An Investigation of Confidence Interval Methods for the Thresholds and Slopes of Psychometric Functions

N. Jeremy Hill
Department of Experimental Psychology, University of Oxford, OX1 3UD, UK.

jeremy.hill@psy.ox.ac.uk

Introduction

Several studies over the last 15 years [1–5] have advocated the use of bootstrap methods rather than asymptotic-theory methods to compute confidence intervals for the threshold and slope of a psychometric function. The bootstrap comes in many forms, however, and one of the aims of the current research was to compare, using Monte Carlo simulation, the coverage properties of some of its variations.

A second aim was to test bootstrap methods under conditions in which *lapse rate* is taken into account, as a constrained variable parameter during the fitting process. The inclusion of this nuisance parameter is necessary in order to avoid biased threshold and slope estimates [6–7].

Method

Whereas a single bootstrap obtains a confidence interval by repeatedly simulating the behaviour of an observer, a Monte Carlo coverage test aims to simulate the entire experiment repeatedly, including the procedures carried out by the *experimenter* to estimate parameters and obtain confidence intervals. This process, described in more detail below, was used to investigate the accuracy of the following confidence interval methods :

Asymptotic methods

- MLE ±2 estimated standard errors from probit analysis,
- Finney's fiducial limits on thresholds, based on probit analysis (see ref [9], p.79).

Bootstrap methods

- MLE ±2 bootstrap standard errors,
- basic bootstrap,
- bootstrap percentile,
- parametric BC_a.

Where bootstrap methods were used, confidence intervals were based on 1999 parametric bootstrap replications, on each the 500 replications of the entire experiment. See Davison and Hinkley [8], chapter 5 for descriptions of the different bootstrap methods.

Monte Carlo Procedure for Testing Coverage

1. Start with a generating psychometric function, which will describe the true behaviour of a simulated observer in a 2AFC psychophysical experiment:

   $$\psi_{gen}(x) = \gamma_{gen} + (1 - \gamma_{gen} - \lambda_{gen})\, F_{gen}(x)$$

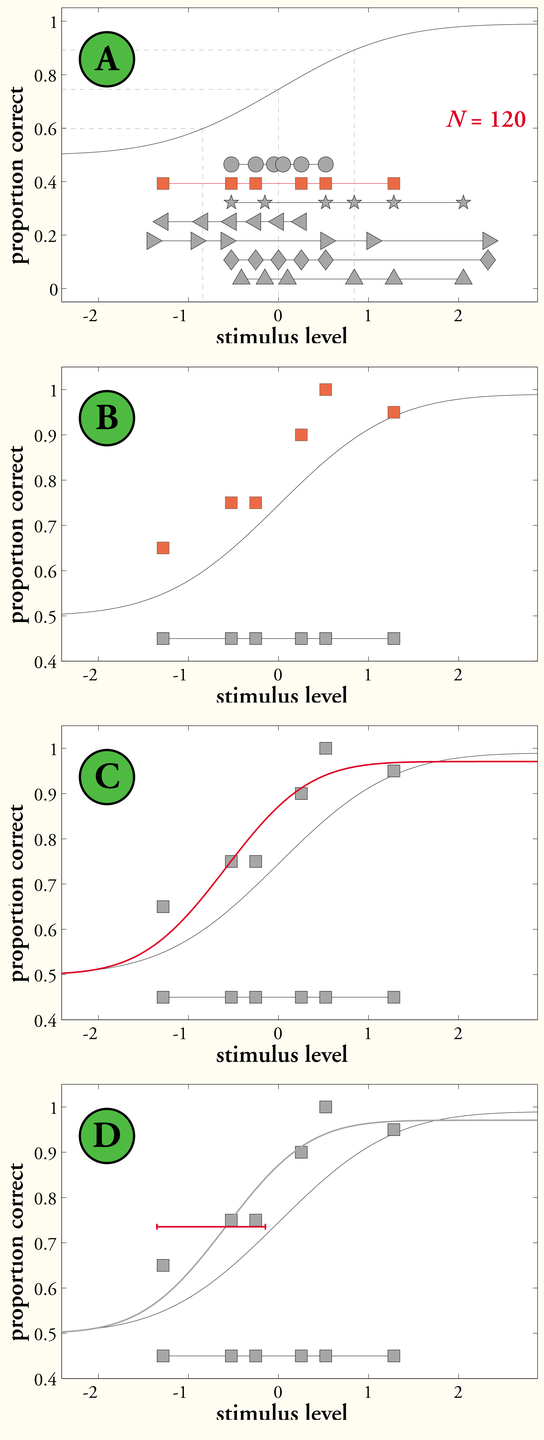
   where $F_{gen}(x)$ is the standard cumulative normal, $\gamma_{gen} = 0.5$, $\lambda_{gen} = 0.01$.
2. Choose a sampling scheme. Seven different schemes were tested, shown in figure A.
3. Choose the total number of trials N (to be divided equally between the six points). The values 120, 240, 480 and 960 were all tested.
4. Repeat 500 times:
   - The simulated observer performs the experiment: the vertical position of each red square in figure B is drawn from the appropriate binomial distribution.
   - The simulated experimenter fits a curve to the data, obtaining estimates of λ and F (red curve in figure C). A maximum-likelihood multi-parameter search method was used, as described by Wichmann and Hill [7]. γ was assumed to be fixed at 0.5, and λ was allowed to vary, but constrained to lie in the range [0, 0.05].
   - The simulated experimenter computes two-tailed 95.4% confidence intervals for the threshold and for the slope of F (based on its own estimate of F, because it does not know F_{gen}). A number of different confidence interval methods were explored (see above).
   - Record whether the true threshold or slope value falls within the interval, or whether it is falsely rejected in the lower tail, or in the upper tail. In figure D, the true threshold value has been rejected in the upper tail.
5. Let P_{LO} be the observed rate of rejections in the lower tail, and P_{UP} be the observed rate of rejections in the upper tail, in the set of 500 repetitions. Compute estimated **coverage**:

