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

Thesis submitted in the University of Oxford
as a partial requirement for the fulfilment of
the degree of Doctor of Philosophy.

Trinity Term, 2001

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.

Various methods for computing confidence intervals and confidence regions for the threshold and slope of the psychometric function were investigated in the context of block-design psychophysical experiments of the sort that are typically carried out with trained adult human observers.

Several variations on the bootstrap method, along with the more traditional methods of probit analysis, were tested using computer simulation, comparing (a) the accuracy of overall coverage, (b) the *balance* of coverage between the two sides of a two-tailed interval, and (c) the *stability* of coverage with regard to variation in the total number of observations and in the distribution of stimulus values. For thresholds, the bootstrap percentile and bias-corrected accelerated ($\mathrm{BC_a}$) methods were the most reliable, and for slopes the $\mathrm{BC_a}$ method was generally the best choice. The differences between methods were greater, and their performance was generally poorer, (a) for slopes than for thresholds, (b) in the two-alternative forced-choice than in the yes-no design, and (c) when the observer's rate of guessing and/or "lapsing" cannot be assumed to be zero and must therefore be estimated. The problem of bias in the initial slope estimate was also exacerbated by the addition of guessing and lapsing rates as nuisance parameters.

Computer-intensive confidence interval methods were also used to assess the relative efficiency of different distributions of stimulus values, with regard to the estimation of threshold and slope. The most efficient sampling patterns shared certain characteristics irrespective of the number of blocks into which they were divided. Certain unevenly spaced sampling patterns were marginally more efficient than evenly spaced ones.

Further simulations illustrated that, given broad assumptions about the way in which stimulus intensities are chosen in realistic experiments, the assumption of fixed stimulus values, which is intrinsic to the bootstrap methods commonly applied to psychometric functions, may lead to low coverage.

Note

This document is intended to be bound along with a CD-ROM whose contents are as follows:

- The directory tree `/simulations/` contains text files in which simulation results are tabulated for many of the tests conducted in this project. The large number of different conditions explored in the simulations meant that it was impossible to describe them all fully: passing references are made in the thesis to some tests for which no summary diagrams are provided. Even for those simulation sets for which figures have been plotted, the large number of simulations would make the inclusion of a hard copy of the raw data on which the figures are based impracticable. Therefore, at several points in the thesis, reference is made to a directory in the "results archive". It is to the CD-ROM that this refers. Results are provided in the hope that they may be useful for further analysis, but they are not essential to the reading of the thesis.

- The directory tree `/software/` contains source files for the fitting software used to obtain parameter estimates in the simulations, written by Jeremy Hill, 1995–2001. The core fitting and simulation routines were written in ANSI C, and implemented so that they could interface with MATLAB versions 5 and up (The MathWorks, Inc, 1997). They are supported by a number of functions written in the MATLAB script language.

- The directory tree `/thesis/` contains an electronic copy of this document, in Adobe Portable Document Format (PDF), along with related papers by Wichmann and Hill (2001).

At the time of writing (September 2001), a version of the software is also available on the World-Wide Web at:

http://users.ox.ac.uk/~sruoxfor/psychofit/pages/download.shtml

Additional resources concerning this project can also be found at the site. However, note that the version of the software used for most of the simulations reported in this thesis, and included on the CD-ROM (version 2.5.2) includes numerous small improvements and fixes, relative to the version posted as at September 2001 (version 2.5.1).

Acknowledgments

I could not have wished for a more knowledgeable, patient or helpful supervisor than Bruce Henning. It has been a privilege and a great pleasure to work with him, and with fellow students Felix Wichmann, Dan Tollin, Stephanie McGuire and Anna Zalevski. Felix got the ball rolling with his ideas on function-fitting, and this project is founded on the software and general approach that resulted from a highly enjoyable collaboration with him over several years.

Many thanks are due to Dr. Mario Cortina-Borja for helping me to understand many of the statistical issues involved, and for his enthusiasm. Thanks also to Dr. John Bithell for allowing me to attend some of his statistics lectures, and to Dr. Peter Clifford for his statistical advice. Dr. Peter McLeod provided invaluable advice during my earlier research on motion perception, and I have also enjoyed the opportunity of working with Dr. Vincent Walsh. I would also like to thank the many people who took the trouble to answer my USENET postings—it was always a pleasant surprise to find that there were so many knowledgeable people out there who seemed to have nothing better to do than answer my questions on statistics, programming and typesetting.

An adequate account of all the things I have to thank Kate Lyons for would take me well over the word limit, so I shall say no more here than: thank you. I am also grateful to Gwyn Lintern, Claus Wisser and Anna Zalevski for day-to-day moral support over the weeks of thesis writing. Finally I would like to thank my parents, Chris and Margaret Hill, for my education, for the freedom that I enjoy, and for their continual encouragement.

Financial support was provided by the Christopher Welch Scholarship Fund, the Maplethorpe Scholarship at St. Hugh's College, and the Wellcome Trust.

Computation and figure plotting were carried out in Matlab 5.2.1 (The MathWorks Inc, 1997) for MacOS. The report was typeset using OzTeX 4.0 (Andrew Trevorrow, 1999), a shareware implementation of LaTeX 2ε for the Macintosh.

Contents

List of figures

List of tables

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 $\mathrm{BC_a}$ 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.

1. Introduction

1.1 Modelling psychophysical data

Psychophysics is concerned with the relation between physical attributes of a stimulus, such as its intensity, and an observer's ability to detect or respond appropriately to it. A typical psychophysical experiment involves repeated presentation of a stimulus in a number of discrete *trials*. In a *yes-no* design, a trial consists of a single presentation of a stimulus, which the observer must categorize as being either the target or non-target stimulus—the response variable is then the proportion of occasions on which the observer gives a positive response, classifying the stimulus as a target. In *forced-choice* designs, a trial consists of two or more stimuli, separated in space or in time, one of which (chosen at random on each trial) is the target. The response variable is the proportion of trials on which the observer can correctly identify the target stimulus within the set.[†]

Stimulus intensity will generally be denoted by x. The number of different stimulus levels in a data set will be denoted by k, and a vector $\boldsymbol{x}$ of length k will contain all the stimulus intensity values for the data set. The vector $\boldsymbol{n}$ will specify the number of trials n_i performed at each stimulus intensity x_i, and the total number of trials, $\sum n_i$, will be denoted by N. The vector

[†] A third kind of design is the *identification* paradigm, in which a single stimulus is presented on each trial, and detection performance is measured by the proportion of trials on which the observer correctly identifies the category to which the stimulus belongs, given a forced choice between a number of mutually exclusive categories defined on a dimension *other than* intensity. For example, the detectability of a luminance grating might be measured by the observer's ability to identify whether it is horizontally or vertically orientated (given equal probability of either, on each trial). Statistically, results from the identification paradigm can be treated identically to those from forced-choice designs. For the purposes of this study, their inclusion in the category of forced-choice experiments will be implicit.

$\boldsymbol{r}$ will denote the number of correct or positive responses at each stimulus intensity, and $\boldsymbol{y}$ will denote the observed proportions of such responses, so that $y_i = r_i/n_i$. For a general explanation of notation conventions, and a glossary of the algebraic terms used in this report, see appendix A.

In order to explain the data and make predictions about future experimental conditions, the experimenter often fits a parametric model $p = M(x, \boldsymbol{\rho}; \boldsymbol{\theta})$ to the data, where p is the underlying value of y that the model predicts, $\boldsymbol{\theta}$ is a vector containing the model's parameters, and $\boldsymbol{\rho}$ is a vector which contains any other explanatory variables besides stimulus intensity x. A particular combination of values $\boldsymbol{\rho}$ designates a single experimental condition, in which the observer's performance is modelled by a single *psychometric function*, $p = \psi(x)$.

Estimates for the parameter values are often found by the method of maximum-likelihood,[1–10] so that estimates $\hat{\boldsymbol{\theta}}$ are those values for which likelihood or log-likelihood is greatest. Expressions for likelihood and log-likelihood are given in appendix B, equations (B.42) and (B.43), respectively. The method assumes that response probabilities p_i are stationary throughout the experiment, and that the observed responses r_i are therefore binomially distributed, each with underlying probability of success p_i.

Some of the parameters $\boldsymbol{\theta}$ will be of theoretical interest to the experimenter, but others may be *nuisance parameters* whose values are of little interest in themselves, but which must be estimated in order to obtain unbiased estimates of the other parameters. In the yes-no design, two examples of nuisance parameters are the observer's *guess rate*, which is the probability with which the observer makes a positive response even when there is no information from sensory mechanisms about the signal's presence (for example, when no signal is present at all),[†] and the *lapse rate*, which is the probabil-

[†] Some analyses of single-interval experiments plot the proportion of *correct* responses, given a mixture of signal and non-signal trials—then the psychometric function is statistically more similar to that of the 2-AFC paradigm. The term "yes-no" is not used here to refer to such an approach, but rather to the formulation in which the response measure is the proportion of positive responses given the presence of a signal—thus the psychometric function ranges from near 0 to near 1. Therefore, in the context of "yes-no" designs, the term "guess rate" should not be construed to mean the probability of giving a correct

(footnote continues ⟶)

ity that the observer fails to report the target, irrespective of its intensity. Under the assumption that guesses and lapses are uncorrelated with x, and that their rates of occurrence are stationary, the psychometric function can be written as

$$\psi(x) = \gamma + (1 - \gamma - \lambda) F(x), \tag{1.1}$$

where γ is the guess rate, λ is the lapse rate, and $F(x)$ is the underlying *detection function*, a function with range 0 to 1 (inclusive or exclusive) that is independent of γ and λ, and which determines the shape of the psychometric function. The shape of $F(x)$ emerges from $M(x, \boldsymbol{\rho}; \boldsymbol{\theta})$ given the relevant set of conditions $\boldsymbol{\rho}$. It is nearly always monotonic (usually monotonically increasing) and often sigmoidal in shape.

The formulation of the psychometric function given in equation (1.1), or a special case of (1.1) in which $\lambda = 0$, is often used.[6–10] It can also be applied in m-alternative forced choice (m-AFC) designs: the lower asymptote γ is fixed at the chance level $1/m$, and need not be estimated as a free parameter;[†] the upper asymptote offset λ, as in the yes-no case, is equal to the rate at which the observer makes stimulus-independent errors.

1.2 The psychometric function

The current study considers the particular case in which the model M deals with only one experimental condition at a time—thus, the parameters $\boldsymbol{\theta}$ are concerned with describing only one psychometric function. Beside the upper and lower bounds defined by λ and γ, two aspects of the psychometric function are of interest: its location and scale along the x-axis. The underlying detection function $F(x)$ therefore has two parameters, which will be referred

answer by chance, as it does for forced-choice designs. Rather, it is used solely to mean the stimulus-independent probability of a positive response.

[†] The use of a fixed lower asymptote equal to the chance performance level assumes that the observer has been trained on the task using appropriate feedback signals, and is motivated to try to maximize the number of correct responses. Such assumptions are implicit in the treatment of 2-AFC designs in the current study.

to as α and β. The psychometric function has four parameters in total:

$$\psi\left(x;\,\boldsymbol{\theta}\right) \;=\; \gamma + \left(1-\gamma-\lambda\right) F\left(x;\,\alpha,\beta\right). \tag{1.2}$$

where the vector $\boldsymbol{\theta}$ is used as a collective shorthand for the four parameters:

$$\boldsymbol{\theta} = \left(\alpha,\beta,\gamma,\lambda\right)^{\mathrm{T}}.$$

1.2.1 The shape of the detection function

Many different two-parameter functions have been employed as the detection function $F(x)$. Examples, from the literature that examines the statistical properties of psychometric functions, include the cumulative normal,[11–13] the logistic[3,7,8,14] and the Weibull function.[2,4–6,9,10] Formulae for these three functions are provided in the appendix, sections B.1.1, B.1.2 and B.1.3, respectively. In each, the parameters α and β together determine the location and scale of the function. Their exact rôles, and their relation to the units in which stimulus is measured, vary according to the shape chosen (for this reason, the standardized measurements of *threshold* and *slope* will be defined in section 1.2.2).

According to the experimental situation, there may be theoretical reasons for choosing one particular function shape over another. Signal detection theory[15] predicts a cumulative normal shape for the yes-no psychometric function, and also predicts that the 2-AFC psychometric function should be a cumulative normal with mean 0 (passing through $x = 0, \psi = 0.5$ at its steepest point). This assumes a linear mapping between the stimulus dimension and the decision axis. Real 2-AFC psychometric functions, on the other hand, tend to accelerate in the region in which $\psi(x)$ is low, and are often better fit by scaling a sigmoidal function (such as the cumulative normal, logistic or Weibull) between 0.5 and $1-\lambda$, as per equation (1.2). Assuming that quantities on the decision axis are normally distributed, such shapes are obtained if one assumes a non-linear (usually accelerating) transformation from the physical stimulus dimension to the decision axis.

The Weibull function in particular provides a very good fit to 2-AFC data from visual contrast detection[16,17] and contrast discrimination[17] experiments. It also has the potentially useful property that probability summation over multiple mechanisms, each with a Weibull response, produces another Weibull function. This property has found a useful rôle in developments of signal detection theory which model an attentional field distributed across multiple channels.[18,19]

For the purposes of the current research, however, it will be assumed that the experimenter merely wants a threshold and slope estimate from a well-fitting function, without (yet) having to worry about a fully-fledged model for the mapping between the physical world and the decision axis. At a later stage of research, when several psychometric functions in several different experimental conditions have been sampled, the experimenter may well want to develop a more general model $M(x, \boldsymbol{\rho})$ that predicts performance across the whole corpus of data, taking into account multiple experimental conditions $\boldsymbol{\rho}$. When such a model is in place, the experimenter would presumably want to adjust the parameters of the whole model to obtain maximum-likelihood estimates, rather than estimating a threshold and slope separately for each individual condition. If one has a good idea of the mapping from stimulus axis to decision axis under different experimental conditions, the artificial intermediate step of using the Weibull function (or any other particular two-parameter function) to model individual conditions becomes redundant, or at best it is relegated to a role in the initial "guessing" phase of the fit.

Therefore, the choice between the logistic, cumulative normal and Weibull functions (or any other similar function) will not be treated as a theoretically important issue. For the purposes of estimating thresholds and slopes (section 1.2.2) and confidence limits for those measures, the choice of different sigmoidal functions generally makes little difference in practice, relative to the experimental variability typically observed in the measures themselves.[10] Nevertheless, in order to verify the lack of importance of one's choice of psychometric function shape, many of the simulations of the current study were repeated with more than one shape. The coverage simulations of chapter 3, for example, use the cumulative normal function for both yes-no and 2-AFC simulations, but the yes-no simulations are repeated with the logistic function, and the 2-AFC simulations with the Weibull function. Although the choice of shape was indeed found to make little difference, the logistic and Weibull results are reported in order to allow direct comparison with previous studies.[4,10,20,21]

1.2.2 Threshold and slope

As previously stated, the two parameters of particular interest in the psychometric function, α and β, determine the location and scale of the detection function. Often, location and scale are reported by quoting the values of α and β as they appear in the formulae of section B.1, or in whatever form the experimenter chooses to express the equation for the detection function. As stated in section 1.2.1, however, the choice of equation for the detection function will not be considered to be of high theoretical importance. Therefore, the terms usually employed to refer to the location and scale in the context of psychometric functions, "threshold" and "slope" are defined here in terms that do not rely on any particular formulation.

Threshold t is defined as the inverse of the detection function F at a particular detection level of interest. The notation t_f will be used to denote the value of x such that $F(x) = f$. For example, $t_{0.5}$ represents the mid-point of the psychometric function, the point at which $F(x) = 0.5$. A threshold measure specifies the location of the psychometric function along the x-axis.

Slope s is defined as the derivative of $F(x)$ with respect to x, evaluated at a particular threshold point. So s_f denotes $\mathrm{d}F/\mathrm{d}x$ evaluated at $x = t_f$, $s_{0.5}$ being the slope of the detection function at its mid-point. Thus, slope is a measure of the rate at which performance improves with increasing stimulus intensity.

When the term "threshold" is used on its own, it should generally be taken to refer to $t_{0.5}$. Similarly, "slope" on its own means $s_{0.5}$.

N.B. In many reports, thresholds and slopes are described in terms of the range of $\psi(x)$ rather than that of $F(x)$. For example, the mid-point of a 2-AFC psychometric function is often referred to as the “75% threshold”, i.e. the point at which $\psi(x) = 0.75$. This convention is not adopted here. Threshold and slopes are instead defined in terms of the underlying detection function $F(x)$, because $F(x)$ reflects those aspects of the psychometric function that are of interest to an experimenter (*viz.* the characteristics of the underlying detection mechanisms) without involving the mechanisms of “guessing” and “lapsing” that are assumed to yield no direct insight into the psychological phenomena of interest. Terminology such as “75% threshold” is therefore somewhat imprecise—75% is often chosen as a standard threshold measure in 2-AFC designs because it is the mid-point of the psychometric function, but this is only the case if $\lambda = 0$. If $\lambda = 0.02$, for example, then the 74% threshold is the mid-point. An additional convenient aspect of defining threshold and slope with respect to $F(x)$ is that a term such as $t_{0.5}$, for example, represents the same point on the detection function, regardless of whether the experiment uses a yes-no, 2-AFC, or other forced-choice design.

Differences in an observer’s sensitivity to stimuli across different experimental conditions are often examined by comparing thresholds and/or slopes across two or more psychometric functions. The comparison can be made in different ways, depending on (a) the detection level(s) at which the experimenter wishes to measure threshold and slope, and (b) the relative importance of information about thresholds and information about slopes. The answers to both issues depend on the experimental context: when considering the former, there may be reasons to examine a particular detection level for comparison with previous experiments that also studied that level; the answer to the latter question depends on the psychophysical phenomenon under study. There has been a general tendency for reports of psychophysical experiments to concentrate on thresholds alone, due in part to the prevalence of adaptive procedures that provide efficient and accurate threshold estimates in a relatively small number of trials, usually at the cost of very imprecise and often biased[22–24] slope estimates. Where slope estimates are taken (as is possible with some adaptive procedures[25,26]) or where full psychometric functions are measured, the experimenter often uses them merely to verify that the psychometric functions are roughly parallel (usually without statistical support) in order to justify using the threshold as the single dimension of performance.

However, there are cases in which psychometric function slopes are important in their own right. Wichmann[17] found that, whereas it was impossible to distinguish statistically between classes of models for contrast discrimination using previously reported data (which consisted mostly of thresholds from adaptive procedures), it *was* possible to make a decisive hypothesis test using a corpus of block-design psychometric functions: the crucial information lay in significant slope differences within the corpus. Differences in slope may indicate changes in an observer's certainty about a stimulus: in signal detection theory,[15] a change in the psychometric function, from a steep slope and high threshold to a shallower slope at a lower threshold, is the signature of a change from the ideal observer for a signal known only statistically to the ideal observer for a signal known exactly.[†] Alternative theoretical approaches predict similar slope changes due to "uncertainty"[18] or "distraction"[19] when multiple channels are attended simultaneously. In certain circumstances there may be a change of slope that is not correlated with a shift in the mid-point of the psychometric function. For example, Tyler[27] presents results from a 2-AFC visual detection task in which the signal is a two-dimensionally amplitude-modulated 4-cycle-per-degree sinusoidal grating: psychometric function slope becomes steeper as the spatial extent of the envelope increases, without any significant change in the $d' = 1$ (76%) performance threshold. A slope estimate might also be useful in its own right as a diagnostic criterion, in cases where there are significant between-subject differences in slope. Such a case was found by Patterson, Foster and Heron,[28] who studied the ability of Multiple Sclerosis patients to detect small-field flashes of light in a yes-no task, relative to that of normal controls: the patient group showed significantly shallower slopes when the flash was to be detected against background light, but at two out of the three background

[†] This is true when performance is plotted against signal-to-noise ratio on semi-logarithmic coordinates, as in the formulation of Green and Swets (1966)[15]—see page 194, *ibid.*

light levels at which a slope difference was found, there was no difference in the 50% threshold between the groups.

Thresholds and slopes will be considered separately in the current simulation studies, except in chapter 4, in which joint confidence regions for threshold and slope are investigated.

1.2.3 Nuisance parameters

The inclusion of λ (and γ, in yes-no designs) as free parameters of the fit is a step recommended by some authors[6–10] in order to avoid bias in the estimation of slope. The problem is illustrated for a 2-AFC psychometric function in figure 1.1.

The upper panel of figure 1.1 is from Wichmann and Hill's first paper on psychometric functions.[9] The blue circles show observed proportions of correct responses from a psychophysical data set with 50 trials per point. The last point ($x = 3.5$) is an exception: here, only 49 trials have yet been completed, with 1 still to run. The maximum-likelihood fit to the data so far, using a Weibull function, is shown by the solid blue curve. Now suppose that, on the very last trial of the last block, the observer happens to press the wrong response button, or alternatively blinks, misses the stimulus presentation interval, is forced to guess the correct answer, and guesses incorrectly. The last block, at 98% correct, is now shown by the yellow triangle.

The solid yellow curve shows the maximum-likelihood Weibull function fit to the revised data assuming the fixed value $\lambda = 0$. Note that, in order to accommodate the block that includes the "lapse", the slope of the psychometric function has become much shallower, yielding a very poor fit to most of the other data points. The slope estimate is biased, as is any threshold estimate other than $t_{0.6}$ (the 80% point, which is roughly where the two solid curves cross). The broken yellow curve, on the other hand, shows the maximum-likelihood fit when λ is allowed to vary. The maximum-likelihood value $\hat{\lambda}$ is roughly 0.014, the curve is a much better fit to the rest of the data points, and the threshold and slope estimates are now very similar to those obtained before the lapse occurred.

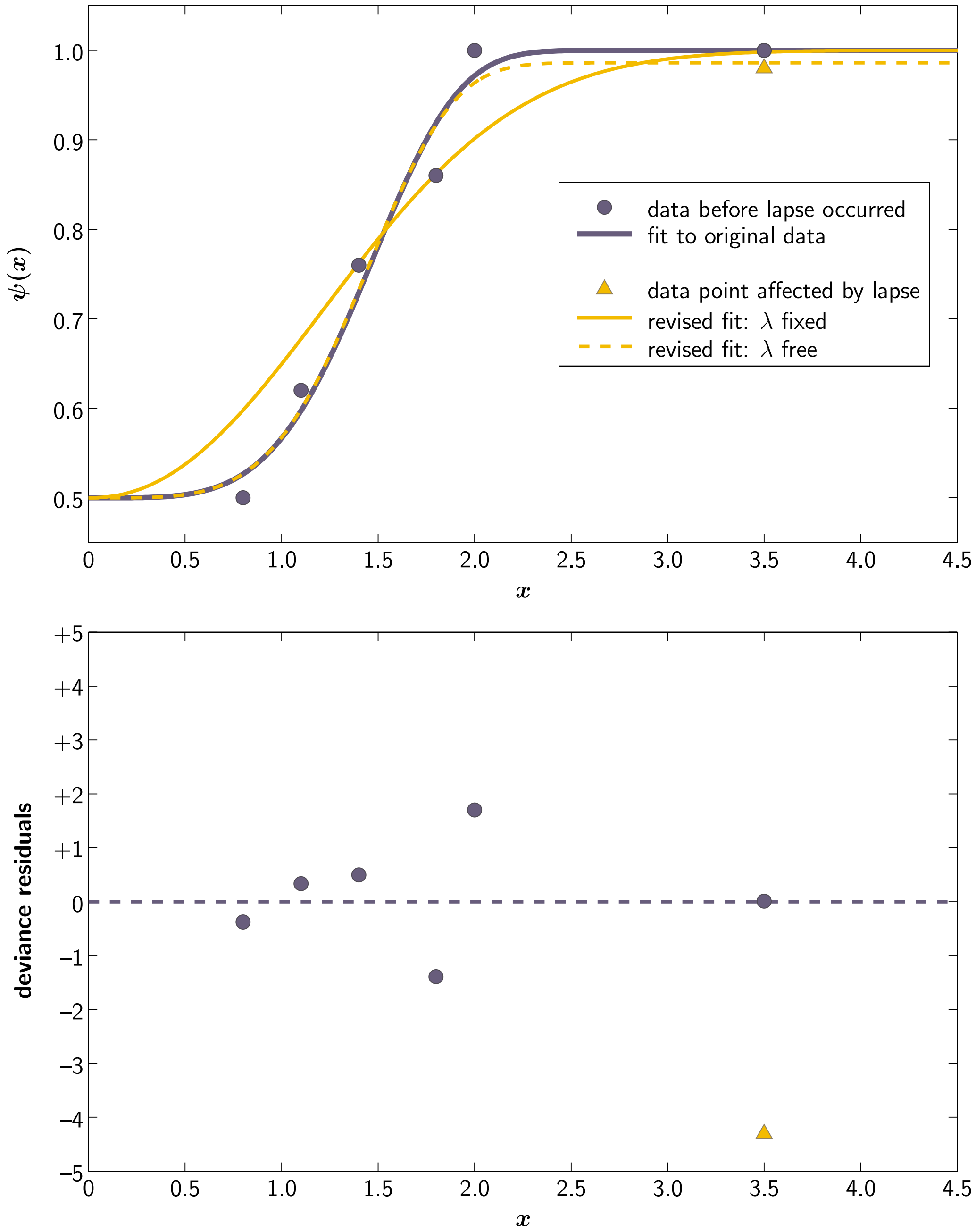


Fig. 1.1: A demonstration of the effect of “lapses” on a maximum-likelihood fit using a 2-AFC Weibull function. See section 1.2.3 for details.

The example shown in figure 1.1 is an exaggerated illustration, but Wichmann and Hill[9] show using Monte Carlo simulation that bias can be significant, for more realistic data sets, whenever an inaccurate fixed value of λ is used. The large influence of the apparently small shift in the position of the last data point is due to the very small variability of the binomial distribution around an expected probability close to 1.0. The original (solid blue) curve is a poor fit to the revised data because it predicts a performance of almost exactly 1.0 at $x = 3.5$. If the probability of correct performance were really 1.0, then an observed probability of 0.98 would be impossible (with a log-likelihood of $-\infty$). In order to provide a likely fit, the predicted probability at $x = 3.5$ must be reduced, and if λ is fixed at 0, the only way of doing so is to reduce the slope of the function drastically.

The lower panel of figure 1.1 demonstrates that the failure of the fixed-λ approach can be seen as a lack of robustness to what is actually an extreme outlier. A maximum-likelihood fit can also be seen to be minimizing the *log likelihood ratio* or *deviance*, D, a quantity which is monotonically related to likelihood:

$$D = 2\sum_{i=1}^{k}\left\{r_i \log\left(\frac{r_i}{n_i p_i}\right) + (n_i - r_i)\log\left[\frac{n_i - r_i}{n_i(1-p_i)}\right]\right\}. \tag{1.3}$$

Just as the sum-squared-error loss function is the sum of squared arithmetic residuals, D can be seen as the sum of squared *deviance residuals*, where each residual d_i is the square root of the deviance value computed for point i alone, signed according to the sign of $y_i - p_i$ (see Wichmann and Hill's paper[9] for more about the analysis of deviance and deviance residuals). The deviance residuals of the data set, given the initial fit (the solid blue curve in the upper panel) are plotted against x in the lower panel of figure 1.1. As before, the yellow triangle indicates the last block after the lapse occurred. Note that it has dropped a long way from the position it occupied before the lapse occurred, relative to the magnitude of the other residuals: in a least-squares sense it is clearly an extreme outlier, and it is not surprising that the fit must change drastically in order to accommodate it.

Note that the same arguments apply to γ in yes-no designs, if it cannot be assumed that the observer's guess rate is 0. The low binomial variability around *low* expected response probabilities has the same effect as that at high probabilities.

1.3 Fitting the psychometric function

The addition of nuisance parameters to the model means that the maximum-likelihood search procedure must search for the global maximum in a three- or four-dimensional parameter space. Performing a search efficiently in more than two dimensions is no trivial task, but the simplex search algorithm of Nelder and Mead[29] is well-suited to the problem. The current simulation studies used software developed by the author, which was also used by Wichmann and Hill.[9,10]

In cases where λ and γ were allowed to vary, they were first fixed at 0.01 (an arbitrarily chosen but plausible value) while a maximum-likelihood grid search was performed to obtain initial "guess" values for α and β. The values of α and β thus obtained, along with the guess values $\lambda = 0.01$ and $\gamma = 0.01$, were then used to initialize the simplex. Over the range of N studied, the imprecision of the simplex search results was found to be negligible relative to the variability of the parameters themselves.

The fitting process was guided by the assumption that the observer's true lapse rate and guess rate would not take large values (greater than, say, 0.05). This guideline was implemented by the use of a rectangular Bayesian prior (see the appendix, section B.2.1) which constrained λ (and γ, where appropriate), to lie within the range $[0, 0.05]$. A constrained multi-parameter maximum-likelihood search of this nature is similar to the approach of Treutwein and Strasburger[8] and identical to that of Wichmann and Hill.[9,10]

1.4 Testing hypotheses about psychometric functions

Psychophysical hypotheses may take a very broad form, such as, for example, “the detection of [some type of stimulus] is mediated by a divisive contrast gain-control mechanism,” or they may concern much simpler observations, such as “observers are more sensitive to the stimulus, at the 50% detection level, under conditions $\boldsymbol{\rho}_1$ than under conditions $\boldsymbol{\rho}_2$”.

The former sort of hypothesis may be examined by formulating the model in question (a divisive gain-control mechanism, in the above example) and applying a statistical test of *goodness-of-fit*. A typical goodness-of-fit test assesses some measure of the dispersion of the observed data around the values predicted by the model, against the distribution of dispersion values obtained under the assumption that the model is correct. The deviance measure of equation (1.3) is one suitable measure of dispersion, which also allows related models to be compared against one another. Hypothesis tests of this sort, in the context of visual contrast gain control, are treated in detail by Wichmann.[17]

The current study concentrates on simpler models that predict single psychometric functions. Goodness-of-fit tests also have an important rôle to play when fitting single psychometric functions, in order to verify whether the data are compatible with the assumed shape of the detection function, and with the assumption that the data truly arise from stationary binomial processes. Wichmann and Hill[9] describe a number of tests that may be applied in this situation, including an analysis of deviance residuals that may indicate whether the observer’s performance changes from one block of trials to the next. Goodness-of-fit tests will not be considered here. Instead, it is the second kind of hypothesis, above, that will be examined.

In order to make statistical comparisons of an observer’s performance under two or more experimental conditions, *confidence intervals* for thresholds and slopes are frequently computed. The statistical significance of an effect may be gauged by comparing the size of the effect to the size of the interval. Broadly speaking, a confidence interval is a numerical range within which the true threshold or slope value can be asserted to lie, with a certain probability or *confidence level*, based on the expected variability of the observed data. For the thresholds and slopes of psychometric functions, probit analysis[30] traditionally offered the most widely accepted set of techniques for the computation of confidence intervals. However, the intervals obtained by probit analysis rely on approximations to the probability distributions of thresholds and slopes, which are only asymptotically correct, as $N \rightarrow \infty$. Consequently, the accuracy of probit estimates of variability has been called into question, and Monte Carlo simulation studies have suggested that they are potentially inaccurate in their application to psychometric functions.[11,13] Recently, *bootstrap* methods,[31,32] which are computationally intensive confidence interval methods that use Monte Carlo simulation to estimate variability, have been proposed as an alternative to the more traditional asymptotic approaches.[4,10,12,13,33–35]

This thesis aims to explore two general aspects of threshold and slope confidence intervals. The first is their accuracy—in other words, the extent to which their *coverage* (the probability that the true value lies within the confidence interval) matches the intended confidence level. The second is their width—narrow confidence intervals are clearly more desirable than wide ones, provided their coverage is accurate, because the statistical test that they represent is more powerful (*power* being the probability of finding a significant effect, given that an effect really exists).

Chapter 3 uses Monte Carlo simulation to examine the former question, simulating experiments repeatedly in order to measure the proportion of occasions on which the true threshold or slope value lies within the confidence intervals computed by a particular method. Chapter 4 then briefly considers the coverage of confidence *regions* that may be used to make inferences about threshold and slope simultaneously. Chapter 5 examines the effect of *sampling scheme* (see below) on the accuracy and precision with which thresholds and slopes are estimated, where the width of the confidence interval is used in order to measure precision. Finally, chapter 6 examines the effect on confidence interval coverage of uncertainty in the placement of stimuli.

Several factors may affect both the width of a confidence interval and the accuracy of the probability with which it covers the true value:

- Experimental design, i.e. the question of whether the experimenter uses a yes-no design, a 2-AFC design, or another forced-choice design. In the coverage tests of chapters 3, 4 and 6, the yes-no and 2-AFC designs will be considered. The simulations of chapter 5 will focus on the 2-AFC design.

- The method used to compute the confidence interval. Coverage tests will compare the performance of the probit methods described in section 2.1, and six different variations on the bootstrap, described in section 2.2.

- Whether the lapse rate λ (and the guess rate γ in yes-no designs) are assumed to be 0, or whether they are included in the fit as unknown nuisance parameters. The former set of assumptions will be referred to as the *idealized* case, and the latter the *realistic* case. Both cases will be considered in the coverage tests of chapters 3, 4 and 6. Chapter 5 will examine the realistic case. In all simulated fits, it will be assumed that λ and γ are low (less than 0.05, consistent with a trained adult observer).

- The total number of trials, N. Again, it will be assumed that the observer is a motivated adult, and therefore capable of performing at least 100, and anything up to about 1000 trials in a single experimental condition. Psychometric functions will be based on 120, 240, 480 and 960 trials.

- Intended coverage. Both the width and coverage of an interval are clearly affected by the confidence level that the interval is intended to have. Two confidence levels will be considered: 68.3% and 95.4%. The former confidence level has the same intended coverage as the "standard error bar", a familiar statistical yardstick which is obtained from the mean ± one standard error, assuming that the quantity in question is normally distributed. The latter, 95.4%, is the same coverage as a standard interval constructed from the mean ± two standard errors, and provides coverage of a comparable level to that used in many statistical tests (90%, 95% and 99% are often used). All intervals will be two-tailed—thus, they aim to cover the *central* 68.3% or 95.4% of the probability distribution of the estimate in question.

- The configuration of the stimulus values $\boldsymbol{x}$, which will be referred to as the *sampling scheme*. Ideally, the coverage of a confidence interval should be independent of the sampling scheme, since the experimenter does not have precise control over the locations in which the stimulus values happen to lie relative to the curve of the true psychometric function. For the coverage tests of chapters 3, 4 and 6, a small number of illustrative sampling schemes (defined in section 1.5) will be used. Chapter 5 will explore sampling schemes in more detail.

1.5 Sampling schemes

Wichmann and Hill[9,10] examined the effect of sampling scheme on the statistical properties of threshold and slope estimates, and noted in particular that some schemes caused slope estimates to be more sensitive to errors in the estimate of λ than others,[9] and that confidence interval width was more sensitive to errors in the initial fit for some schemes than for others. To illustrate some of the possible effects, they used a set of seven schemes, whose definitions are reproduced in section 1.5.1.

The seven schemes are by no means intended to represent an exhaustive inventory of possible sampling patterns (for one thing, they all have the same number of points). They were used by Wichmann and Hill, and will be used in some of the current simulations, because the computationally intensive nature of Monte Carlo techniques makes it impossible to explore exhaustively the effects of variations in sampling scheme. Wichmann and Hill found, in their second paper,[10] that their examples highlighted some of the important differences that the choice of sampling scheme can produce. So, while chapter 5 will extend Wichmann and Hill's work, and explore the accuracy and precision of a wider range of sampling schemes, the coverage tests of chapters 3 and 4, which are even more computationally demanding, will fall back once again on these seven examples.

Note that Wichmann's 7 sampling schemes are designed specifically for forced-choice experimental designs, and in particular the 2-AFC design. Many of them are asymmetric about the mid-point of the curve, to investigate and take advantage of the asymmetry in the variability about different points on a 2-AFC psychometric function (see section 5.3). As such they do not constitute a reasonable set of examples for the effects of sampling scheme variation in yes-no designs. Section 1.5.2 therefore defines a different set of examples for the purposes of the yes-no coverage simulations performed in chapter 3 and in section 4.3.3.

1.5.1 Seven sampling schemes for the 2-AFC design

Figure 1.2 shows the seven sampling schemes devised by Wichmann and Hill[9,10] as illustrative examples of the effect of sampling placement in 2-AFC experiments. The figure is similar to Figure 1 from Wichmann and Hill's second paper.[10] Although $\boldsymbol{\theta}$ is different in the figure, the values of $F(x)$ and $\psi(x)$ for each sampling scheme, which are listed in table 1.1, are identical to those used in the paper.

In pilot simulations, sampling schemes were designed by hand in order to explore a number of ways in which the pattern of stimulus values can vary: bias towards the low or high end of the function, wide or narrow spacing, and clustering towards or away from the mid-point. Differences in these attributes were found to yield significantly different results in terms of the bias and precision of threshold and slope estimation. The seven schemes of figure 1.2 were selected in order to provide a set of sampling schemes

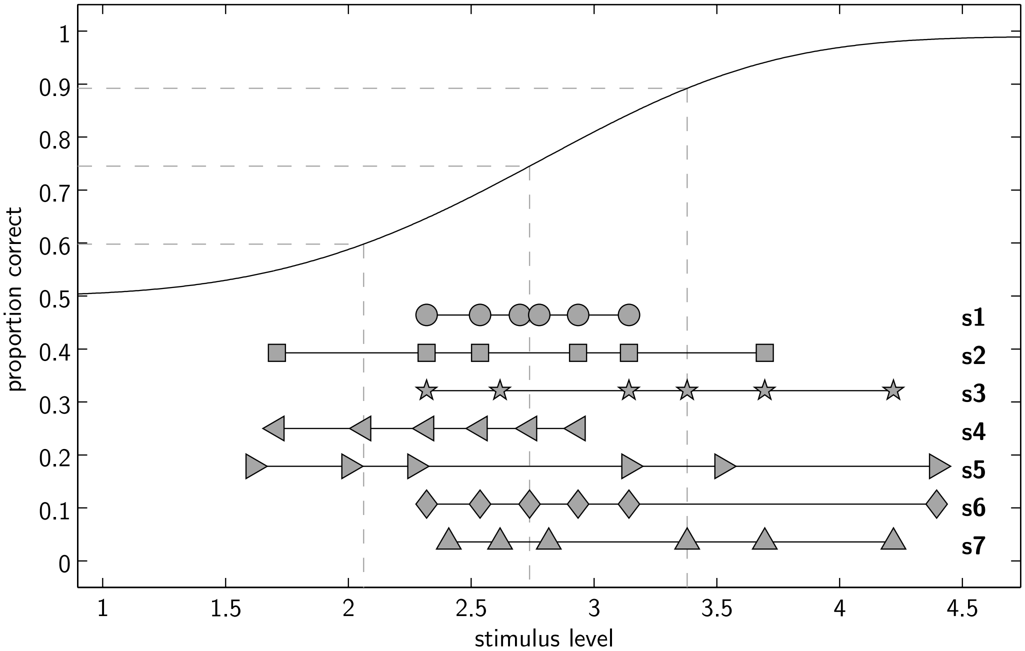


Fig. 1.2: Seven sampling schemes (each represented as a chain of symbols) after Wichmann and Hill.[9,10] A 2-AFC Weibull psychometric function with $\boldsymbol{\theta} = (3, 4, 0.5, 0.01)^{\mathrm{T}}$ is plotted to show the location of each sample relative to the underlying function. The dotted lines show $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ for this function, corresponding to performance levels of 0.598, 0.745 and 0.892, respectively.

ψ **values (**f **values)**

s1 ●	s2 ■	s3 ★	s4 ◀	s5 ▶	s6 ⧫	s7 ▲
0.647 (0.30)	0.549 (0.10)	0.647 (0.30)	0.549 (0.10)	0.539 (0.08)	0.647 (0.30)	0.667 (0.34)
0.696 (0.40)	0.647 (0.30)	0.716 (0.44)	0.598 (0.20)	0.588 (0.18)	0.696 (0.40)	0.716 (0.44)
0.735 (0.48)	0.696 (0.40)	0.843 (0.70)	0.647 (0.30)	0.637 (0.28)	0.745 (0.50)	0.765 (0.54)
0.755 (0.52)	0.794 (0.60)	0.892 (0.80)	0.696 (0.40)	0.843 (0.70)	0.794 (0.60)	0.892 (0.80)
0.794 (0.60)	0.843 (0.70)	0.941 (0.90)	0.745 (0.50)	0.916 (0.85)	0.843 (0.70)	0.941 (0.90)
0.843 (0.70)	0.941 (0.90)	0.980 (0.98)	0.794 (0.60)	0.985 (0.99)	0.985 (0.99)	0.980 (0.98)

Table 1.1: For each of the sampling schemes of figure 1.2, the f values (by which the sampling schemes are defined) are given in parentheses after the ψ values that are produced when that sampling scheme is used with $\gamma = 0.5$ and $\lambda = 0.01$ (this particular combination of parameters is used in section 5.4 and for many of the tests of chapter 3). Each column corresponds to one sampling scheme.

that might occur in practice, but which were maximally illustrative of such effects.[†]

1.5.2 Seven sampling schemes for the yes-no design

As the sampling schemes of section 1.5.1 were specifically designed for use with 2-AFC psychometric functions, a new set of seven was chosen for to explore the yes-no design. Their values are given in table 1.2 and plotted in figure 1.3.

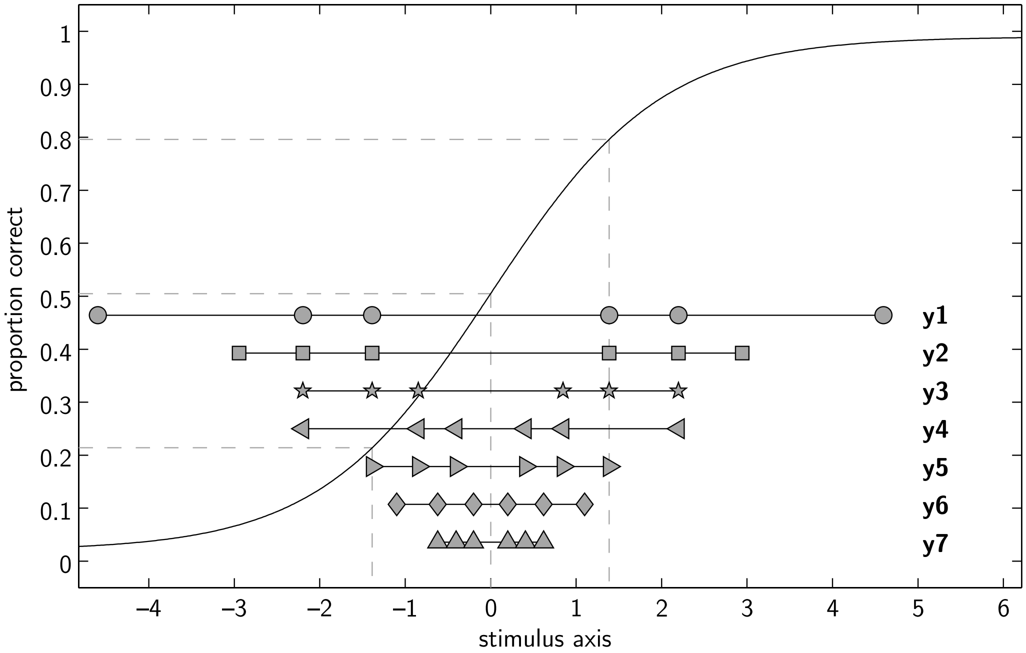


Fig. 1.3: Seven sampling schemes, each represented as a chain of symbols. A logistic psychometric function with $\boldsymbol{\theta} = (0, 1, 0.02, 0.01)^{\mathrm{T}}$ is plotted for comparison. The dotted lines show $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ for this function, corresponding to performance levels of 0.214, 0.505 and 0.796 respectively.

[†] Note that in all the simulations of this report, the true psychometric function of the simulated observer is constant, and unaffected by any psychological influence of the range or order of presentation of the stimulus levels.

ψ values (f values)

y1 ●	y2 ■	y3 ★	y4 ◀	y5 ▶	y6 ⧫	y7 ▲
0.030 (0.01)	0.069 (0.05)	0.117 (0.10)	0.117 (0.10)	0.214 (0.20)	0.263 (0.25)	0.359 (0.35)
0.117 (0.10)	0.117 (0.10)	0.214 (0.20)	0.311 (0.30)	0.311 (0.30)	0.359 (0.35)	0.408 (0.40)
0.214 (0.20)	0.214 (0.20)	0.311 (0.30)	0.408 (0.40)	0.408 (0.40)	0.456 (0.45)	0.456 (0.45)
0.796 (0.80)	0.796 (0.80)	0.699 (0.70)	0.602 (0.60)	0.602 (0.60)	0.553 (0.55)	0.553 (0.55)
0.893 (0.90)	0.893 (0.90)	0.796 (0.80)	0.699 (0.70)	0.699 (0.70)	0.650 (0.65)	0.602 (0.60)
0.980 (0.99)	0.942 (0.95)	0.893 (0.90)	0.893 (0.90)	0.796 (0.80)	0.748 (0.75)	0.650 (0.65)

Table 1.2: For each of the sampling schemes of figure 1.3, the f values (by which the sampling schemes are defined) are given in parentheses after the ψ values that are produced when that sampling scheme is used with $\gamma = 0.02$ and $\lambda = 0.01$ (the parameters used in section 3.2.2). Each column corresponds to one sampling scheme.

