Extended Abstract

Testing Hypotheses About Psychometric Functions

An investigation of some confidence interval methods, their validity, and their use in the assessment of optimal sampling strategies.

N. Jeremy Hill, St. Hugh's College, University of Oxford, UK.

D. Phil. Thesis, Trinity Term 2001.

A *psychometric function* describes the relation between the physical intensity of a stimulus and an observer's ability to detect or respond correctly to it. The performance dimension is expressed as the probability of a positive or correct response, and measurements are based on a number of discrete trials at a number of different stimulus intensities. Each trial consists of a single stimulus presentation (in subjective or "yes-no" designs) or a set of stimulus presentations of which one is the target stimulus (in "forced choice" designs) followed by a response that can be represented as a single binary value: a "yes" or "no" in yes-no designs, or a correct or incorrect response in forced-choice designs. The psychometric function usually increases monotonically with stimulus intensity, and sigmoidal functions such as a logistic, cumulative normal or Weibull function are commonly fitted to the data, usually by the method of maximum likelihood.

To compare sensitivity across different stimulus conditions, *thresholds* are often compared, a threshold being the stimulus value that corresponds to a certain performance level, and which therefore specifies the location of the psychometric function along the stimulus axis. In many circumstances, the *slope* of the psychometric function is also of interest, indicating the rate at which performance increases with increasing stimulus intensity. In addi-

tion, one or two nuisance parameters may need to be estimated: the upper asymptote offset λ, which is related to the rate at which the observer makes stimulus-independent errors or "lapses", and (in yes-no designs) the lower asymptote γ, which is the rate at which the observer guesses that the stimulus is present even in its absence.

Statistical inference about the estimated threshold and slope of a psychometric function often involves the estimation of confidence intervals for those measures. Traditionally, probit analysis[1] offered the most widely accepted method of doing so. However, the confidence intervals thus obtained are only asymptotically correct, as the total number of trials N tends toward infinity. At the low values of N typically encountered in psychophysical experiments, probit methods have been shown to be potentially inaccurate,[2,3] particularly in two-alternative forced choice (2-AFC) designs. In the last 15 years the computationally intensive alternative offered by *bootstrap* resampling methods[4,5] has been advocated in the context of psychometric functions.[3,6–11]

Bootstrap methods come in many forms, some of which are potentially more accurate estimators of confidence interval boundaries than others.[4,5,12] The current research aims to compare the performance of a range of confidence interval methods, including probit methods and several different variations on the bootstrap. The different confidence interval methods are introduced in chapter 2. Their accuracy will be examined empirically by Monte Carlo simulation, in the context of psychometric functions obtained from psychophysical experiments on adults, and in particular in the situation in which nuisance parameters must be estimated. An additional aim is to follow up and extend the work of Wichmann and Hill[11,13] in using computationally intensive methods to assess the relative efficiency of different distributions of stimulus intensities in the estimation of psychophysical thresholds and slopes.

Monte Carlo tests of confidence interval coverage were carried out for a number of different confidence interval methods applied to the threshold and to the slope of a psychometric function. The confidence interval methods studied included five parametric bootstrap methods: the bootstrap standard error method, the basic bootstrap, the bootstrap-t method incorporating a parametric Fisher-information estimate for the Studentizing transformation, the bootstrap percentile method, and the bootstrap BC_a method in which a least-favourable direction vector for each measure of interest was obtained by parametric methods. In addition, standard-error confidence intervals were obtained from probit analysis, and fiducial intervals for the threshold were computed using the method described by Finney.[1]

The results are reported in chapter 3. In general, most of the confidence interval methods were more accurate for thresholds than for slopes, better in yes-no than in 2-AFC designs, and better under idealized conditions (in which there were no nuisance parameters) than under realistic conditions (in which there was a small non-zero rate of “guessing” or “lapsing” that the experimenter must also estimate).

