameters of psychometric functions. *Percept. Psychophys.* **47**, 127–134.

McKee, S. P., Klein, S. A. and Teller, D. Y. (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Percept. Psychophys.* **37**, 286–298.

Taylor, M. M. and Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *J. Acoust. Soc. Am.* **41**, 782–787.