References for chapter 1

[1] Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data.* London: Chapman and Hall., second edition.

[2] Watson, A. B. (1979). Probability summation over time. *Vision Research*, **19**: 515–522.

[3] 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.

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

[5] Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, **47**(1): 87–91.

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

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

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

[9] 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 .

[10] 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 .

[11] 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.

[12] 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.

[13] 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.

[14] Simpson, W. A. (1988). The method of constant stimuli is efficient. *Perception and Psychophysics*, **44**(5): 433–436.

[15] Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

[16] Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**(2): 215–223.

[17] Wichmann, F. A. (1999). *Some Aspects of Modelling Human Spatial Vision: Contrast Discrimination.* PhD thesis, University of Oxford, UK.

[18] Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America*, **2**: 1508–1532.

[19] Kontsevich, L. L. & Tyler, C. W. (1999). Distraction of attention and the slope of the psychometric function. *Journal of the Optical Society of America*, **A16**: 217–222.

[20] 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.

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

[22] Wetherill, G. B. (1963). Sequential estimation of quantal reponse curves. *Journal of the Royal Statistical Society, Series B*, **25**(1): 1–48.

[23] Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**(6): 1763–9.

[24] Leek, M. R, Hanna, T. E. & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception and Psychophysics*, **51**(3): 247–256.

[25] King-Smith, P. E. & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, **37**(12): 1595–1604.

[26] Kontsevich, L. L. & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, **39**(16): 2729–2737.

[27] Tyler, C. W. (1997). Why we need to pay attention to psychometric function slopes. *Vision Science and its Applications, Technical Digest*, **1**: 240–243.

[28] Patterson, V. H, Foster, D. H. & Heron, J. R. (1980). Variability of visual threshold in Multiple Sclerosis. *Brain*, **103**: 139–147.

[29] Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, **7**(4): 308–313.

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

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

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

[33] 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.

[34] 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.

[35] 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.

2. Confidence interval methods

Sections 2.1 and 2.2 describe methods of obtaining confidence intervals for the threshold and slope of a psychometric function. If u denotes the measure of interest (threshold or slope), then the confidence limits $[u_{\text{LO}}, u_{\text{UP}}]$ denote the lower and upper endpoints of a two-tailed interval which is designed to have *coverage* equal to $1 - 2\eta$, i.e. to contain the true underlying value u_{gen} on a proportion of occasions equal to $1 - 2\eta$. The tails of the interval should be balanced, so that the probability of rejecting the true value in each tail is η.

Section 2.1 describes the confidence interval methods based on probit analysis. Section 2.2 describes the basic principles of the bootstrap and six different bootstrap confidence interval methods.

2.1 Probit methods

Probit analysis has a long history in psychology which can be traced back to Fechner's experiments on weight discrimination in 1860. An account is given by Finney,[1] who details the methods and their application with specific reference to the dose-response curves of medicine and toxicology.

The term "probit" is a contraction of "probability unit" which reflects the approach of treating the response variable on a transformed scale $y' = \Upsilon^{-1}(y)$, where Υ is the "link" function.[2] It provides a method of fitting the psychometric function by reducing it to a linear function $y' = \beta_{\text{pr}}(x - \alpha_{\text{pr}})$. The term was originally coined with the cumulative normal function specifically in mind as the link function, the "probit of y" being defined as $\Phi^{-1}(y) + 5$ where $\Phi^{-1}(\cdot)$ is the inverse cumulative of the standard normal distribution, but the theory can be applied to other shapes. When the logistic function is used, the units are often called “logits”. For the purposes of the current study, “probit” will be treated as a generic term, not specifically tied to any particular psychometric function shape, which can embrace “logits”, “normits” (a term which is also sometimes applied to analyses that use the cumulative normal function) and other kinds of transformed response scale.

Fitting a function by probit analysis involves an iterative procedure in which the location and scale parameters α_{pr} and β_{pr} are estimated by weighted linear regression. The weight on each block, W_i, is updated after each iteration—it is always proportional to n_i but also takes into account the expected binomial variability around each the probability value $\Upsilon(\hat{\beta}_{\text{pr}}\,[x_i-\hat{\alpha}_{\text{pr}}])$, and the slope of the function, $\mathrm{d}\Upsilon/\mathrm{d}x$. The estimates $\hat{\alpha}_{\text{pr}}$ and $\hat{\beta}_{\text{pr}}$ converge to their maximum-likelihood values. However, when there are additional parameters in the fit (for example, when $\Upsilon(\cdot)$ scales the function between unknown lower and upper bounds γ and $1-\lambda$) the iterative procedure is not necessarily the most efficient or reliable way of maximizing likelihood. Since a maximum-likelihood procedure has already been defined in section 1.3, probit analysis will not be used to obtain parameter estimates. Rather, the interesting aspect of probit analysis for current purposes is that it provides asymptotic formulae for the variance of α_{pr} and β_{pr}, and for confidence limits on α_{pr}.

The formulae provided by Finney[1] are adapted here to yield the variance of the threshold $t_{0.5}$ and slope $s_{0.5}$. The function $\Upsilon(\cdot)$ need not be formulated, except to state that $\Upsilon(\beta_{\text{pr}}\,[x_i-\alpha_{\text{pr}}])$ is identical to $\psi(x;\,\hat{\boldsymbol{\theta}}_0)$ when $\alpha_{\text{pr}}=\hat{t}_{0.5}$ and $\beta_{\text{pr}}=\kappa\,\hat{s}_{0.5}$. The constant κ depends on the shape of the psychometric function, but it is eliminated in the derivation of the following.

A generalized formula for the probit weights is

$$W_i=\frac{n_i\,\dot{p}_i^2}{p_i\,(1-p_i)}\quad,\quad\text{where}\;\;\dot{p}_i=\left.\frac{\mathrm{d}\psi}{\mathrm{d}x}\right|_{x_i}\quad. \tag{2.1}$$

When the weights are evaluated at the MLE, so that each p_i is equal to $\psi(x_i;\hat{\boldsymbol{\theta}}_0)$, they can then be used to obtain the standard error se_t for the estimated threshold $\hat{t}_{0.5}$:

$$\mathrm{se}_t^2 = \frac{1}{\sum W_i} + \frac{\left(\hat{t}_{0.5} - \bar{x}\right)^2}{\sum W_i \left(x_i - \bar{x}\right)^2} \ , \tag{2.2}$$

and standard error se_s for the estimated slope $\hat{s}_{0.5}$:

$$\mathrm{se}_s^2 = \frac{\hat{s}_{0.5}^2}{\sum W_i \left(x_i - \bar{x}\right)^2} \ , \tag{2.3}$$

where $\bar{x} = \sum W_i \, x_i / \sum W_i$, and the sum is taken over blocks $i = 1 \ldots k$.

Standard error estimates (2.2) and (2.3) may be used to compute standard confidence intervals of coverage $1 - 2\eta$ as follows:

$$[u_{\mathrm{LO}}, \quad u_{\mathrm{UP}}] = \hat{u}_0 \mp \mathrm{se}_u \ \Phi^{-1}(1 - \eta). \tag{2.4}$$

Probit standard error estimates were examined for the threshold and slope of a yes-no psychometric function by Foster and Bischof.[3] At lower values of N, they were found to be less accurate and precise than the bootstrap standard error method (section 2.2.1), particularly for slopes.

In addition to the normal-theory limits of equation (2.4), probit analysis provides confidence limits for the threshold based on fiducial bands around the estimated psychometric function. The probit fiducial limits[†] for $\hat{t}_{0.5}$ are given by

$$[t_{\mathrm{LO}}, \quad t_{\mathrm{UP}}] = \hat{t}_{0.5} + \frac{g}{1-g}(\hat{t}_{0.5} - \bar{x}) \mp \frac{z}{1-g}\sqrt{\frac{1-g}{\sum W_i} + \frac{\left(\hat{t}_{0.5} - \bar{x}\right)^2}{\sum W_i \left(x_i - \bar{x}\right)^2}} \ , \tag{2.5}$$

[†] The term “fiducial” refers to one of two ways of looking at the construction of a confidence interval, both of which yield the same results in most situations. It is not proposed to labour the distinction here: the potentially controversial word “fiducial” is used purely as a label to refer to method (2.5), because this is the term used by Finney.[1] See Edwards (1992)[4] for a discussion of the fiducial argument.

where $z = \Phi^{-1}(1 - \eta)$ for an interval of overall coverage $1 - 2\eta$, and

$$g = \frac{z^2}{\sum W_i \left(x_i - \bar{x}\right)^2} \quad .$$

Probit fiducial limits for psychophysical thresholds were investigated in 2-AFC designs by Teller[5] and by McKee, Klein and Teller.[6] The latter study noted that the confidence limits tended to be fairly accurate (as compared with limits obtained from Monte Carlo simulation for a known psychometric function) when $N \geq 100$, although they made appreciable errors (particularly in the lower confidence limit) at lower values of N.

2.2 Bootstrap methods

Bootstrap techniques were introduced in 1979 by Bradley Efron, since which time they have undergone rapid development and have been applied in a wide variety of situations. Good tutorial accounts and reviews of the statistical literature can be found in the books by Efron and Tibshirani[7] and by Davison and Hinkley.[8]

The name "bootstrap" arises from the expression "to pull oneself up by one's bootstraps", after the feat performed by the mythical character Baron Munchausen to save himself from sinking into a swamp. Thus, it has the connotation of performing an impossible task, getting something for nothing or, in a statistical context, apparently obtaining information from nowhere. In fact, a bootstrap technique is no more magical in this sense than the use of an ordinary statistical "plug-in" estimate of variability: in both, the probability distribution of estimates around a true (unknown) value u_{gen} is approximated by using the experimenter's estimate $\hat{u}_0$ in place of u_{gen}. The difference is that the bootstrap technique uses Monte Carlo simulation to approximate the probability distribution rather than relying on parametric assumptions about its shape—assumptions that are often only asymptotically correct.

Thus, bootstrap confidence limits for an estimate $\hat{u}$ are based on the *bootstrap distribution* which is estimated by taking a large number R of Monte Carlo estimates $\hat{u}_1^* \ldots \hat{u}_R^*$. Each bootstrap estimate $\hat{u}_i^*$ is derived in exactly the same way as the initial estimate, but from a *simulated* set of responses $\boldsymbol{r}_i^*$. Bootstrap techniques may be parametric or non-parametric, a distinction which refers to the way in which simulated data are generated. The non-parametric method is to sample with replacement from the original data. For psychophysical data this means that, in the j th block of the i th simulated data set, the number of correct responses r_{ij}^* is the total number of ones in a set of size n_j drawn with replacement from the original set of r_j ones and $n_j - r_j$ zeros. Thus, $r_{\cdot j}^*$ is binomially distributed with probability of success y_j in n_j trials.[†] Foster and Bischof[9] applied this form of the bootstrap to the psychometric function, and report its accuracy to be greater than that of an asymptotic method based on the "combination of observations", particularly when N is low. The non-parametric bootstrap was also recommended by Treutwein and Strasburger.[14]

The parametric bootstrap, on the other hand, uses a parametric model to obtain the generating probability in each block. Wichmann and Hill[12] point out that, when a function has already been fitted to psychophysical data in order to estimate threshold and slope, a parametric model has already been assumed. Generating probability values are therefore available without the need for any extra assumptions. In the parametric bootstrap, $r_{\cdot j}^*$ is binomially distributed with probability of success $p_j = \psi(x_j; \hat{\boldsymbol{\theta}}_0)$. This is the form in which the bootstrap is generally applied to psychometric functions,[3,10,12,13,15]

[†] Though this approach is conventionally referred to as non-parametric, it is not without assumptions about how the data are generated. In their simplest form, both non-parametric and parametric bootstrap methods assume that the individual observations of a block are independent and identically distributed. Under such an assumption, bootstrap resampling from block of Bernoulli trials will automatically yield a binomially distributed total. Though this assumption is made in the simulations presented here and elsewhere,[3,9–13] its validity is questionable given that serial correlations between an observer's responses may in fact occur, leading to supra-binomial variability. Alternative non-parametric resampling techniques, such as a block-based resampling system *within* each block of trials, may be suitable to address this issue. In order to test the effectiveness of such a method, a model for the observer's trial-to-trial behaviour would have to be assumed in simulation. This is beyond the scope of the current study, which tests confidence interval methods under the assumption of independent identically distributed responses.

and which will be used in the current study.

Provided the distribution $(\hat{u}_i^* - \hat{u}_0)$ of bootstrap estimates around the initial estimate is similar to the distribution of estimates around the true value $(\hat{u}_0 - u_{\mathrm{gen}})$ the accuracy of the bootstrap method is limited only by the number of simulations R. The shape of the distribution is estimated directly by simulation, rather than by a method whose accuracy relies on a large enough N, which is why bootstrap methods are often found to be more accurate than asymptotic methods when N is low. However, the bootstrap does rely on the shape of the bootstrap distribution around $\hat{u}_0$ being sufficiently similar to the true distribution around u_{gen}. The strength of this assumption (i.e. the validity of the "plug-in" principle,[7] or as Wichmann and Hill[12] call it, the "bootstrap bridging assumption") usually relies on numerical closeness of $\hat{u}_0$ to u_{gen}, which *does* tend to be better at higher N.

There are many different ways of using the bootstrap distribution to obtain confidence limits, some of which use more information from the bootstrap distribution to correct for errors in the bridging assumption than others. Six variations are described in the following sub-sections. Depending on the application, the accuracy of the different variations may differ. The bootstrap standard error, basic bootstrap and bootstrap percentile methods are generally found to be *first-order accurate*, which means that, to a first approximation, the error in their coverage probability is proportional to $N^{-\frac{1}{2}}$. Various improvements to the bootstrap method have been developed, including the bootstrap-t method (section 2.2.3) and the $\mathrm{BC_a}$ method (section 2.2.5), which can be shown to be *second-order accurate* (error in coverage probability $\propto N^{-1}$) in many applications.[†]

With the exception of the bootstrap standard error method, all the methods described below are based on estimated quantiles of the bootstrap dis-

[†] Although two methods may be shown to have the same *order* of accuracy, their absolute accuracy may differ depending on the context in which they are applied: coverage error might be proportional to (say) $N^{-\frac{1}{2}}$ for both methods, but with different constants of proportionality. Each new application of bootstrap methods requires a new theoretical analysis and new simulation studies (of the kind reported in chapters 3, 4 and 6) to compare the performance of different bootstrap variations.

tribution $\hat{u}^*$. A subscript in parentheses will be used to denote an estimated quantile.[†] For example, the η quantile of the distribution of R bootstrap values $\hat{u}^*$ is written as $\hat{u}^*_{(\eta)}$, and is estimated by taking the $(R+1)\eta$ $^{\text{th}}$ ordered value of the distribution. Where $(R+1)\eta < 1$ or $(R+1)\eta > R$, the result is undefined, and where it is not a whole number, the result is linearly interpolated between the nearest two values. The reverse process will be known as the *cumulative probability estimate* or CPE, given by

$$\text{CPE}\{u; u^*\} = \frac{1}{R+1} \sum_{i=1}^{R} \mathrm{I}\{u_i^* \leq u\}, \tag{2.6}$$

where $\mathrm{I}\{\cdot\}$ is the indicator function. Note that, assuming there are no repeated values in u^*, the two processes are exact inverses of one another for a value $u \in u^*$, so that $\text{CPE}\{u^*_{(\eta)}; u^*\} \equiv \eta$ when $(R+1)\eta$ is an integer.

It is recommended[8] that R be at least 999 when confidence levels of the order of 0.95 and 0.99 are to be considered. The current study uses $R = 1999$.

Sections 2.2.1–2.2.5 detail the bootstrap standard error, basic bootstrap, bootstrap-t, bootstrap percentile and BC_a methods. The equations are adapted from Davison and Hinkley[8] sections 5.2 and 5.3, with appropriate substitution and rearrangement of the notation for the current purpose. Section 2.2.6 describes and illustrates the expanded bootstrap method suggested by Wichmann and Hill.[12]

2.2.1 The bootstrap standard error method

The bootstrap distribution can be used in order to estimate the standard deviation of the true distribution, which is then used as a standard error estimate in the construction of the standard interval of overall coverage $1-2\eta$:

$$[u_{\text{LO}},\quad u_{\text{UP}}] = \hat{u}_0 \mp \hat{\text{se}}_u\ \Phi^{-1}(1-\eta), \tag{2.7}$$

[†] "Quantiles" will be discussed rather than "percentiles", so that the subscript value is always in the range $(0, 1)$ and the superfluous factor 100 can be omitted. However, the name "bootstrap percentile" will be retained as it is the conventionally recognized name for the method described in section 2.2.4.

where $\Phi^{-1}(\cdot)$ is the inverse cumulative of the standard normal distribution and $\hat{\mathrm{se}}_u$ is the standard deviation the bootstrap values $\hat{u}^*_1 \ldots \hat{u}^*_R$.

The bootstrap standard error method was applied to the threshold and slope of the psychometric function by Foster and Bischof,[3,9,10] and was found to be a better estimator of the true standard error at low N than either the probit method of equation (2.4) or the "incremental" method.[16]

As confidence intervals, bootstrap standard intervals (2.7) are generally found to be first-order accurate. They are often less accurate than other first-order accurate methods, because their accuracy is limited by the extent to which the true distribution is normal, and therefore symmetrical—Efron and Tibshirani[7] point out that "the most serious errors made by standard intervals are due to their enforced symmetry" (page 180).

2.2.2 The basic bootstrap method

The essence of the bootstrap is to take $\hat{u}^* - \hat{u}_0$ (the distribution of bootstrap estimates around the initial estimate) as an approximation to $\hat{u}_0 - u_{\mathrm{gen}}$ (the distribution of estimates around the true underlying value).

By taking such a step, confidence limits can be computed directly, without any parametric assumptions about the form of the distribution $(\hat{u}_0 - u_{\mathrm{gen}})$, as follows. Consider the low tail of a two-tailed confidence interval of overall coverage $1 - 2\eta$. The limit ϱ is the value such that

$$\Pr(u_{\mathrm{gen}} \leq \varrho) = \eta,$$

which is equivalent to

$$\Pr(\hat{u}_0 - u_{\mathrm{gen}} \geq \hat{u}_0 - \varrho) = \eta.$$

The bootstrap step is to make the substitution on the left hand side of the inequality, so that

$$\Pr(\hat{u}^* - \hat{u}_0 \geq \hat{u}_0 - \varrho) = \eta,$$

which can then be re-arranged as

$$\Pr(\varrho \geq 2\hat{u}_0 - \hat{u}^*) = \eta.$$

This is satisfied by taking

$$\varrho = 2\hat{u}_0 - \hat{u}^*_{(1-\eta)}.$$

Thus, the basic bootstrap limits for a two-tailed interval of coverage $1 - 2\eta$ are given by:

$$[u_{\text{LO}}, \;\; u_{\text{UP}}] = \left[2\hat{u}_0 - \hat{u}^*_{(1-\eta)}, \;\; 2\hat{u}_0 - \hat{u}^*_{(\eta)}\right]. \tag{2.8}$$

Note the reversal of the quantiles of $\hat{u}^*$: the lower limit is computed using the higher quantile of the bootstrap distribution, and vice versa. This may seem counter-intuitive at first, but it is appropriate when considering asymmetric distributions. If, for example, the bootstrap distribution has a long upper and short lower tail, and this is an accurate estimate of the shape of the true distribution, then it is entirely appropriate for the confidence interval to have a long *lower* arm and short *upper* arm: the estimator is likely to have made a larger positive error, to arrive at the observed value from below (if the true value lies below) than the negative error it is likely to have made in order to arrive at the observed value from above (if the true value lies above).

Hall[17,18] refers to (2.8) as the "bootstrap percentile" method, a classification which will *not* be used here. The nomenclature of Efron and Tibshirani[7] and Davison and Hinkley[8] is preferred, in which "bootstrap percentile" refers to the use of the quantiles of $\hat{u}^*$ without reversal (see section 2.2.4, below).

In the limit as $R \to \infty$, the basic bootstrap is correct for distributions of arbitrary shape, *provided* that the bootstrap distribution is of identical shape to the true distribution. In practice, this is rarely the case, and such an assumption generally leads to confidence intervals whose coverage converges fairly slowly to the desired level. Basic bootstrap intervals are generally found to be first-order accurate.

2.2.3 The bootstrap-t method

The Studentized bootstrap or bootstrap-t method is a development of the basic bootstrap. It is found to be second-order accurate in many applications, including the logistic regression design treated by Lee,[19] which is analogous to a yes-no psychometric function without nuisance parameters.

The bootstrap-t method relies on a separate approximation for $\hat{v}$, the variance of u, for which any one of a number of methods may be used. The method considered here uses the asymptotic approximation to the parameter covariance matrix $\hat{\boldsymbol{V}}$ provided by the inverse of the expected Fisher information matrix $\hat{\boldsymbol{I}}$ (see the appendix, section B.2). The vector $\dot{\boldsymbol{u}}$ of derivatives of u with respect to each the parameters (see the appendix, section B.1) is used in conjunction with $\hat{\boldsymbol{V}}$ to obtain an approximation to the variance $\hat{v}$ of u

$$\hat{v} = \dot{\boldsymbol{u}}\ \hat{\boldsymbol{V}}\ \dot{\boldsymbol{u}}^{\mathrm{T}}. \tag{2.9}$$

(Thus if u is one of the parameters, θ_i, then $\hat{v}$ is simply the appropriate element $\hat{V}_{ii}$ on the diagonal of the covariance matrix.)

The initial estimated variance $\hat{v}_0$ is obtained using the initial parameter estimate $\hat{\boldsymbol{\theta}}_0$ to evaluate $\hat{\boldsymbol{I}}$ and $\dot{\boldsymbol{u}}$. In addition, a bootstrap estimate $\hat{v}^*_i$ is obtained on each simulation using the bootstrap parameter set $\hat{\boldsymbol{\theta}}^*_i$. The intention is to create an approximately pivotal distribution $\hat{z}^*$, where

$$\hat{z}^*_i = N^{-\frac{1}{2}}\ \hat{v}^{*\,-\frac{1}{2}}_i\ (\hat{u}^*_i - \hat{u}_0). \tag{2.10}$$

The appropriate quantiles of $\hat{z}^*$ are then transformed back onto a meaningful scale by multiplying by the square root of the initial variance estimate, to yield confidence limits analogous to (2.8), as follows:

$$[u_{\mathrm{LO}},\ \ u_{\mathrm{UP}}] = \left[\hat{u}_0 - N^{\frac{1}{2}}\ \hat{v}_0^{\ \frac{1}{2}}\ \hat{z}^*_{(1-\eta)},\quad \hat{u}_0 - N^{\frac{1}{2}}\ \hat{v}_0^{\ \frac{1}{2}}\ \hat{z}^*_{(\eta)}\right]. \tag{2.11}$$

Alternative methods for the computation of $\hat{v}$ include the non-parametric delta method (see sections 2.7.2, 3.2.1 and 5.2.2 of Davison and Hinkley[8]), and various forms of jackknife method.[7,20] The simulations of chapter 3 will use (2.9) because the additional level of iteration required by the non-parametric methods would make repeated Monte Carlo simulation of the entire bootstrap process impracticable. It is sometimes recommended[8,21] that the *observed* rather than expected Fisher information be used to obtain $\hat{\boldsymbol{V}}$. The expected information matrix will be used here, however, in order to allow direct comparison with the simulation studies of Lee[22] and of Swanepoel and Frangos,[20] both of which used (2.9) based on $\hat{\boldsymbol{V}} = \hat{\boldsymbol{I}}^{-1}$ in the context of logistic regression.

A two-dimensional version of the bootstrap-t method will also be investigated in chapter 4: following the method of Hall,[17] the covariance matrix $\hat{\boldsymbol{V}}$ is used instead of the scalar variance estimate $\hat{v}$ in (2.10) and (2.11)—see page 129.

2.2.4 The bootstrap percentile method

The bootstrap percentile method is similar to the basic bootstrap, in that quantiles of the bootstrap distribution are used without any kind of non-linear transformation. It differs from the basic bootstrap, however, in that the quantiles are used unswapped, so that

$$[u_{\text{LO}}, \quad u_{\text{UP}}] = \left[\hat{u}^*_{(\eta)}, \quad \hat{u}^*_{(1-\eta)}\right]. \tag{2.12}$$

Hall[18] is critical of (2.12), calling it the "backwards bootstrap" because of its failure to reverse the quantiles of $\hat{u}^*$ as the logic of section 2.2.2 would demand, and he applies the name "bootstrap percentile" to equation (2.8). Nonetheless, Davison and Hinkley[8] provide an ingenious theoretical justification, which is to suppose that there is some monotonic function $\zeta(\hat{u}^*)$ that normalizes the bootstrap distribution. The quantiles of the transformed distribution may be freely reversed, or not, because the distribution is symmetrical. Thus, lower and upper confidence limits may be obtained on the transformed scale by $\zeta(\hat{u}^*)_{(\eta)}$ and $\zeta(\hat{u}^*)_{(1-\eta)}$, respectively. As $\zeta(\cdot)$ is monotonic, transformation back onto the u axis, $\zeta^{-1}\left[\zeta(\hat{u}^*)_{(\eta)}\right]$ and $\zeta^{-1}\left[\zeta(\hat{u}^*)_{(1-\eta)}\right]$, yields limits that are equal to $\hat{u}^*_{(\eta)}$ and $\hat{u}^*_{(1-\eta)}$, respectively.

The bootstrap percentile method is only first-order accurate in theory, but Efron and Tibshirani[7] observe that it is often more reliable than the bootstrap-t method in practice, because the latter often tends be more heavily influenced by a few outliers.

Bootstrap percentile intervals were applied in the context of 2-AFC psychometric functions by Maloney[13] and by Wichmann and Hill.[12,15]

2.2.5 The $\mathrm{BC_a}$ method

The *bias-corrected accelerated* or $\mathrm{BC_a}$ method was introduced by Efron[23] as a second-order accurate adjustment to the bootstrap percentile method. It is also described by Efron and Tibshirani[7] and by Davison and Hinkley.[8] The following account is adapted from Davison and Hinkley[8] (section 5.2.3 on pages 203–207 and problem 7 on page 249).

Confidence limits are computed as for the bootstrap percentile method, using quantiles of the bootstrap distribution directly:

$$[u_{\mathrm{LO}},\ \ u_{\mathrm{UP}}] = \left[\hat{u}^*_{(\tilde{\varepsilon}_{\mathrm{LO}})},\ \ \hat{u}^*_{(\tilde{\varepsilon}_{\mathrm{UP}})}\right], \tag{2.13}$$

where $\tilde{\varepsilon}_{\mathrm{LO}}$ and $\tilde{\varepsilon}_{\mathrm{UP}}$ are bias-corrected accelerated versions of the unadjusted confidence levels $\varepsilon_{\mathrm{LO}} = \eta$ and $\varepsilon_{\mathrm{UP}} = 1 - \eta$, respectively. The adjustment is

$$\tilde{\varepsilon} = \Phi\left(w + \frac{\Phi^{-1}(\varepsilon) + w}{1 - \xi\left(\Phi^{-1}(\varepsilon) + w\right)}\right),$$

where w is the bias correction term, and ξ is the *acceleration* or skewness correction factor. The bias correction term is defined by

$$\Phi(w)\ =\ \Pr(\hat{u}_0 < u_{\mathrm{gen}}\,|\,u_{\mathrm{gen}})\ \approx\ \Pr(\hat{u}^* < \hat{u}_0\,|\,\hat{u}_0),$$

which leads to the following expression based on the bootstrap distribution:

$$w = \Phi^{-1}\left(\frac{1}{R+1}\sum_{i=1}^{R}\mathrm{I}\{\hat{u}^*_i \le \hat{u}_0\}\right), \tag{2.14}$$

where I$\{\cdot\}$ is the indicator function. Estimation of the skewness correction factor ξ is somewhat more involved. One method is to use the non-parametric delta method mentioned in section 2.2.3. The current study uses the parametric approximation

$$\xi \;=\; \frac{\mathrm{E}^*\{\dot{\ell}^{*}_{\mathrm{LF}}{}^{3}\}}{6\ \mathrm{var}^*\left\{\dot{\ell}^{*}_{\mathrm{LF}}\right\}^{\frac{3}{2}}} \quad .$$

The i $^{\mathrm{th}}$ value in the bootstrap distribution $\dot{\ell}^{*}_{\mathrm{LF}}$ is the derivative of log-likelihood, evaluated at the bootstrap parameter estimate $\hat{\boldsymbol{\theta}}^{*}_{i}$, taken in the *least favourable direction* in the parameter space. Thus, acceleration is equal to one sixth of the ratio between the mean of the cubed bootstrap derivatives and the cube of the standard deviation of the bootstrap distribution of derivatives. Each bootstrap log-likelihood derivative value is given by

$$\dot{\ell}^{*}_{\mathrm{LF}i} = \left.\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\hat{\boldsymbol{\theta}}^{*}_{i}} \cdot\, \hat{\boldsymbol{\delta}}_{u},$$

where $\hat{\boldsymbol{\delta}}_u$ is a unit vector denoting the least favourable direction in parameter space for the purposes of inference about the measure of interest u. It is given by

$$\hat{\boldsymbol{\delta}}_u = \frac{\hat{\boldsymbol{I}}^{-1}\dot{\boldsymbol{u}}}{\left\|\hat{\boldsymbol{I}}^{-1}\dot{\boldsymbol{u}}\right\|} \quad ,$$

where $\hat{\boldsymbol{I}}$ is the expected Fisher information matrix and $\dot{\boldsymbol{u}}$ is the vector of derivatives of u with respect to each of the parameters $\boldsymbol{\theta}$. For further details on the computation of $\partial\ell/\partial\boldsymbol{\theta}^{\mathrm{T}}$ and $\hat{\boldsymbol{I}}$, see the appendix, section B.2. For the computation of $\dot{\boldsymbol{u}}$, see section B.1.

The BC$_{\mathrm{a}}$ method was recommended for use in the context of psychometric functions by Wichmann and Hill,[12] based on reports that it yielded improvements in coverage accuracy in other applications, but without theoretical or empirical support for its application in psychophysics.

2.2.6 Expanded bootstrap intervals

The expanded method was proposed by Wichmann and Hill[12] as a method of examining the sensitivity of the confidence limits to error in the initial fit, and providing conservative confidence limits. From the initial estimate ${}^0\hat{\boldsymbol{\theta}}_0$, the initial estimate of the measure of interest ${}^0\hat{u}_0$ is computed, and a confidence interval $[{}^0u_{\mathrm{LO}}, {}^0u_{\mathrm{UP}}]$ is obtained by one of the above bootstrap methods. The next step is to examine how the interval expands as ${}^0\hat{\boldsymbol{\theta}}_0$ changes, in order to obtain an indication of how the interval might be affected by error in the initial estimate (in other words, to assess the extent to which the bootstrap bridging assumption fails). The method is designed to test possible errors in a number of different directions in parameter space. To this end, a confidence *region* (see chapter 4) is computed, expressing likely variation in all the parameters simultaneously, and new parameter sets ${}^1\hat{\boldsymbol{\theta}}_0, \ldots {}^8\hat{\boldsymbol{\theta}}_0$ are chosen at eight different locations on the boundary of the region. The bootstrap is then re-run eight times, substituting each of the new parameter sets ${}^i\hat{\boldsymbol{\theta}}_0$ for the original estimate ${}^0\hat{\boldsymbol{\theta}}_0$, to obtain a new pair of limits $[{}^iu_{\mathrm{LO}}, {}^iu_{\mathrm{UP}}]$. The expanded limits are then given by

$$[u_{\mathrm{LO}},\ \ u_{\mathrm{UP}}] = \left[{}^0\hat{u}_0 - \max_{i=0\ldots8}\left\{{}^i\hat{u}_0 - {}^iu_{\mathrm{LO}}\right\},\quad {}^0\hat{u}_0 + \max_{i=0\ldots8}\left\{{}^iu_{\mathrm{UP}} - {}^i\hat{u}_0\right\}\right]. \qquad (2.15)$$

Thus the expanded interval represents a “worst-case” scenario, encompassing the confidence limits furthest removed from the initial estimate, given that it is not known which of the parameter sets ${}^0\hat{\boldsymbol{\theta}}_0, \ldots {}^8\hat{\boldsymbol{\theta}}_0$ yields the best bootstrap approximation to the true probability distribution. Depending on how widely the secondary parameter sets ${}^1\hat{\boldsymbol{\theta}}_0, \ldots {}^8\hat{\boldsymbol{\theta}}_0$ are spaced (which in turn depends on the *expansion level*, i.e. the coverage of the underlying region) the method therefore offers some security against possible failure of the bootstrap bridging assumption.

Wichmann and Hill[12] took the ratio of the width of the expanded interval to the width of the ordinary bootstrap interval as an index of the sensitivity of a particular sampling scheme to likely error in the initial fit, “likely error” being estimated by the confidence region on which the expanded method is based. They used a crude method for constructing the region, simultaneously asserting a 68.3% bootstrap percentile confidence interval in α and another in β, to produce a rectangle. Parameter sets ${}^{1}\hat{\boldsymbol{\theta}}_0, \ldots {}^{8}\hat{\boldsymbol{\theta}}_0$ were located at the corners and on each side of the rectangle, with all the ${}^{i}\hat{\lambda}_0$ equal to ${}^{0}\hat{\lambda}_0$. The rectangle was generally found to contain roughly 50% of the original bootstrap points in a two-dimensional projection, although this depended on the data, as some data sets introduced more covariance between α and β than others, an effect which a rectangular region does not take into account. The bootstrap percentile method was used to obtain confidence limits at each of the nine parameter sets.

The current study uses two refinements of the method. First, the $\mathrm{BC_a}$ method is used to obtain confidence limits for each parameter set. Second, the underlying region is obtained by the bootstrap deviance method described in section 4.2, and thus reflects likely error in *all* the parameters, including nuisance parameters. The parameter sets ${}^{1}\hat{\boldsymbol{\theta}}_0, \ldots {}^{8}\hat{\boldsymbol{\theta}}_0$, being chosen to lie on the region boundary, have equal likelihood given the original data. The implementation of the method includes an algorithm for choosing the eight points such that they explore the largest deviations in all the parameters while being spread out as much as possible in the α–β plane. Two examples are shown in figure 2.1—each example is the result of a 2-AFC logistic fit to a different data set, with λ allowed to vary in the range $[0, 0.05]$. The expansion level is equal to 0.5 in both examples.

In each panel of figure 2.1, the red triangle marks the initial estimate ${}^{0}\hat{\boldsymbol{\theta}}_0$. The light blue points mark the parameter sets from the first bootstrap distribution ${}^{0}\hat{\boldsymbol{\theta}}^{*}$ that lie within the region, and the dark blue points mark those that lie outside. The error bars close to the axes show the central 68.3% (inner box) and 95.4% (outer bar endpoints) of the bootstrap distribution, separately in the α and β dimensions. Note that, particularly in the lower panel, there are some dark blue points apparently near the centre of the distribution, "underneath" the light blue points. This is because the plot is a two-dimensional projection of a space that actually includes a third dimension λ, so the confidence "region" is in fact a volume. Nevertheless the exploratory parameter sets ${}^{1}\hat{\boldsymbol{\theta}}_0, \ldots {}^{8}\hat{\boldsymbol{\theta}}_0$, marked by the yellow triangles, are placed so that they are spread out with respect to the two parameters that are of particular interest, α and β.

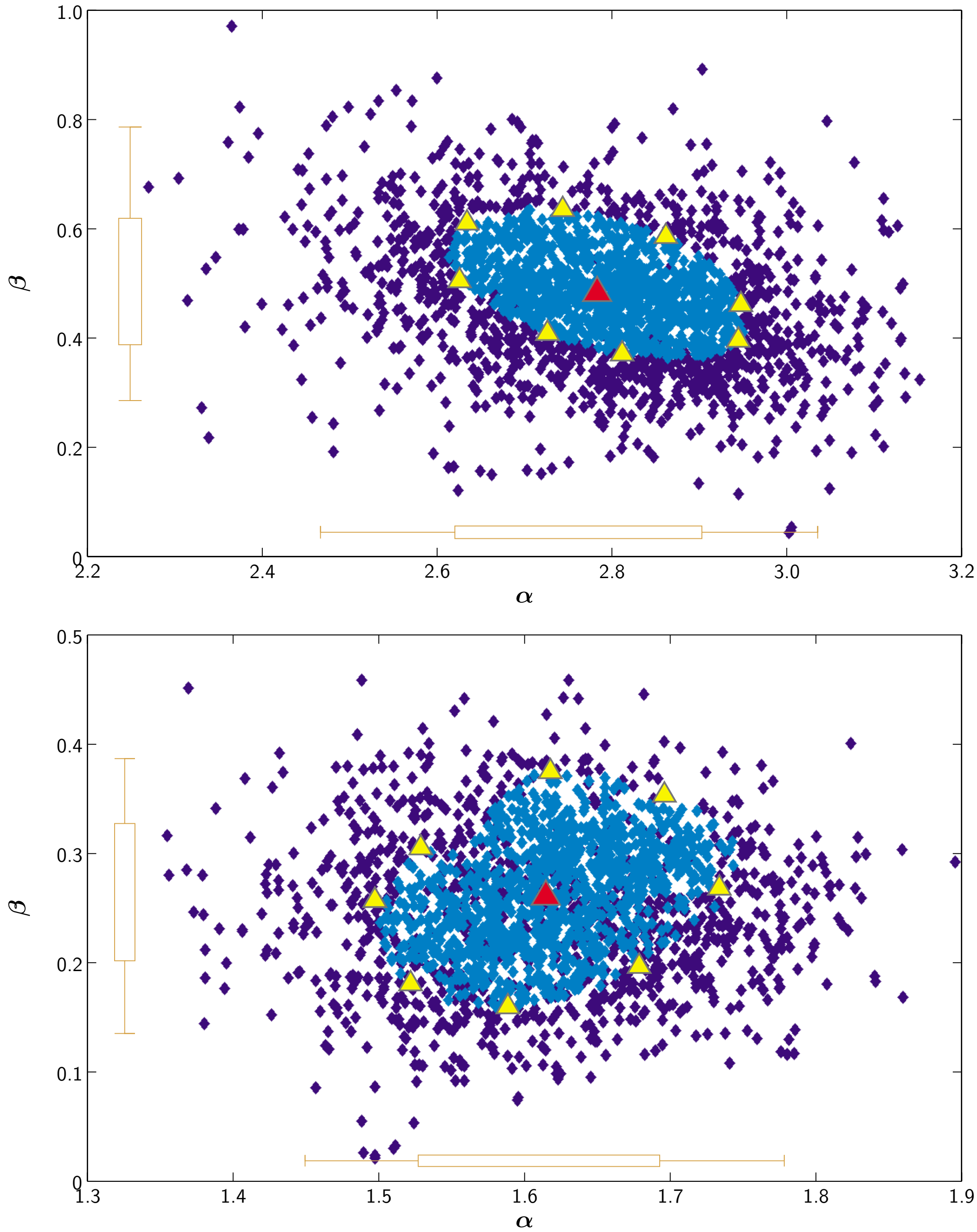


Fig. 2.1: The upper and lower panels illustrate the expanded bootstrap method for logistic function fits to two different 2-AFC data sets. See section 2.2.6 for details.

References for chapter 2

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

[2] MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall, second edition.

[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] EDWARDS, A. W. F. (1992). *Likelihood.* Baltimore: Johns Hopkins University Press. Expanded Edition.

[5] TELLER, D. Y. (1985). Psychophysics of infant vision: Definitions and limitations. In GOTTLIEB, G. & KRASNEGOR, N (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life: a Methodological Overview.* Norwood, NJ: Ablex.

[6] 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.

[7] EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

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

[9] 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.

[10] 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.

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

[12] 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 .

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

[14] 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.

[15] 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.

[16] Foster, D. H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biological Cybernetics*, **53**: 189–194. An erratum is given on page 412 of the same volume.

[17] Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**(3): 481–493.

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

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

[20] 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.

[21] Hinkley, D. V. (1978). Likelihood inference about location and scale parameters. *Biometrika*, **65**(2): 253–261.

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

[23] Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, **82**: 171–200.

3. Testing the coverage of confidence intervals

3.1 Introduction

3.1.1 General methods

In order to assess the accuracy of a method for obtaining confidence intervals, a Monte Carlo coverage test is performed. Where Monte Carlo methods for obtaining confidence intervals use a computer to simulate, over and over again, the behaviour of an observer, a Monte Carlo coverage test takes this principle one stage further by simulating, over and over again, the steps carried out by an experimenter.

The procedure used in this chapter is as follows: a psychometric function shape ψ_{gen} and a parameter vector $\boldsymbol{\theta}_{\text{gen}}$ are chosen. Together they describe the true psychometric function for an observer on a particular task. A sampling scheme is chosen, consisting of a set of f-values (see section 1.5) which, when transformed through the inverse of $F(x; \alpha_{\text{gen}}, \beta_{\text{gen}})$ determine the stimulus values $\boldsymbol{x}$ at which observations are to be taken, and which also, via the linear transformation $p = \gamma_{\text{gen}} + (1 - \gamma - \lambda)f$, determine the corresponding generating probability values $\boldsymbol{p}_{\text{gen}}$. The number of observations to be taken at each point is given by a vector $\boldsymbol{n}$ (the vectors $\boldsymbol{x}$, $\boldsymbol{p}_{\text{gen}}$ and $\boldsymbol{n}$ all have length k). For each of a large number of repetitions C, a simulated data set $\boldsymbol{r}_i$ is then generated, in which each value r_{ij} is a simulated performance score drawn from the binomial distribution $\text{Bi}(n_j, p_j)$. Each data set $\boldsymbol{r}_i$ is then passed to a simulated "experimenter", which is a subroutine that does not know the true parameter vector $\boldsymbol{\theta}_{\text{gen}}$, but which uses the data set to obtain a maximum-likelihood parameter estimate, and then uses that parameter estimate to obtain a confidence interval $[u_{i(\text{LO})}, u_{i(\text{UP})}]$ for a measure of interest u. The method for obtaining such confidence intervals might be, for example, a bootstrap method.

The result, a set of C confidence intervals, is a simulation of what would happen if an experimenter were to perform the same experiment C times, *including* the processes of fitting and confidence interval estimation (thus, if a bootstrap method is being tested, the test is extremely computationally intensive because each of the C repeats requires R psychometric function fits). If the method under examination exhibits perfect coverage, then the expected proportion of confidence intervals that contain the correct underlying value u_{gen} (computed from $\boldsymbol{\theta}_{\text{gen}}$) is equal to the nominal coverage of the confidence intervals (for example, 0.683 or 0.954).

The coverage estimate $\hat{c}$ for a given confidence interval method, then, is equal to the proportion of simulated confidence intervals of u that contain the true value u_{gen}. As the true value can be either inside (with estimated probability $\hat{c}$) or outside of a confidence interval, the standard error of $\hat{c}$ is estimated by the plug-in binomial standard error formula $\widehat{\text{se}}_c = \sqrt{\hat{c}(1-\hat{c})/C}$.

A second concern, besides accurate coverage, is that a method should produce intervals of correctly *balanced* coverage. For a two-tailed hypothesis test of 95% coverage, for example, the true value should occur 2.5% of the time in the upper tail, and 2.5% of the time in the lower tail. If it appears 5% of the time in the upper tail but not at all in the lower tail, then the interval has the correct overall coverage probability, but it is unbalanced and will lead to hypothesis testing errors. Note that this sense of the word "balance" is a special case of the commonly-used definition provided by Beran,[1,2] which is that the coverage probabilities of two or more simultaneously asserted confidence statements are asymptotically equal, so that the two statements are "treated fairly". Lee[3] applied the bootstrap method to logistic regression problems using Beran's approach, showing the method to be asymptotically balanced for a simultaneous confidence region whose constituent statements are of the general form $R_{\boldsymbol{u}}(\boldsymbol{\theta}) < d_{\boldsymbol{u}}$ where $R_{\boldsymbol{u}}(\cdot)$ is an appropriate transformation (or "root" function) of any linear combination of the intercept and slope parameter, weighted by a vector of coefficients $\boldsymbol{u}$. In a subsequent paper[4] Lee's Monte Carlo simulations suggested that the bootstrap is better balanced than the asymptotic methods of Bonferroni and of Scheffe, although he measured the balance between coverage of intercepts and coverage of slopes, rather than the balance of coverage between the two sides of a two-tailed one-dimensional confidence interval. Tail imbalance for a single measure of interest can be fitted into the same framework, as a two-tailed confidence interval is just a pair of simultaneously asserted statements $u > u_{\text{LO}}$ and $u < u_{\text{UP}}$, which can be expressed in the notation of Beran and Lee using one root indexed by a certain vector $\boldsymbol{u}$ and another indexed by $-\boldsymbol{u}$.