   $$c = 1 - P_{LO} - P_{UP}$$

   and estimated **imbalance**:

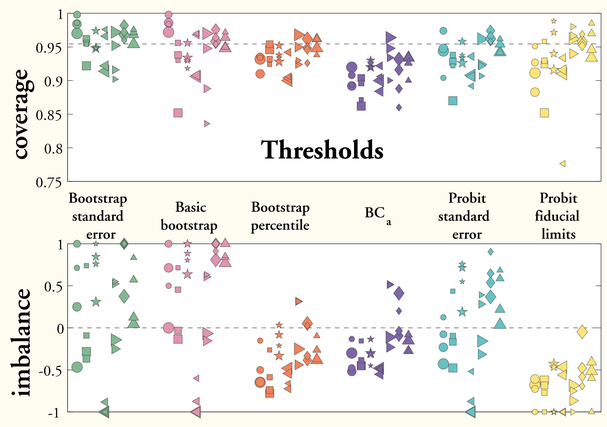
   $$a = (P_{LO} - P_{UP}) \,/\, (P_{LO} + P_{UP})$$

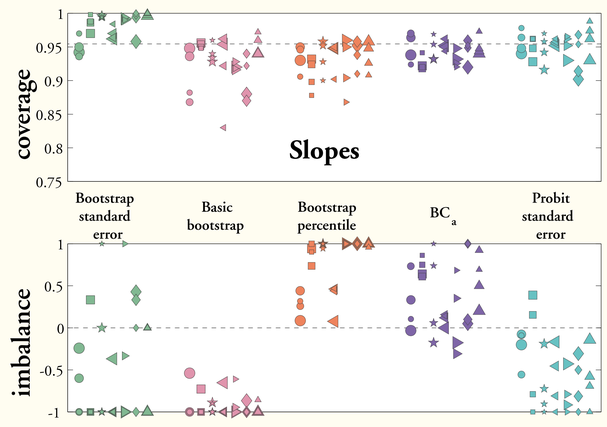
   If the confidence interval method is perfect, then the expected values are $c = 0.954$ and $a = 0$. If $a = -1$, interval limits are being set too low, to the extent that all rejections occur in the upper tail and none in the lower. If $a = +1$, limits are being set too high, with all rejections occurring in the lower tail.



Results

In the results figures, symbol shape denotes which of the seven sampling schemes was used, as per figure A, and symbol size denotes N (the smallest size corresponds to 120, then 240, 480 and 960). Note that the effects of increasing N vary according to sampling scheme and confidence interval method. However, it is not necessary to look closely at size and shape differences in order to appreciate the overall differences between confidence interval methods. The standard error of c is roughly 0.01 for most of the points.





Results for threshold confidence intervals

All methods were prone to variation in both coverage and imbalance, depending on sampling scheme and on N. The bootstrap percentile and BC_a methods were the least afflicted by such variations. Of the two, the BC_a method was better balanced, but had a tendency to produce confidence intervals whose coverage was too low.