In many cases, confidence interval coverage was found to be inaccurate even though the true value of the relevant measure (threshold or slope) lay within the interval on roughly the correct proportion of occasions: despite accurate *overall* coverage, two-tailed intervals sometimes failed to be properly *balanced* with equal proportions of false rejections occurring in the two tails. An example is the probit fiducial method for thresholds in simulated 2-AFC experiments. Previous studies[2,3] have suggested that probit methods are accurate when the total number of trials N exceeds about 100. However, while the current study found that the coverage of two-tailed 95.4% intervals was very accurate overall, it was also found that coverage in the lower part of the interval was too high, compensating for low coverage in the upper part.

Under the best conditions (thresholds in the idealized yes-no case) all the confidence interval methods performed in a very similar manner. For slopes in the idealized yes-no case, there was also little to choose between the best bootstrap methods and the probit method: the bootstrap-t method was found to be accurate, as Swanepoel and Frangos[14] also found, yet in the range of N studied by Swanepoel and Frangos and in the current study ($120 \leq N \leq 960$), the probit method was equally accurate (there is reason to believe that bootstrap methods may be more accurate than the probit method at lower N, however[3]). In other conditions, where the performance of all confidence interval methods generally deteriorated, some methods were better than others. The bootstrap percentile and $\mathrm{BC_a}$ methods were found to be the most accurate methods for thresholds, and although still far from perfect, the $\mathrm{BC_a}$ method was the best choice for slopes. The $\mathrm{BC_a}$ method was found to be particularly effective in the idealized 2-AFC case, in that it was able to produce balanced confidence intervals for thresholds at different performance levels on the psychometric function: thus it was less sensitive to asymmetric placement of the stimulus values relative to the threshold of interest. The bootstrap percentile method, by contrast, was only balanced when the performance level corresponding to threshold was close to 75%. In 2-AFC, bootstrap methods were generally found to be considerably better than probit methods in the range of N studied.

One of the observed differences between confidence interval methods was their *stability*, i.e. their sensitivity to variation in N and in the *sampling scheme* or distribution of stimulus values on the x-axis. The bootstrap standard error and basic bootstrap methods, for example, tended to produce very different coverage results depending on sampling scheme, whereas the $\mathrm{BC_a}$ method was generally the most stable. Some previous approaches, in which stimulus values are chosen randomly and independently in each Monte Carlo run,[15–17] may mask such differences between confidence interval methods.

In all the simulations, a change in the mathematical form of the psychometric function had little effect. In order to allow direct comparison with a range of existing literature, yes-no simulations were carried out using the logistic function, and 2-AFC simulations were carried out using the Weibull function. All the simulations were repeated using the cumulative normal function, and one set of 2-AFC simulations was repeated using the logistic function. In none of the cases did a change in the form of the psychometric function produce any qualitative or appreciable quantitative alteration to the observed effects of different confidence interval methods, sampling schemes, and values of N.

Under realistic assumptions, the estimation of the upper asymptote offset λ (and also the lower asymptote γ in yes-no designs) presents a problem. It has previously been noted[13,18,19] that the maximum-likelihood estimates of these "nuisance parameters" of the psychometric function are correlated with the slope estimate, and that therefore any mis-estimation of γ or λ may lead to mis-estimation of slope. A particular example of such an effect occurs when an observer makes stimulus-independent errors or "lapses", but when the experimenter assumes idealized conditions in which the observer never lapses, so that λ is fixed at 0 during fitting. In such a case, the slope of the psychometric function is under-estimated, and the same is true whenever the estimated or assumed value of λ is too low. The converse effect, a tendency to *over*-estimate slope, can be observed when the estimate of λ is too high, and such an error exacerbates the natural tendency, which has previously been noted,[7,18,20] for the maximum-likelihood method to overestimate slope even in idealized conditions.