Both Beran[2] and Lee[4] quantify imbalance simply by taking the absolute difference between two coverage probabilities (or the largest absolute difference between any pair of coverage probabilities). A slightly different metric will be used here, because using the difference between tail probabilities would make it difficult to compare imbalance values between two tests whose overall coverage is different: the range of possible values $[0, 1-\hat{c}]$ would change. Instead, a standardized metric will be used, and the direction of the imbalance will be preserved: imbalance will be defined as the difference between the conditional probabilities, given that a false rejection of the true value has occurred, of its occurrence in the lower tail and its occurrence in the upper tail, i.e.:

$$a = \frac{P_{\text{LO}} - P_{\text{UP}}}{P_{\text{LO}} + P_{\text{UP}}} \quad \forall \, \{P_{\text{LO}} + P_{\text{UP}} > 0\}, \tag{3.1}$$

where P_{LO} and P_{UP} are the false rejection probabilities in the lower and upper tail respectively: $P_{\text{LO}} = \Pr(u_{\text{gen}} < u_{\text{LO}})$ and $P_{\text{UP}} = \Pr(u_{\text{gen}} > u_{\text{UP}})$. The imbalance metric is similar to the confidence interval "shape" metric used by Efron and Tibshirani,[5] except that it is a fractional difference of rejection probabilities, rather than a fractional difference of lengths of the two sides of the interval. Each coverage test yields an estimated imbalance $\hat{a}$ based on the estimates $\hat{P}_{\text{LO}}$ and $\hat{P}_{\text{UP}}$ obtained from the set of C confidence intervals. The ideal result is $\hat{a} = 0$, indicating that the the confidence interval is perfectly balanced. The result $\hat{a} = -1$ would indicate that the confidence interval bounds are set too low, to such an extent that all the false rejections occur in the upper tail and none in the lower tail. Conversely, $\hat{a} = +1$ indicates that the confidence interval bounds are set too high, with all the false rejections occurring in the lower tail and none in the upper tail. In the special case $P_{\text{LO}} = P_{\text{UP}} = 0$, a is defined as 0, because the interval is balanced at least in the sense that the same number of false rejections (zero) occurs in each tail.

The standard error se_a of a can be calculated from the following variance formula:

$$\begin{aligned}\text{se}_a^2 \;=\; & \left[\sum_{l=0}^{C}\sum_{h=0}^{C-l} \tfrac{n!}{l!h!(n-l-h)!} {P_{\text{LO}}}^l {P_{\text{HI}}}^h (1 - P_{\text{LO}} - P_{\text{HI}})^{n-l-h} {a_{(l,h)}}^2\right] \\ - \; & \left[\sum_{l=0}^{C}\sum_{h=0}^{C-l} \tfrac{n!}{l!h!(n-l-h)!} {P_{\text{LO}}}^l {P_{\text{HI}}}^h (1 - P_{\text{LO}} - P_{\text{HI}})^{n-l-h} a_{(l,h)}\right]^2 .\end{aligned} \tag{3.2}$$

where $a_{(l,h)}$ is defined as 0 when $l = h = 0$, and $(l - h)/(l + h)$ otherwise. Standard errors for $\hat{a}$ are obtained using the plug-in estimates $\hat{P}_{\text{LO}}$ and $\hat{P}_{\text{UP}}$ in place of P_{LO} and P_{UP}.

Swanepoel and Frangos[6] used the Monte Carlo method to measure $\hat{c}$ for 95% confidence intervals for the slope parameter of a logistic function with $\gamma = \lambda = 0$. They used two symmetrical sampling schemes: one with $k = 5$ which they tested at $N = 100$, 150 and 200, and another of similar spread but with $k = 10$, which they tested at $N = 200$, 300 and 400. Their confidence intervals were obtained by the bootstrap-t method using four different techniques for estimating $\hat{v}$ (see section 2.2.3): a parametric method using the Fisher approximation (considered here) and three jackknife methods. Of these, the parametric method is by far the least computationally intensive, and Swanepoel and Frangos found its performance to be second best, with $\hat{c} = 0.94 \pm 0.01$ for nearly all cases.[†] Lee[4] also used the parametric method to

[†] The winning procedure was a third-order-corrected jackknife method. While it was roughly

(footnote continues ⟶)

obtain Studentized bootstrap parameter distributions for a logistic function (again, only the idealized yes-no situation, with $\gamma = \lambda = 0$, was considered). The coverage results for the slope parameter were somewhat more variable than those of Swanepoel and Frangos, although this may be attributable to Lee's cruder implementation of the bootstrap-t method, in which all confidence intervals were symmetrical about the MLE.

The current chapter aims to use the statistics $\hat{c}$ and $\hat{a}$ to compare the performance of a number of different confidence interval methods:

- bootstrap standard error,
- basic bootstrap,
- bootstrap-t (using the parametric Fisher approximation),
- bootstrap percentile,
- bootstrap $\mathrm{BC_a}$,
- expanded bootstrap $\mathrm{BC_a}$ (at various levels of expansion),
- probit standard errors,
- probit fiducial limits on thresholds.

In order to replicate and extend Swanepoel and Frangos's observations,[6] one set of tests will be performed with the logistic function and with γ and λ fixed at 0. In the main, however, tests will focus on realistic psychophysical conditions, as psychophysicists are likely to be more interested in the case in which it cannot be assumed that $\gamma = \lambda = 0$. The results of the idealized

equivalent to the parametric method in terms of its estimated coverage, Swanepoel and Frangos preferred it because it produced significantly shorter confidence intervals. As the two procedures had the same $\hat{c}$ but different mean lengths, it seems likely that they had different values of $\hat{a}$, but the authors do not report the extent to which intervals were balanced. It would therefore be interesting to test this. Such a test will not be attempted here: the current chapter aims to explore a broader set of variables, and Monte Carlo tests of bootstrap methods are computationally demanding enough without the extra level of iteration required by the jackknife.

condition will therefore be interesting to the extent that they contrast with the second set of tests, in which the true γ and λ are small non-zero values that the experimenter must estimate—in this case, using a constrained maximum-likelihood search. Results of such a comparison will be reported in section 3.2. The remaining majority of tests will concentrate on the 2-AFC case, with γ fixed at 0.5. Again, an idealized and a realistic case will be compared. In the former, it will be assumed that lapses do not occur, so λ can be safely fixed at 0. In the latter, λ will be a free parameter of the model, constrained to lie within the neighbourhood of its small non-zero underlying value.

Performance may be affected by k, by N, and by the exact placement of the k sample points. These factors create a large space which it is impractical to explore thoroughly.[†] Therefore, exploration will be confined, in sections 3.2 and 3.3, to a set of 7 sampling schemes, all of which have $k = 6$, at $N =$ 120, 240, 480 and 960.

Unless otherwise stated, each of the tests comprised $R = 1999$ bootstrap runs on each of $C = 500$ simulations of an experiment. Also, unless otherwise stated, the correct psychometric function form was always used to fit simulated data sets (thus, if the underlying function was the Weibull function, the Weibull was used for fitting). Thresholds and slopes were computed at $f = 0.2$, $f = 0.5$ and $f = 0.8$. Confidence intervals were calculated for target coverage probabilities of 0.683 and 0.954. The target will be identified where necessary using a subscript on the results; for example, $\hat{c}_{68}$ will denote the measured coverage of a confidence interval whose target coverage was 0.683, and $\hat{a}_{95}$ will denote the measured imbalance of a confidence interval whose target coverage was 0.954.

[†] The software allowed most of the confidence interval methods to be tested simultaneously, with two exceptions: only a single level of expansion could be tested at once, and the extra time needed to calculate $\hat{v}^*$ on each bootstrap replication made it impractical to test bootstrap-t and expanded methods simultaneously. Running on a Macintosh G3 at 350 MHz, a single test set (consisting of 7 different sampling schemes at 4 different values of N) required roughly five days of continuous processing if the test included an expanded method, or two days if it included bootstrap-t.

3.1.2 Freeman-Tukey transformation of coverage probability estimates

Estimated coverage probabilities $\hat{c}$ will be transformed using the normalizing and variance-stabilizing function of Freeman and Tukey:[7]

$$\vartheta_C(c) = \sin^{-1}\sqrt{\frac{c\,C}{C+1}} \;+\; \sin^{-1}\sqrt{\frac{c\,C+1}{C+1}}. \tag{3.3}$$

Whereas the standard error of $\hat{c}$ itself is dependent on the estimated value, $\vartheta_C(\hat{c})$ is a monotonically increasing transformation of $\hat{c}$ which yields values whose standard error is asymptotically independent of $\hat{c}$, being approximately equal to $(C + \frac{1}{2})^{-\frac{1}{2}}$. In practice, "approximately" means that for $C \leq 500$, and for all observable coverage estimates except 0 and 1 (i.e. for all values $\{1, 2, \ldots C-1\}/C$) the exact standard error of $\vartheta_C(\hat{c})$ does not deviate from the asymptotic value by more than about 3% of the asymptotic value itself. (For $\hat{c} = 0$ or $\hat{c} = 1$ the estimated standard error is naturally 0.)

Thus, by plotting $\vartheta_C(\hat{c})$ instead of $\hat{c}$ on figures such as figure 3.1, the statistical significance of differences in coverage can be more easily assessed by eye. Assuming $C = 500$, the standard error bars for every point would be of extent ± 0.045.

Note that the Freeman-Tukey function depends on C and, because of occasional invalid results within the test, the effective value of C was less than 500 in some rare cases. Naturally, the standard error is somewhat higher for these points, according to the formula $(C + \frac{1}{2})^{-\frac{1}{2}}$. However, there is another slight complication: in order to make the estimates from different tests directly comparable to one another, and to the axis on which they are plotted, the same monotonic function will be applied to all, *viz.* the function for $C = 500$. The effect of such a generalization is to perturb the variance-stabilizing effect of the transform somewhat: when $C = 350$, for example, the standard error values at $c < 0.03$ and $c > 0.97$ can rise by up to 6% of the asymptotic value, and when $C = 100$, the standard error values at $c < 0.07$ and $c > 0.93$ can rise by up to 14% of the asymptotic value. Such effects are sufficiently small, and values of C less than about 350 occur sufficiently rarely, that such errors do not substantially affect the interpretation of any of the results presented here.

When $C \leq 500$, the Freeman-Tukey transform can be shown to be appreciably superior to certain other commonly-used variance-stabilizing formulae such as $\Phi^{-1}(c)$ or $2 \sin^{-1} \sqrt{c}$. The latter is actually a special case of equation (3.3), being the limit of $\vartheta_{C}(c)$ as $C \rightarrow \infty$.

3.1.3 Graphical representation of results

In reporting the results, figures will generally follow the format of figure 3.1. The upper panel shows estimated coverage probability $\hat{c}$ on the ordinate, transformed using the normalizing and variance-stabilizing transform $\vartheta_{500}(\cdot)$ of section 3.1.2. The lower panel shows the estimated imbalance $\hat{a}$ from equation (3.1) on the ordinate. In both panels, the results are grouped along the abscissa according to the confidence interval method used, and each symbol refers to a single coverage test (i.e. 500 iterations of an experiment using one particular sampling scheme at one particular value of N).

The symbols can be viewed “from a distance” to show the general trends that are characteristic of a particular confidence interval method. At the risk of overloading the diagrams with information, the shape of the symbols will further denote which sampling scheme was used, and their size will denote the value of N. In most cases it is the general trend that will be interesting, so the reader rarely need worry about trying to decipher symbol shapes and sizes. There are a few instances, however, in which variations in sampling scheme and/or N explain a specific discernible trend in the results, and attention will be drawn to such trends in the text. Otherwise, the symbols at least serve to illustrate the significant point that there *are* very few simple trends that can be attributed solely to sampling scheme or to N.

Symbol shapes will correspond to different sampling schemes. In the 2-AFC simulations of section 3.3, they will refer to Wichmann’s 7 sampling schemes in the manner illustrated in figure 1.2. As Wichmann’s sampling schemes are only really appropriate for forced-choice designs, symbols in the

the yes-no simulation results of section 3.2 will correspond to the alternative set of 7 sampling schemes shown of figure 1.3.

Error bars have not been drawn on the points. As discussed in section 3.1.2, the standard error of a transformed estimated coverage probability $\vartheta_{500}(\hat{c})$ is roughly constant, being approximately 0.045 when $C = 500$. As previously mentioned, there were certain rare cases in which the number of valid simulations dropped below 500. For bootstrap methods, C never dropped below 350, the corresponding standard error being roughly 0.053. Probit fiducial limits for two particular tests (the poorly sampled 2-AFC schemes ● and ◀ at the lowest N value, $N = 120$), proved calculable only about 100 times out of the 500: the corresponding standard error of roughly 0.1 can be taken to be the absolute worst case.

The standard error of an estimated imbalance value, computed using equation (3.2), depends on both $\hat{c}$ and $\hat{a}$ as well as C. As a rough guide, sample values are given in table 3.1 for $C = 500$.

Z	$\hat{c}$	$\vartheta_{500}(\hat{c})$	$\widehat{\text{se}}_a$			
			$\hat{a} = 0$	$\hat{a} = \pm 0.5$	$\hat{a} = \pm 0.9$	$\hat{a} = \pm 0.99$
0.5	0.383	1.335	0.057	0.049	0.025	0.008
1	0.683	1.944	0.080	0.069	0.035	0.011
1.5	0.866	2.391	0.123	0.107	0.054	0.017
2	0.954	2.707	0.214	0.186	0.093	0.030
2.5	0.988	2.910	0.446	0.386	0.198	0.076
3	0.997	3.021	0.718	0.659	0.503	0.445

Table 3.1: Equivalent $\pm Z$ scores (first column) are given for a number of example coverage probability estimates $\hat{c}$ (second column). The Freeman-Tukey[7] transformed value $\vartheta_{500}(\hat{c})$ is given in the third column. The standard error associated with the transformed value is approximately 0.045 assuming $C = 500$. The standard error $\widehat{\text{se}}_a$ of the imbalance statistic $\hat{a}$ depends on both $\hat{c}$ and $\hat{a}$. Example values are given in the remaining columns of the table, assuming $C = 500$. The standard error for $-a$ is the same as that for $+a$.

In most cases, only the results for 95.4% confidence intervals will be shown. With few exceptions, trends in the 68.3% results followed those of

95.4%, although they were usually less pronounced; for example, both the absolute mean imbalance value, and the spread of imbalances around the mean, tended to be smaller, for a given set of tests. Where trends differ at the two different target confidence levels, results will be presented for both.

3.1.4 Criteria for judging results

Ideally, coverage c should be exactly equal to the target coverage of the desired interval (0.683 or 0.954) and the imbalance a should be be exactly 0. Inevitably, however, some error occurs, not only because c and a are estimated using a finite number of simulations C, but because confidence interval methods themselves are not perfect, and may be flawed in ways which depend both on sampling scheme and on N.

Since, in a real experiment, the true parameters $\boldsymbol{\theta}_{\text{gen}}$ are unknown, one can never be sure where one's chosen stimulus values will fall relative to the true curve. In other words, a pre-specified sampling scheme can never be recreated accurately. Thus, the perfect confidence interval method will produce results which do not depend on sampling scheme; using it, we would be able to be confident that our intervals have certain coverage properties, even if we had inadvertently sampled the psychometric function too narrowly, or too tightly, or with too much of a bias to one side or the other.

An experimenter will also want to obtain results efficiently, which means keeping N as low as possible. However, the precise value of N will vary depending on the experimental context. Thus our hypothetical perfect confidence interval method should also produce results which hold for all commonly encountered values of N.

Thus $\hat{c}$ and $\hat{a}$ values should exhibit as little variation with sampling scheme and with N as possible. This quality of *stability* is the first criterion.

The observed values of $\hat{c}$ and $\hat{a}$ are also important, but they are of secondary importance. A low $\hat{c}$ value, for example, might not be a serious problem so long as it is *reliably* low, as it might be compensated for simply by inflating the target coverage level (an instance of this can be seen in figure 3.3, where the $\hat{c}_{95}$ values for the BC_a method cluster tightly around their mean of 88.4%: one could say, therefore, that the $\mathrm{BC_a}$ method gives reliable 88.4% confidence intervals if one aims for 95.4%, so it is possible that 95.4% might be reliably achieved by aiming higher). Another way of compensating for low coverage might be to employ an expanded bootstrap method such as the one described in section 2.2.6 and tested in section 3.3.5.

Non-zero imbalance values might also be corrected, by adjusting the two tail rejection probabilities separately, although imbalance values of $+1$ and -1 are particularly undesirable as they give no indication of the appropriate size of such a correction. Non-zero imbalance should probably be considered more serious than inaccurate overall coverage, as any putative correction for the latter is a little more straightforward.

3.2 Performance of bootstrap methods for yes-no designs

The logistic function with $\alpha_{\mathrm{gen}} = 0$, $\beta_{\mathrm{gen}} = 1$ was taken as the true underlying function under "idealized" and "realistic" conditions.

In the idealized case, the underlying values of γ and λ were set to 0, and γ and λ were fixed at 0 during all fitting processes. The results are filed in the results archive under

- `simulations/coverage/yesno/g0f0l0f0/logistic/with_bootstrap_t/`

and are shown in figures 3.1 and 3.2. They are discussed in section 3.2.1.

In the more realistic case, psychophysically plausible underlying values were chosen: $\gamma = 0.02, \lambda = 0.01$. Both parameters were treated as unknowns, so they were allowed to vary both in the initial fit to each data set generated from the true function, and in any fits to bootstrap data sets generated from each estimated function. In all fits, both parameters were constrained, using a flat Bayesian prior, to lie within the range $[0, 0.05]$. The results are filed in the results archive under

- `simulations/coverage/yesno/g02l01/logistic/with_bootstrap_t/`

and are shown in figures 3.3 and 3.4. They are discussed in section 3.2.2.

For both conditions, four values of N were tested (N = 120, 240, 480 and 960) with each of the seven sampling schemes defined in section 1.5.2.

Both sets of tests were repeated with the cumulative normal psychometric function shape ($\alpha_{\mathrm{gen}} = 0$, $\beta_{\mathrm{gen}} = 1$), and probit analysis was used to obtain asymptotic-theory confidence intervals—see section 3.4.1 for results.

3.2.1 Idealized yes-no case (zero guess- and lapse-rates assumed)

Results are shown in figures 3.1 and 3.2 for 95.4% confidence intervals on thresholds and slopes, respectively.

With regard to thresholds, results from the bootstrap percentile and $\mathrm{BC_a}$ methods are in close agreement, both of them exhibiting very good coverage that is relatively independent of sampling scheme and N. For the $\mathrm{BC_a}$ method, the distribution of the 28 estimates of $\hat{c}_{95}$ is 0.952 ± 0.010 and the distribution of $\hat{c}_{68}$ estimates (not shown) is 0.687 ± 0.020—note that in both cases the standard deviation of the set of estimates is roughly equal to the standard error of an individual estimate from the set, as set out in table 3.1. The distribution of bootstrap percentile estimates is indistinguishable from that of the $\mathrm{BC_a}$ estimates.

The bootstrap standard error and basic bootstrap methods produce results that are more spread out in both coverage and imbalance, with a tendency for the more narrowly sampled schemes (⧫ and ▲ in particular) to generate over-conservative confidence intervals. The bootstrap-t method seems to over-correct for this effect: at lower N values, ⧫ and ▲ produce confidence intervals that are too small.

A slightly different picture emerges when slopes (figure 3.2) are considered. Coverage is generally good for the bootstrap percentile method, but it has a positive imbalance (confidence limits are set too high), which increases as the psychometric function is more widely sampled (▲ → ●). The bootstrap standard error and basic bootstrap methods give widely spread

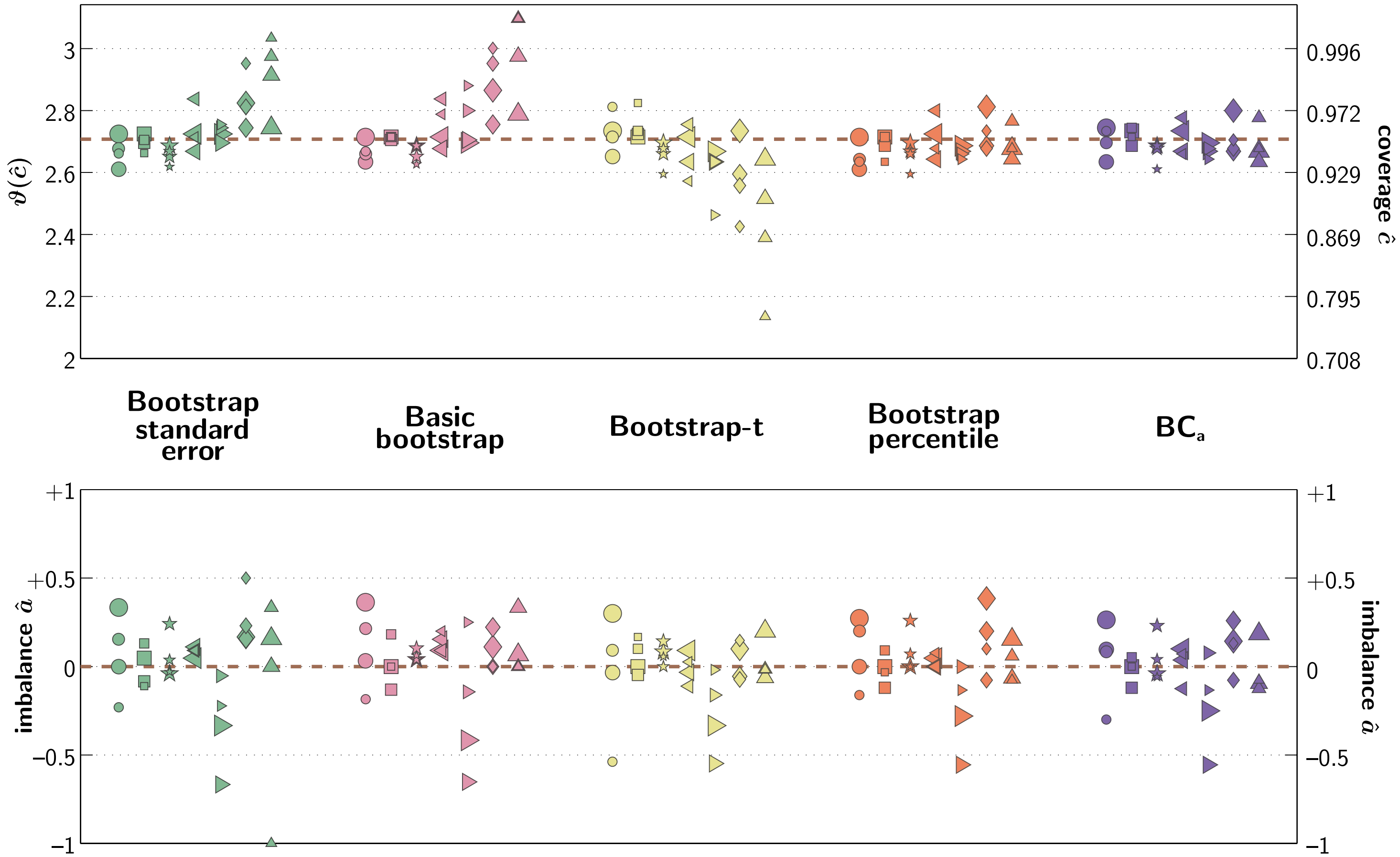


Fig. 3.1: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the logistic function in the idealized yes-no case ($\gamma = \lambda = 0$). Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.2.1 for details.

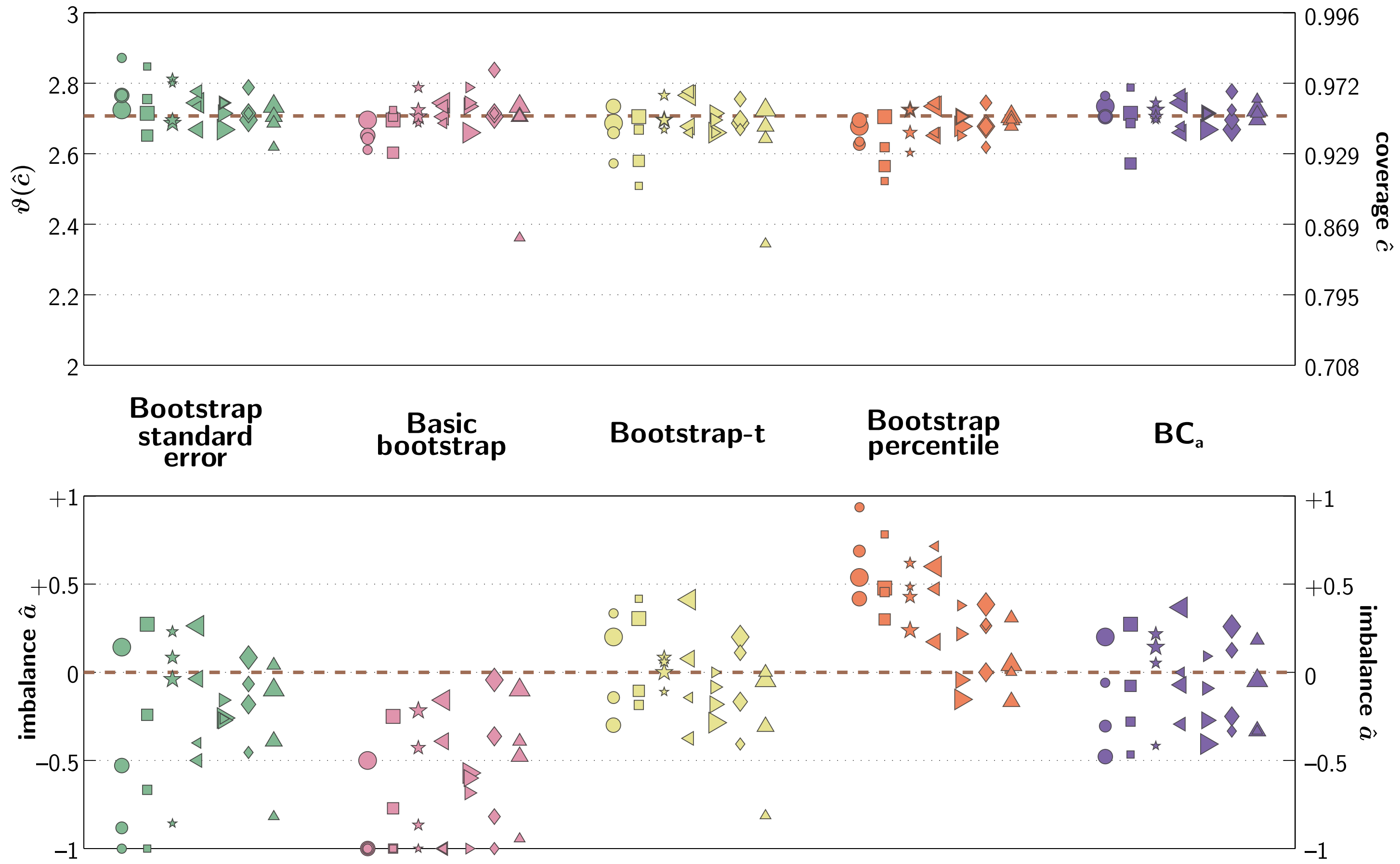


Fig. 3.2: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the logistic function in the idealized yes-no case ($\gamma = \lambda = 0$). Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.2.1 for details.

imbalance values with a strong tendency towards the negative side, particularly at low N. The $\mathrm{BC_a}$ and bootstrap-t methods are the best, with narrow groups of points centred on the target values. The bootstrap-t method is slightly inferior to the $\mathrm{BC_a}$ because of low coverage on ▲ at $N = 120$, but generally the performance of the bootstrap-t method at the 95.4% target ($\hat{c}_{95} = 0.948 \pm 0.014$ if that one outlier is excluded) is consistent with the findings of Swanepoel and Frangos.[6]

3.2.2 Realistic yes-no case

Results are shown in figures 3.3 and 3.4 for 95.4% confidence intervals on thresholds and slopes, respectively.

The addition of the constrained free parameters γ and λ has made a noticeable difference to all the results, and has further differentiated the bootstrap methods from each other.

For thresholds, there is a general tendency towards negative imbalances: all methods produced confidence intervals which were slightly biased towards lower threshold values. As the sampling schemes themselves all have symmetrically distributed f-values, this tendency must be due to the asymmetry in ψ created by the differing values of γ_{gen} and λ_{gen}.[†] The bootstrap percentile and bootstrap standard error methods both exhibit good coverage, with coverage probabilities closest to target and fairly well clustered together (i.e. relatively unaffected by sampling scheme or by N). Performance is still worse than in the idealized yes-no case, however. Bootstrap standard error coverage has dropped slightly below target, and bootstrap percentile imbalance values drift more to the negative side of the 95.4% target. The $\mathrm{BC_a}$, and in particular the basic bootstrap and bootstrap-t methods, suffer from low coverage, which seems to be worse for tighter distributions of sampling points (▶, ⧫ and ▲).

For slopes, overall performance has also worsened, as compared with the

[†] A repeat test confirmed this. When the true parameter values are reversed ($\gamma_{\mathrm{gen}} = 0.01$, $\lambda_{\mathrm{gen}} = 0.02$), all methods tend to produce slight *positive* imbalances of a similar magnitude (results are not shown).

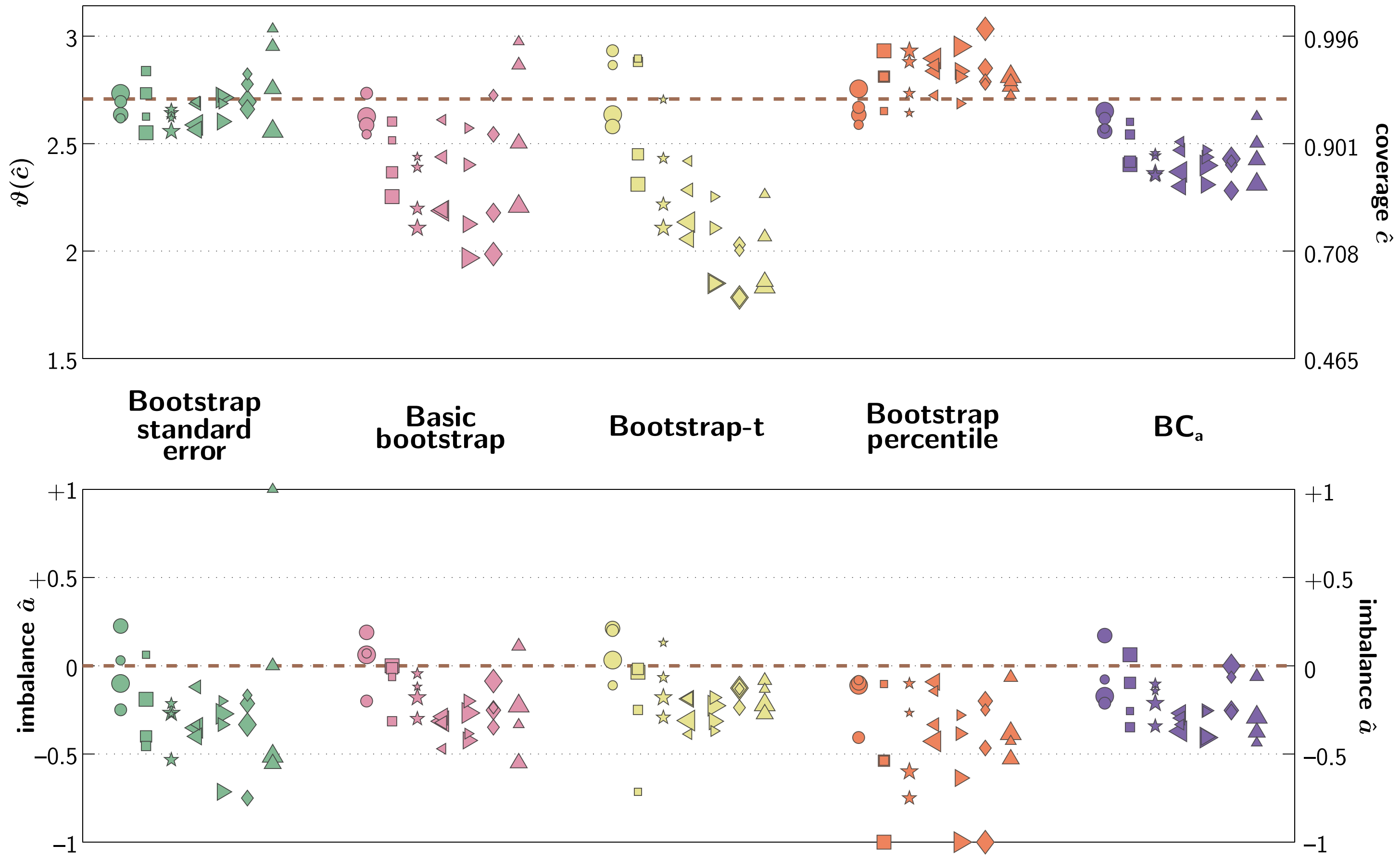


Fig. 3.3: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the logistic function in the realistic yes-no case. Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.2.2 for details.

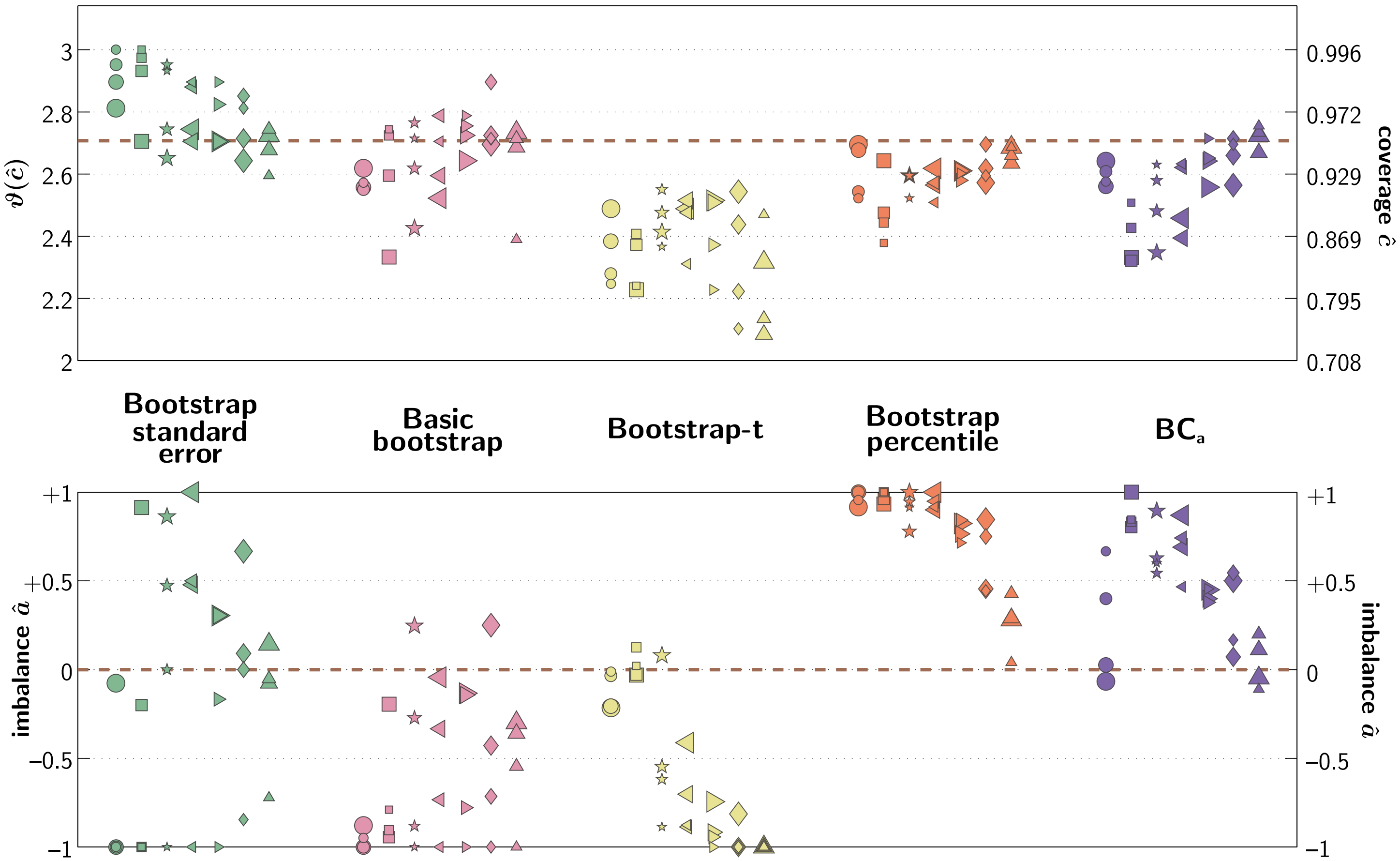


Fig. 3.4: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the logistic function in the realistic yes-no case. Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.2.2 for details.

idealized case: coverage has generally decreased below target levels, and there is a wider spread of imbalance values for all methods. The $\mathrm{BC_a}$ method drifts towards positive imbalance values, and coverage is low for many sampling schemes. The bootstrap-t method produces negative imbalances, and coverage is even lower. The bootstrap percentile method suffers less from low coverage than the other methods, but drifts even more towards positive imbalances than it did in the idealized yes-no context, so that for many sampling schemes, there is complete imbalance: no high slope values are ever rejected. The bootstrap standard error method exceeds the coverage targets, but there is very wide spread in its imbalance values (with a tendency towards positive imbalance), and very large differences between the highly inflated coverage of ● and ■ and the more accurate coverage of the more widely sampled schemes. The best of the bootstrap methods, although its coverage is somewhat too low and negatively unbalanced, is arguably the basic bootstrap, which has the smallest variation across different sampling schemes and values of N.

3.3 Performance of bootstrap methods in 2-AFC designs

A Weibull function with $\alpha_{\mathrm{gen}} = 3$ and $\beta_{\mathrm{gen}} = 4$ was used to define the true psychometric function in both an "idealized" and a "realistic" case of a 2-AFC experiment. In both cases, $\gamma_{\mathrm{gen}} = 0.5$, and γ was fixed at 0.5 for all fits. Tests of the bootstrap standard error, basic bootstrap, bootstrap-t, bootstrap percentile and $\mathrm{BC_a}$ methods were carried out, using Wichmann's 7 sampling schemes (see section 1.5.1) at $N = 120$, 240, 480 and 960.

The idealized case aimed to simulate conditions in which the observer makes no stimulus-independent errors (hence $\lambda_{\mathrm{gen}} = 0$), and the experimenter knows the assumption $\lambda = 0$ to be safe (thus λ was fixed at 0 for all fits). The results are filed in the results archive under

- `simulations/coverage/2AFC/l0f0/weibull/with_bootstrap_t/`

and are shown in figures 3.5 and 3.6. They are discussed in section 3.3.1.

The realistic case allowed for stimulus-independent errors. The generating value $\lambda_{\mathrm{gen}} = 0.01$ was chosen, and λ was treated as an unknown nuisance parameter in all fits, including bootstrap fits. Its value was constrained, using a flat Bayesian prior, to lie within the range $[0, 0.05]$. The realistic test set was repeated using the logistic function with $\boldsymbol{\theta}_{\mathrm{gen}} = (2.737, 0.494, 0.5, 0.01)^{\mathrm{T}}$—parameters that were chosen so that the threshold and slope values at $f = 0.5$ were the same as for the Weibull tests with $\alpha_{\mathrm{gen}} = 3$, $\beta_{\mathrm{gen}} = 4$). The results are filed in the results archive under

- `simulations/coverage/2AFC/l01/weibull/with_bootstrap_t/`
- `simulations/coverage/2AFC/l01/logistic/with_bootstrap_t/`

There were no substantial differences between the results from the two different function shapes, so only the Weibull results will be shown here. They are shown in figures 3.7 and 3.8 for thresholds and slopes, respectively. They are discussed in section 3.3.2. The realistic Weibull tests were repeated three more times, to investigate the performance of the expanded $\mathrm{BC_a}$ method—see section 3.3.5 for results.

The idealized and realistic test sets were further repeated using the cumulative normal psychometric function shape ($\alpha_{\mathrm{gen}} = 5$, $\beta_{\mathrm{gen}} = 0.2$), and probit analysis was used to obtain asymptotic-theory confidence intervals—see section 3.4.2.

3.3.1 Idealized 2-AFC case

Results are shown in figures 3.5 and 3.6 for 95.4% confidence intervals on thresholds and slopes, respectively.

In general, the results are similar to those obtained in the idealized yes-no case (section 3.2.1), although many of the trends are more pronounced.

The bootstrap percentile and $\mathrm{BC_a}$ methods are the best methods for thresholds. Their performance is roughly equal, and roughly equivalent to their performance in the idealized yes-no case. The bootstrap standard error

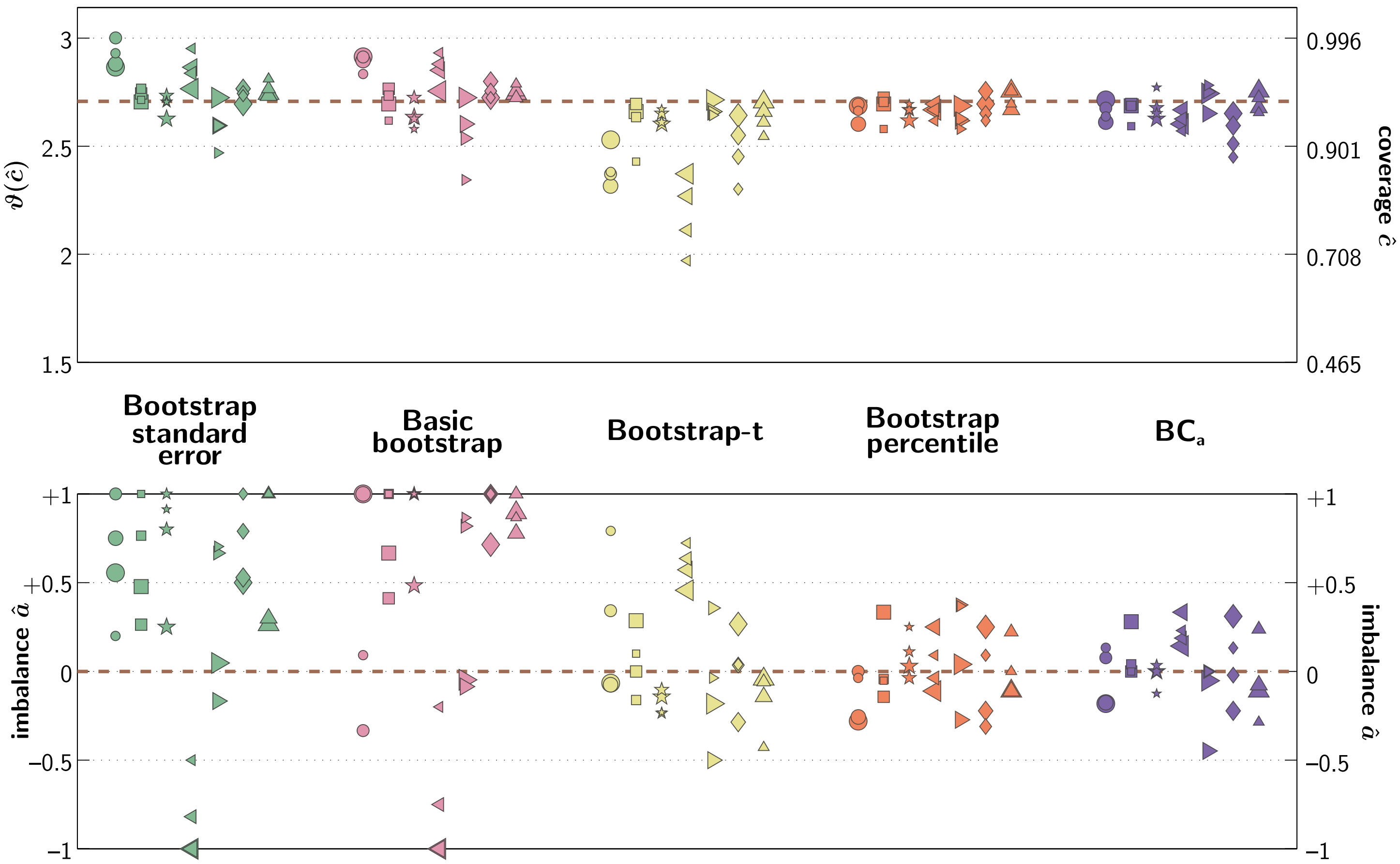


Fig. 3.5: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the Weibull function in the idealized 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.1 for details.