The probit standard error method was more variable than the bootstrap percentile and BC_a methods, and yielded larger imbalance values. It was, however, as good as the two cruder bootstrap methods (bootstrap standard deviation and basic bootstrap), if not better, in this regard. Finney's fiducial threshold limits should probably be ranked worst, being highly imbalanced on the negative side.

Results for slope confidence intervals

All methods performed worse on slopes than on thresholds: they all tended to be imbalanced in one direction or the other, and in many cases the imbalance was large (often +1 or -1).

The BC_a method is less unbalanced than the others, but suffers slightly from low coverage, as was also the case for thresholds. The probit standard error method is almost as good, but it has a higher proportion of large imbalance values.

Concluding remarks

A good confidence interval method should be well balanced, and should minimize the differences in coverage and imbalance that are associated with variations in sampling scheme and in N. It should also display an accurate level of coverage, although inaccurate coverage may not be a fatal flaw provided the other two requirements are satisfied: if the method is well balanced, the experimenter could compensate for low coverage simply by increasing the target coverage level, or by employing a technique such as Wichmann and Hill's [10] expanded bootstrap method.

By these criteria, the BC_a method emerges as the best of the methods studied, for the calculation of 95% confidence intervals. However, performance on slopes was somewhat unsatisfactory for all methods, and the difference between the BC_a and the much less computationally expensive probit method was not large relative to the differences caused by variations in sampling scheme. This poor performance is partly ascribable to the inclusion of λ as a nuisance parameter: if $\lambda_{gen} = 0$, and λ is fixed at 0 during fitting, all results improve, with the BC_a overcoming its low coverage problem and emerging as clearly superior to probit methods on both thresholds and slopes (results not shown). Unfortunately, in experiments with real observers, it is necessary to take account of λ in order to avoid bias.

If slope is an important parameter in one's analysis, work currently in progress suggests that the use of two-dimensional bootstrap confidence *regions*, rather than separate one-dimensional threshold and slope intervals, greatly improves results, in particular reducing the differences between sampling schemes.

Software

Software for fitting psychometric functions, and for obtaining confidence intervals for thresholds and slopes, is available on the web. The source code can be compiled on MacOS, Windows9x, Linux and most other UNIX systems, to create either a standalone program, or a MEX file that interfaces with MATLAB®. Supporting MATLAB® m-files are also supplied, to aid user input and graphical visualization. Precompiled MEX-files are available for MacOS and Windows, and precompiled executables are available for MacOS, DOS and Digital UNIX.

http://users.ox.ac.uk/~sruoxfor/psychofit/

References

[1] **McKee, S. P, Klein, S. A. & Teller, D. Y.** (1985). Statistical properties of forced-choice psychometric functions: Implications of probit analysis. *Perception and Psychophysics*, **37**(4): 286–298.

[2] **Foster, D. H. & Bischof, W. F.** (1987). Bootstrap variance estimators for the parameters of small-sample sensory-performance functions. *Biological Cybernetics*, **57**(4–5): 341–7.

[3] **Maloney, L. T.** (1990). Confidence intervals for the parameters of psychometric functions. *Perception and Psychophysics*, **47**(2): 127–134.

[4] **Foster, D. H. & Bischof, W. F.** (1991). Thresholds from psychometric functions: Superiority of bootstrap to incremental and probit variance estimators. *Psychological Bulletin*, **109**(1): 152–159.

[5] **Foster, D. H. & Bischof, W. F.** (1997). Bootstrap estimates of the statistical accuracy of the thresholds obtained from psychometric functions. *Spatial Vision*, **11**(1): 135–139.

[6] **Swanson, W. H. & Birch, E. E.** (1992). Extracting thresholds from noisy psychophysical data. *Perception and Psychophysics*, **51**(5): 409–422.

[7] **Wichmann, F. A. & Hill, N. J.** (2001). The psychometric function I: Fitting, sampling and goodness-of-fit. *Perception & Psychophysics* (in press). A pre-print is available online at: http://users.ox.ac.uk/~sruoxfor/psychofit/pages/papers.shtml

[8] **Davison, A. C. & Hinkley, D. V.** (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

[9] **Finney, D. J.** (1971). *Probit Analysis*. Cambridge University Press, 3rd edition.

[10] **Wichmann, F. A. & Hill, N. J.** (2001). The psychometric function II: Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics* (in press). A pre-print is available online at: http://users.ox.ac.uk/~sruoxfor/psychofit/pages/papers.shtml

Acknowledgments

For their support, thanks to the Christopher Welch fund, to St Hugh's College, Oxford, and to the Wellcome Trust.