The nuisance parameters λ and γ themselves can be difficult to estimate accurately, a problem which was previously noted by Green[21] and illustrated by Treutwein and Strasburger.[19] The bias in the estimation of λ, for example, depends on the true underlying value of λ itself. When the true value is 0.01, as it was in most of the current simulations, there is a tendency, over the range of N-values studied, for the maximum-likelihood estimate $\hat{\lambda}$ to be larger than 0.01. This leads to overestimation of slope, and inaccuracy in the coverage of confidence intervals for both threshold and slope. In particular, slope coverage probability dropped below target for the bootstrap-t and $\mathrm{BC_a}$ methods, which were the methods that relied on the asymptotic approximation to the parameter covariance matrix given by the inverse of the Fisher information matrix. In the $\mathrm{BC_a}$ method, coverage probability for thresholds also dropped, an effect which was found to change according to the underlying value of λ and the consequent accuracy with which λ could be estimated.

In addition to the one-dimensional methods listed above, four bootstrap methods were applied, in chapter 4, to the problem of computing likelihood-based joint confidence *regions* which allow inferences to be made about threshold and slope simultaneously. The basic bootstrap, bootstrap-t and bootstrap percentile methods were tested, along with a method that used bootstrap likelihood values directly. The last of these proved to be exceptionally accurate, if somewhat conservative—however, it could not separate inferences about threshold and slope from the effects of nuisance parameters. The coverage of the other bootstrap methods was in some cases better and in some cases worse than the performance of the corresponding one-dimensional interval method. All four methods suffered to some extent from bias in the estimation of slope, and were consequently imperfectly balanced in their coverage of slope values above and below the maximum-likelihood estimate.

Further simulations in chapter 5 examined the question of the optimal placement of stimulus values, in order to achieve maximum efficiency and minimal bias in the estimation of thresholds and slopes from a 2-AFC psychometric function.

When efficiency of threshold estimation is the important criterion, probit analysis predicts that, for finite N, the optimal distribution of sample points about the threshold to be estimated has a certain *non*-zero spread, depending on the number of observations and on the confidence level desired. This is at odds with the asymptotic assumption voiced by several authors, and widely followed as a guideline for stimulus placement in adaptive procedures, that optimally efficient estimation of thresholds is to be achieved by placing all observations as close to the threshold as possible. Monte Carlo simulation confirmed the probit predictions: despite the fact that probit intervals tend to be poorly balanced in their coverage (chapter 3) in 2-AFC, and have previously been shown to be inaccurate,[2,3] the predictions of probit analysis were found to be qualitatively correct, in that probit interval widths were highly consistent with Monte Carlo simulations in predicting the *relative* threshold estimation efficiency of different sampling schemes.

The mean and spread of sample points proved to be a fairly good predictor of sampling efficiency with regard to thresholds, and the even spacing of samples proved to be an efficient strategy, assuming that optimal mean location and spread could be achieved. However, there were notable cases in which certain *uneven* sampling patterns were found to be more efficient: in particular, one highly efficient strategy proved to be to place a small number of trials at very a high performance level, and then concentrate on levels closer to threshold than the optimal spread would otherwise indicate. The gain in efficiency, relative to evenly spaced sampling, was nevertheless quite small.

The relationship between efficiency of slope estimation and sampling scheme was not so straightforward, and was not fully explained by the mean and spread of stimulus locations. Predictions from probit analysis were also less consistent with the results of Monte Carlo simulation in the slope results than in the threshold results. The simulations concentrated on the realistic 2-AFC case, with the underlying value of λ set to 0.01: as mentioned above, this condition is particularly prone to bias, and nearly all the sampling schemes studied overestimated the slope of the psychometric function by a considerable amount.

Within the range of N studied, there was an appreciable change in the optimal spread of stimulus values as N increased: for thresholds, the optimally efficient sampling scheme became narrower, converging towards the asymptotic ideal of zero spread. For slopes, optimal spread converged towards the asymptotically predicted (non-zero) value.