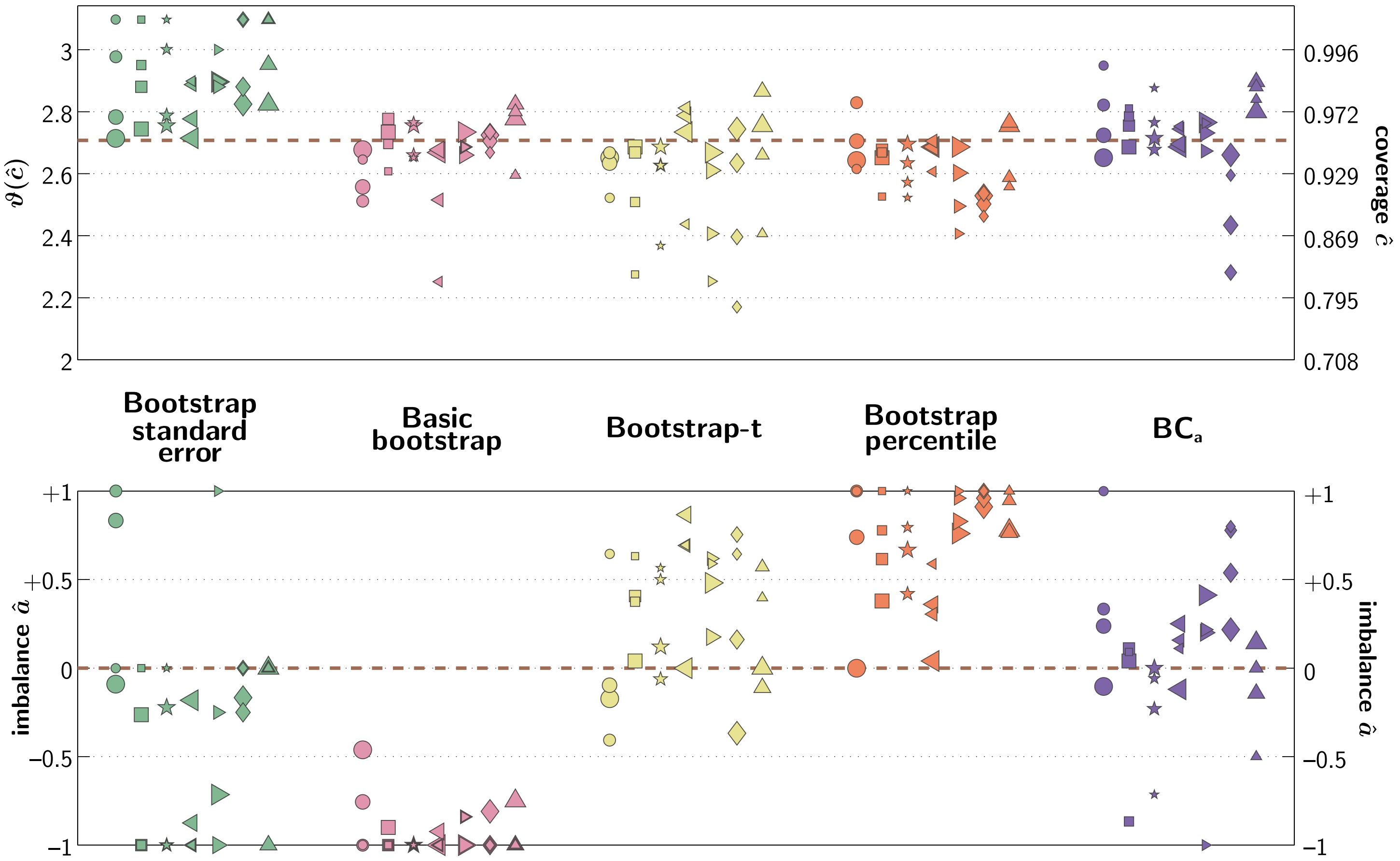


Fig. 3.6: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the Weibull function in the idealized 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.1 for details.

and basic bootstrap methods are the worst, with a range of imbalance values even wider than that observed in the yes-no condition. The bootstrap-t method is intermediate, with a somewhat wider range of imbalance values than the bootstrap percentile or BC$_{\mathrm{a}}$ methods, and lower coverage, particularly for the sampling schemes with no sample points at high expected performance levels, ● and ◀.

For slopes, the bootstrap-t and BC$_{\mathrm{a}}$ methods are the best, although they nevertheless exhibit quite a wide range of imbalance values. The bootstrap percentile method shows good overall coverage but a positive imbalance, and the bootstrap standard error and basic bootstrap methods are highly negatively unbalanced.

3.3.2 Realistic 2-AFC case

Results are shown in figures 3.7 and 3.8 for 95.4% confidence intervals on thresholds and slopes, respectively.

With regard to thresholds, the BC$_{\mathrm{a}}$ and bootstrap percentile methods are both good, in that the points are tightly grouped: both methods therefore have the desirable characteristic of being relatively insensitive to differences in sampling scheme and N. Note that, while it produces the smallest range of coverage values of all the methods, and the smallest imbalance values in either direction, coverage for the BC$_{\mathrm{a}}$ method is consistently too low.

By contrast, both the bootstrap standard error and basic bootstrap methods suffer from a spreading out of imbalance values, which depends on sampling scheme. For both methods, there is an overall tendency towards positive imbalance values, with ★ and ⧫ highly positively unbalanced, but ◀ is extremely negatively unbalanced. Coverage probabilities also suffer: $\hat{c}_{95}$ for ▶ is low, and $\hat{c}_{95}$ estimates for both ● and ◀ break away from the group, rising as N decreases.

The performance of the bootstrap-t method on thresholds is intermediate, showing greater spread than the BC$_{\mathrm{a}}$ or bootstrap percentile methods, although not as great as that of the basic bootstrap or bootstrap standard error method. Sensitivity to differences in sampling scheme is also interme-

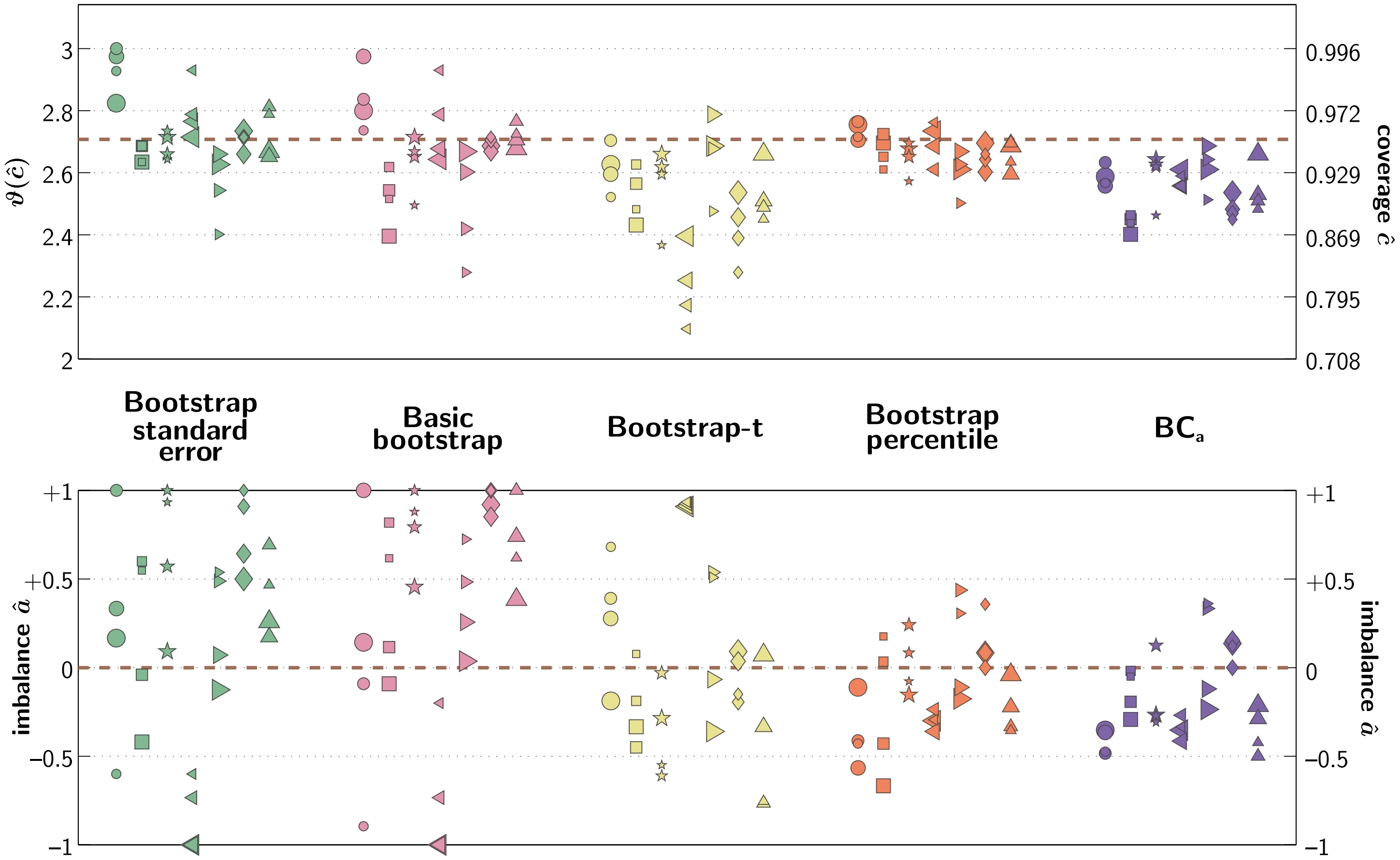


Fig. 3.7: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the Weibull function in the realistic 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.2 for details.

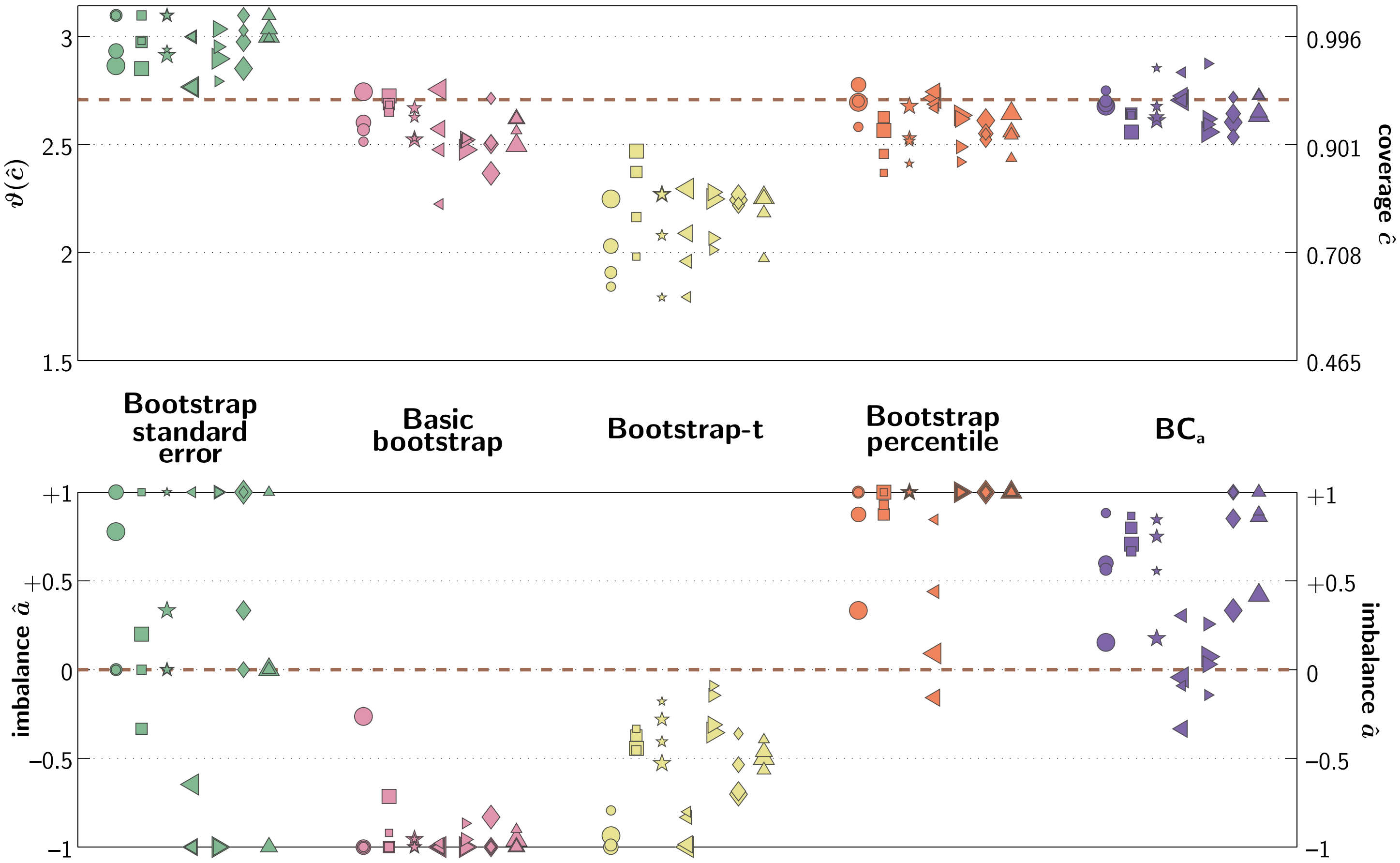


Fig. 3.8: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the Weibull function in the realistic 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.2 for details.

diate. There are few apparent differences between sampling schemes, except that the imbalance of ◀, seen in the basic bootstrap results, has been overcompensated-for (◀ now has a large *positive* imbalance). The bootstrap-t method produces coverage estimates that are consistently too low, in a similar manner to, but to a greater extent than, the $\mathrm{BC_a}$ method.

For slopes, the $\mathrm{BC_a}$ method is the best performer, with the tightest groups, and smallest absolute mean imbalance. Its coverage is slightly too low (although less so than was the case for thresholds), but its performance is generally worse on slopes than on thresholds because of the wider range of imbalance values, with a tendency towards the positive side. Nevertheless, the $\mathrm{BC_a}$ method appears potentially most useful because all the other methods exhibit less manageable problems. The bootstrap standard error method, echoing its performance on thresholds, produces very widely spread results that are dependent on sampling scheme and on N. The basic bootstrap method suffers from low coverage and extreme negative imbalance. The bootstrap-t method suffers from very low coverage and high negative imbalances. Finally the bootstrap percentile method suffers from high positive imbalances, and a tendency towards low coverage.

The bootstrap percentile method emerges as the best method for threshold confidence intervals, and the $\mathrm{BC_a}$ method as the best for slopes. Performance has deteriorated noticeably on all measures, by comparison with the idealized 2-AFC case.

3.3.3 Dependence of threshold coverage on cut level

Another way of viewing the accuracy with which a confidence interval method copes with variations in psychometric function slope is to look separately at threshold performance at more than one detection level. The results from sections 3.3.1 and 3.3.2 are re-plotted in figures 3.9 and 3.10 (the idealized and realistic 2-AFC conditions respectively). This time, the plots are slightly more complicated: for each confidence interval method, the graph is divided into three strips to show results for threshold confidence intervals at three different detection levels: $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$.

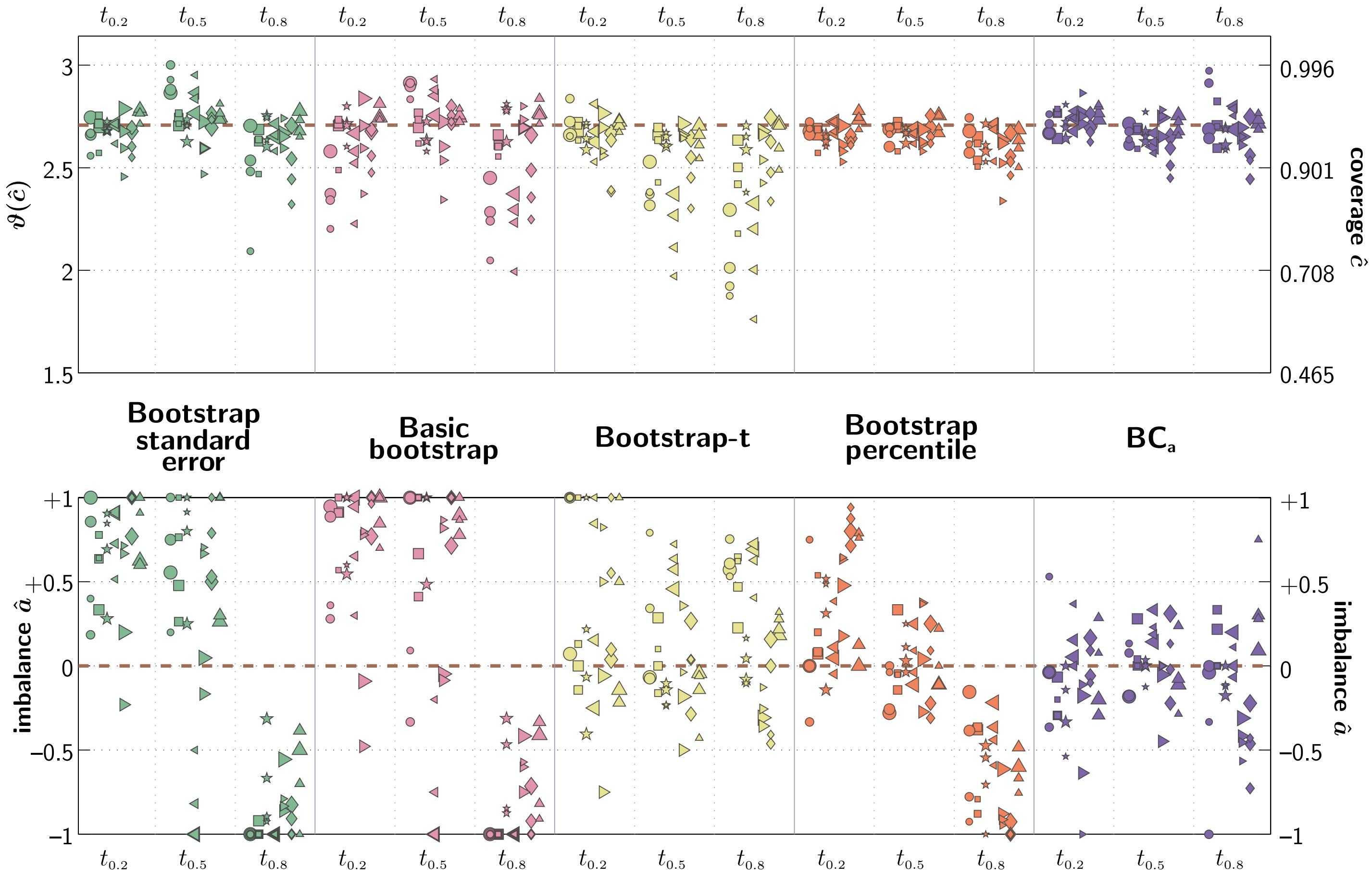


Fig. 3.9: Results of Monte Carlo coverage tests for confidence intervals on thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ obtained from a Weibull psychometric function in the idealized 2-AFC case ($\lambda_{\text{gen}} = 0$, known). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.3 for details.

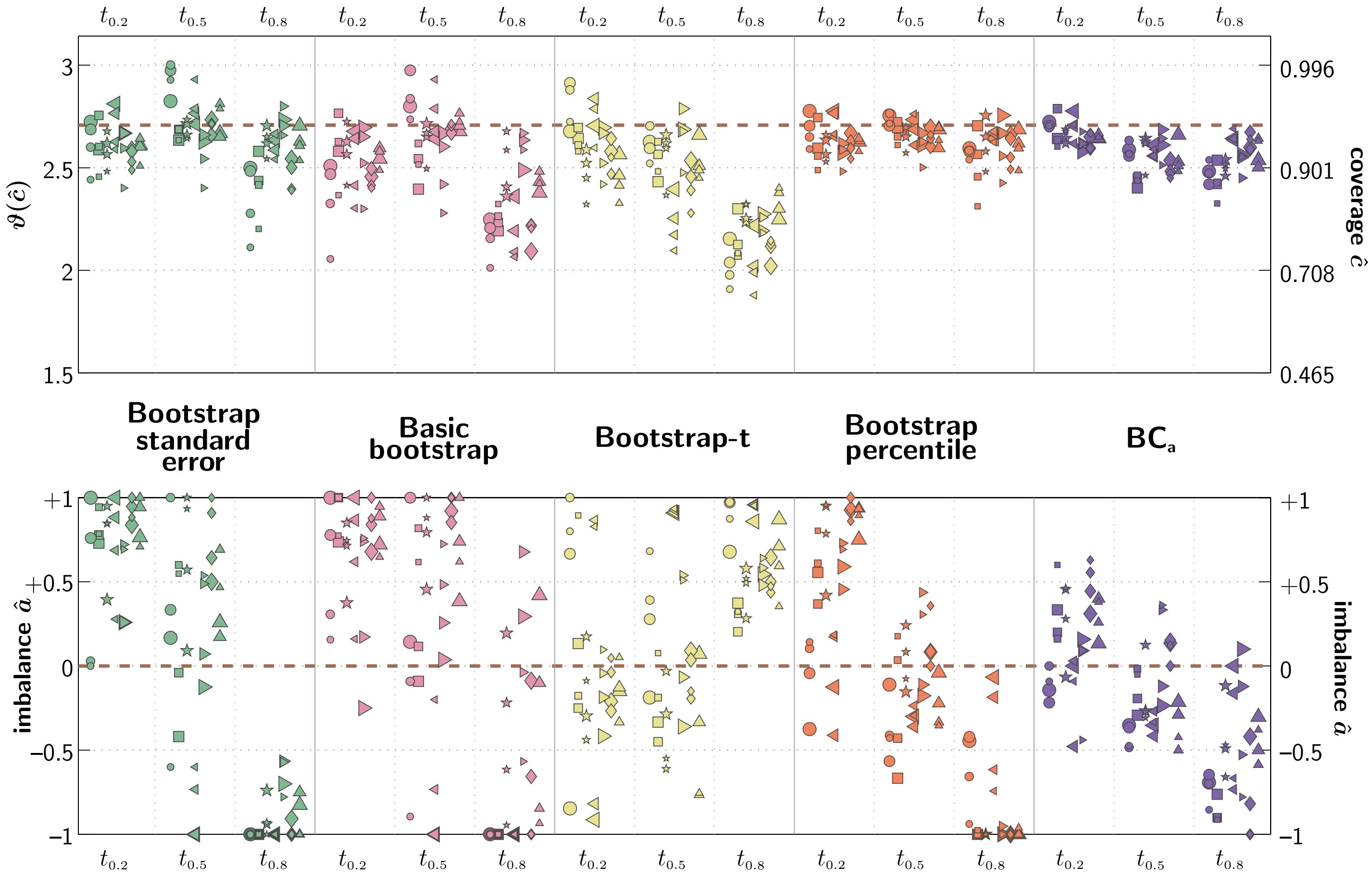


Fig. 3.10: Results of Monte Carlo coverage tests for confidence intervals on thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ obtained from a Weibull psychometric function in the realistic 2-AFC case ($\lambda_{\text{gen}} = 0.01$, unknown). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.3 for details.

In both the idealized and realistic cases, we can see that the bootstrap standard error, basic bootstrap, and bootstrap percentile methods all behave in a similar way: coverage remains roughly independent of the f value at which threshold is measured (with admirably small variation in the case of the bootstrap percentile method), but imbalance is highly positive for $f = 0.2$ and equally highly negative for $f = 0.8$. It seems likely that this effect is due to the bias in sampling that occurs when one's estimate $\hat{\boldsymbol{\theta}}_0$ differs from $\boldsymbol{\theta}_{\text{gen}}$, so that the position of sample points $\boldsymbol{x}$ relative to the curve changes.

Let us consider the bootstrap standard error method first, as the simplest case. Assume that points $\boldsymbol{x}$ are centred on the true threshold $t_{0.5}$, and that they yield an unbiased estimate $\hat{t}_{0.5}$. When $\hat{t}_{0.5} < t_{0.5}$, points $\boldsymbol{x}$ are at higher detection levels on the estimated curve. This means that they are closer to the estimated threshold at the higher detection level, $\hat{t}_{0.8}$, than they would be if the estimate $\hat{t}_{0.5}$ were correct. As a result, bootstrap values $\hat{t}^{*}_{0.8}$ are constrained more tightly than they should be, and the bootstrap estimate of the standard error $\widehat{\text{se}}_{0.8}$ is too low. Therefore the upper confidence limit $\hat{t}_{0.8} + 2\,\widehat{\text{se}}_{0.8}$ is set too low, and the probability of rejection in the upper tail of the confidence interval is too great. When $\hat{t}_{0.5} > t_{0.5}$, points $\boldsymbol{x}$ fall further away from $\hat{t}_{0.8}$, whose variability is therefore *over*-estimated, so that the lower confidence limit $\hat{t}_{0.8} - 2\,\widehat{\text{se}}_{0.8}$ is *also* set too low, and that probability of rejection in the lower tail is too small. The result is a negative imbalance, and similar logic predicts a positive imbalance for the lower threshold estimate $\hat{t}_{0.2}$.

The basic bootstrap method corrects for the extent to which the distribution of bootstrap thresholds $\hat{t}^{*}$ is non-normal, by computing confidence limits from percentiles of $\hat{t}^{*}$ rather than a fixed multiple of $\widehat{\text{se}}$, but it does not correct for the sampling bias, and so suffers the same problem. The bootstrap percentile method reverses the lengths of the two halves of the confidence interval, so that the non-normality of $\hat{t}^{*}$ works against the sampling bias effect; however, it is clear from figure 3.9 that this is not sufficient in either the idealized or realistic 2-AFC cases: the problem is still present, with reduced magnitude relative to the basic bootstrap method, but in the same direction.

The bootstrap-t method aims to correct the bias problem by correcting the second moment of $\hat{t}^*$. In figure 3.9 we see that the overall downward trend in $\hat{a}_{95}$ has been flattened out by the bootstrap-t method, although the results are still poor because of the wide spread of imbalance values, and because there is now a downward trend in $\hat{c}_{95}$ with increasing detection level. The BC$_\mathrm{a}$ method, which aims to correct for bias in $\hat{t}^*$ and skew in $\dot{\ell}(\hat{t}^*)$, was more successful, in that there is no overall trend in either $\hat{c}_{95}$ or $\hat{a}_{95}$.[†]

However, neither the bootstrap-t method nor the BC$_\mathrm{a}$ method performs well in the realistic case (figure 3.10). Note that both methods rely on a sufficiently accurate estimate of parameter covariance in order to correct for sampling bias—an estimate which, in its turn, relies on sufficiently accurate parameter estimates. If there is systematic bias in the simulated experimenter's first estimates of λ, for example, this will be correlated with a bias in slope estimates, but the amount of correlation, which the BC$_\mathrm{a}$ method estimates from $\boldsymbol{I}^{-1}$, will also be mis-estimated. This is illustrated in section 3.3.4.

3.3.4 Performance of confidence interval methods depends on the unknown value $\lambda_{\mathbf{gen}}$

The problems associated with an unknown lapse rate, such as inaccurate coverage for slopes or the inability of the BC$_\mathrm{a}$ method to correct for sampling bias as observed in figure 3.10, arise because the upper asymptote offset λ, whose value reflects lapse rate, is difficult to estimate; if lapses are rare (occurring less than, say, 5% of the time) and their effect is only measurable at high observer performance levels (below about 90% the effect of a lapse is swamped by ordinary binomial variability) then there are very few trials per data set that give any information about the lapse rate. Worse still, the magnitude of such effects changes depending on the actual value of the true

[†] Thus the BC$_\mathrm{a}$ method could be used to produce well-balanced simultaneous confidence intervals, taking a wider sense of Beran's[1] definition of *balance*: confidence intervals at the three levels have similar coverage properties, so each would each be fairly represented if all three were asserted simultaneously.

unknown λ, as illustrated by the following simulations: four sets of tests were run, using the cumulative normal 2-AFC psychometric function ($\alpha_{\text{gen}} = 5$, $\beta_{\text{gen}} = 0.2$, $\gamma_{\text{gen}} = 0.5$) as the true psychometric function in conjunction with four different values for λ_{gen}: 0, 0.01, 0.02 and 0.03. The bootstrap standard error, basic bootstrap, bootstrap percentile and BC$_\text{a}$ methods were tested.

The four sets of results are filed in the results archive as:

- `simulations/coverage/2AFC/l0/cumnorm/`
- `simulations/coverage/2AFC/l01/cumnorm/`
- `simulations/coverage/2AFC/l02/cumnorm/`
- `simulations/coverage/2AFC/l03/cumnorm/`

Results are plotted for the BC$_\text{a}$ method in figure 3.11, and for the bootstrap percentile method in figure 3.12. Again, as in section 3.3.3, results are shown side-by-side for thresholds at three different detection levels: $f = 0.2$, $f = 0.5$ and $f = 0.8$. For the BC$_\text{a}$ method, it can be seen that trend in tail-imbalance between the three threshold confidence intervals is reduced as λ_{gen} increases. The other bootstrap methods did not show such an effect; their behaviour is exemplified by the results of the bootstrap percentile method, in which the imbalance for $t_{0.5}$ has a tendency to rise from a negative value towards 0 as λ_{gen} increases, but there is little change in the trend in imbalance values between different threshold levels, as imbalance for $t_{0.2}$ remains highly positive and imbalance for $t_{0.8}$ remains highly negative at all the tested values of λ_{gen}.

The imbalance results from the BC$_\text{a}$ method follow a similar pattern to the trend in estimation bias for the three threshold levels. Each set of 500 simulated experiments in the coverage tests yielded a distribution of 500 $\hat{t}_{0.2}$'s, $\hat{t}_{0.5}$'s, and $\hat{t}_{0.8}$'s. The median bias was computed by taking the difference between the median $\hat{t}_f$ value for each distribution and the true value t_f. Median bias values were then standardized by dividing by $\frac{1}{2}$WNPI$_{68}$ for each distribution.[†] The results are shown in figure 3.13. It can be seen that there

[†] The median is used here instead of the mean, and $\frac{1}{2}$WNPI$_{68}$ is used instead of standard deviation, to guard against undue influence from extreme values, as the $\hat{t}$ distributions were often highly non-normal (see page 272 for the definition of WNPI).

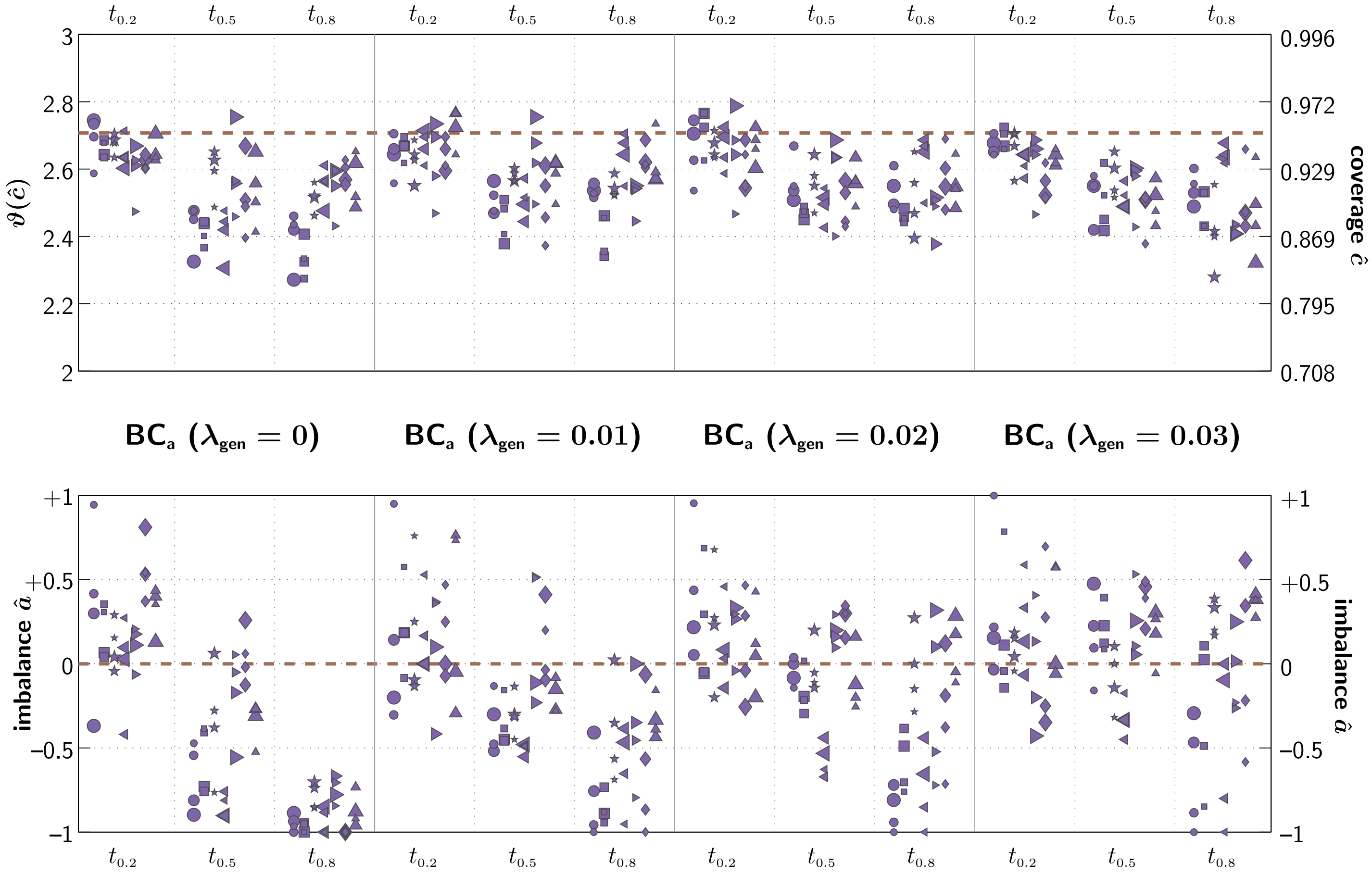


Fig. 3.11: Results of Monte Carlo coverage tests for BC$_a$ confidence intervals on thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ obtained from a cumulative normal psychometric function in four different realistic 2-AFC cases: λ is always unknown, with underlying values 0, 0.01, 0.02 or 0.03. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.4 for details.

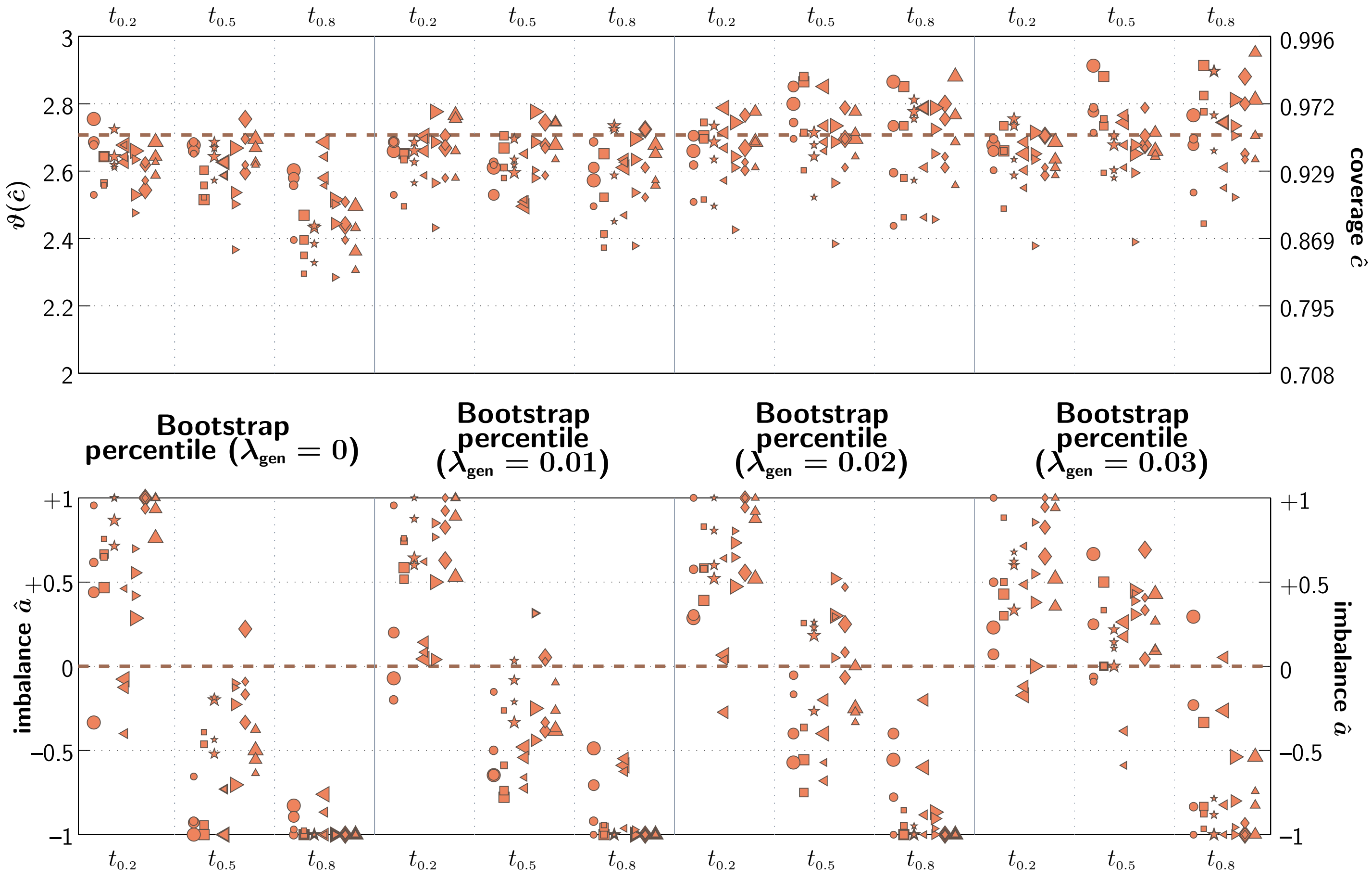


Fig. 3.12: Results of Monte Carlo coverage tests for bootstrap percentile confidence intervals on thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ obtained from a cumulative normal psychometric function in four different realistic 2-AFC cases: λ is always unknown, with underlying values 0, 0.01, 0.02 or 0.03. Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.3.4 for details.

are differences in estimation bias between the three threshold levels, and that such differences are most pronounced at $\lambda_{\text{gen}} = 0$, diminishing as λ_{gen} increases. The pattern closely mirrors that of the imbalances in figure 3.11, suggesting that it is inaccuracy in the initial estimates that prevents the BC_a method from correcting fully for the imbalance induced by sampling bias. The pattern of estimation bias might itself be explicable by the difficulty of obtaining an accurate estimate of λ. Figure 3.14, which plots the mean error of λ estimates for each of the generating values, supports this idea: generally, more accurate mean estimates are obtained when λ_{gen} is larger.[†]

It is unfortunate that different values of the unknown parameter λ_{gen} can affect the coverage results in such a way, because such behaviour violates one of the criteria suggested in section 3.1.4, *viz.* that, where coverage results are inaccurate, they should at least be *reliably* inaccurate. To summarize: both the bootstrap percentile and BC_a methods seemed promising, because they yielded the narrower ranges of $\hat{c}_{95}$ and $\hat{a}_{95}$ than other methods. However, the imbalance of the bootstrap percentile method differs for thresholds at different detection levels, suggesting that the method is sensitive to sampling bias—that is to say, the balance of the method's coverage depends on how close one's sample points lie relative to the true threshold value. The BC_a method can correct for sampling bias, but the correction is dependent on an accurate estimate for λ. When λ is known exactly (as in the idealized case) the BC_a correction works well, but when λ is unknown it fails to more or less of an extent, depending on the unknown value λ_{gen}. Failure allows sampling bias to affect both imbalance and overall coverage. The unknown value λ_{gen} affects the imbalance of threshold ($t_{0.5}$) confidence intervals in both methods.

N.B: Problems occur when λ is mis-estimated, not just because λ has become a free parameter. It has been suggested (A. Derrington, personal

[†] Treutwein and Strasburger[8] observe a similar effect, which can be seen in the top right panel of their figure 10. The continuation of the trend (gross underestimation of the higher true values of λ) is not relevant—as the authors explain, a Bayesian prior was employed, and underestimation occurs simply because the true value lies outside the likely range of values as defined by the prior. Note, however, that the first three or four columns of the panel show a similar trend to that of figure 3.14.

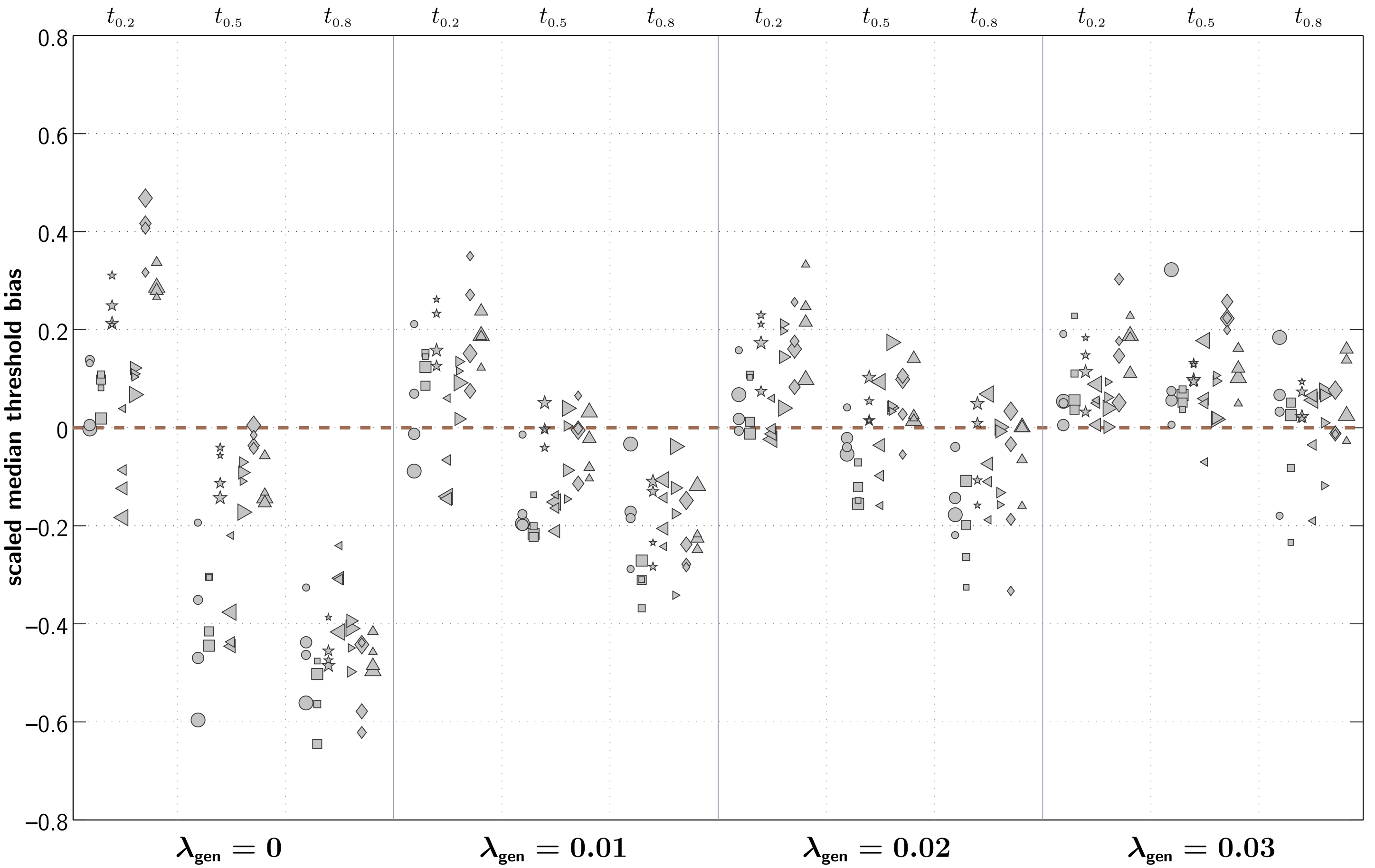


Fig. 3.13: Median estimation bias for thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$, scaled by $2/\text{WNPI}_{68}$. Points are based on 500 simulated experiments using a cumulative normal psychometric function in four realistic 2-AFC cases: λ is always unknown, and could take underlying values 0, 0.01, 0.02 or 0.03. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 3.3.4 for details.

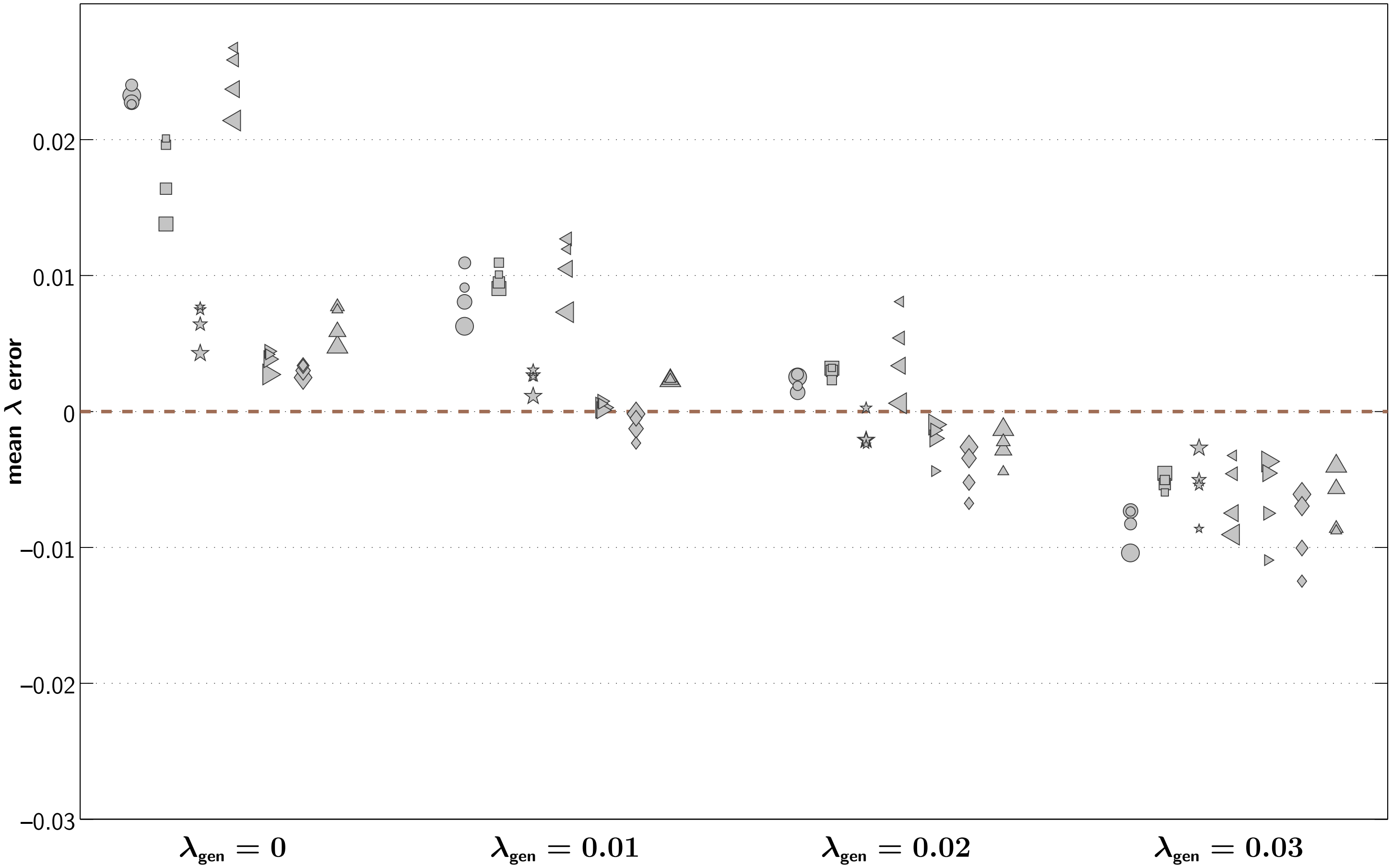


Fig. 3.14: Mean error in the estimation of λ. Points are based on 500 simulated experiments using a cumulative normal psychometric function in four realistic 2-AFC cases: λ is always unknown, and could take underlying values 0, 0.01, 0.02 or 0.03. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 3.3.4 for details.

communication) that taking a fixed *non*-zero value of λ, such as 0.01 or 0.02, might alleviate the bias in threshold slope measurement associated with the assumption of a fixed zero λ. However, Wichmann and Hill[9] found that the bias in slope estimation depended on the magnitude and direction of the difference between the true value λ_{gen} and the assumed fixed value, rather than on any particular fixed value. Three sets of coverage tests (results not plotted) support similar conclusions where coverage and imbalance are concerned. The first two test sets compared two complementary 2-AFC cases: one in which $\lambda_{\text{gen}} = 0$ but a fixed value of 0.01 is assumed during fitting, and another in which $\lambda_{\text{gen}} = 0.01$ but a fixed value of 0 is used for fitting. As has been the case in the other simulations of this chapter, the BC_a and bootstrap percentile methods were found to be the best in both situations. However, they suffered from equal and opposite imbalance problems: positive for thresholds and negative for slopes when the fixed value of λ underestimated the true value, and vice versa when the fixed value overestimated the true value. A third set of simulations found that the results of overestimating a λ_{gen} of 0.01 using a fixed value of 0.02 were almost identical to the results of overestimating a λ_{gen} of 0 using a fixed value of 0.01. The results can be found in the archive under

- `simulations/coverage/2AFC/l0f01/cumnorm/`
- `simulations/coverage/2AFC/l01f0/cumnorm/`
- `simulations/coverage/2AFC/l01f02/cumnorm/`

3.3.5 Results for the expanded BC_a methods in the realistic 2-AFC case

The realistic 2-AFC simulations of section 3.3.2 were repeated, except that the expanded BC_a method (see section 2.2.6) was tested instead of the bootstrap-t method. The three repetitions used regions of coverage 0.5, 0.25 and 0.125, and the results are filed as:

- `simulations/coverage/2AFC/l01/weibull/with_expanded05/`

- simulations/coverage/2AFC/101/weibull/with_expanded025/
- simulations/coverage/2AFC/101/weibull/with_expanded0125/

The effects of the three different levels of expansion are shown in figures 3.15 and 3.16. The format of the figures is slightly different from the usual representation. First of all, results for 68.3% target coverage (lighter symbols) are plotted along with those for 95.4% target coverage (darker symbols). Second, different quantities are plotted in the upper and lower parts of the figure. The upper panel shows $\vartheta_{500}(\hat{P}_{\mathrm{UP}})$, with $\vartheta_{500}(0)$ towards the top. The lower part of the figure shows $\vartheta_{500}(\hat{P}_{\mathrm{LO}})$, with $\vartheta_{500}(0)$ towards the bottom. Thus it is possible to imagine the MLE in the centre of the figure, with the points spreading upwards and downwards away from the centre. If they reach, and cross, the broken line, then target coverage has been achieved in that direction.

If the results were plotted in the standard format, we would see that, relative to the $\mathrm{BC_a}$ method, the expanded method increases $\hat{c}_{68}$ and $\hat{c}_{95}$ without any great change in spread, which was the aim, but unfortunately it greatly increases the spread of $\hat{a}_{68}$ and $\hat{a}_{95}$—this effect can be seen by the differences in the spread of points between the upper and lower panels of the figures. It is more interesting, however, to plot the coverage of the upper and lower halves of the confidence interval separately, because the expanded method is a conservative measure intended to compensate for the shortcomings of other methods, and thereby achieve minimum standards of coverage. We are interested in knowing at what expansion level such standards can be achieved. From figure 3.15, it appears that an expansion level of 0.5 guarantees, for all the sampling schemes studied in the chosen range of N, that coverage is sufficient in both the upper and lower half of the confidence interval, at both target coverage levels. At a level of 0.25, a few of the points are still more than one standard error away from their targets, in both the upper and lower parts of the confidence interval, and at both 68.3% and 95.4%. A similar picture emerges for slopes in figure 3.16.

Expanded methods have a tendency to exaggerate the sensitivity of some sampling schemes. Though it is true that schemes ● and ◀ are more suscep-

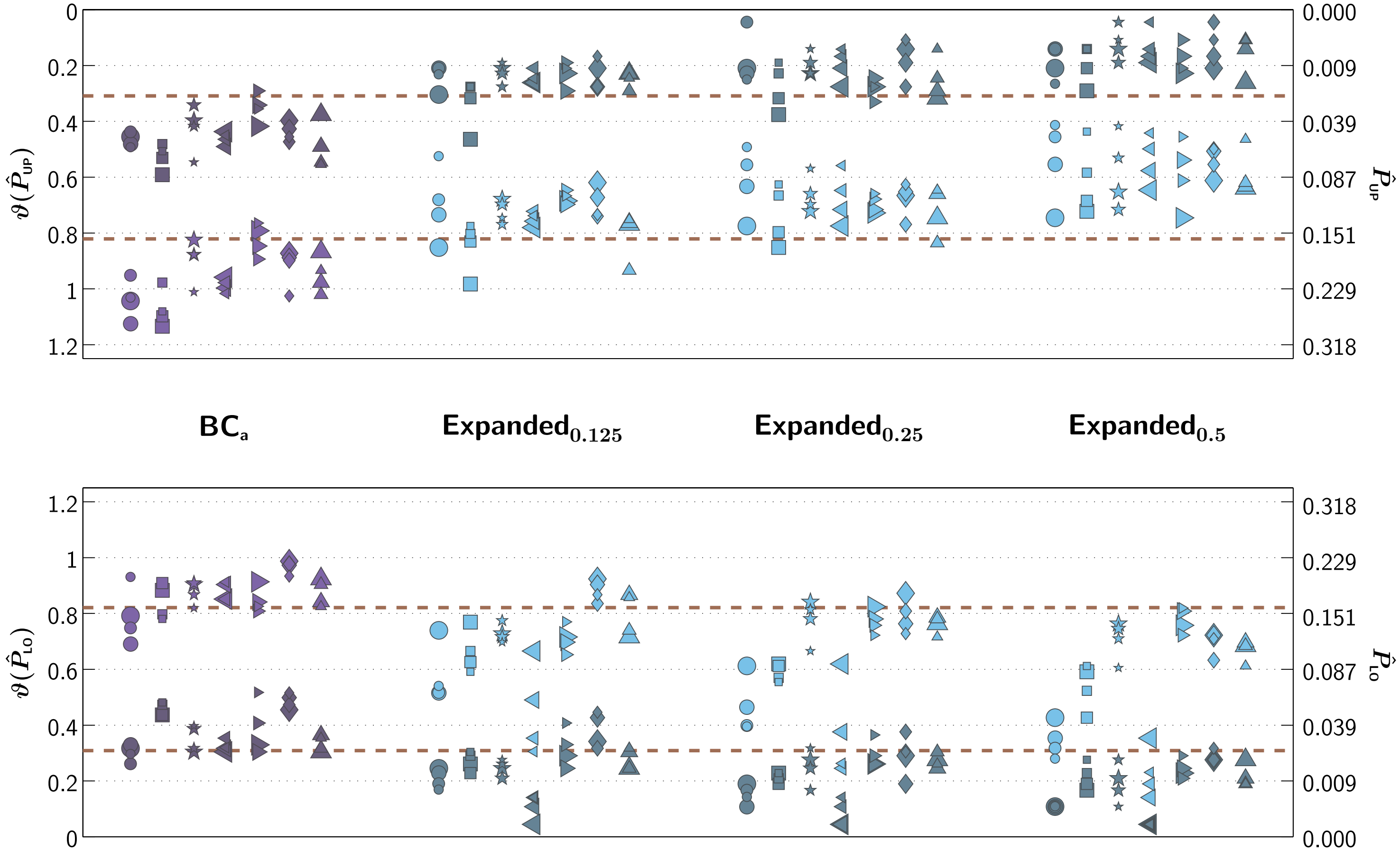


Fig. 3.15: Results of Monte Carlo coverage tests for expanded bootstrap BC$_{\mathrm{a}}$ threshold intervals obtained from a Weibull psychometric function in the realistic 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 3.3.5 for details.

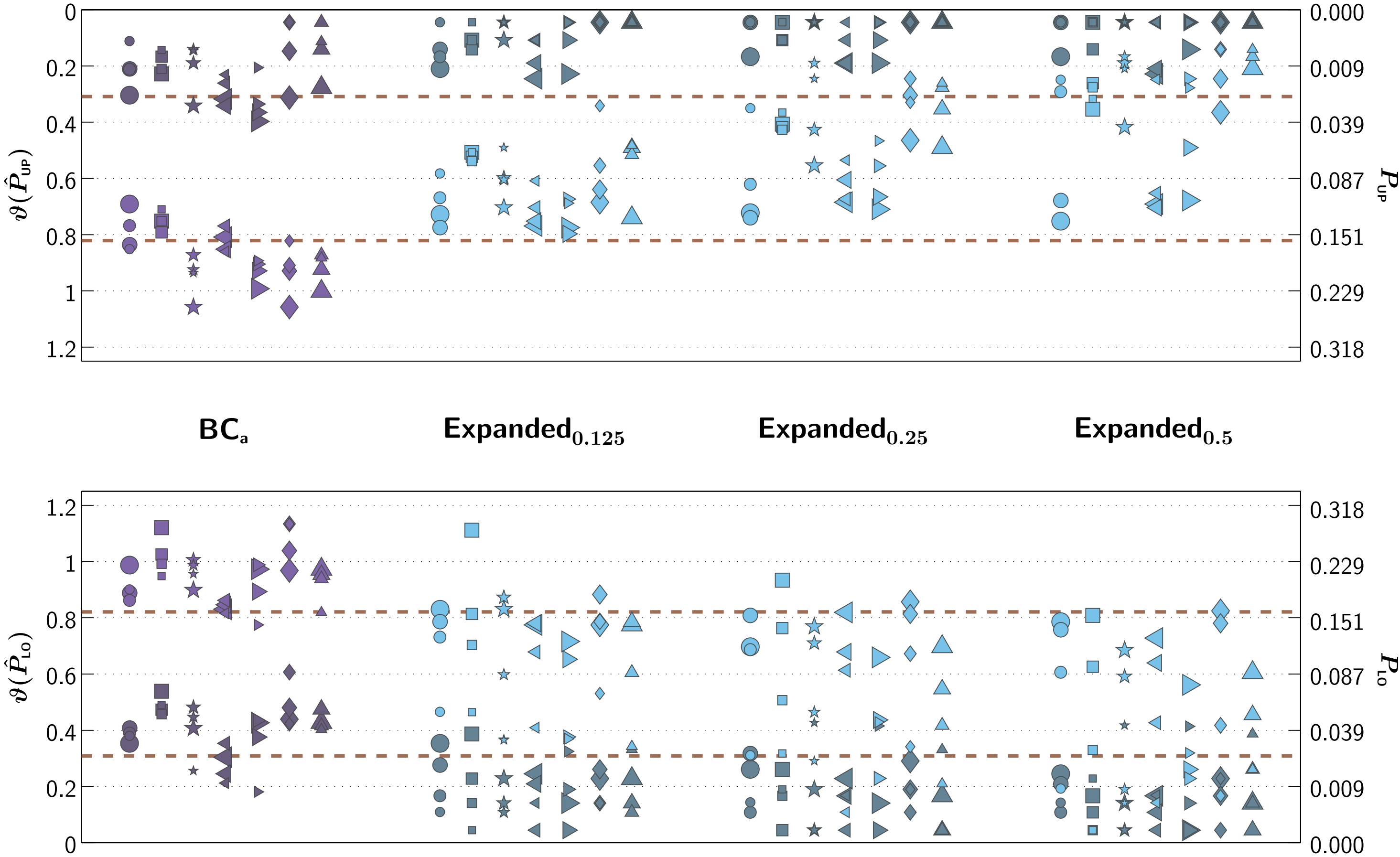


Fig. 3.16: Results of Monte Carlo coverage tests for expanded bootstrap BC_a slope intervals obtained from a Weibull psychometric function in the realistic 2-AFC case. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 3.3.5 for details.

tible to bootstrap error, the expanded method, even at a level of 0.125, has a tendency to over-correct for this fact. Their greater apparent sensitivity, as measured by the expanded method, is illustrated by the fact that they are over-covering the true underlying values to a greater extent than other sampling schemes. Therefore Wichmann and Hill,[10] who used expanded confidence interval lengths as an index of a scheme's susceptibility to bootstrap error, may have unfairly exaggerated the disadvantages of such schemes.[†]

There is, of course, room for potential improvements in the expanded confidence interval algorithm, as it is largely heuristic in nature (see section 2.2.6). Perhaps a different number of repetitions (other than 8), at points distributed in parameter space according to slightly different criteria, might reduce the differences between different sampling schemes' estimated coverage and imbalance values. However, as it stands, the expanded $\mathrm{BC_a}$ method at a level of 0.5 seems as if it might at least offer a reliable way of ensuring minimum levels of coverage.

3.4 Comparison with probit methods

The following two sub-sections compare the performance of probit methods (see section 2.1) with that of the $\mathrm{BC_a}$ bootstrap method, in both idealized and realistic conditions. Other studies of probit methods have generally used the cumulative normal psychometric function, so the bootstrap simulations of sections 3.2.1–3.2.2 and 3.3.1–3.3.2 were repeated using the cumulative normal for greater congruence with previously published work. Section 3.4.1

[†] This is still a debatable point, however. Whereas the simulations of the current chapter define sampling schemes relative to the true underlying psychometric function, the simulations reported in chapter 5, and by Wichmann and Hill, take a different hypothetical starting point: they are designed to compare widths of the confidence intervals that an experimenter might actually measure, as a function of sampling scheme, so sampling schemes are defined relative to the experimenter's *estimate* of the psychometric function. The results of such a test show an experimenter which sampling scheme to aim for—the positions that the data points should occupy relative to the estimated curve. However, as the true psychometric function is unknown, there will be some error involved in achieving that sampling scheme, and an expanded bootstrap method provides a rough indication of the risk associated with such error.

deals with yes-no designs, and section 3.4.2 deals with 2-AFC designs.

3.4.1 Probit methods in yes-no designs

The simulated yes-no experiments of sections 3.2.1 and 3.2.2 were repeated, using the cumulative normal function with $\alpha_{\mathrm{gen}} = 0$ and $\beta_{\mathrm{gen}} = 1$ instead of the logistic function. Bootstrap-t intervals were not calculated, but probit intervals were calculated. The results are filed under

- `simulations/coverage/yesno/g0f0l0f0/cumnorm/with_expanded0125/`
- `simulations/coverage/yesno/g02l01/cumnorm/with_expanded0125/`

For the bootstrap standard error, basic bootstrap, bootstrap percentile and BC$_{\mathrm{a}}$ methods, results were very similar to those obtained with the logistic function, in both the idealized and realistic cases. These results will not therefore be shown, except for the BC$_{\mathrm{a}}$ method which, being generally the best of the bootstrap methods, will provide a convenient standard against which to compare the performance of the probit interval methods.

Figures 3.17 and 3.18 show results for 95.4% threshold and slope confidence intervals, respectively. In each figure, probit results from both the idealized and realistic cases are shown, alongside BC$_{\mathrm{a}}$ results for comparison.

The first thing to note is that, in the idealized case with λ and γ fixed at 0, probit methods perform just as well as bootstrap methods. Probit fiducial limits on thresholds, and the probit standard method for slopes, both produce coverage and imbalance estimates that are as closely grouped around the target values as those for the BC$_{\mathrm{a}}$ method. Therefore, while the results of section 3.2.1 support Swanepoel and Frangos' finding[6] that the bootstrap-t method provides good coverage, the current results suggest that its coverage is no better than that of much less computationally expensive asymptotic methods, at least for $120 \le N \le 960$.

The addition of λ and γ as constrained free parameters is detrimental to the performance of probit methods as well as bootstrap methods. Note that, where performance of the BC$_{\mathrm{a}}$ method is now poor, performance of

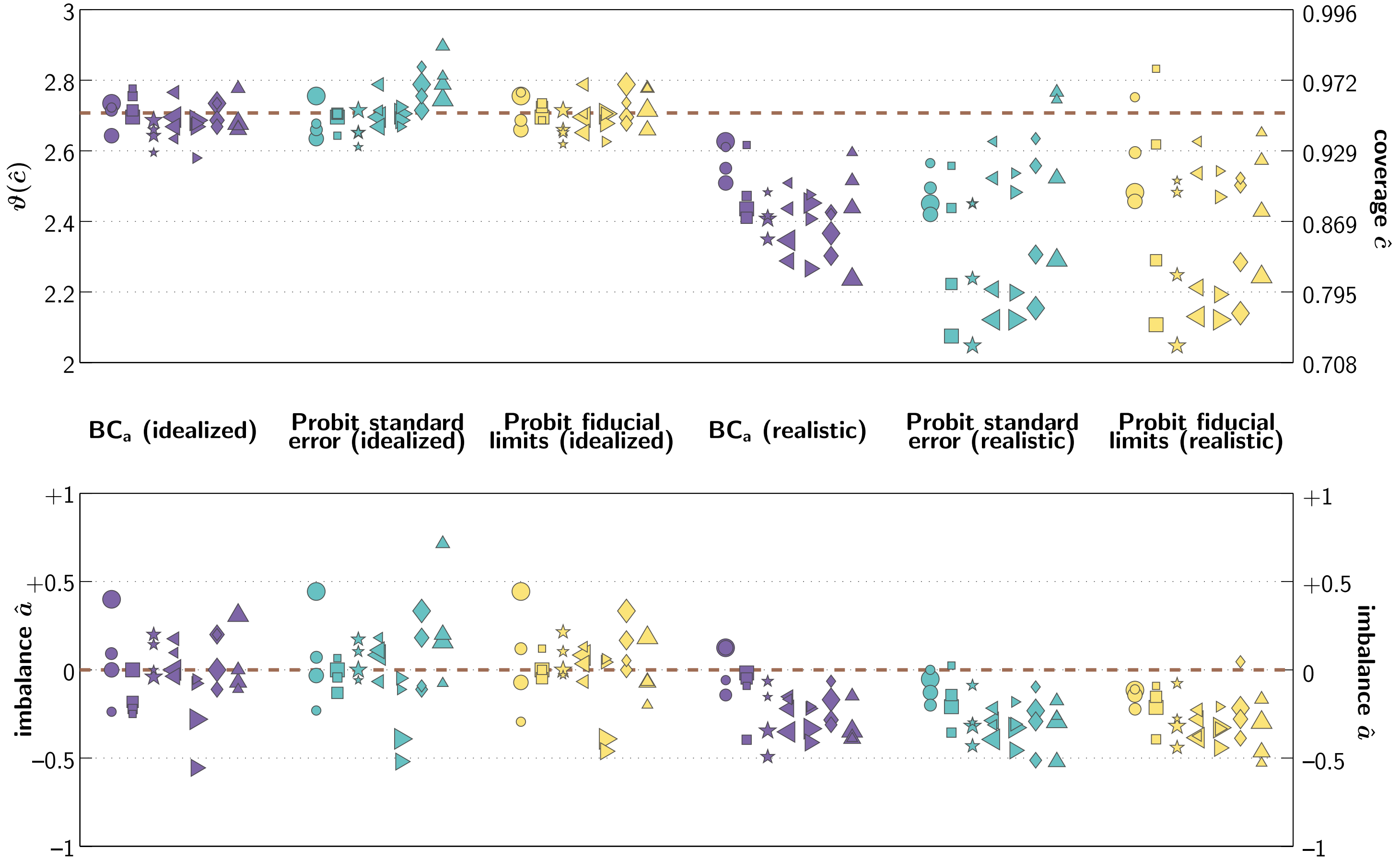


Fig. 3.17: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in a simulated yes-no experiment. Results are shown for both the idealized case (first three groups) and the realistic case (second three groups). Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.4.1 for details.

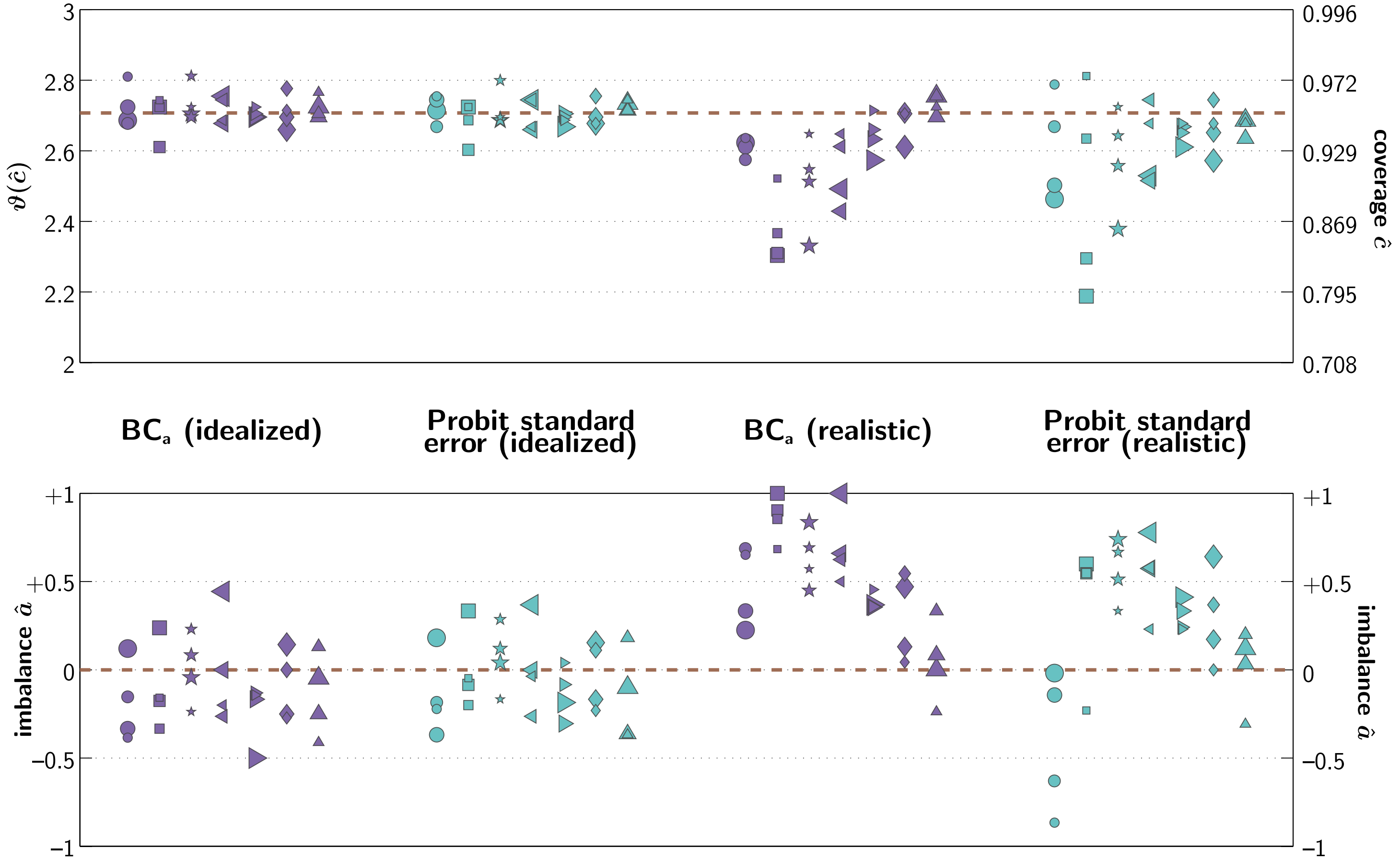


Fig. 3.18: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in a simulated yes-no experiment. Results are shown for both the idealized case (first two groups) and the realistic case (second two groups). Symbol shapes denote the seven sampling schemes of figure 1.3. See sections 3.1.3 and 3.4.1 for details.

the probit methods has become even worse, in that coverage estimates are more widely spread out for thresholds, and both coverage and imbalance are more widely spread out for slopes, when probit methods are used. (However, despite its wide spread of imbalance values for slopes, the probit standard error method does not suffer from such extreme values as the BC$_\mathrm{a}$ method, which is even completely unbalanced for schemes ■ and ◀ at $N = 960$.)

3.4.2 Probit methods in 2-AFC designs

The simulated 2-AFC experiments of section 3.3 were repeated, using the cumulative normal function with $\alpha_{\mathrm{gen}} = 5$ and $\beta_{\mathrm{gen}} = 0.2$ instead of the Weibull function. Bootstrap-t intervals were not calculated, but probit intervals were.

The results are filed under

- `simulations/coverage/2AFC/l0f0/cumnorm/`
- `simulations/coverage/2AFC/l01/cumnorm/`

for the idealized and realistic cases respectively.

Results were found to be qualitatively different for 68.3% intervals and 95.4% intervals. Therefore, results will be shown separately at the two coverage levels. Figures 3.19 and 3.20 show results for thresholds, at 68.3% and 95.4%, respectively. Figures 3.21 and 3.22 will show results for slopes in the same way. In each figure, probit results from both the idealized and realistic cases are shown, alongside the BC$_\mathrm{a}$ results for comparison.

Bootstrap results were very similar to those obtained with the Weibull and logistic functions (section 3.3). The majority of bootstrap results will not therefore be shown, with the exception of the BC$_\mathrm{a}$ method which provides a standard against which to compare the performance of the probit interval methods.

As in the yes-no case, the change from idealized to realistic assumptions (the addition of the parameter λ as an unknown with a non-zero underlying value) has had detrimental effects on all the methods, most notably a drop in mean overall coverage for both thresholds and slopes. At 68.3%, the probit

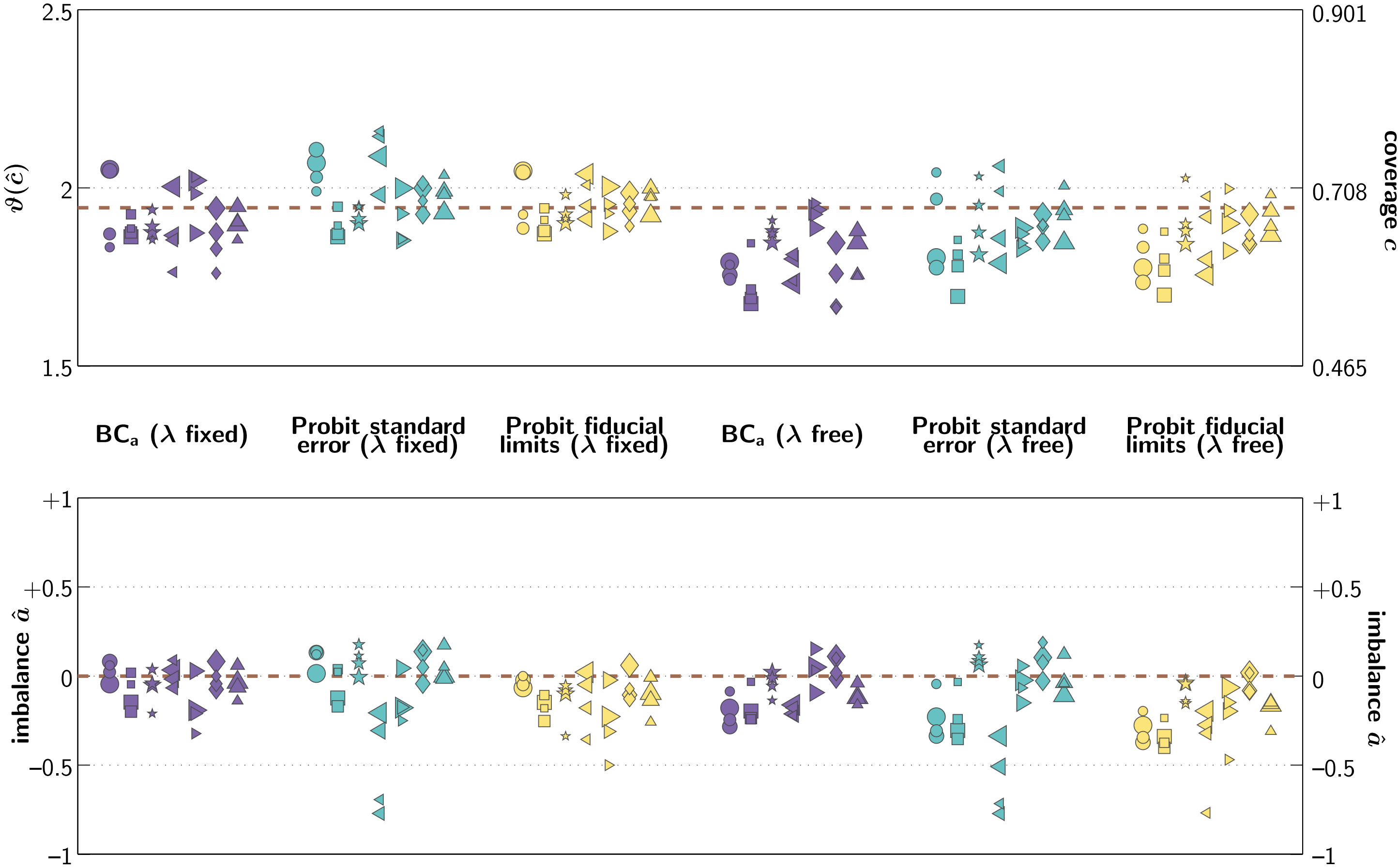


Fig. 3.19: Results of Monte Carlo coverage tests for 68.3% threshold confidence intervals obtained from the cumulative normal function in a simulated 2-AFC experiment. Results are shown for both the idealized case (first three groups) and the realistic case (second three groups). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.4.2 for details.

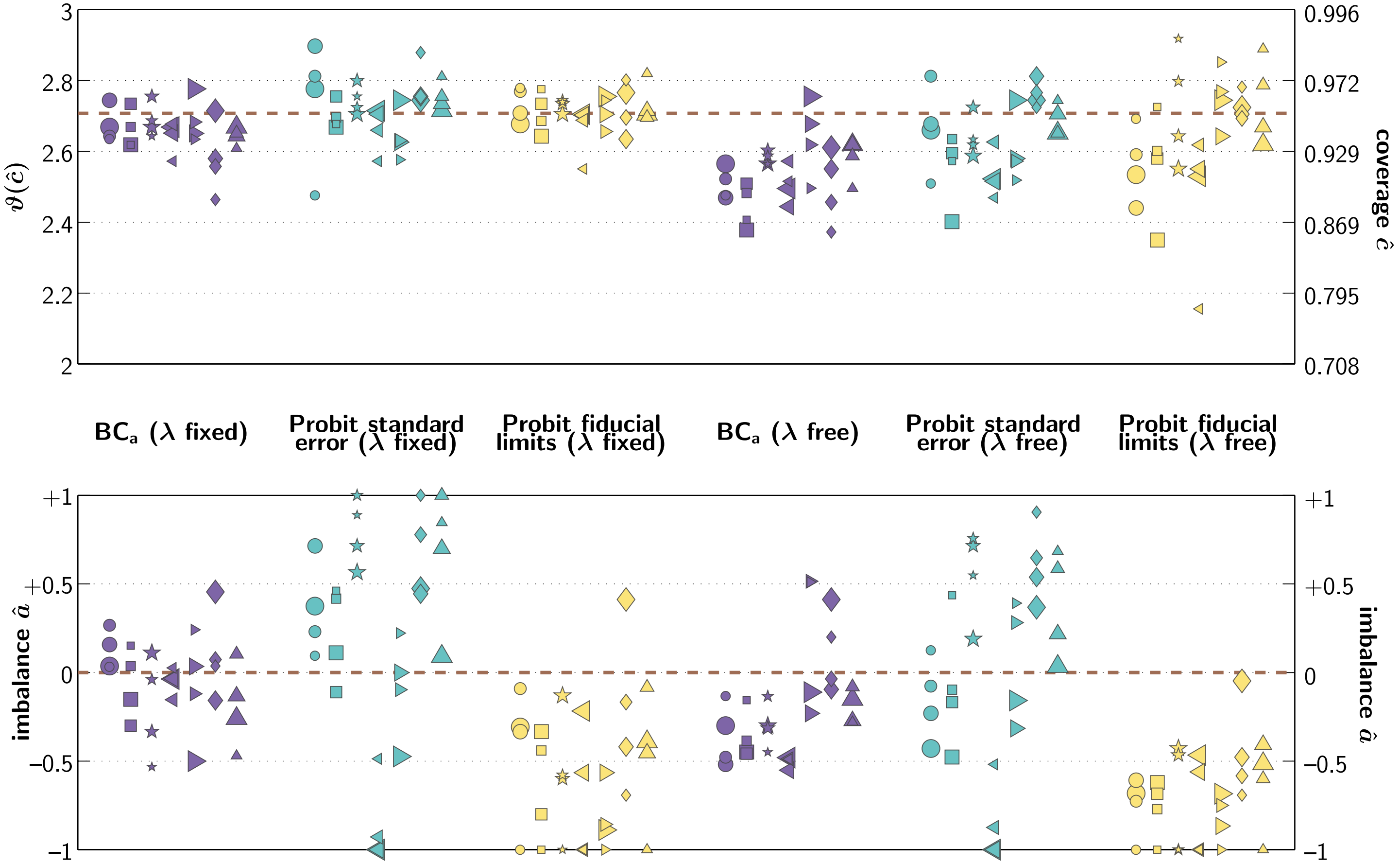


Fig. 3.20: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in a simulated 2-AFC experiment. Results are shown for both the idealized case (first three groups) and the realistic case (second three groups). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.4.2 for details.

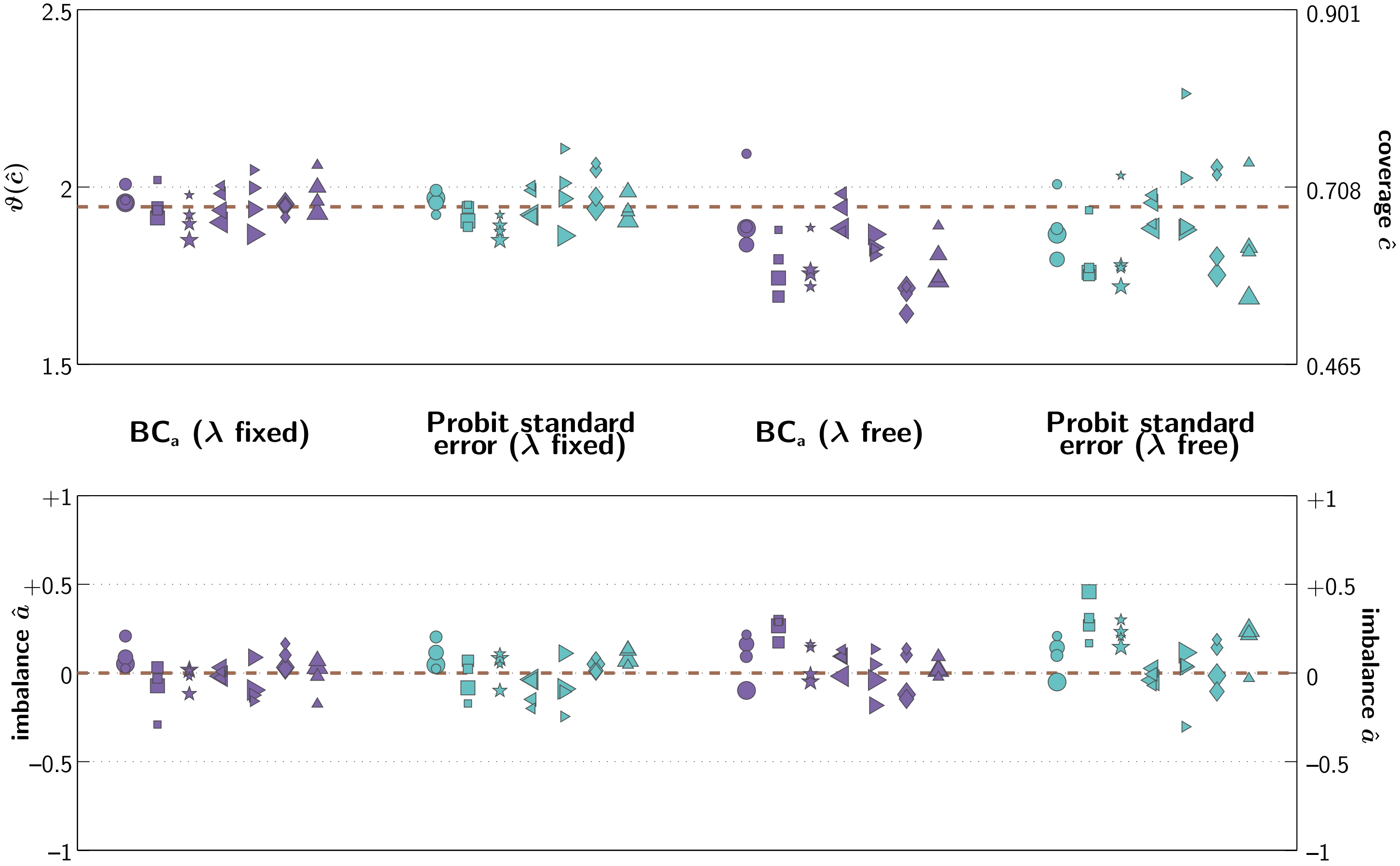


Fig. 3.21: Results of Monte Carlo coverage tests for 68.3% slope confidence intervals obtained from the cumulative normal function in a simulated 2-AFC experiment. Results are shown for both the idealized case (first two groups) and the realistic case (second two groups). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.4.2 for details.

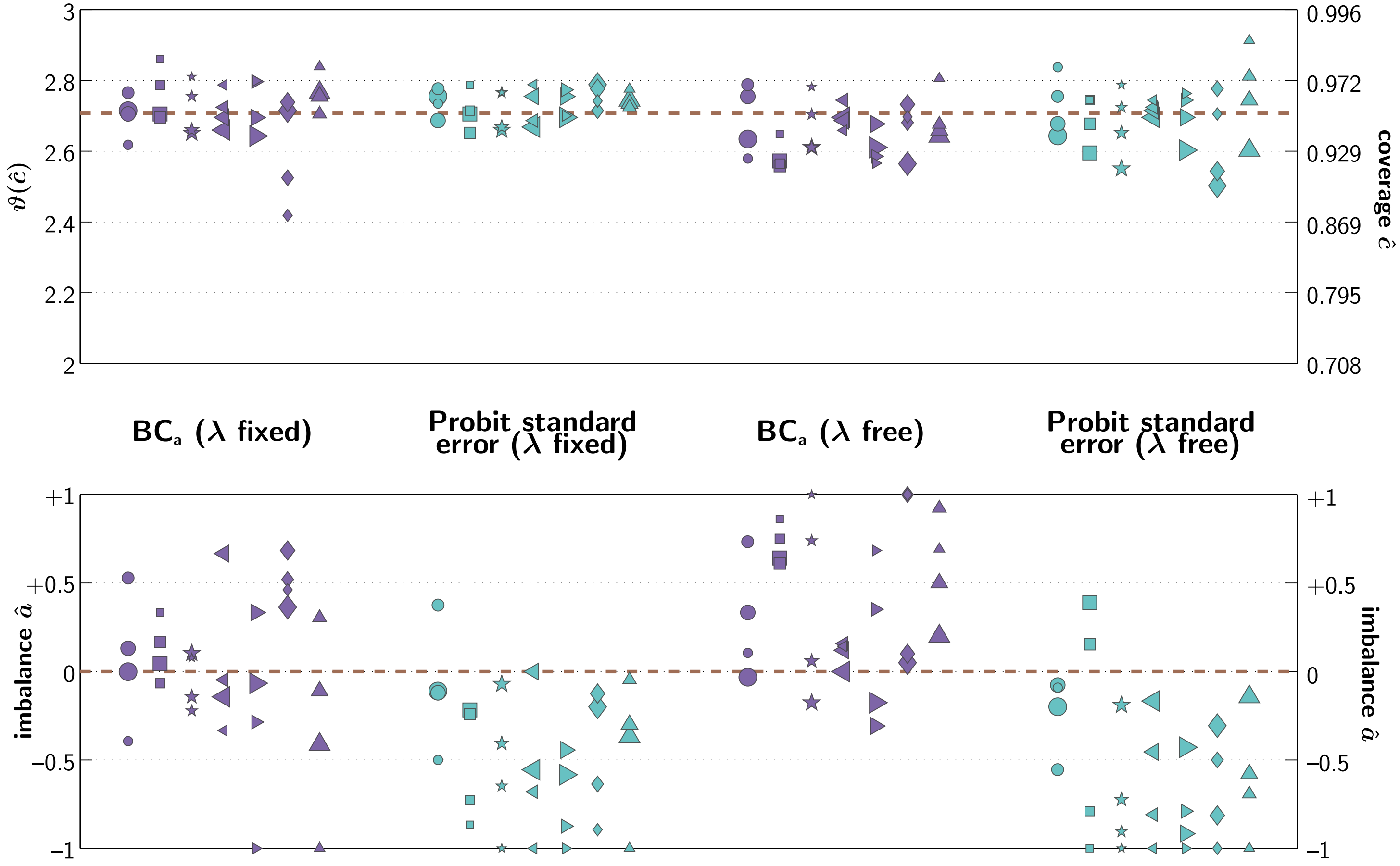


Fig. 3.22: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in a simulated 2-AFC experiment. Results are shown for both the idealized case (first two groups) and the realistic case (second two groups). Symbol shapes denote the seven sampling schemes of figure 1.2. See sections 3.1.3 and 3.4.2 for details.

methods are very little different from the BC_a in either the idealized or realistic case. At 95.4%, as in the yes-no case, probit methods performed worse than the BC_a: imbalance values for the probit standard error method are more spread out for thresholds, and a strong negative imbalance was found in both probit fiducial limits for thresholds and probit standard errors for slopes. The pattern of results differs from that found in the yes-no simulations, however, in that probit methods exhibit such problems at 95.4% even in the *idealized* case. It seems, therefore, that probit methods are potentially suitable for yes-no experiments, and are good for estimating short confidence intervals (of the order of ±1 standard error) even in 2-AFC experiments. They should not, however, be used in 2-AFC experiments to compute larger confidence intervals (of the order of ±2 standard errors) even if it assumed that the problem of unknown λ can be safely disregarded.