With regard to thresholds, there was little or no effect of k, the number of blocks into which the N observations were divided: the mean and spread of the optimally efficient sampling scheme were not affected, nor was the distribution of bias and efficiency scores measured outside the optimal region. For slopes, there was little effect when k exceeded 5, although there was a discernible advantage to sampling with smaller numbers of blocks ($k = 3$ and $k = 4$): the simulations imposed a minimum spacing between blocks, and the 3- and 4-point schemes were able to concentrate more closely on the two asymptotically optimal sampling points.

The simulations of chapter 5 addressed the question of what the optimally efficient sampling schemes look like, *without* addressing the question of how such sampling is to be achieved relative to an unknown psychometric function. In practice, a larger k will be useful from the point of view of sequential estimation, as it allows a greater number of opportunities to re-position the stimulus value according to the current best estimate of the optimal location. Sequential stimulus selection has so far been ignored in the application of bootstrap methods to psychometric functions.[3,6,7,11] However, it can be presumed to occur to some extent in many experimental designs (including many that are described as "constant stimuli" experiments) whether the stimuli are selected "by eye" or by a formally specified adaptive procedure. The simulations of chapter 6 suggest that the assumption of fixed stimuli can lead bootstrap methods to produce confidence intervals whose coverage is too low. Furthermore, sequential selection introduces an increasing relationship between threshold coverage and N, a fact which may undermine one of the principal advantages of the bootstrap, namely that it is less sensitive to error than asymptotic methods when N is low. It is recommended that future developments of bootstrap methods in psychophysics should concentrate on formal specification of the algorithm for stimulus selection, and that bootstrap replications of the experiment should include simulation of the stimulus selection process, using the same algorithm as that employed by the experimenter.

References

[1] Finney, D. J. (1971). *Probit Analysis.* Cambridge University Press, third edition.

[2] 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.

[3] 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.

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

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

[6] 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.

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

[8] 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.

[9] Hill, N. J. & Wichmann, F. A. (1998). A bootstrap method for testing hypotheses concerning psychometric functions. Presented at CIP98, the Computers In Psychology meeting at York University, UK.

[10] Treutwein, B. & Strasburger, H. (1999). Assessing the variability of psychometric functions. Presented at the 30th European Mathematical Psychology Group Meeting in Mannheim, Germany, August 30–September 2 1999.

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

[12] Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, **16**(3): 927–953.

[13] Wichmann, F. A. & Hill, N. J. (2001). The psychometric function I: Fitting, sampling and goodness-of-fit. *Perception and Psychophysics* (in

press). A pre-print is available online at: http://users.ox.ac.uk/~sruoxfor/psychofit/pages/papers.shtml .

[14] Swanepoel, C. J. & Frangos, C. C. (1994). Bootstrap confidence intervals for the slope parameter of a logistic model. *Communications in Statistics: Simulation and Computation*, **23**(4): 1115–1126.

[15] Gong, G. (1986). Cross-validation, the jackknife, and the bootstrap: Excess error estimation in forward logistic regression. *Journal of the American Statistical Association*, **81**(393): 108–113.

[16] Lee, K. W. (1990). Bootstrapping logistic regression models with random regressors. *Communications in Statistics: Theory and Methods*, **19**(7): 2527–2539.

[17] Lee, K. W. (1992). Balanced simultaneous confidence intervals in logistic regression models. *Journal of the Korean Statistical Society*, **21**(2): 139–151.

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

[19] Treutwein, B. & Strasburger, H. (1999). Fitting the psychometric function. *Perception and Psychophysics*, **61**(1): 87–106.

[20] O'Regan, J. K. & Humbert, R. (1989). Estimating psychometric functions in forced-choice situations: Significant biases found in threshold and slope estimations when small samples are used. *Perception and Psychophysics*, **46**(5): 434–442.

[21] Green, D. M. (1995). Maximum-likelihood procedures and the inattentive observer. *Journal of the Acoustical Society of America*, **97**(6): 3749–3760.