3.5 Summary and discussion

The current chapter used Monte Carlo simulation to investigate the accuracy of coverage of various bootstrap methods that can be used to compute two-tailed confidence intervals for psychophysical thresholds and slopes. Yes-no experiments and 2-AFC experiments were simulated, and in each paradigm an idealized case, in which it is known that the observer never guesses or lapses, was compared with a more realistic case in which λ (and γ in the yes-no case) took a small non-zero value and was treated as an unknown nuisance parameter. Seven sampling schemes were tested, each at four different values of N. Several bootstrap methods were tested, and their performance was compared with that of asymptotic-theory confidence intervals from probit analysis. Overall coverage was estimated for two-tailed confidence intervals, as well as imbalance between coverage in the two sides of the interval.

Previous studies by Swanepoel and Frangos[6] and by Lee[4] used Monte Carlo simulation to test the coverage of confidence intervals of the parameters of psychometric functions. Good results were reported for the bootstrap-t method. However, they considered only the yes-no experimental paradigm, only the logistic function, only the idealized case, and only the bootstrap-t method. Results from section 3.2.1 are consistent with the findings of both studies, in that the bootstrap-t method was found to be fairly accurate for slopes in the idealized yes-no case, although it suffered from low coverage on thresholds when the range of the sampling scheme was narrow—the $\mathrm{BC_a}$ method performed better in this case.[†] However, at the range of N values studied here, which very nearly encompasses the values studied by Swanepoel and Frangos,[6] no bootstrap method performed better than probit analysis under the same conditions (section 3.4.1). In yes-no experimental designs, it was only in the realistic case that bootstrap methods were superior to the probit method.

Thus, when it is safe to assume $\gamma = \lambda = 0$ in a yes-no context, it is advisable to use the far less computationally expensive methods of probit fiducial limits (for thresholds) and probit standard errors (for slopes), rather than bootstrap methods. This conclusion is unlikely to hold for very low values of N, however. Another Monte Carlo simulation study by Foster and Bischof[11] compared the bootstrap standard error and probit standard error methods using the cumulative normal psychometric function. Again, only the idealized yes-no case was addressed. They found the accuracy of bootstrap standard errors to be relatively insensitive to changes in the value of N, whereas probit standard errors became much less accurate, particularly for slopes, when N dropped below about 50 (psychophysical experiments rarely use block designs with $N < 100$, although there is a temptation towards such designs in some clinical settings where psychometric function slope is important but it is difficult to take large numbers of observations). Probit standard errors were found by Foster and Bischof to be more accurate than bootstrap standard errors for the range of N considered in this chapter. The simulations of sections 3.2.1 and 3.4.1 supported this finding, in that probit

[†] Note that the performance of the bootstrap-t method may vary according to the method used within it to estimate $\hat{v}$. The current study used a parametric method, relying on the asymptotic Fisher approximation to the parameter covariance matrix. It may therefore have been more vulnerable to error at low values of N and in sampling schemes which included samples at very high or low performance values.

methods were found to be at least as good as the best of the bootstrap methods tested, and the bootstrap standard error method was not the best (the bootstrap percentile and $\mathrm{BC_a}$ methods were better for thresholds, and the bootstrap-t and $\mathrm{BC_a}$ methods were better for slopes). Foster and Bischof were concerned with the accuracy of standard error estimates rather than the coverage accuracy for larger confidence intervals, but the current simulations found the results to be comparable: there was little difference between the trends at 68.3% and those at 95.4% in the idealized yes-no case.[†]

In an earlier paper,[13] Foster and Bischof also investigated the bootstrap standard error method in the idealized 2-AFC case, for which they found the results to be less accurate than those of the idealized yes-no case. The bootstrap method was found to be vastly superior to an "incremental" method (based on "combination of observations" and described by Foster[14]). In section 3.3.1, simulation results show that the bootstrap standard error method to be inferior to other bootstrap methods such as the bootstrap percentile and $\mathrm{BC_a}$. Probit methods were found to be inferior to the $\mathrm{BC_a}$ method in 2-AFC experiments, being suitable for short confidence intervals of the order of ±1 standard error, where their performance was similar to that of the $\mathrm{BC_a}$ but unsuitable for longer confidence intervals of the order of ±2 standard errors, where they were poorly balanced. This is consistent with the findings of McKee *et al.*[15] who demonstrated that probit methods produced poor approximations to the standard error of thresholds from a known 2-AFC psychometric function, particularly when $N < 100$ or when the function is poorly sampled. The poor performance of probit methods in the 2-AFC case was an exception to the general trend, in that their performance was poor in *both* the idealized and the realistic case. For most of the simulations, the addition of unknown nuisance parameters γ and/or λ caused coverage accu-

[†] Note that the accuracy of the standard error estimate is a direct indicator of the accuracy of confidence interval *length* ("correctness"), but not a direct indicator of the accuracy of confidence interval *coverage*, so Foster and Bischof's results may not be directly comparable to those presented here. Coverage relies not only on mean confidence interval length but also on the magnitude and direction of the correlation between confidence interval length and estimation error $|\hat{u} - u|$. Hall[12] provides a more detailed theoretical treatment of this distinction.

racy for most confidence interval methods to deteriorate. The simulations of section 3.3.4 suggest that this is at least partly because of the difficulty of estimating small values of λ and γ accurately.

Some test sets used different psychometric function shapes from others. In the yes-no simulations, the logistic function was used in order to allow more direct comparison with previous Monte Carlo coverage studies. In the 2-AFC simulations, the Weibull function was tested because of its widespread use in current vision research employing 2-AFC methods. However, when one set of 2-AFC tests was repeated using the logistic function, and when both the yes-no and the 2-AFC tests were repeated using the cumulative normal, results were qualitatively indistinguishable. This is reassuring, to the extent that it suggests that the properties of confidence interval methods are not dependent on a particular mathematical form of the psychometric function, but rather on the observer performance levels predicted by the function.

The use of a set of fixed sampling schemes contrasts with the approach of Lee,[4] who used a probabilitistically generated sampling scheme on each Monte Carlo replication of the experiment. Lee's method has the advantage of realism, in that it emulates the inevitable randomness of stimulus placement on an unknown function, but it has the disadvantage of averaging out variations in coverage accuracy that might occur from one experiment to the next due to random variation in sampling scheme. Much of the variability in the coverage and imbalance estimates of the current chapter was associated with differences in sampling scheme and in N. Some methods (for example, the $\mathrm{BC_a}$ method) were more stable in this regard than others (such as the basic bootstrap method). The relationships between sampling schemes, N and coverage accuracy were not straightforward, however. In many cases there was no clear trend in the results as N increased from 120 to 960. Where a trend was visible, its direction could change depending on sampling scheme or on confidence interval method. For example, in the bootstrap standard error and basic bootstrap clusters in the upper panel of figure 3.5, $\hat{c}_{95}$ for ◀ drops towards the target level from above as N increases, whereas ▶ rises towards the target level from below. In the upper panel of figure 3.1, $\hat{c}_{95}$

for ▲ drops towards the target level from above in the bootstrap standard error cluster, but rises towards the target level from below in the bootstrap-t cluster. In most cases, where a trend was visible, it was usually towards better coverage accuracy as N increased, as in all the forgoing examples. However, this was not always the case; for example, in figure 3.3 some sampling schemes, (including ▲ in the $\mathrm{BC_a}$ cluster, and most of the schemes in the basic bootstrap cluster) show the opposite trend, with more accurate coverage at *lower* values of N.

Coverage accuracy, and in particular the consistency of results across different sampling schemes and different values of N, was better for some bootstrap methods than for others. In general, the bootstrap standard error and basic bootstrap methods performed worst, the bootstrap-t method slightly better, and the bootstrap percentile and $\mathrm{BC_a}$ methods were the best. Theoretical treatments[5,12,16] generally predict that the $\mathrm{BC_a}$ and bootstrap-t methods should be the best, as they display the properties of second-order accuracy and second-order correctness that the others lack. The bootstrap-t method may fall short in the current study because of its reliance on the parametric approximation to $\hat{v}$. It would be instructive to compare some of the other methods that can be employed within the bootstrap-t—for example, the more computationally intensive jackknife-based methods explored by Swanepoel and Frangos.[6]

The $\mathrm{BC_a}$ method generally produced the most consistent results across different sampling schemes and values of N. For this reason, it is the fairest available method with which to compare the efficiency of sampling schemes, and it will be used for such a purpose in chapter 5. Also, the $\mathrm{BC_a}$ method was generally found to be the most balanced of the methods studied. However, simulations in section 3.3.4 suggest that the balance of threshold confidence intervals from the $\mathrm{BC_a}$ method may depend on the value of λ_{gen},[†] because the accuracy of estimation of lambda varies, so no general correction for $\mathrm{BC_a}$ coverage can be proposed. For thresholds the bootstrap percentile method

[†] Naturally one would anticipate similar problems with both γ *and* λ in the realistic yes-no case.

often has a more accurate overall coverage level than the $\mathrm{BC_a}$, and results are nearly as tightly grouped together as those of the $\mathrm{BC_a}$. However, the bootstrap percentile method is slightly more prone to imbalance, and its poor balance between confidence intervals for threshold at different detection levels, even in the idealized case (see section 3.3.3), serves as a warning that its good performance is likely to be contingent on fortuitously managing to centre one's sample points close to the true threshold.

When realistic psychophysical conditions were assumed (i.e. when λ and/or γ took non-zero values and were treated as unknowns), the coverage of the $\mathrm{BC_a}$ method was consistently too low. Such problems may be compensated for by inflating one's target coverage when computing confidence intervals (for example, aiming for 0.99 when 0.954 is desired), or by using an expanded interval to obtain highly conservative estimates of confidence interval bounds.[†] Given that the performance of the $\mathrm{BC_a}$ method (reflected in the balance of rejection probabilities in the two tails of a confidence interval) depends on an unknown parameter, the latter solution would be more advisable. For the expanded method, the simulations of section 3.3.5 indicate that an expansion level of 0.5 is advisable to ensure adequate coverage in both sides of the confidence interval. Another approach might be to forgo the use of one-dimensional confidence intervals and instead using a two-dimensional confidence *region*—this is dealt with in chapter 4.

[†] Note that no such solution adds power for free to the hypothesis test associated with the confidence interval. Rather, inflation of the target coverage level, or the use of the expanded method, naturally increases the width of the confidence interval relative to that of the ordinary $\mathrm{BC_a}$ interval, the intention being to *correct* for the fact that the width and coverage of the ordinary interval are lower than they should be. Without correction, and if the interval were assumed to have reached its target coverage, the test would appear to be more powerful than it actually is, leading to a higher type I error rate than intended.

Unfortunately, no definite rule can be given for the required magnitude of the correction—section 3.3.4 shows that, using current methods, this would depend on the (unknown) true value of λ.

References for chapter 3

[1] Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, **83**(403): 679–686.

[2] Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, **85**(410): 417–426.

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

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

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

[6] 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.

[7] Freeman, M. F. & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**(4): 607–611.

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

[9] 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` .

[10] 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 .

[11] 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.

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

[13] 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.

[14] Foster, D. H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biological Cybernetics*, **53**: 189–194. An erratum is given on page 412 of the same volume.

[15] 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.

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

4. Coverage properties of joint confidence regions

4.1 Introduction

Psychophysical studies in general, and research from the psychophysical literature concerning bootstrap methods in particular, have tended to focus on one-dimensional confidence interval methods. Possibly this is because many psychophysical hypotheses are formulated in terms of a single dimension (usually threshold) although it may also be due in part to the scarcity or relative novelty of accessible two-dimensional methods. There are advantages to a two-dimensional approach, however: assuming that there is theoretical justification for both threshold and slope to vary according to the parameters of a psychophysical experiment, then for the psychometric function as for any other multivariate setting, a joint hypothesis test may be more powerful and more informative than a test based on a single dimension of variability, or on two separate dimensions. An example of the application of two-dimensional methods might be the assessment of whether an observer's behaviour is more consistent with the ideal observer for a signal known exactly, or with the ideal observer for a signal known only statistically: the former predicts both a lower threshold and a shallower slope than the latter, according to the formulation of Green and Swets.[1]

In two or more dimensions, a confidence interval becomes a confidence *region*—an area, volume or hypervolume within which the true value of a parameter vector can be asserted to lie with a given confidence level c. Many interval methods, including all of the bootstrap methods examined in chapter 3, can be generalized to multiple dimensions, according to either of two methods:

- **Simultaneous confidence intervals** are defined using multiple confidence statements of the form $R_{\boldsymbol{u}}(\boldsymbol{\theta}) < \varrho_{\boldsymbol{u}}$, where $R_{\boldsymbol{u}}(\cdot)$ is a scalar value computed by applying some function (a "root" function) $R(\cdot)$ to a linear combination of parameters $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\theta}$. For each direction of interest $\boldsymbol{u}$, a limit $\varrho_{\boldsymbol{u}}$ is computed, and the resulting statements are asserted simultaneously. The number and distribution of directions of interest determine the shape of the interval—the set $\{\boldsymbol{u}\}$ may be finite, in which case the resulting region has straight sides, or infinite, in which case it is smooth. Bootstrap methods for obtaining $\varrho_{\boldsymbol{u}}$, using a Studentizing transformation for $R(\cdot)$, are detailed by Beran[2,3] for the general case, and adapted for the purposes of logistic regression by Lee.[4,5] The bootstrap method was shown to be better balanced than the older asymptotic alternatives of Scheffe, of Tukey and of Bonferroni, both in theory[2–4] and by Monte Carlo simulation.[5]

- **Likelihood-based confidence regions** are regions in parameter space which are bounded by a contour of constant likelihood (or log-likelihood) defined by $\ell(\boldsymbol{\theta}) = \varrho$ (thus, a point $\boldsymbol{\theta}$ is included in the region if $\ell(\boldsymbol{\theta}) \geq \varrho$). Asymptotically, as $N \to \infty$, likelihood distributions become normal (see Kendall and Stuart,[6] page 59ff.) so the asymptotic form of the confidence region is always symmetrical about the MLE in any given direction, its boundary being described by an ellipse or ellipsoid. The size, shape and orientation of the ellipse are determined by the parameter covariance matrix $\boldsymbol{V}$, which may be estimated as the inverse of the expected or observed Fisher information matrix or by other means. The method can easily be applied to the logistic or other regression estimator,[7] and has been used in the context of psychometric functions by Hawley.[8,9]

  However, the symmetry of the ellipses may be the major drawback of the asymptotic approach—as Efron and Tibshirani[10] warn, "the most

> serious errors made by standard intervals are due to their enforced symmetry" (page 180). This is potentially true when N is of the order of magnitude encountered in psychophysical experiments, as illustrated by Jennings[11] who noted and quantified the mismatch between normal-theory ellipses and actual likelihood contours in logistic regression for $N = 400$. A potentially good alternative is the bootstrap. Hall[12] describes a bootstrap approach to the generation of likelihood-based confidence regions, showing that the bootstrap-t method can produce asymptotically second-order correct region boundaries. The likelihood-based bootstrap approach will be adopted here.

The requirements for a good confidence region are similar to those for a good one-dimensional confidence interval. Besides the need for accuracy of the region's overall coverage level c, it is also desirable that the region be balanced (see section 3.1.1). In general terms, this means that the probability that the true parameter set $\boldsymbol{\theta}_{\mathrm{gen}}$ lies inside the region should be independent of the direction in which $\boldsymbol{\theta}_{\mathrm{gen}}$ lies relative to the estimate $\hat{\boldsymbol{\theta}}_0$, for all directions of interest. For simultaneous confidence intervals, balance simply means that the coverage of each component statement $R_{\boldsymbol{u}}(\boldsymbol{\theta}) < \varrho_{\boldsymbol{u}}$ is equal, and imbalance can be quantified by taking the difference between the measured coverage probabilities of two such components,[5] or the maximum difference between any two components.[3] For the likelihood-based regions used in the current chapter, the two-dimensional space of threshold and slope estimates will be divided into a limited number of sectors, and the rejection probabilities in each sector will be compared graphically.

As was the case for one-dimensional intervals, the coverage and balance of a confidence region should ideally be stable, in the sense of being independent of sampling scheme and of N. In order to test the stability of bootstrap confidence region methods, the same set of tests was applied as in chapter 3: each of seven sampling schemes was tested at four different values of N.

4.2 Methods

The general method is the same as that described in section 3.1.1: from the true psychometric function $\psi_{\text{gen}}(x; \boldsymbol{\theta}_{\text{gen}})$, $C = 500$ simulated data sets are generated, and a function is fitted to each one. Each fitted function is used to obtain a confidence region by one of four bootstrap methods described below, and the estimated coverage score $\hat{c}$ is equal to the observed proportion of occasions on which the generating parameter set $\boldsymbol{\theta}_{\text{gen}}$ lies within the confidence region.

The principle of the likelihood-based bootstrap method is to find an approximation to the likelihood function $\ell(\boldsymbol{\theta})$ and to obtain a bootstrap distribution of likelihood values $\ell^*_1 \ldots \ell^*_R$. The likelihood contour value ϱ for a region of target coverage c is then simply equal to the $(1-c)$ quantile of the distribution ℓ^*. Thus, the region contains the highest $c\,R$ bootstrap likelihood values. A point $\boldsymbol{\theta}$ lies inside the region iff $\ell(\boldsymbol{\theta}) \geq \varrho$, a condition which can be restated as

$$\text{CPE}\{\ell(\boldsymbol{\theta}); (\ell^*_1 \ldots \ell^*_R)\} \geq 1 - c. \tag{4.1}$$

Therefore, all that need be recorded for each simulated experiment is the cumulative probability estimate of $\ell(\boldsymbol{\theta}_{\text{gen}})$ in ℓ^*. Estimated coverage $\hat{c}$ is then equal to the proportion of such values that equal or exceed one minus the target coverage level c.

A number of different bootstrap methods can be applied within this framework, differing only in the way in which $\ell(\boldsymbol{\theta})$ is obtained. Computationally the simplest way is to evaluate the log-likelihood of equation (B.43) at each of the bootstrap parameter estimates, given the original data. Thus $\ell(\boldsymbol{\theta}; \boldsymbol{r}_0)$ is compared against the distribution of $\ell(\hat{\boldsymbol{\theta}}^*_1; \boldsymbol{r}_0) \ldots \ell(\hat{\boldsymbol{\theta}}^*_R; \boldsymbol{r}_0)$, where $\boldsymbol{r}_0$ is the observed data set. This is equivalent to a bootstrap version of the well-known likelihood ratio (or "deviance") method (see Davison and Hinkley, 1997,[13] page 234), and will therefore be referred to as the bootstrap deviance method in the following. It is also used in the fitting software as a basis for the "expanded" bootstrap method (see section 2.2.6). For the purposes of the expanded method it is useful in that the region reflects likely estimation error in all the parameters, including nuisance parameters. However, when the method is used directly to find a confidence region boundary for the purposes of hypothesis testing, while it has the desirable property of transformation-invariance, its inability to separate the dimensions of interest (α and β) from any nuisance parameters (γ and/or λ) is its principal disadvantage.[†]

Alternative methods estimate $\ell(\boldsymbol{\theta})$ using a smoothed density estimate of the distribution of bootstrap parameter sets. This approach has two advantages. First, density estimation can be carried out after the bootstrap points are "flattened" with respect to the dimensions of any nuisance parameters, so the resulting region reflects only variation in the parameters of interest. Second, the coordinates of the bootstrap points can be transformed in other ways before density estimation, allowing, in principle, any of the bootstrap confidence interval methods of section 2.2 to be adapted for two or more dimensions. The bootstrap percentile method, for example, uses the bootstrap parameter estimates $\hat{\boldsymbol{\theta}}^*_1 \ldots \hat{\boldsymbol{\theta}}^*_R$ without transformation. The basic bootstrap method reflects them about the initial estimate to obtain a distribution of $2\hat{\boldsymbol{\theta}}_0 - \hat{\boldsymbol{\theta}}^*_i$. The bootstrap-t method performs the reflection after Studentization using the estimated parameter covariance matrix $\hat{\boldsymbol{V}}$, so that the density estimate is based on $\hat{\boldsymbol{\theta}}_0 - \hat{\boldsymbol{V}}_0{}^{\frac{1}{2}} \hat{\boldsymbol{V}}^*_i{}^{-\frac{1}{2}} (\hat{\boldsymbol{\theta}}^*_i - \hat{\boldsymbol{\theta}}_0)$. This is very similar to the bootstrap-t method presented by Hall,[12] except that Hall's method performs the density estimation on the pivotal distribution of $N^{-\frac{1}{2}} \hat{\boldsymbol{V}}^*_i{}^{-\frac{1}{2}} (\hat{\boldsymbol{\theta}}^*_i - \hat{\boldsymbol{\theta}}_0)$, obtains a region boundary, and *then* transforms the boundary coordinates back into appropriately scaled parameter space with pre-multiplication by

[†] Note that, although a hypothesis test that uses the bootstrap deviance method makes a decision by comparing the deviance of the MLE parameter set, given the observed data, against a simulated distribution of deviance values, it should not be confused with a Monte Carlo goodness-of-fit test based on deviance, of the sort advocated by Wichmann and Hill.[14] The former compares $\ell(\boldsymbol{\theta}; \boldsymbol{r}_0)$ for some $\boldsymbol{\theta}$ against the bootstrap distribution of $\ell(\hat{\boldsymbol{\theta}}^*_i; \boldsymbol{r}_0)$, and the power of the test depends upon N. The latter is a relatively weak test for over- or under-dispersion that compares $\ell(\hat{\boldsymbol{\theta}}_0; \boldsymbol{r}_0)$ against the Monte Carlo distribution of $\ell(\hat{\boldsymbol{\theta}}^*_i; \boldsymbol{r}^*_i)$, and whose result does not depend on N, but rather on the explanatory accuracy of the model and on whether the variability of the process that generated the data was truly binomial around truly stationary generating probability values.

$N^{\frac{1}{2}}\hat{\boldsymbol{V}}_0{}^{\frac{1}{2}}$ and reflection about the initial estimate. Such a procedure is not followed here because the existence of the nuisance parameter λ (and, in the yes-no case, γ) means that pre-multiplication by $\hat{\boldsymbol{V}}_0{}^{\frac{1}{2}}$ is a 3- (or 4-) dimensional transformation, whereas we are only interested in defining a boundary in 2 dimensions. Transformation back into meaningful parameter space might compromise the contiguity of the two-dimensional projection of the region.

Density estimation is therefore carried out in actual (α, β) space, after transformation. A two-dimensional Gaussian kernel is used in order to obtain a smoothed density estimate, so that

$$\ell(\alpha, \beta) = g \sum_{i=1}^{R} \exp\left\{ -\frac{1}{2} \left[\left(\frac{\alpha - \hat{\alpha}_i^*}{h_\alpha} \right)^2 + \left(\frac{\beta - \hat{\beta}_i^*}{h_\beta} \right)^2 \right] \right\}, \tag{4.2}$$

where g is a constant which can be disregarded, and h_α and h_β are the smoothing parameters of the kernel. These are proportional to the standardizing denominator values for the distributions of $\hat{\boldsymbol{\alpha}}^*$ and $\hat{\boldsymbol{\beta}}^*$, respectively, for which the half-width of the central 68.3% non-parametric percentile interval will be used (see page 272). Thus $h_\alpha = h\ \textsc{wnpi}_{68}\{\hat{\boldsymbol{\alpha}}^*\}/2$ and $h_\beta = h\ \textsc{wnpi}_{68}\{\hat{\boldsymbol{\beta}}^*\}/2$. Hall,[12] operating on data that were assumed already to have been standardized by the Studentizing transform, recommends a value of h between 0.6 and 0.7 when $R = 2000$. Hall chose these values "by eye" in order to provide smooth and convex contours—they were a little larger than the value of 0.5 indicated by the cross-validation method of Bowman.[15] In pilot simulations for the current study, the value $h = 0.65$ did indeed seem to give smooth, convex results in a broad random selection of simulated experiments. Therefore, $h = 0.65$ was chosen.

The basic bootstrap, bootstrap-t, bootstrap percentile and bootstrap deviance methods were all tested using a 2-AFC logistic function as the generating psychometric function, with $\alpha_{\text{gen}} = 2.737$ and $\beta_{\text{gen}} = 0.494$. Realistic experimental conditions were assumed—thus the observer's lapse rate was assumed to be unknown, and the true value λ_{gen} was set at 0.01. Each of the seven 2-AFC sampling schemes from section 1.5.1 was tested at $N = 120$, 240, 480 and 960. Regions of target coverage 68.3% and 95.4% were measured, based on $R = 1999$ bootstrap simulations on each of $C = 500$ repetitions.

4.3 Results

4.3.1 Coverage

The graphical conventions of section 3.1.3 are used in figure 4.1 to show the overall coverage of the four confidence region methods: the shape of each symbol corresponds to one of the seven sampling schemes of figure 1.2, and the size of each symbol corresponds to the total number of trials N. Coverage estimates $\hat{c}$ are transformed using the Freeman-Tukey transform of section 3.1.2, assuming $C = 500$.[†] The standard error of the transformed value $\vartheta_{500}(\hat{c})$ is approximately 0.045 at all values of $\hat{c}$ (except $\hat{c} = 1$, which does not occur at all in the results shown in figure 4.1).

Lighter symbols show the coverage of regions whose target coverage was 68.3%, and darker symbols show the results for regions of 95.4% target coverage. The two target levels are indicated by the red broken lines. The bootstrap deviance method performs particularly well, if a little too conservatively: the coverage of the region exceeds both target levels for nearly all the tests, and the points are fairly well grouped together, indicating that the coverage of the region is not greatly susceptible to variation due to different sampling schemes and values of N. The other methods perform less well: their coverage estimates are more spread out from one sampling scheme or value of N to another, and the general tendency is for coverage to be too low. The performance of the bootstrap-t method at 95.4% is a promising exception, however—coverage is more accurate, with relatively little variation.

Table 4.1 compares the overall coverage of the one- and two-dimensional version of each method. The one-dimensional results are taken from the set

[†] The maximum number of invalid results on any of the 28 tests was 6 out of 500, which corresponds to a fractional increase in asymptotic standard error of 6%, and a negligible perturbation of the variance-stabilizing effect of the transform.

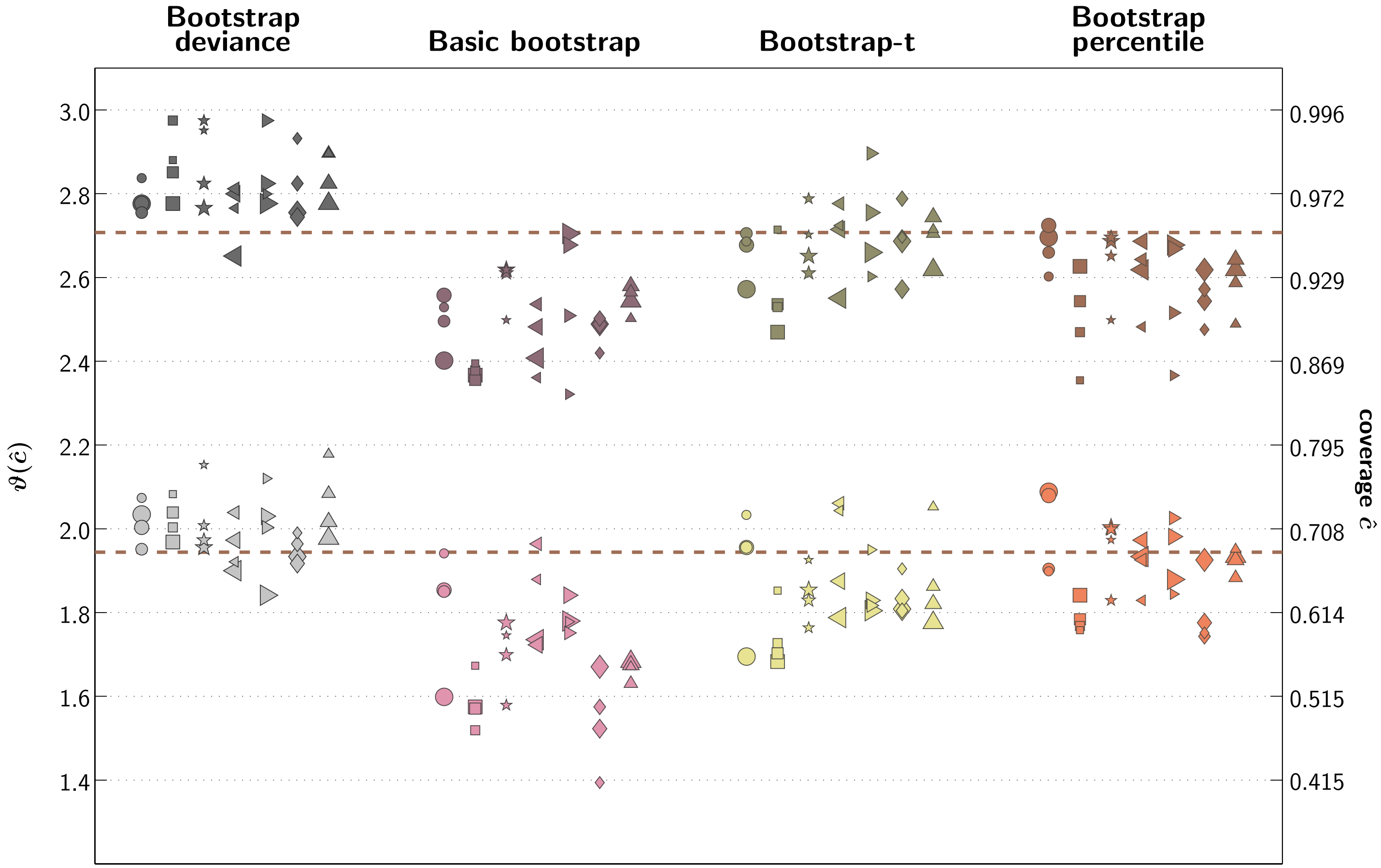


Fig. 4.1: Results of Monte Carlo coverage tests for 68.3% and 95.4% joint confidence regions obtained from the logistic function in the realistic 2-AFC case, using four different likelihood-based bootstrap methods. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 4.3.1 for details.

of logistic simulations mentioned in section 3.3,[†] which is stored in the archive under:

- simulations/coverage/2AFC/l01/logistic/with_bootstrap_t/ .

The current two-dimensional simulation results are stored under:

- simulations/coverage/2AFC/l01/logistic/regions/ .

The table shows that neither the basic bootstrap nor the bootstrap percentile method performs better in two dimensions than it did in the dimensions of threshold and slope separately. For the bootstrap percentile method, coverage is lower, less accurate and less stable. The basic bootstrap method may be very slightly more stable in its two-dimensional form, but the mean coverage level is lower and further away from its target. The bootstrap-t method, on the other hand, shows great improvement at 95.4% relative to its one-dimensional performance: the target coverage probability is attained with greater accuracy and greater precision when the two-dimensional method is used than when either dimension is considered separately.

4.3.2 Imbalance

Besides the requirement of accurate overall coverage, there is also the question of a confidence region method's *balance*. Figures 4.2–4.5 show results separately for each of the four region methods, and in order to given an idea of the magnitude and direction of any imbalance in the regions, the space in which threshold and slope can vary is divided into sectors. Threshold is on the horizontal, and slope is on the vertical axis of each figure, as indicated by the "compass" in the centre. The sectors are the two-dimensional equivalent of the "tails" of a two-tailed test, and the ideal result is that all the sectors have equal unconditional false-rejection probability. For each of the 500 estimated parameter sets, a position vector consisting of a threshold and slope

[†] These logistic simulation results were not plotted in section 3.3, because they were very similar to the Weibull results already shown in figures 3.7 and 3.8. The logistic results are used here, however, in order to provide the fairest possible comparison with the region method simulations, which also used the logistic function.

	Basic bootstrap	Bootstrap-t	Bootstrap percentile
(68.3% target)	1.94	1.94	1.94
1-D thresholds	1.93 ± 0.14	1.77 ± 0.12	1.92 ± 0.06
1-D slopes	2.04 ± 0.15	1.74 ± 0.07	1.91 ± 0.08
2-D regions	1.70 ± 0.13	1.86 ± 0.11	1.90 ± 0.10
(95.4% target)	2.71	2.71	2.71
1-D thresholds	2.70 ± 0.17	2.46 ± 0.18	2.68 ± 0.06
1-D slopes	2.56 ± 0.12	2.28 ± 0.17	2.60 ± 0.10
2-D regions	2.50 ± 0.10	2.67 ± 0.09	2.59 ± 0.10

Table 4.1: For three different bootstrap methods at two different target coverage levels, the accuracy and precision of overall coverage is compared between one-dimensional and two-dimensional cases. For each target coverage probability, the corresponding Freeman-Tukey-transformed value is given, followed on successive rows by the mean ± the standard deviation of the group of 28 transformed coverage estimates from one-dimensional threshold confidence intervals, one-dimensional slope confidence intervals, and two-dimensional confidence regions.

value, $(\hat{t}_{0.5}, \hat{s}_{0.5})_i$, was computed. The vector was then subtracted from the vector corresponding to the true psychometric function, $(t_{0.5}, s_{0.5})_{\mathrm{gen}}$, to obtain the difference vector $(\Delta t, \Delta s)_i$ which indicates the direction in which the true values lie relative to the estimate. The distribution of Δt's was then standardized by dividing by $\frac{1}{2}\textsc{wnpi}_{68}$, (see page 272) and the distribution of Δs's was likewise standardized by its own $\frac{1}{2}\textsc{wnpi}_{68}$. The simulations were then binned into 8 sectors of equal angular width, according to the angles φ_i of their standardized difference vectors,[†] and the observed unconditional coverage probability in each of the sectors is transformed by the Freeman-Tukey function and plotted radially. Effectively, the format is analogous to that of figures 3.15 and 3.16 in that the estimate $\hat{\boldsymbol{\theta}}_0$ can be imagined in the centre of the figure, with coverage probability spreading outwards. Results for 68.3% regions are shown, which means that the target coverage probability in each sector is 0.960. For all the region methods, the pattern of results for 95.4% intervals was found to be very similar to that for 68.3%. The 95.4% results are less informative, however, because the very small expected number of false rejections in each sector (2.8 as opposed to 19.8) meant that, very often, none were observed. Darkened symbols indicate the cases in which no false rejections were observed in a given sector ($\hat{c} = 1$).

All the methods show the same general trend: an imbalance between the upper and lower sectors. Coverage tends to be too high when the true slope value lies above the estimate, and too low when it lies below. The imbalance is most pronounced for the percentile method, and least pronounced for the deviance method, with the basic bootstrap and bootstrap-t being intermediate, and very similar to one another. Note that, despite its lower overall coverage, the basic bootstrap method may not actually be any worse than the bootstrap-t: in those sectors where coverage is too low, the two methods

[†] Although it is a somewhat imprecise approach to compute the region in (α, β) space and then report the balance results in $(t_{0.5}, s_{0.5})$ space, the latter representation was chosen because, as previously discussed in section 1.2.2, it provides a standard means of expression that does not rely on the chosen mathematical form of the psychometric function, and which also allows more intuitively straightforward comparison with the results of chapter 3. When the angle φ is computed from α and β instead of $t_{0.5}$ and $s_{0.5}$, the results are appear qualitatively indistinguishable from those presented.

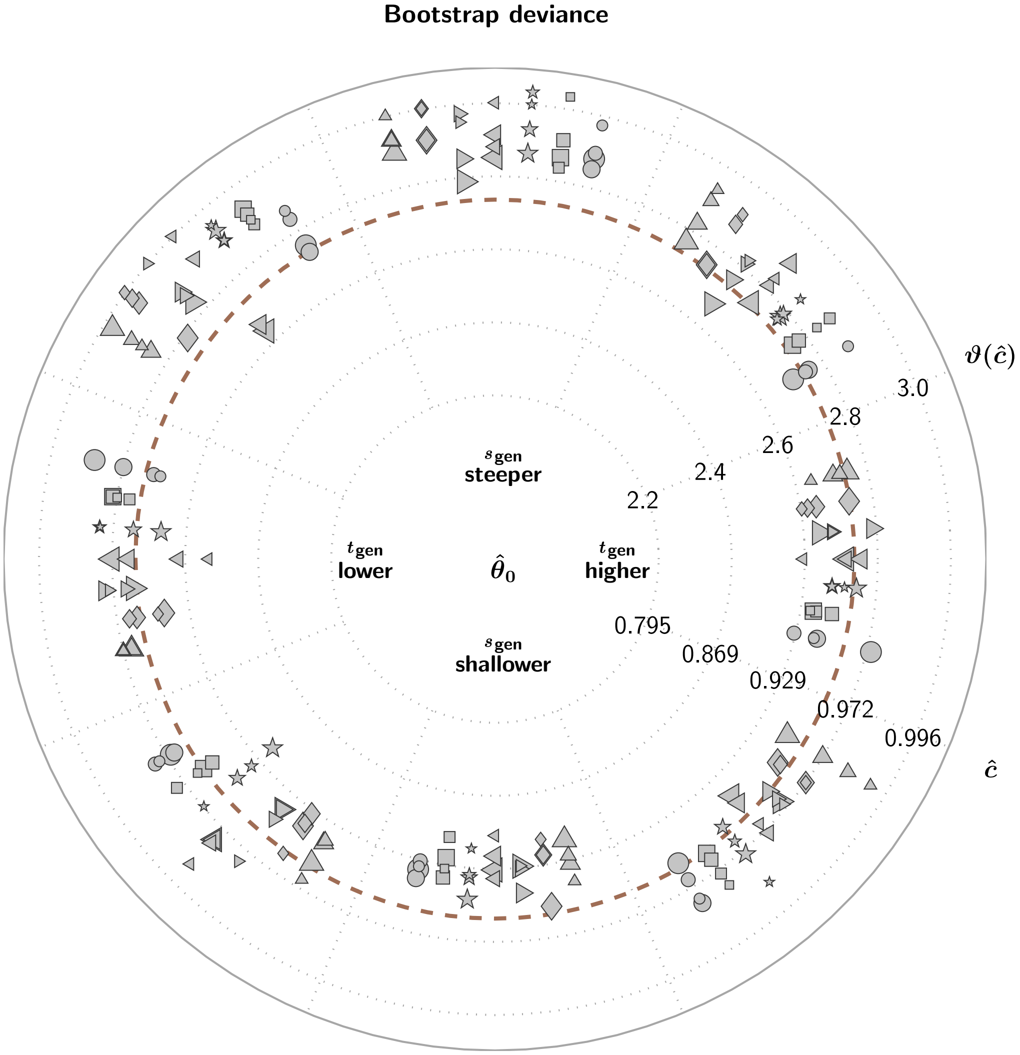


Fig. 4.2: Results of a Monte Carlo coverage test for a 68.3% likelihood-based joint confidence region obtained by the bootstrap deviance method. Rejection probabilities for each of 8 sectors is plotted radially to indicate the balance of the region's coverage with regard to the direction of the estimation error. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 4.3.2 for details.

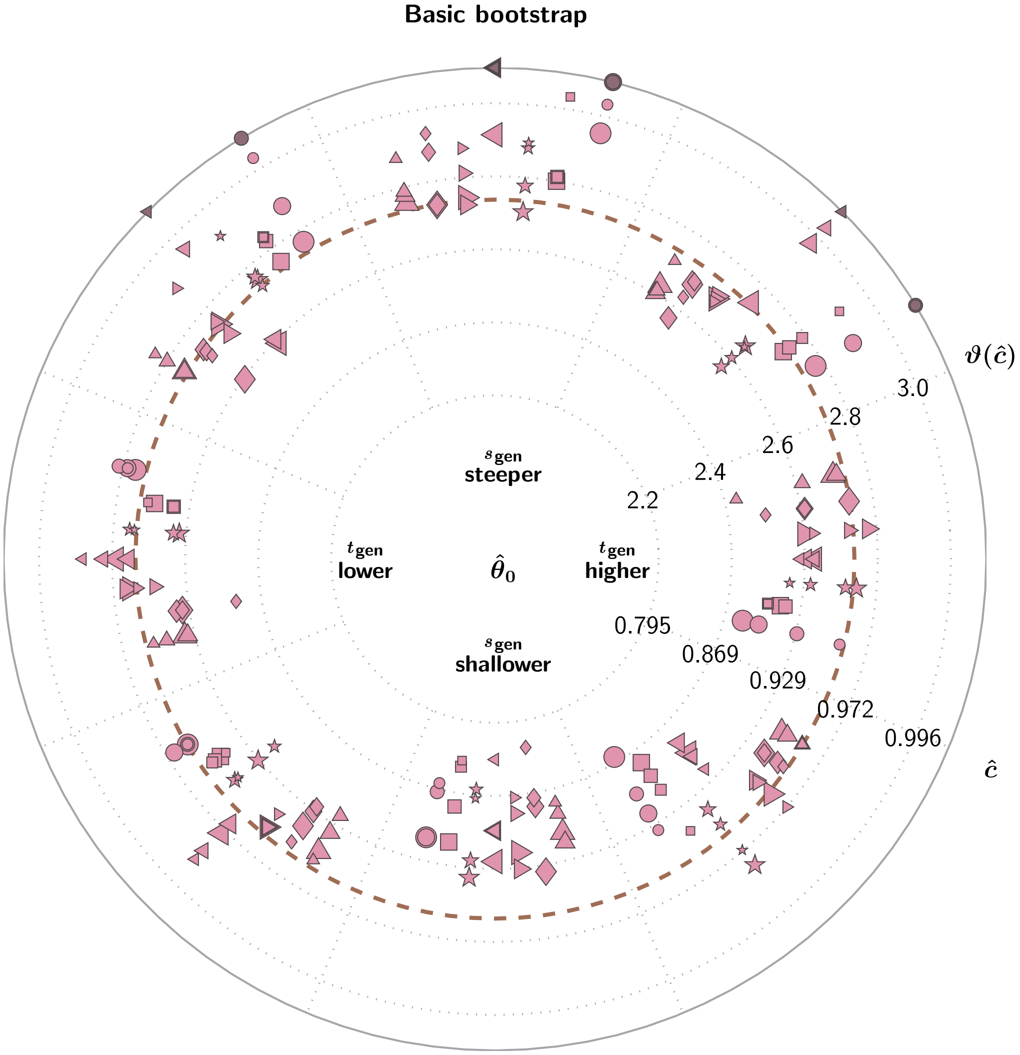


Fig. 4.3: Results of a Monte Carlo coverage test for a 68.3% likelihood-based joint confidence region obtained by the basic bootstrap method. Rejection probabilities for each of 8 sectors is plotted radially to indicate the balance of the region's coverage with regard to the direction of the estimation error. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 4.3.2 for details.

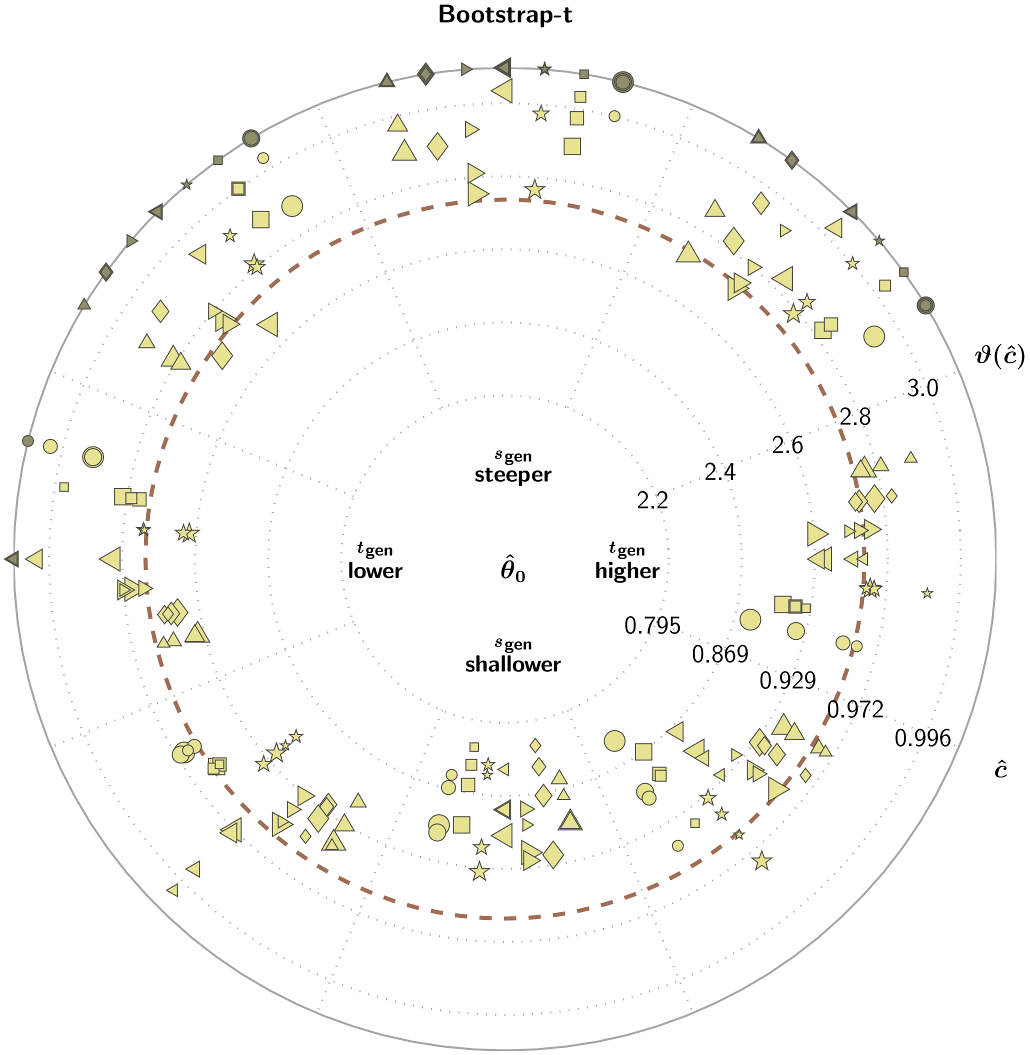


Fig. 4.4: Results of a Monte Carlo coverage test for a 68.3% likelihood-based joint confidence region obtained by the bootstrap-t method. Rejection probabilities for each of 8 sectors is plotted radially to indicate the balance of the region's coverage with regard to the direction of the estimation error. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 4.3.2 for details.

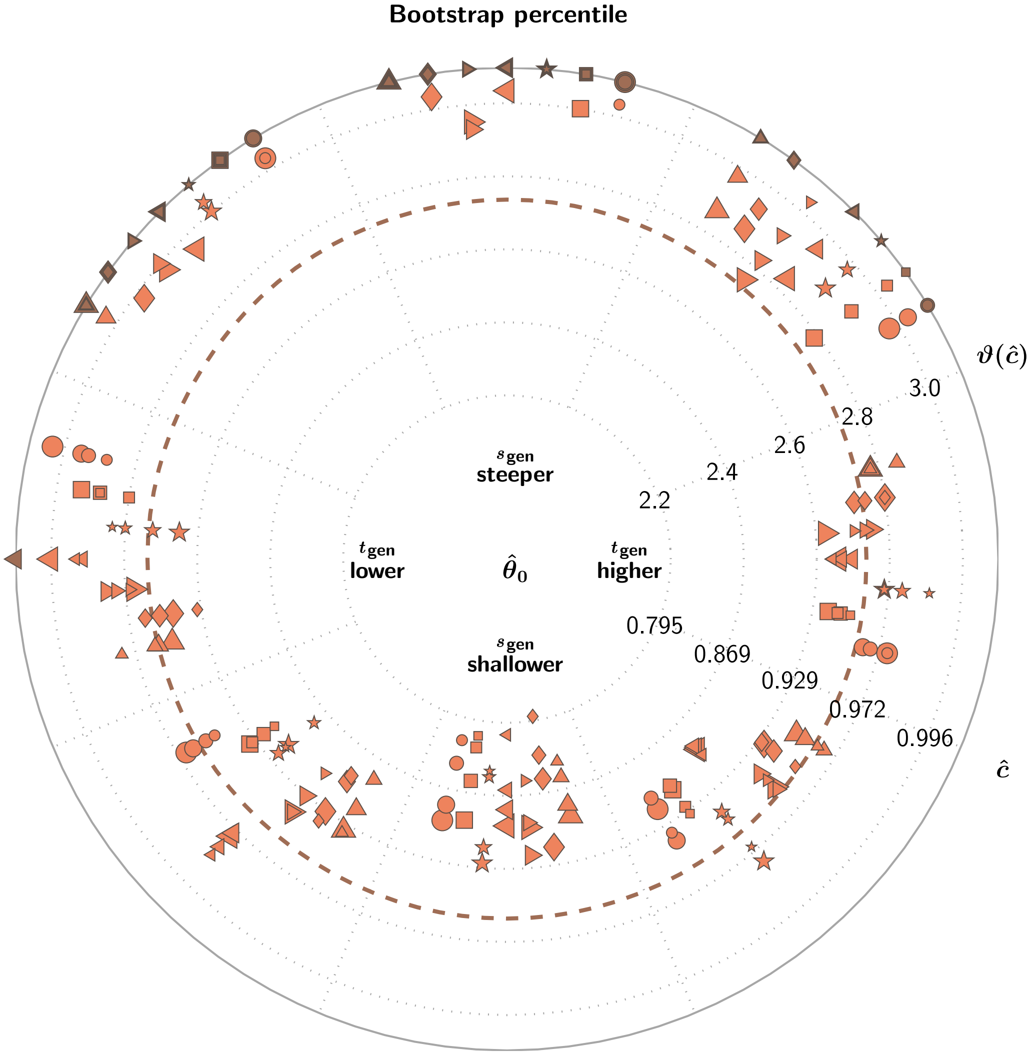


Fig. 4.5: Results of a Monte Carlo coverage test for a 68.3% likelihood-based joint confidence region obtained by the bootstrap percentile method. Rejection probabilities for each of 8 sectors is plotted radially to indicate the balance of the region's coverage with regard to the direction of the estimation error. Symbol shapes denote the seven sampling schemes of figure 1.2. See section 4.3.2 for details.

are very similar, but where the bootstrap-t brings the overall coverage level back up by over-covering in other sectors, the basic bootstrap is simply more accurate. Using either method, the experimenter would have to remember that the confidence region probably does not extend far enough towards low slope values. Using the bootstrap-t, it is also probably the case that the region extends too far up towards higher slope values.

4.3.3 Bootstrap deviance and experimental design

So far only the realistic 2-AFC case has been considered. Being the experimental context of principal interest in the current research, and the one in which statistical inference in one dimension has thus far proved most difficult, only this case was selected in order to test the bootstrap methods that relied on density estimation. The bootstrap deviance method, is very easy to perform if a bootstrap distribution of parameter estimates has already been obtained. It was therefore practicable to combine a test of the bootstrap deviance method with each of the sets of simulations reported in chapter 3.

Figure 4.6 shows overall coverage for four of the sets, reflecting four different combinations of experimental design and experimental assumptions: ideal yes-no (γ and λ both fixed at 0); realistic yes-no ($\gamma_{\text{gen}} = 0.02$, $\lambda_{\text{gen}} = 0.01$, must be estimated); ideal 2-AFC (λ fixed at 0); and realistic 2-AFC ($\lambda_{\text{gen}} = 0.01$, must be estimated). All four sets used the cumulative normal psychometric function. Some of their one-dimensional results have already been reported (section 3.4). They are stored as:

- `simulations/coverage/yesno/g0f0l0f0/cumnorm/with_expanded0125/`
- `simulations/coverage/yesno/g02l01/cumnorm/with_expanded0125/`
- `simulations/coverage/2AFC/l0f0/cumnorm/`
- `simulations/coverage/2AFC/l01/cumnorm/`

There is little noticeable difference between results for the four cases. Between idealized and real cases there is a slight decrease in stability, but the direction of the trend for individual results is generally upwards—in other words, the addition of unknown nuisance parameters tends to make the test slightly more conservative. The effect is very small, however. Differences in balance between the four cases (not shown) were also small—all four cases produced a pattern very similar to that of figure 4.2. This is perhaps surprising, given the rather marked effect on one-dimensional slope coverage observed in chapter 3 when nuisance parameters were added. However, it should be borne in mind that such effects were primarily due to the fact that λ and γ are correlated with slope, and are themselves difficult to estimate accurately or precisely, leading to bias and imprecision in the slope estimates. The bootstrap deviance method, being based on the single dimension of likelihood without any attempt to remove λ and γ from the hypothesis test, does not separate the parameters and thus largely avoids such problems. Naturally this limits its use somewhat in the realistic cases, because hypotheses cannot be formulated that are wholly independent of the nuisance parameter values.

4.4 Summary and concluding remarks

For the realistic 2-AFC case, the bootstrap deviance method provides consistent and conservative confidence regions, and is well balanced with respect to the dimensions of threshold and slope. It is therefore highly suitable for the purposes of the expanded bootstrap method (section 2.2.6), in which it is used to estimate the likely error in all the estimated parameters. When it is used as a confidence region method in its own right, the bootstrap deviance method shows excellent coverage properties for both yes-no and 2-AFC experiments in the idealized case (i.e. when there are no nuisance parameters), and unlike one-dimensional methods, its performance was not seriously degraded in 2-AFC relative to yes-no. Under more realistic assumptions, the overall coverage and balance of the method are still good, but great care must be taken that an apparently significant difference between psychometric functions is not simply due to a difference in the estimated nuisance parameters

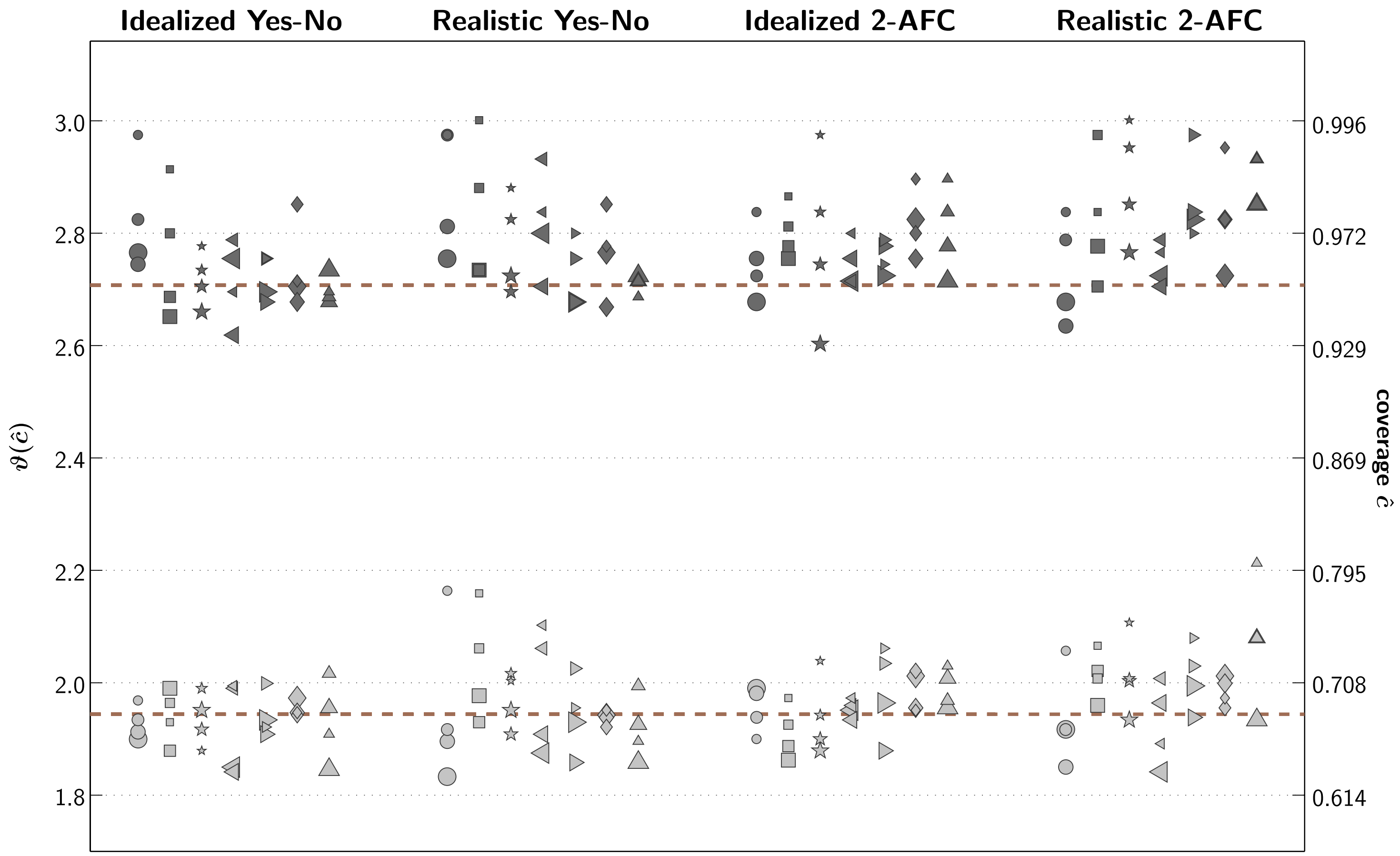


Fig. 4.6: Results of Monte Carlo coverage tests for 68.3% and 95.4% bootstrap deviance confidence regions obtained from the cumulative normal function in the idealized yes-no, realistic yes-no, idealized 2-AFC and realistic 2-AFC cases. Symbol shapes denote the sampling schemes of figures 1.3 (yes-no) and 1.2 (2-AFC). See section 4.3.3 for details.

γ and/or λ. A potentially valuable future development of the method might be to compute the coordinates of the region boundary explicitly, and then "flatten" them with respect to the dimensions of γ and λ.

Three other likelihood-based bootstrap methods were assessed in the realistic 2-AFC case: the basic, bootstrap-t and percentile methods. All three used non-parametric density estimation in the α–β plane, and thus could potentially separate the effects of threshold and slope differences from the effects of nuisance parameters. The bootstrap-t was a modified form of the method presented by Hall,[12] which he demonstrated to have better asymptotic coverage properties than the basic bootstrap method, for general purposes. In the current application, its overall coverage was indeed better than that of the other two methods, appearing if anything slightly too conservative. However, its apparent superiority to the basic bootstrap proved to be because its under-coverage of shallow slope values was compensated by over-coverage of steep slopes.

The general trend towards positive imbalance with regard to slopes was shared by all four methods studied, to a greater or lesser extent. This probably reflects the general tendency of psychometric function slopes to be overestimated (see section 5.1.1), which is exacerbated in the realistic 2-AFC case because of the difficulty of obtaining an accurate estimate of λ when $\lambda_{\text{gen}} = 0.01$ (see also sections 3.3.4 and 5.5.5).

Nevertheless, the two-dimensional version of the bootstrap-t method performed better than the one-dimensional version, being less susceptible to variation in coverage between different sampling schemes and values of N. It would be interesting to know whether further improvements in performance of the bootstrap-t might be possible, using different (non-parametric) methods to estimate $\boldsymbol{V}^*$ rather than the asymptotic Fisher approximation. It would also be interesting to adapt the BC_a method to the problem of generating two-dimensional confidence regions.

References for chapter 4

[1] Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

[2] Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, **83**(403): 679–686.

[3] Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, **85**(410): 417–426.

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

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

[6] Kendall, M. K. & Stuart, A. (1979). *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.* New York: Macmillan, fourth edition.

[7] Chang, Y. C. I. & Martinsek, A. T. (1992). Fixed size confidence regions for parameters of a logistic regression model. *Annals of Statistics*, **20**(4): 1953–1969.

[8] Hawley, M. L. (1990). Comparison of adaptive procedures for obtaining psychophysical thresholds using computer simulation. Master's thesis, Boston University.

[9] Hawley, M. L. & Colburn, H. S. (1995). Application of confidence intervals and joint confidence regions to the estimation of psychometric functions. *Journal of the Acoustical Society of America*, **97**: 3277.

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

[11] Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, **81**(394): 471–476.

[12] Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**(3): 481–493.

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

[14] 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 .

[15] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**: 353–360.

5. Towards optimal sampling of the 2-AFC psychometric function

The bias and precision of one's estimator of a threshold or slope are affected significantly by the "sampling scheme" one uses, that is to say the particular arrangement of sample points on the stimulus axis, and the performance levels at which those samples consequently strike the psychometric function. This fact is what motivated the development of adaptive psychophysical procedures, which aim to place stimuli at points which maximize the efficiency of threshold estimation, but it is equally true of the set of stimulus values used in a "method of constant stimuli" or block-design experiment. This chapter will define the relevant measures of efficiency by which a block-design experiment can be assessed, and explore the ways in which such measures are affected by sampling scheme.

Several studies have used Monte Carlo simulation to explore the way in which the placement of trials may affect the bias and efficiency of threshold and slope estimation. Lam and colleagues[1–3] investigated the differences in sampling precision for threshold and slope estimates that depend on sampling placement, for a limited set of 2-, 3- and 4-point sampling schemes. They found[2] that careful placement of 4 blocks of trials yielded more efficient slope estimates than a more conventional even spread of 11 blocks, for a comparable total number of trials, and also[3] that the optimal spread of stimulus values changed according to the number of blocks k, the total number of trials N and the experimental design (yes-no, 2-AFC, 3-AFC or 4-AFC). Teller[4] and McKee, Klein and Teller[5] examined the effect of the mid-point and extent of the stimulus range on the efficiency of threshold estimation for evenly spaced 2- to 5-point schemes. Wichmann and Hill[6] examined the seven hand-picked 6-point sampling schemes shown in figure 1.2, and demonstrated significant differences among them with regard to the efficiency of estimation of both thresholds and slopes. Some of the differences they found depended on aspects of the *uneven* distribution of stimulus values in the schemes, which goes against the common assumption, expressed by Watson and Fitzhugh,[7] that "the method of constant stimuli has only two important parameters... the number of sample points... [and] the step in strength between sample points."

Wichmann and Hill's results will be replicated and reported in a revised and extended form in section 5.4. A wider range of sampling schemes will then be explored in section 5.5, with the aim of expressing the effect of stimulus distribution independently of the effects of the number of blocks k and the total number of observations N. The results should be useful in providing generalized rules, and guidelines for building suitable algorithms, for efficient stimulus placement.

5.1 Criteria for scoring sampling schemes

The way in which a sampling scheme is evaluated depends on which particular measures are of interest: the particular detection levels at which thresholds are measured, and the relative importance of threshold and slope measurements (see section 1.2.2). In the simulations of this chapter, thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$, and the slope $s_{0.5}$ were measured. Generally, only results for $t_{0.5}$ and $s_{0.5}$ will be reported. Results will be reported separately for threshold and slope confidence intervals, without attempting to combine the two scores according to any assumptions about their relative importance.

For every estimate of interest, an experimenter will want to know both how accurately a given sampling scheme estimates the correct value, and how precisely or efficiently. Definitions of bias and efficiency are given in the following subsections.

5.1.1 Bias

Bias is often reported in terms of the numerical difference between the expected value of a simulated estimate and the true underlying value. It should be noted, however, that the magnitude of the bias in one's estimate of a quantity u is only meaningful in terms of the variability of that same estimator (or in terms of other estimators that the experimenter will use to obtain values for comparison with u). A small numerical mean or median bias might still be unacceptable if the estimator's variability is even smaller, as this would lead to hypothesis-testing errors either of type I (finding apparently significant differences between conditions where none exist) or type II (failing to find a true difference between experimental conditions).

"Bias", in the results of this chapter, will therefore be defined in the sense of equation (2.14): the probability (estimated by Monte Carlo simulation) of obtaining an estimate less than or equal to the true value is passed through the inverse of a cumulative normal function with zero mean and unit variance, so that the bias score w is expressed as the equivalent number of standard deviations of the standard normal distribution. Efron and Tibshirani[8] recommend, as a rough guide, that a good estimator should be biased by no more than a quarter of its own standard deviation (hence $-0.25 \leq w \leq +0.25$). A negative value of w indicates over-estimation, and a positive value indicates under-estimation.

O'Regan and Humbert,[9] examining the logistic function, report that maximum-likelihood estimation can often produce biased estimates of psychometric function slope, particularly when the number of trials is small. They note that slope bias is greater in 2-AFC than in yes-no (an effect also noted in the adaptive procedure study of Leek, Hanna and Marshall[10]), and also report biases in the estimation of threshold.[†] Swanson and Birch[11] and Maloney[12] also note the tendency to overestimate the β parameter of a 2-AFC

[†] The biases found by O'Regan and Humbert were reported to be "significant" and indeed they were, in the sense that they were significantly unlikely to have occurred purely because of the inherent randomness of Monte Carlo simulation. However, for most of the sampling schemes they tested, the bias was of small magnitude relative to the estimator's own variability.

Weibull function, the latter reporting in addition that the Monte Carlo distribution of β is noticeably skewed when N is below about 500. The added difficulty associated with the accurate estimation of λ (see sections 1.2.3 and 3.3.4) can be expected to exacerbate the problem of slope bias.

The relationship between w and N will be of interest. Some biased sampling schemes may be acceptable for small N, where variability is high, but they may be significantly biased at larger N if variability decreases relative to the magnitude of the error of estimation. For others, the magnitude of error might decrease along with variability, providing a stable estimate for all N.

5.1.2 Efficiency

Efficiency can be defined as precision per unit effort. Taylor and Creelman[13] defined "sweat factor" $K = N\sigma^2$ as a metric for comparing the efficiency of sampling strategies, once a measure of variability σ has been obtained by simulation (in their paper, σ is the standard error of a threshold estimate). The relative efficiency of two estimators is simply the ratio of their sweat factors, usually expressed as a percentage. Taylor and Creelman further define the *ideal* sweat factor $K_{\min}$, which provides the numerator for the calculation of an absolute measure of the efficiency of threshold estimation performance of a sampling scheme S (which coincides with the standard definition of efficiency):

$$\begin{aligned} \text{efficiency}_S &= 100\,\frac{K_{\min}}{K_S}, && (5.1)\\ \text{where } K_{\min} &= \frac{p(1-p)}{[\mathrm{d}p/\mathrm{d}x]^2}, && (5.2) \end{aligned}$$

and where p is the performance value $\psi(x)$, and $\mathrm{d}p/\mathrm{d}x$ the derivative $\dot{\psi}(x)$, evaluated at the threshold $x = t$. The derivation of $K_{\min}$ for thresholds effectively assumes that all N trials are positioned at exactly the correct threshold value—under which assumption the above expression for $K_{\min}$ can be obtained from equation (2.2). Thus it represents an unattainable lower bound on K, and all measured efficiencies will be less than 100%. Taylor[14] notes that $K_{\min}$ is also the asymptotic sweat factor of a realizable process (the Robbins-Monro[15] stochastic approximation method) and is thus a *greatest* lower bound for K (representing, in Taylor's words, "not just an upper bound, but a least upper bound on the performance of realizable techniques".)

The computation of an ideal sweat factor for slopes is less straightforward, as the ideal approach is clearly not to place all one's trials at the same stimulus value. In equation (2.3), variance is minimized by maximizing the weighted variance of stimulus values about their weighted mean, where the weighting for each point depends on its predicted performance value and performance gradient. The exact spacing of points for ideal slope measurement therefore depends on the functional form $F(x)$ one chooses for the psychometric function. A working value of $K_{\min}$ for slopes will be obtained using a heuristic method, following the two-point approach of Wetherill[16] and O'Regan and Humbert[9]—see section 5.3.1.

Using the normal-theory assumption that variance decreases linearly with N, the sweat factor metric is useful because, in theory, it will be constant for a given sampling scheme, independent of N. As with all normal-theory approximations, however, we cannot guarantee that its assumptions hold sufficiently accurately for real psychophysical data. In fact, as we shall see, there are differences, depending on one's sampling scheme, not only in the coefficient that relates the widths of confidence intervals to N, but also in the exponent of N (where normal theory would predict an exponent of -0.5) which means that the efficiency score of equation (5.1) is not invariant with respect to N. So, the relative efficiency of two sampling schemes might change according to how many observations are taken, one being better for small N and another better for large N. This can be observed even using asymptotic methods to measure K, and was illustrated by Finney[17] for thresholds in the yes-no context. His table 8.1 lists values of a metric similar to sweat factor, for a number of 2- and 3-point sampling schemes, and shows that narrower sampling schemes are favoured, relative to wider schemes, to a greater extent at higher than at lower values of N.

5.2 Aims

The current chapter aims to investigate the same issue as that explored by Finney's[17] table 8.1, *viz.* the way in which the relative merits of different sampling schemes change with k and with N. The 2-AFC experimental situation will be considered here, however, whereas the situation examined by Finney was equivalent to a yes-no experiment in which it is assumed that $\gamma = \lambda = 0$.

The 2-AFC psychometric function has received relatively little attention in the statistical literature. This is unfortunate, given its wide use in psychophysical research, and in any case the 2-AFC experimental design is particularly interesting from the statistical point of view, because it poses a more subtle question of sample placement than the yes-no design due to its asymmetric nature. Expected binomial variability decreases from the bottom of the function to the top, and so it is often more efficient, and "safer" in terms of susceptibility to bootstrap error, to place one's sample points with a shift towards higher performance levels. Given this, exactly how far is it safe to go with such a shift? To what extent are improved efficiency and reduced sensitivity to bootstrap error offset by the risk of estimation bias that may arise due to the asymmetry of one's samples?

Teller and colleagues[4,5] have examined in some detail the effect of sampling scheme on threshold interval length in the idealized (known λ) 2-AFC situation, concentrating on $N \leq 150$. Both papers categorize sampling schemes according to the number of points k, the width of the stimulus range $x_{\max} - x_{\min}$, and the mean stimulus value $\bar{x}$. For both probit intervals[4] and Monte Carlo percentile intervals[5] it is reported that a wider stimulus range is "safer" in that it tends equalizes the expected confidence interval widths at different values of $\bar{x}$ (if the stimulus values are tightly grouped, then they must be closer to the threshold to avoid the width of the interval becoming very large). Teller[4] also reports a main effect of range on interval width, although McKee, Klein and Teller[5] report that the effect is small if not insignificant over the set of range values studied, when Monte Carlo intervals rather than probit intervals are considered. Probit intervals were generally found to be inaccurate at smaller values of N (they were too large, particularly on the lower side, when compared with the Monte Carlo intervals). However, McKee *et al.* state that probit intervals can generally be considered approximately correct for $N \geq 100$. This is somewhat at odds with the findings in chapter 3, where probit intervals were found to be inaccurate in the 2-AFC case even at higher values of N—they were unbalanced, with too few rejections in the lower tail.[†]

The current chapter aims to explore the issues of sampling scheme bias and efficiency in greater detail, taking a greater range of values of k and including larger values of N, such as are more regularly used in adult 2-AFC psychophysics ($120 \leq N \leq 960$). It aims to obtain bias and efficiency measures for slopes as well as for thresholds, and also aims to test the difference between results from the Monte Carlo approach and the bootstrap approach (see section 5.3.2, below). Results from chapter 3 indicate that when one takes the bootstrap approach, $\mathrm{BC_a}$ intervals are more accurate than either probit intervals or intervals based on unadjusted Monte Carlo percentiles in the 2-AFC case, particularly for intervals whose coverage is of the order of 95%, which is the coverage level considered by Finney,[17] Teller[4] and McKee *et al.*[5] Therefore the relative efficiency of slope estimation of different sampling schemes may differ according to whether one takes the $\mathrm{BC_a}$ results or the Monte Carlo results, the $\mathrm{BC_a}$ results being the more relevant.

Simulations will use psychometric functions whose upper asymptote is slightly less than 1.0, and must be estimated—this more realistic situation is more interesting to the psychophysicist, because it allows for the inevitable tendency of observers to "lapse". Since the maximum-likelihood estimates of slope and of the lapse-rate parameter λ co-vary, it is to be expected that

[†] One potential reason for the discrepancy is that a coverage test tends to involve somewhat tougher tests of a confidence interval method than a Monte Carlo test of interval length, in that poorer sampling schemes may arise in the former than are usually chosen for testing in the latter. A related reason is that coverage accuracy not only relies on the correctness of the interval bounds, but the relationship between interval bound correctness and the error of one's threshold estimate—see the footnote on page 119.

the inclusion of λ as a (constrained) free parameter will affect the relative efficiency of different sampling schemes with regard to slope.

A wider range of sampling schemes will be explored than those considered by Teller and her colleagues.[4,5] Values of k up to 12 will be considered, which poses the challenge of categorizing sampling schemes in a meaningful way—as each of k points may be positioned independently, sampling schemes may theoretically vary in k dimensions. The two simple dimensions of stimulus range and mean stimulus value considered by Teller and colleagues may be inadequate to explain some of the differences between sampling schemes. Note in particular the results of section 5.4, in which Wichmann's 7 sampling schemes (defined in section 1.5.1) are tested: despite having similar mean stimulus values and stimulus range, slight differences in the distribution of stimulus values within that range yield qualitatively and significantly different results for the sampling schemes ★, ⧫ and ▲. After consideration of the seven illustrative examples in section 5.4, a more thorough investigation of sampling schemes will be conducted.

5.3 Methods

5.3.1 Simulation

To test a given sampling scheme, a psychometric function shape F_{gen} is chosen, along with a set of generating parameters $\boldsymbol{\theta}_{\text{gen}}$. The Weibull function will be used here, with parameters $\alpha = 3, \beta = 4, \gamma = 0.5, \lambda = 0.01$.[†] The sampling scheme is used to determine the vector of stimulus values $\boldsymbol{x}$ and the vector of block sizes $\boldsymbol{n}$. Simulated data sets are then generated from the curve—in each data set, the number of correct responses at stimulus value x_i is drawn from the binomial distribution $\text{Bi}\left[n_i, \psi(x_i; \boldsymbol{\theta}_{\text{gen}})\right]$. A curve is fitted

[†] The parameters $\alpha = 3$ and $\beta = 4$ were chosen, here and in section 3.3, purely because they provided a good fit to a particular set of masked grating detection data that was often used as an example when testing the software. Pilot simulations indicated that a change in parameter values, and even a change in psychometric function shape, had no appreciable effect on the bias or efficiency results (see also page 213).

to the simulated data, using the same psychometric function shape F_{gen} to obtain parameter estimates $\hat{\boldsymbol{\theta}}^*$, from which threshold and slope values are calculated. The process is repeated R times ($R = 1999$ for the purposes of this chapter), yielding distributions of simulated threshold and slope values t^* and s^*, from which confidence intervals of coverage 68.3% and 95.4% are computed.

Confidence interval widths are then scaled to be applicable to a psychometric function whose slope is equal to 1 at $F(x) = 0.5$. The variance of a slope estimate is expected to increase in proportion to ${s_{0.5}}^2$, and the variance of a threshold estimate is proportional to ${s_{0.5}}^{-2}$. So, all threshold confidence interval widths are multiplied by $s_{0.5}(\boldsymbol{\theta}_{\text{gen}})$, and all slope confidence interval widths are divided by $s_{0.5}(\boldsymbol{\theta}_{\text{gen}})$ before being reported.[†]

Bias and efficiency scores for the sampling scheme are then measured, as described in sections 5.1.1 and 5.1.2. Efficiency scores for a measure of interest u will be based on the width of a confidence interval (WCI) computed from the distribution of u^*. Sweat factors for 68.3% intervals will be given by $K = N\left(\frac{1}{2}\text{WCI}_{68}\right)^2$, and sweat factors for 95.4% intervals by $N\left(\frac{1}{4}\text{WCI}_{95}\right)^2$. Efficiency is then obtained by dividing K into the ideal sweat factor K_{min}.

For a threshold t_f, K_{min} is given by equation (5.2), in which $p = \gamma + (1-\gamma-\lambda)f$ and $\mathrm{d}p/\mathrm{d}x = (1-\gamma-\lambda)s_f$. Bearing in mind that confidence intervals are scaled to correspond to $s_{0.5} = 1$, scaled slope values for the Weibull function and $\boldsymbol{\theta}_{\text{gen}} = (3, 4, 0.5, 0.01)^{\mathrm{T}}$ are $s_{0.2} = 0.684$, $s_{0.5} = 1$ and $s_{0.8} = 0.752$. This yields ideal sweat factors of 2.14, 0.79 and 0.71 for thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$, respectively.

A working value of K_{min} for the slope value $s_{0.5}$ was obtained by examining all possible pairs of stimulus values whose performance values $\psi(x)$ were drawn from $\{0.51, 0.515 \ldots 0.975, 0.98\}$, and computing the sweat factor $N\,\text{se}_s^2$ for each using equation (2.3). For a 2-AFC Weibull function with parameters $\boldsymbol{\theta}_{\text{gen}} = (3, 4, 0.5, 0.01)^{\mathrm{T}}$, and assuming n_i to be constant across

[†] An experimenter who wishes to apply the results to his own experimental situation could therefore simply divide the threshold confidence interval widths quoted here (and multiply the slope confidence interval widths) by his own $s_{0.5}(\hat{\boldsymbol{\theta}}_0)$—provided, of course, that the shape of the psychometric function is sufficiently similar.

blocks, the best 2-point sampling scheme was found to consist of performance values $\boldsymbol{p} = \{0.61, 0.965\}$, corresponding to a sweat factor of $8.03{s_{0.5}}^2$. As the slope intervals from sampling scheme tests will be standardized to correspond to a generating slope of 1, the value $K_{\min} = 8.03$ is therefore appropriate.[†]

5.3.2 Confidence interval types

There are at least three possible approaches to the calculation of the confidence interval, depending on one's interpretation of $\boldsymbol{\theta}_{\text{gen}}$:

1. $\boldsymbol{\theta}_{\text{gen}}$ represents the true psychometric function: In this case, an interval is calculated using the relevant percentiles of the Monte Carlo distribution u^*, which are (as $R \to \infty$) true confidence limits for the estimates, assuming that the sample values have been positioned correctly on an *already known* psychometric function. The experimenter could therefore only quote these values if $\boldsymbol{\theta}$ were known already, a situation which would never occur in practice.

2. $\boldsymbol{\theta}_{\text{gen}}$ represents the experimenter's maximum-likelihood estimate of some unknown parameter set: In this interpretation, u^* is treated as a bootstrap distribution, and a confidence interval is computed using bootstrap methods. Since the ψ values of a certain sampling scheme S refer to values on the MLE psychometric function rather than an unknown function, the bias and efficiency measures obtained from the simulation are more useful to an experimenter: if it is reported in this chapter that a certain sampling scheme S has a good bootstrap confidence interval width score, then it is a good idea for the experimenter to aim to reproduce that sampling scheme, choosing stimulus intensities $\boldsymbol{x}$ which will produce the same ψ values as S when transformed through the experimenter's *own* maximum-likelihood estimate of the psychometric function. The method used to obtain bootstrap

[†] The performance values and sweat factor thus obtained are similar to those given by O'Regan and Humbert[9] for the logistic function (allowing for differences between the shape of the logistic and the shape of the Weibull).

confidence intervals will be the $\mathrm{BC_a}$ method, based on its performance in the coverage simulations of chapter 3. The low absolute coverage coverage values of the $\mathrm{BC_a}$ method on many measures will not be a cause for concern in the current chapter: the tests of chapter 3 showed that the $\mathrm{BC_a}$ method was generally the fairest for comparing sampling schemes against each other, because it showed least variability in coverage and imbalance estimates across sampling schemes and between different values of N.

3. $\boldsymbol{\theta}_{\mathrm{gen}}$ represents a curve at which the experimental results may arrive, more or less accurately: In this interpretation, the experimenter wishes to reproduce a certain sampling scheme S, but cannot do so reliably because the estimate must be built up as the experiment proceeds. For example, the experimenter's approach might be to take some pilot data, using a small number of trials or perhaps an adaptive procedure, to obtain a working estimate of the parameter set. Using this working estimate, the experimenter chooses $\boldsymbol{x}$ values so as to replicate the pattern S as closely as possible, and then begins taking data in earnest. However, there will be a discrepancy between the working estimate and the final estimate on which the final confidence interval is based, so the stimulus values $\boldsymbol{x}$ will have effectively "slipped" relative to the curve. The expanded bootstrap method of section 2.2.6 explores the neighbourhood of a parameter set, in a way that reflects the likelihood of "slipping". It therefore provides an approximate way of reflecting the "risk" of trying to reproduce a certain S.

All three interval types were measured, with WCI scores being based on equal-tailed 68.3% and 95.5% confidence intervals. Expanded confidence intervals used $m = 8$ repeats of the bootstrap run, placing the generating parameter sets on the boundary of a confidence region of coverage 0.5 (see section 2.2.6). Results from the expanded confidence intervals should be considered along with the caveat that they are probably too conservative, over-emphasizing the effect of N (see section 3.3.5). In the main, results from approach 2 (bootstrap $\mathrm{BC_a}$ results) will be considered.

5.4 The performance of Wichmann's 7 sampling schemes

Wichmann and Hill[6] demonstrated the differences in efficiency and sensitivity between the sampling schemes described in section 1.5.1, noting in particular that the placement of one or more samples above 95% greatly reduced the sensitivity of confidence interval width to errors in the MLE. Their observations are repeated and extended here, the major differences being as follows:

- The $\mathrm{BC_a}$ method is now also used to obtain confidence intervals, where Wichmann and Hill used the bootstrap percentile method: the results of chapter 3 suggest that the $\mathrm{BC_a}$ method allows for fairer comparison between sampling schemes on slopes.

- A likelihood-based joint confidence region in α and β is now used for obtaining expanded confidence intervals, as described in section 2.2.6, whereas Wichmann and Hill originally used rectangular confidence regions of less accurate coverage.

- A wider range of values of N is explored, including lower values than those examined by Wichmann and Hill, in order to examine the changes in bias and efficiency that occur with N.

- The entire set of simulations was repeated 10 times in order to obtain an indication of the intrinsic variability of the bootstrap method.

Each of the sampling schemes was tested using $R = 1999$ simulation runs as described above, and the test was repeated 10 times at each of nine different values of N: $N = 24$, 48, 72, 96, 120, 168, 480 and 960.

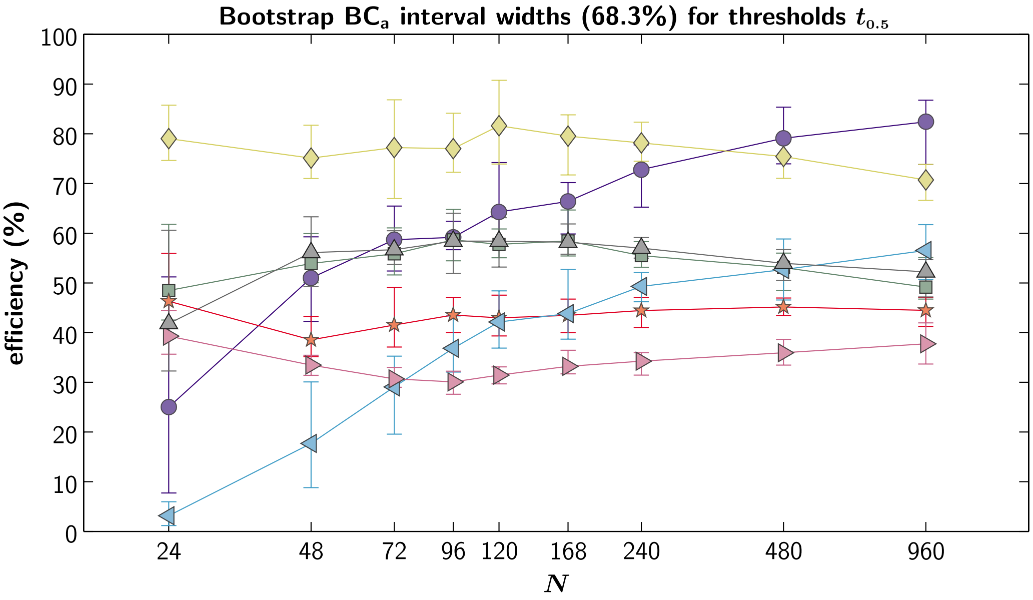
The figures of this section will all use the same format. Bias or efficiency for the seven sampling schemes will be plotted on the ordinate, against N on a log-scale abscissa. Symbol shapes denote the different sampling schemes in the manner shown in figure 1.2. Symbol positions will indicate the mean bias, or the efficiency corresponding to the mean confidence interval width, obtained over 10 repetitions. The error bars around each point show the full range of values measured in the 10 repetitions.

5.4.1 Threshold results

The upper panel of figure 5.1 shows the efficiency scores for the seven sampling schemes on thresholds $t_{0.5}$, using 68.3% $\mathrm{BC_a}$ confidence intervals as the measure. The different sampling schemes have different efficiency scores, as Wichmann and Hill[6] previously found. For most of the values of N studied, the most efficient scheme by a large margin is ⧫, which has five points clustered close to the threshold and a single point at a very high performance level (98.5% correct). It seems to have struck a good compromise between sampling near to the target value, and making use of the low variability at the high end of the function. The scheme with the widest spacing of sample points, ▶, generally performs worst.

Most of the schemes have a fairly constant efficiency score for all studied values of N, but it is important to note that the behaviour of ● and ◀ is not constant. These two sampling schemes are well clustered close to the threshold point (particularly ●), but they neglect to have a single point at high performance values. The result is that they are only efficient when N is large.

The optimal sampling strategy changes as N increases, just as Finney[17] observed in his table of yes-no sampling schemes. In the upper panel of figure 5.1, ● is the most efficient sampling scheme at the highest values of N ($N \geq 480$). The changeover between ● and ⧫, which are very similar except for the one high point in ⧫, happens at roughly $N = 360$—it seems, therefore, that it is worth taking no more than about 60 trials at the high point before concentrating on stimulus values closer to the threshold. On the other hand, at the very lowest value of N ($N = 24$), ● is one of the least efficient schemes, or at best it is highly unreliable. This has implications for



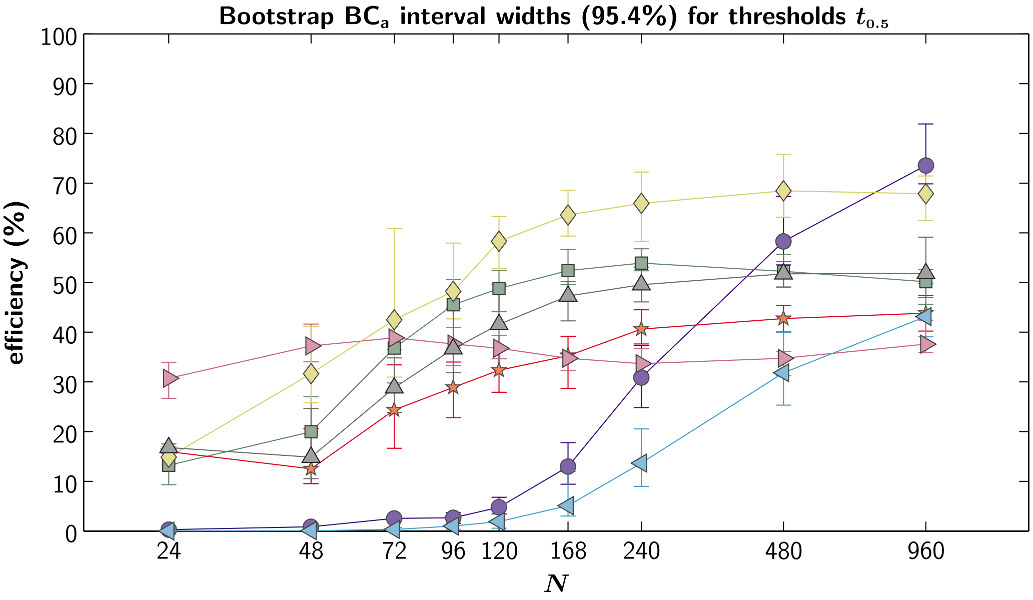


Fig. 5.1: Taylor-Creelman efficiency of estimation of threshold $t_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% bootstrap BC$_a$ intervals, respectively. See section 5.4 for further details.

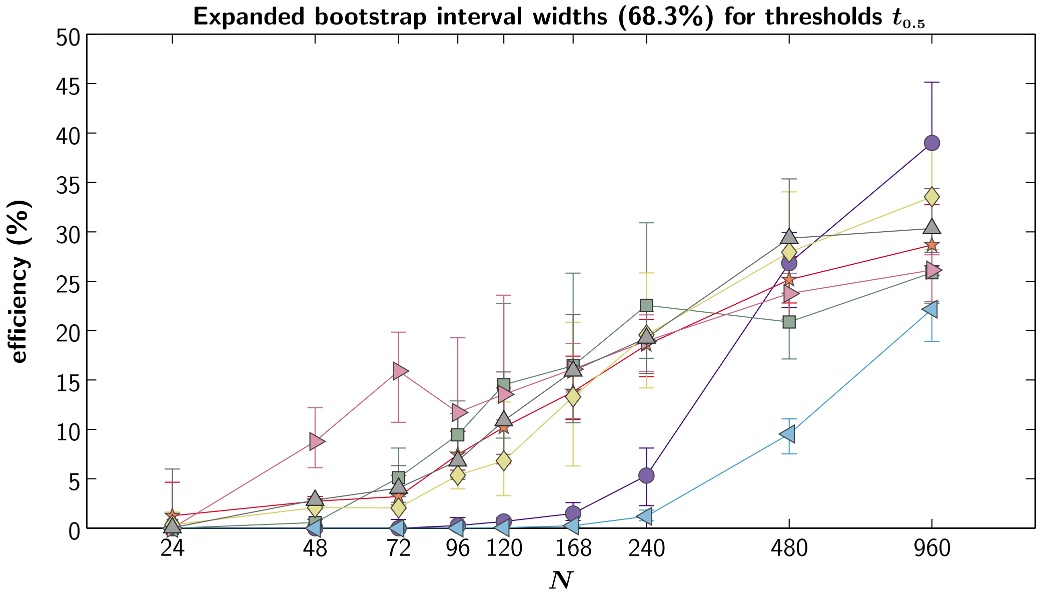
the use of most classical adaptive procedures, which aim to replicate exactly this situation: a small number of trials concentrated as closely as possible around threshold.[†]

The lower panel of figure 5.1 shows the results for 95.4% bootstrap intervals, in which the effect of N is even more pronounced. All sampling schemes, except for ▶ with its very wide spread of stimulus values, decrease in efficiency when N is less than about 100–200. The effect on ● and ◀, the two schemes without a single point above 90% correct, is very pronounced. Now, at all values of N lower than 240, they are very poor relative to the other schemes, and although ● eventually rises to be the most efficient, the crossover with ⧫ occurs later, at around $N = 750$. Thus, if confidence intervals of large coverage are desired, it is highly advisable to place at least one sample point at a high expected performance level (above 90%, or preferably above 95%).

Monte Carlo results for $t_{0.5}$ (interpretation 1 in section 5.3.2) were found to be almost indistinguishable from the $\mathrm{BC_a}$ results, and so will not be shown separately. Expanded bootstrap results (interpretation 3) are instructive. The upper and lower panels of figure 5.2 show the results for 68.3% and 95.4% expanded intervals, respectively. The effect of N is enhanced even further, and is very pronounced even for 68.3% intervals (the results at the two confidence are in fact very similar). As was the case for $\mathrm{BC_a}$ intervals, ● and ◀ are much worse than the other schemes, but now the other schemes are closer in performance to one another. The exception is ▶, which performed badly on the Monte Carlo and bootstrap $\mathrm{BC_a}$ measures, but is now the best scheme for lower values of N: its wide spread of stimulus values is clearly an effective "hedge" against possible stimulus placement error.

All schemes are relatively free of threshold estimation bias at all the values

[†] Adaptive procedures are generally more efficient than we might infer from this.[7,13,18] Sensible experimental procedure is to start at a high stimulus level and work down towards the target threshold, lest the observer become disheartened early on by the difficulty of the task. As a result adaptive runs usually have more observations above threshold than below, and their distribution may be more akin to that of ▲ or ★ than ●. Another contributing factor may be that higher performance levels than 75%, where expected variability is smaller, are often targeted.



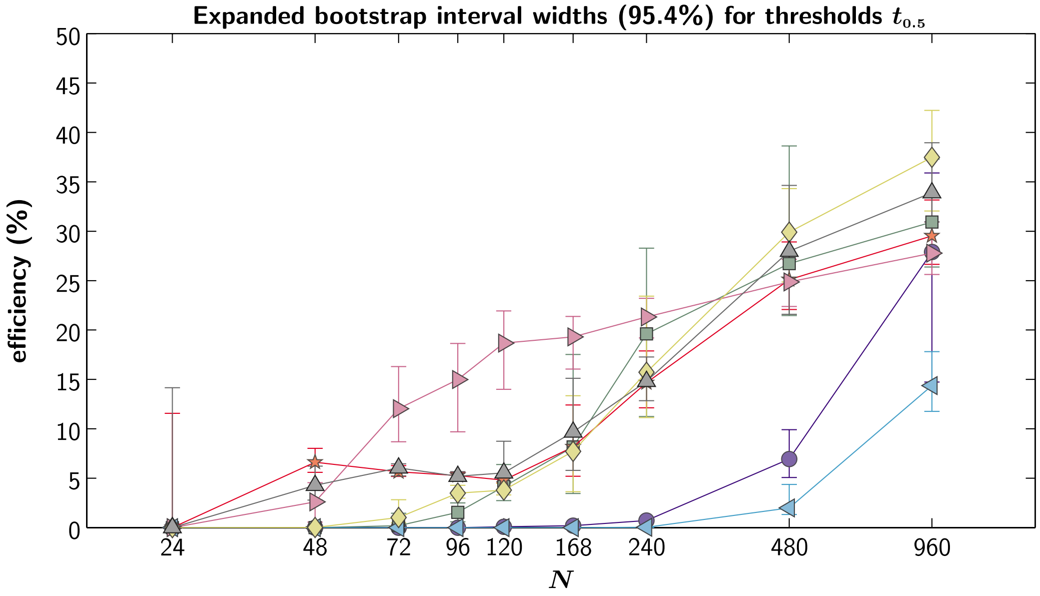


Fig. 5.2: Taylor-Creelman efficiency of estimation of threshold $t_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% expanded bootstrap intervals, respectively. See section 5.4 for further details.

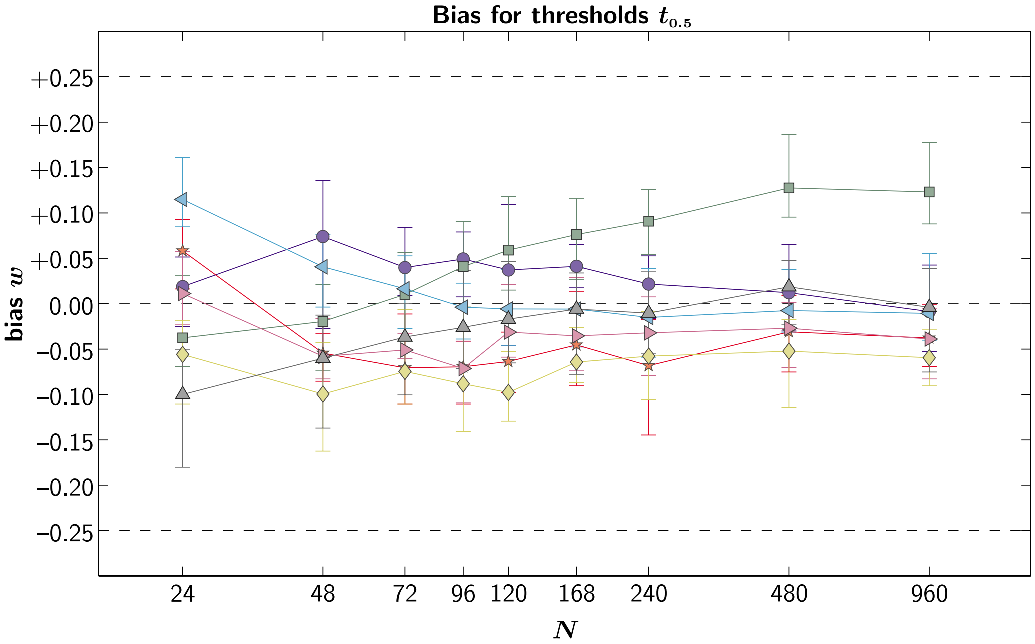


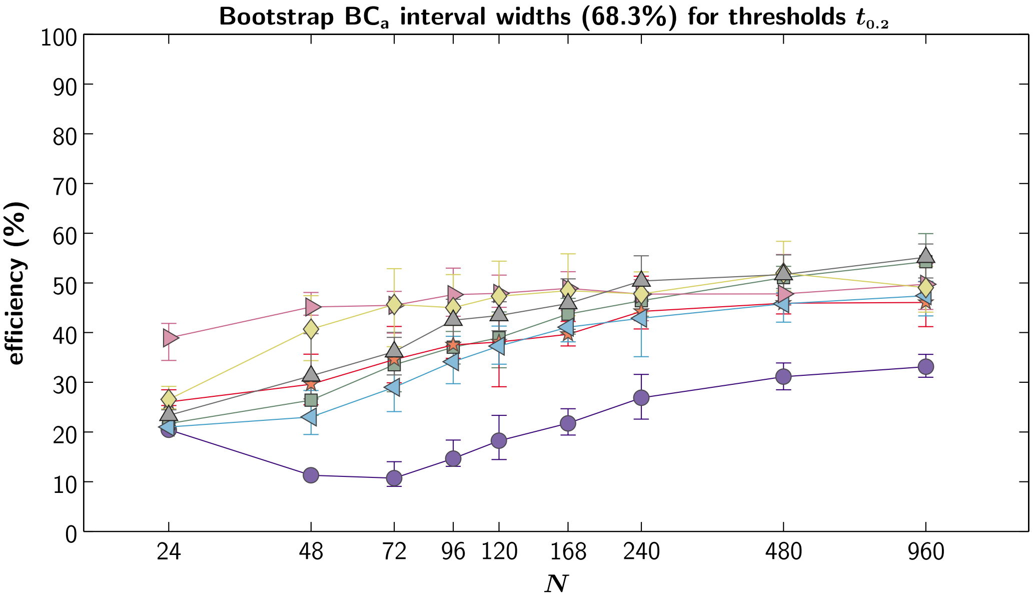
Fig. 5.3: Bias in the estimation of threshold $t_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. See section 5.4 for further details.

of N studied: as figure 5.3 shows, the bias term w remains well within the range ±0.25.

An additional, perhaps somewhat surprising example will help to reinforce the point that the most efficient estimator of a threshold is not necessarily the one that places stimulus values closest to the threshold value. Figure 5.4 shows the efficiency scores for bootstrap $\mathrm{BC_a}$ intervals on the lower performance threshold $t_{0.2}$. Note that ◀ is not efficient, either with regard to 68.3% or 95.4% intervals, despite being the scheme most closely centred around the target threshold point. In fact, only ● is less efficient. By contrast, ⧫ gives consistently smaller confidence intervals, despite having *all* its sample points above the target level; it yields low variability at all threshold levels because it simultaneously constrains both threshold and slope tightly, thus pinning down the entire psychometric function. However, its bias in slope estimation (see section 5.4.2) is also reflected in its bias in the estimation of $t_{0.2}$, as figure 5.5 shows. To strike a reasonable balance of low bias and high efficiency, the widely spread schemes ▶ and ■ are probably the best choices, depending on the number of trials.

5.4.2 Slope results

A rather different pattern emerges in the bootstrap interval results for slopes, shown in figure 5.6. Again, the upper panel shows efficiency scores based on WCI_{68}, and the lower panel shows scores based on WCI_{95}. In the upper panel, ● and ◀ now stay constant, and remain the least efficient schemes at all the values of N studied, whereas the more efficient schemes decline in efficiency as N increases. Note that the schemes which produce the smallest intervals are, once again, the ones with at least one point at very high performance levels ($>$ 95%—remember also that the "ideal" slope sampling scheme for this psychometric function had one of its two points at the 96.5% performance level). Some measurements, notably those from ★ at $N = 24$, occasionally exceeded 100% efficiency. This serves as a reminder that bootstrap measurements of variability are themselves only estimates, and prone to error, particularly when a small number of observations has been taken.



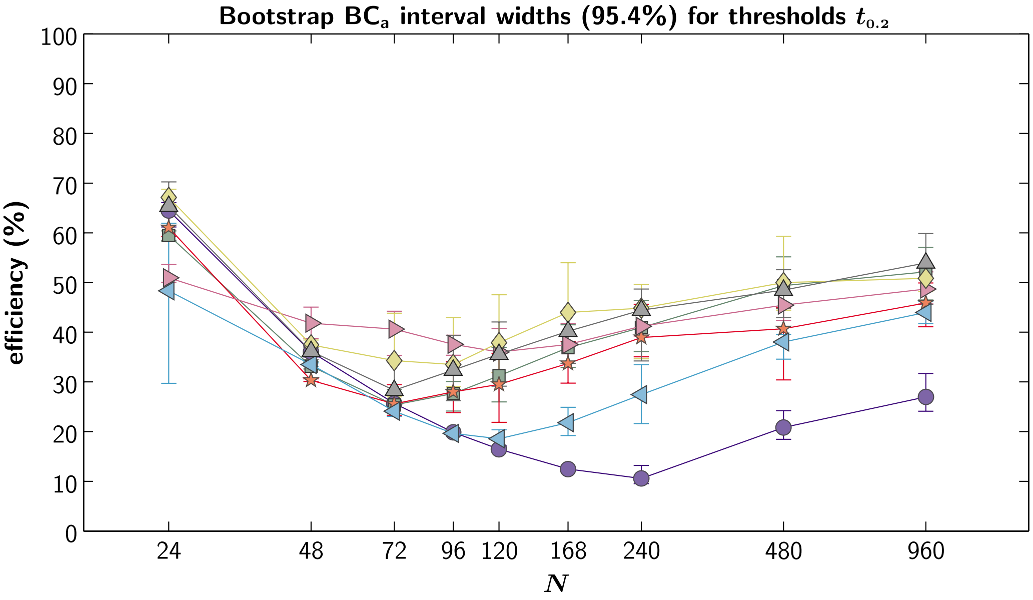


Fig. 5.4: Taylor-Creelman efficiency of estimation of threshold $t_{0.2}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% bootstrap BC$_a$ intervals, respectively. See section 5.4 for further details.

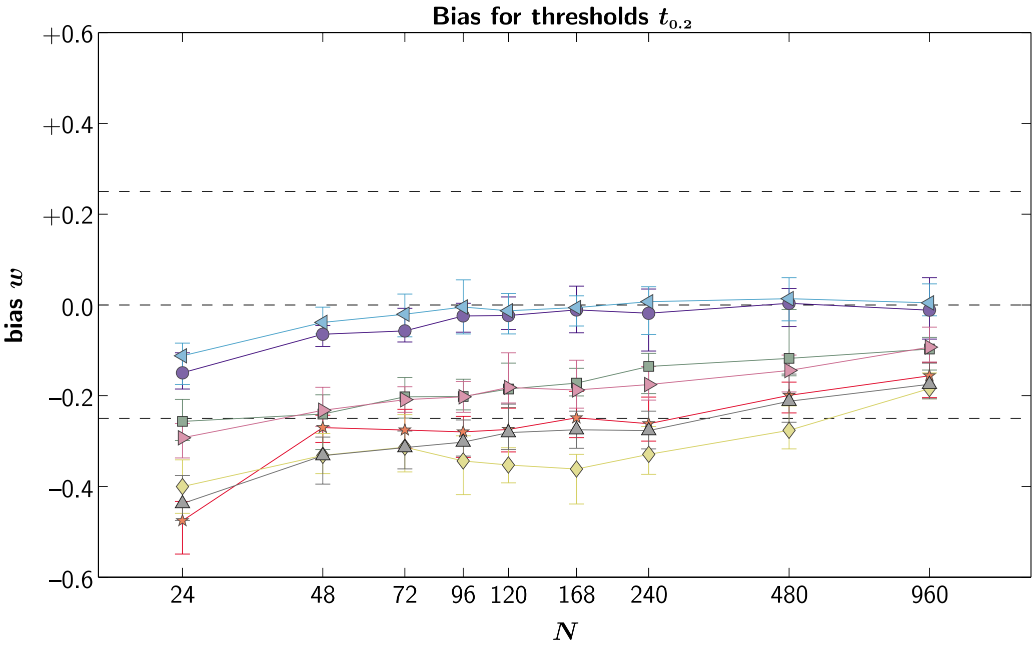
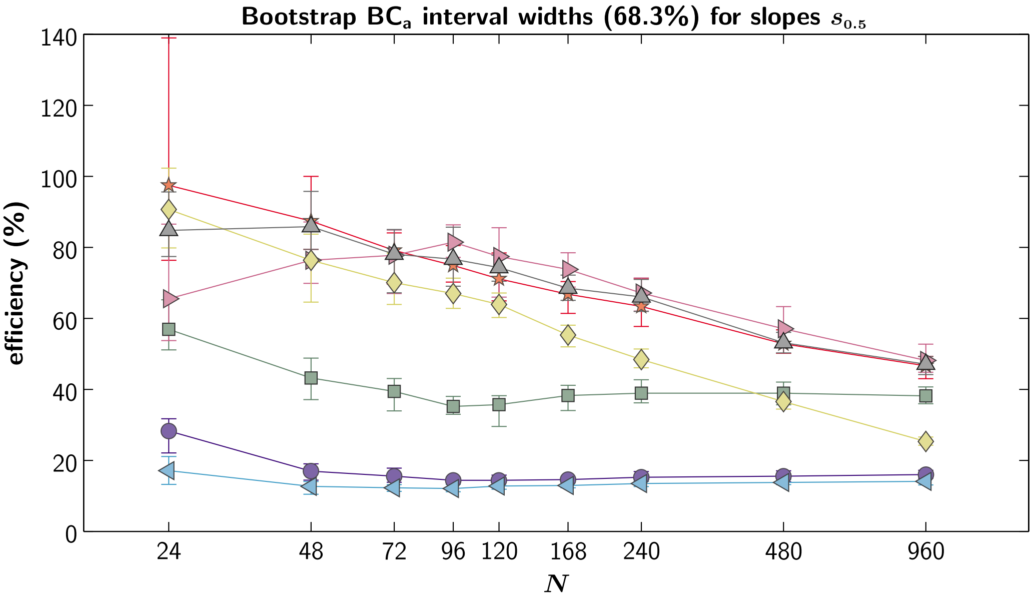


Fig. 5.5: Bias in the estimation of threshold $t_{0.2}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. See section 5.4 for further details.



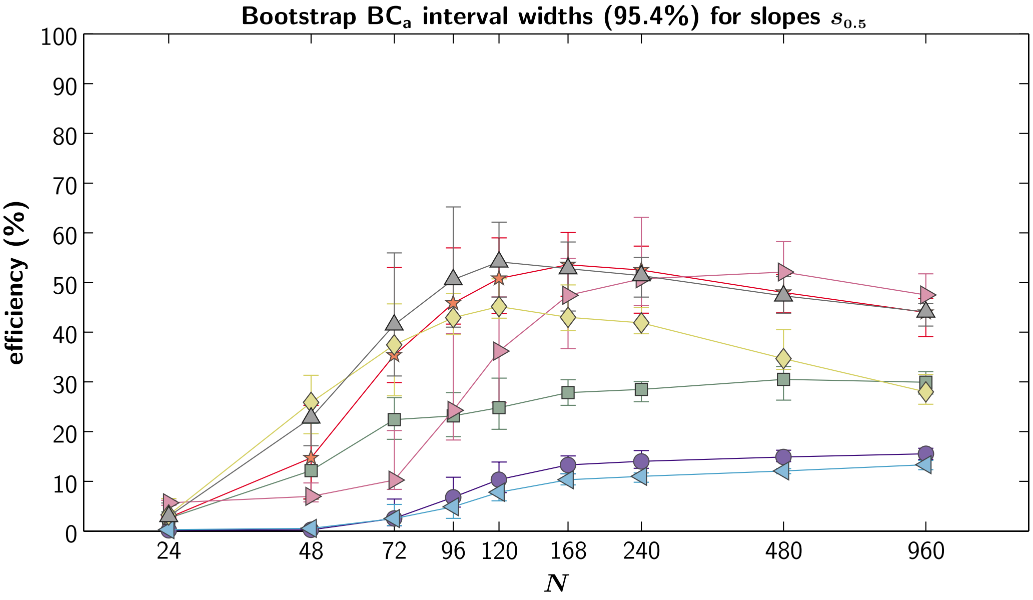


Fig. 5.6: Taylor-Creelman efficiency of estimation of slope $s_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% bootstrap $\mathrm{BC_a}$ intervals, respectively. See section 5.4 for further details.

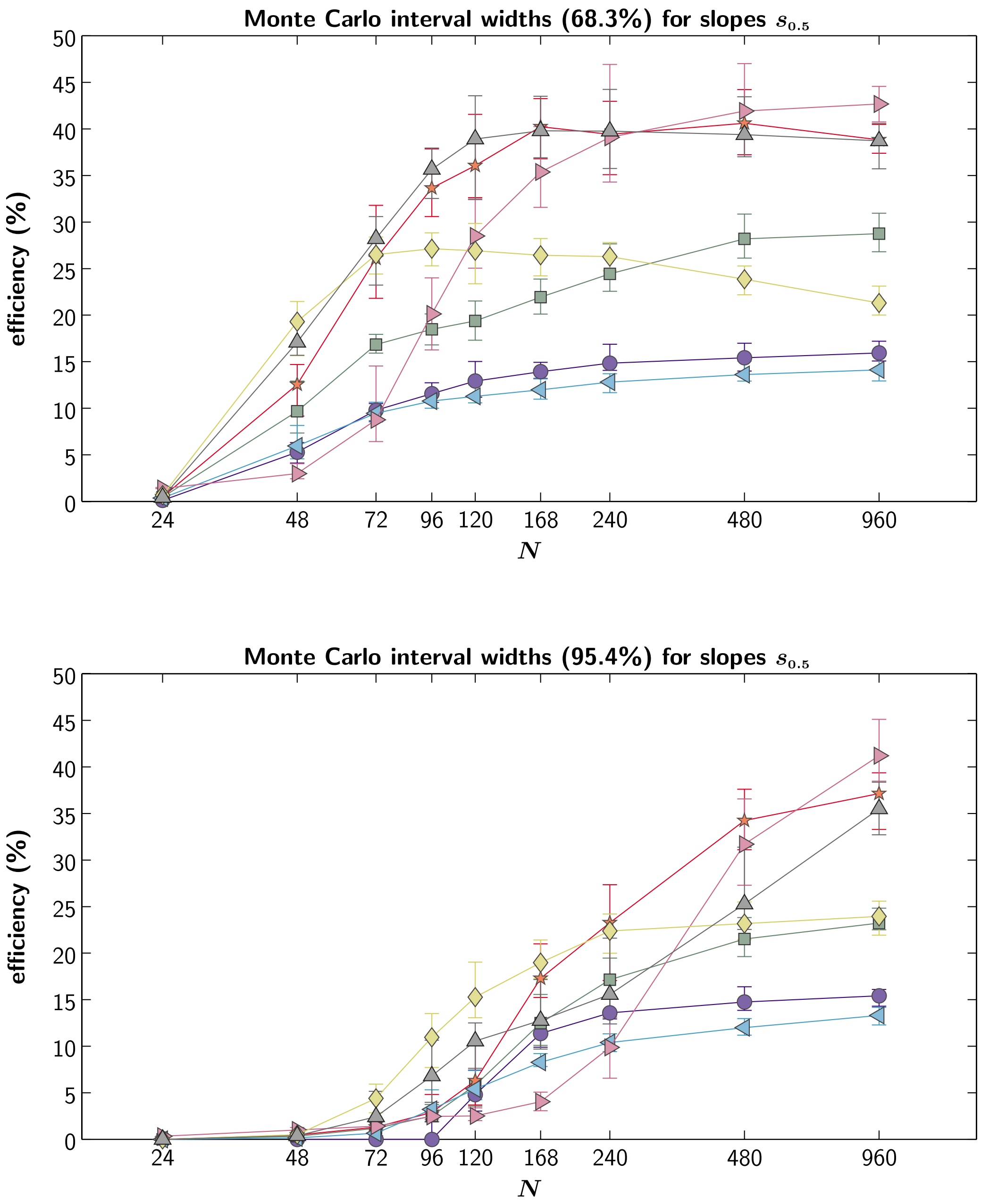
The lower panel of figure 5.6 shows that, when efficiency is measured by the width of 95.4% confidence intervals, N has a much greater effect on all the results. Efficiency generally increases with N, and none of the schemes reaches peak efficiency until around $N = 120$. The most widely sampled scheme, ▶, is a late developer: while it ends up as the most efficient scheme it does not overtake the others until N exceeds about 240, by which time ⧫ has already begun to decline.

Unlike the threshold results of section 5.4.1, the slope results show considerable differences between bootstrap intervals and Monte Carlo percentile intervals.[†] The latter are shown in figure 5.7. The 68.3% interval results from Monte Carlo intervals, shown in the upper panel, look very similar to the 95.4% bootstrap results of figure 5.6, and the 95.4% Monte Carlo results show an even greater positive influence of N. The fact that the Monte Carlo and bootstrap results differ qualitatively indicates that one's criteria of efficiency may apply very differently depending on which method is used to compute intervals. Clearly one should be careful in one's choice of interval method if wishing to compare the efficiency of slope estimation of different sampling strategies. The results of chapter 3 suggest that, while neither method is very good for slopes in a realistic context, the $\mathrm{BC_a}$ method is the more accurate bootstrap method of the two, in both idealized and realistic conditions (see figures 3.6 and 3.8).

Figure 5.8 shows the expanded bootstrap interval results for slopes. As in the threshold results, many of the differences between sampling schemes have dissolved, and N has taken over as the most important factor affecting efficiency: the higher the value of N, the more efficient all the sampling schemes are at estimating slope, if we take the expanded confidence interval widths as our criterion for efficiency.

The bias of slope estimation must be considered as well as efficiency, however. Figure 5.9 shows that all the schemes that produced the smallest slope

[†] The Monte Carlo intervals are constructed from percentiles of the distribution of simulated estimates without adjustment or reversal. This method can also be applied as a bootstrap technique, in which context it is called the bootstrap percentile method (section 2.2.4).



Monte Carlo interval widths (68.3%) for slopes $s_{0.5}$

Monte Carlo interval widths (95.4%) for slopes $s_{0.5}$

Fig. 5.7: Taylor-Creelman efficiency of estimation of slope $s_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% Monte Carlo percentile intervals, respectively. See section 5.4 for further details.

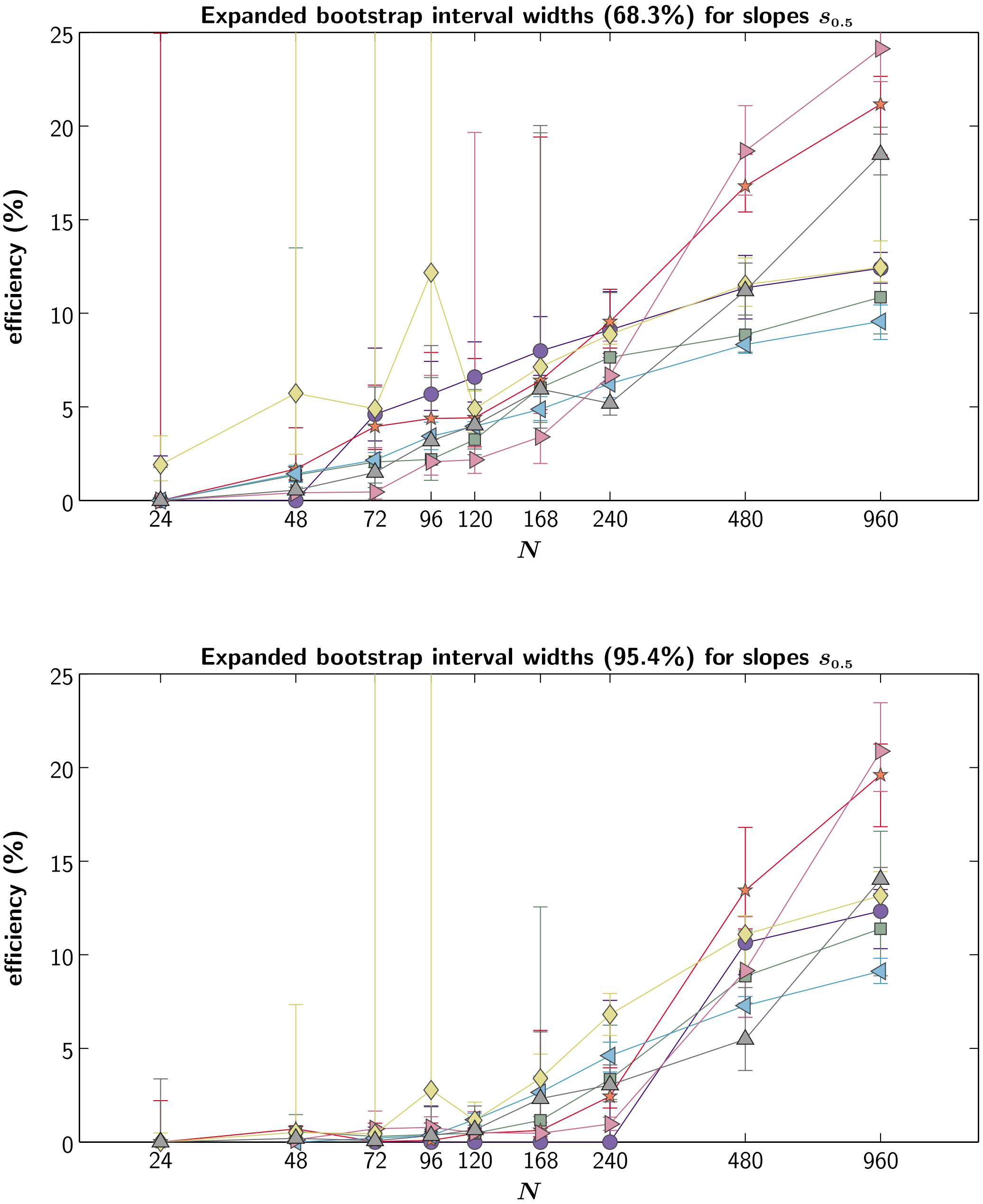


Fig. 5.8: Taylor-Creelman efficiency of estimation of slope $s_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. The upper and lower panels show efficiency scores based on 68.3% and 95.4% expanded bootstrap intervals, respectively. See section 5.4 for further details.

interval widths also display considerable bias, consistently overestimating the slope of the psychometric function. Bias generally decreases as N increases, but most of the schemes only come within the ± 0.25 range at $N = 960$. The bias seen here is the same phenomenon as that discussed in section 3.3.4. The consistent over-estimation of slope can also be seen in figure 3.13 ($t_{0.2}$ is over-estimated, and $t_{0.8}$ is under-estimated). Figure 3.13 also suggests that the magnitude of the problem varies depending on the unknown value of λ_{gen}. Unless more reliable methods can be developed for dealing with lapses (a more robust loss function than the regular binomial likelihood formula, for example), then the best policy for obtaining accurate slope estimates will sometimes be, paradoxically, to use the sampling schemes which would, in an idealized situation in which λ_{gen} were known, be the *least* suitable schemes for slope measurement. The two schemes that produced the largest slope intervals, ● and ◀, show relatively little bias in figure 5.9, and this is a combination of two factors: the first is that, lacking sample points at high performance values, they are less influenced by λ than the other schemes and therefore tend to produce smaller numerical discrepancies between the true slope value and the observed value; the second second factor is that, precisely because their confidence intervals are so much larger than those of other schemes, they render the same numerical discrepancy less significant than would a more efficient scheme.

The decision to use a strategy of inefficient estimation or "efficient mis-estimation" must depend on the magnitude of the psychophysical slope effect under study, and the consequent desired confidence interval width. For example, suppose an experimenter takes $N = 275$ trials, and aims to use a scheme similar to ●, which is more appropriate than ▶ because of the relatively large bias of the latter. However, at this number of trials the relative efficiency of ▶ to ●, computed using 95% bootstrap confidence intervals, is roughly 3.5. Suppose that, in order to be appropriate for the experimental situation, the slope confidence interval needs to be smaller by a factor of more than $\sqrt{3.5}$. Therefore, N must be increased by a factor of more than 3.5, to over 960 (in figure 5.6, the efficiency curve for ● is flat in this region, indicating that confidence interval width is roughly proportional to $N^{-\frac{1}{2}}$). The experimenter continues taking more samples, aiming for the performance levels dictated by ●. At $N = 960$, there is still a little way further to go, but at this large number of trials the bias of ▶ has been reduced to a much smaller value, which the experimenter might have deemed acceptable, and its efficiency is still roughly 3 times that of ●. Therefore, if the experimenter had chosen ▶ instead of ●, the experiment would already be over. In this example, the optimum strategy might be to begin sampling at low or central performance levels in the manner of ● or ◀, and then to spread out the sample points in the manner of ▶ after a certain number of trials has been taken. The more accurate slope estimates of the early samples will (on average, across a number of similar experiments) be beneficial for positioning the later samples (if slope were over-estimated during the early samples, the later ones would be less spread out than the experimenter intended). The details of how many trials to take, whether to spread out, when to start doing so and how far to spread, will depend on the experimental situation and the particular level of precision it demands.

5.4.3 Summary

As Wichmann and Hill[6] found, some sampling schemes are more efficient than others, and the current simulations suggest that such differences are significantly greater than the variation in estimated confidence interval width that occurs between successive repeats of the bootstrap process. For thresholds, the most efficient sampling patterns are not necessarily those that are the most closely concentrated around the target threshold value. The inclusion of one or more points at high performance levels (above 90%, or better still above 95%) greatly decreases the widths of confidence intervals for both thresholds and slopes. While this is certainly an advantage for threshold estimation, for which bias is generally small, it may lead to significant bias in slope estimation, which occurs because of the difficulty of obtaining accurate estimates of λ, with which slope co-varies. For all the sampling schemes studied, slope bias tended to decrease as N increased, but sampling schemes

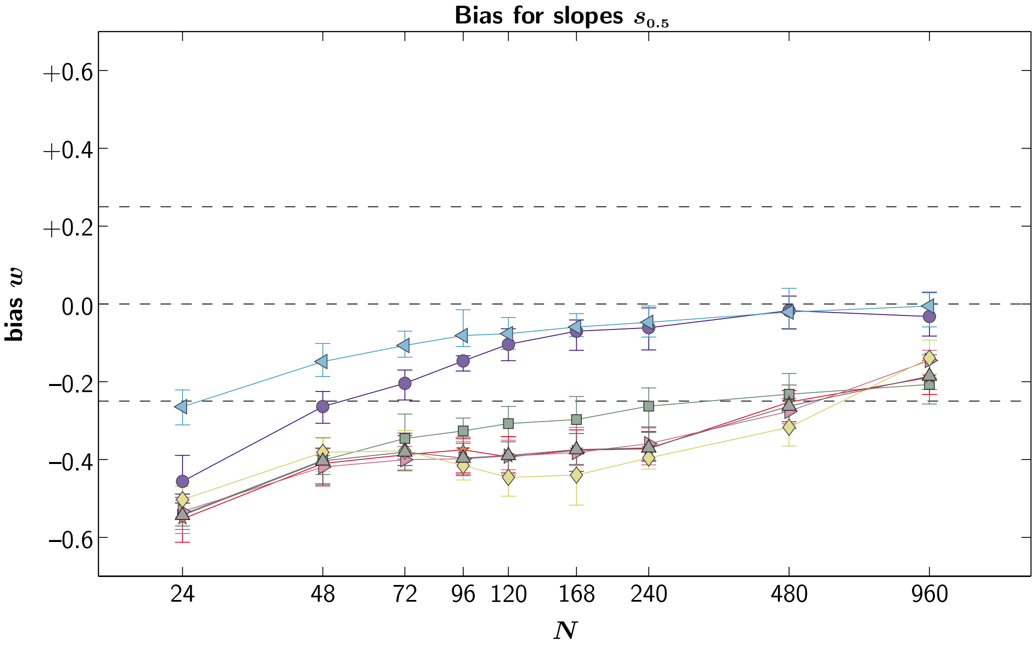


Fig. 5.9: Bias in the estimation of slope $s_{0.5}$ for seven 2-AFC sampling schemes is plotted against N, which is on a logarithmic scale. See section 5.4 for further details.

with points at high performance levels were slower to reduce to acceptable levels of bias than those without.

The relationship between confidence interval width and N is not necessarily the familiar inverse-square relationship that normal theory predicts. This fact can be identified from that fact that efficiency scores may change as N increases, and the manner in which they change depends on sampling scheme. So, the optimal choice of sampling scheme can be different at different values of N. For example, the sampling schemes which are most closely clustered around target threshold values become more efficient threshold estimators, relative to other sampling schemes, as N increases. This is consistent with the observations made by Finney,[17] who also found that the relative efficiency of different 2- and 3-point sampling schemes in yes-no experiments also changed with N. The optimal choice of sampling scheme also depends on whether one looks at short intervals of the order of ±1 standard deviation, or longer intervals of the order of ±2 standard deviations or more. Consideration of longer rather than shorter intervals increases the extent to which efficiency is correlated with N or, to put it another way, tends to increase the penalty associated with small values of N for all sampling schemes. Consideration of expanded bootstrap intervals also enhances the reliance of efficiency on N, and also favours those sampling schemes that are more widely spread.

5.5 Wider exploration of possible sampling schemes

If some sampling schemes are better than others, how can we ensure that we use a good one? Wichmann's 7 sampling schemes are good for illustrating the existence of differences, but they are only isolated examples in an infinite set of possible sampling schemes. They suggest that it is a good idea to bias one's placement of sample points towards higher performance levels, but do not tell us exactly how far to go. They also leave one important dimension entirely unexplored, because they all have 6 points—we still want to know whether, for a constant total number of observations, it is better to group those observations in a small number of sample points or to spread them over a larger number.[†] If the exact placement of stimuli makes such a difference, is it even possible to make a fair comparison between, say, a sampling scheme with $k = 3$ points and another with $k = 8$—might any effect of k be confounded by the question of exactly *where* the extra 5 points are placed?

The simulations described in this section aimed to explore a wide range of possible sampling schemes in order to study the effect of k and N, and any interactions between k and N, on bias and efficiency of threshold and slope estimation. The values $k = 3$, 4, 5, 6, 8, 10 and 12 were investigated, at $N = 120$, 240 and 480. In order to provide a basis for comparison of sampling schemes that had different numbers of points, a parametric and a non-parametric system were developed for generating sampling schemes. These are described in the following two sub-sections.

5.5.1 Parametric sampling scheme generation

A method was developed for the purposes of parametrizing the differences in a subset of possible sampling schemes, whose f-values were defined using a modified beta distribution. The beta distribution, which is given by

$$B(f;\, b_1, b_2) = \frac{f^{(b_1-1)}(1-f)^{(b_2-1)}}{\int_0^1 t^{(b_1-1)}(1-t)^{(b_1-1)}\,\mathrm{d}t}, \quad 0 \le f \le 1, \tag{5.3}$$

can, depending on its parameters, approximate many functions including normal, uniform, linear increasing, linear decreasing and parabolic distributions, as well as positively and negatively skewed distributions—see figure 5.10 for a few examples. As the beta function is so versatile, and because its domain is

[†] This question has a psychological side to it as well as the straightforward statistical one: the greater the number of observations performed at a single stimulus level, the more practice the observer gets at that level, but the greater the probability that the observer's attention might wane. For the purposes of this investigation the question will be considered from only the statistical angle, which can be answered by examining a computer-simulated observer free of any such effects.

$[0, 1]$, it is very useful as a parametric function for defining sampling schemes in terms of their f-values.

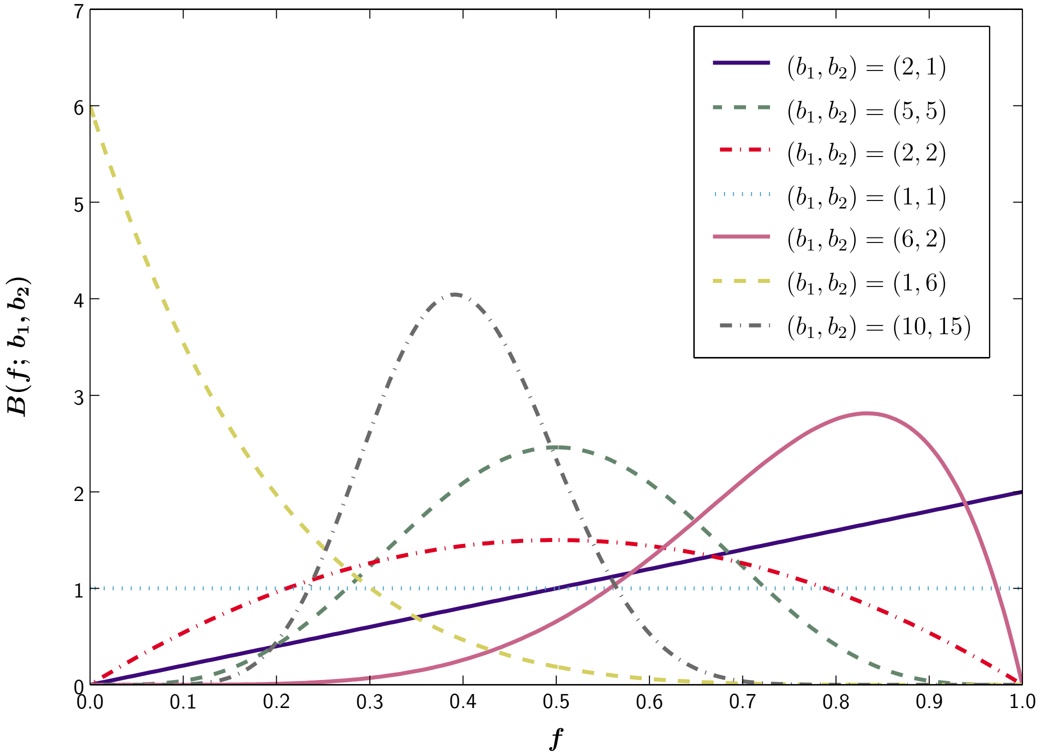


Fig. 5.10: Examples of beta distributions, using various different parameter pairs (b_1, b_2). See equation (5.3), page 174.

One way of using the beta function to generate sampling schemes would be to define a fairly large number of sample points (say, $k = 20$) at the fixed values $\boldsymbol{f} = (1, 2 \ldots k)/(k + 1)$, and then determine the relative proportions of the total number of trials to be taken at each f-level by the relative heights of the distribution at each point, for a chosen pair of parameter values: thus, define $\tilde{n}_i = B(f_i; b_1, b_2)$, then take $n_i = N\tilde{n}_i / \sum_{j=0}^{k} \tilde{n}_j$, rounded to the nearest integer. This method will be referred to as the unequal block distribution method.

A second method, the *equal* block distribution method, asymptotically (as $k \to \infty$ and $N \to \infty$) distributes individual trials with the same pattern as the unequal method, but it allows the number of trials per block to be constant. Block sizes are therefore given by $n_i = N/k$, but the locations of the blocks are given by the inverse cumulative of the beta distribution, evaluated at points $(1, 2 \ldots k)/(k+1)$. Thus, the f-points mark out equal areas under the distribution curve.

Graphical examples of the two block distribution methods will be given, but first a refinement will be added to the generating equation. Using either of the above block distribution methods in conjunction with the beta distribution defined in equation (5.3), a wide variety of sampling schemes might be generated: depending on the chosen values of b_1 and b_2, trials might be concentrated at the high or low end of the psychometric function, concentrated in the middle close to $t_{0.5}$, or more evenly spread over the whole psychometric function. However, the beta distribution allows only very limited scope for bimodal distributions of trials. Since bimodality might very probably be a characteristic of a number of very efficient sampling schemes (particularly where slopes are concerned), a third "asymmetry" parameter will be added to the distribution, as follows:

$$B_3(f;\, b_1, b_2, b_3) = \frac{1+b_3}{2} B(f;\, b_1, b_2) + \frac{1-b_3}{2} B(f;\, b_2, b_1). \qquad (5.4)$$

Thus, when the asymmetry parameter b_3 is equal to 1, equation (5.4) reduces to equation (5.3). When $b_3 = 0$, the resulting distribution is always symmetrical, as it is an equally weighted sum of the plain beta distribution $B(f;\, b_1, b_2)$ and a copy of the plain distribution reflected about $f = 0.5$. Intermediate values of b_3 result in unequal weighting between the plain beta distribution and the reflected copy.

Figure 5.11 shows examples of sampling schemes generated using equation (5.4). The upper panel shows a B_3 function with parameters $(9, 2, 0.3)$. Note that the use of an asymmetry parameter less than 1 has resulted in a bimodal distribution with unequal peaks. The open triangles represent a scheme that was generated from the curve using the unequal block distribution method with $k = 20$ and $N = 480$. The size of the points is related to the number of trials to be taken at each f-level, determined by the height of the curve. The filled triangles represent the results of the equal block distribution method with $k = 8$. The areas marked out by the broken lines under the curve are equal.

In the lower panel, the triangles show how the two schemes from the upper panel translate into stimulus units, using the Weibull function with $\boldsymbol{\theta} = (3, 4, 0.5, 0.01)^{\mathrm{T}}$ which is also shown for comparison. Broken lines mark the positions of $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$, corresponding to performance levels of 0.598, 0.745 and 0.892, respectively. Additional examples of sampling schemes generated by the equal block distribution method, using a number of different parameter triplets (b_1, b_2, b_3), are represented by the filled circles.

The result is that, with three parameters, a very wide range of sampling schemes can be defined. When $b_1 = b_2$, the scheme is always symmetrical, and the width of the range of stimulus values depends on the parameter value: values less than 1 produce a U-shaped distribution in which the points tend to cluster at the extreme upper and lower ends of the function, $b_1 = b_2 = 1$ produces a uniform distribution of expected performance values, and as the values increase above 1, the distribution becomes narrower and tends more towards the normal. Symmetric bimodal distributions (with less of an extreme separation between the peaks than the U-shaped function) can be achieved when $b_1 \neq b_2$ and $b_3 = 0$, and a variety of asymmetric sampling schemes can be generated with $b_1 \neq b_2$ and $b_3 \neq 0$.

The space defined by the three parameters (b_1, b_2, b_3) was explored by taking 20 values for each parameter, and generating sampling schemes from each of the 8000 combinations. Possible values for b_3 were linearly spaced in the range $[0, 1]$, and possible values for b_1 and b_2 were geometrically spaced in the range $[0.8, 10]$. Some of the sampling schemes corresponding to the extreme corners of the parameter space thus defined, such as $(10, 0.8, 1)$ and $(10, 10, 0)$, can be seen in the lower panel of figure 5.11.

Using the equal block distribution method, 8000 sampling schemes were tested for each combination of k and N (21 combinations, for a total of 168,000 sampling scheme tests, which involved a total of 3.36×10^8 fits to simulated data sets). Using the unequal block distribution method, 8000

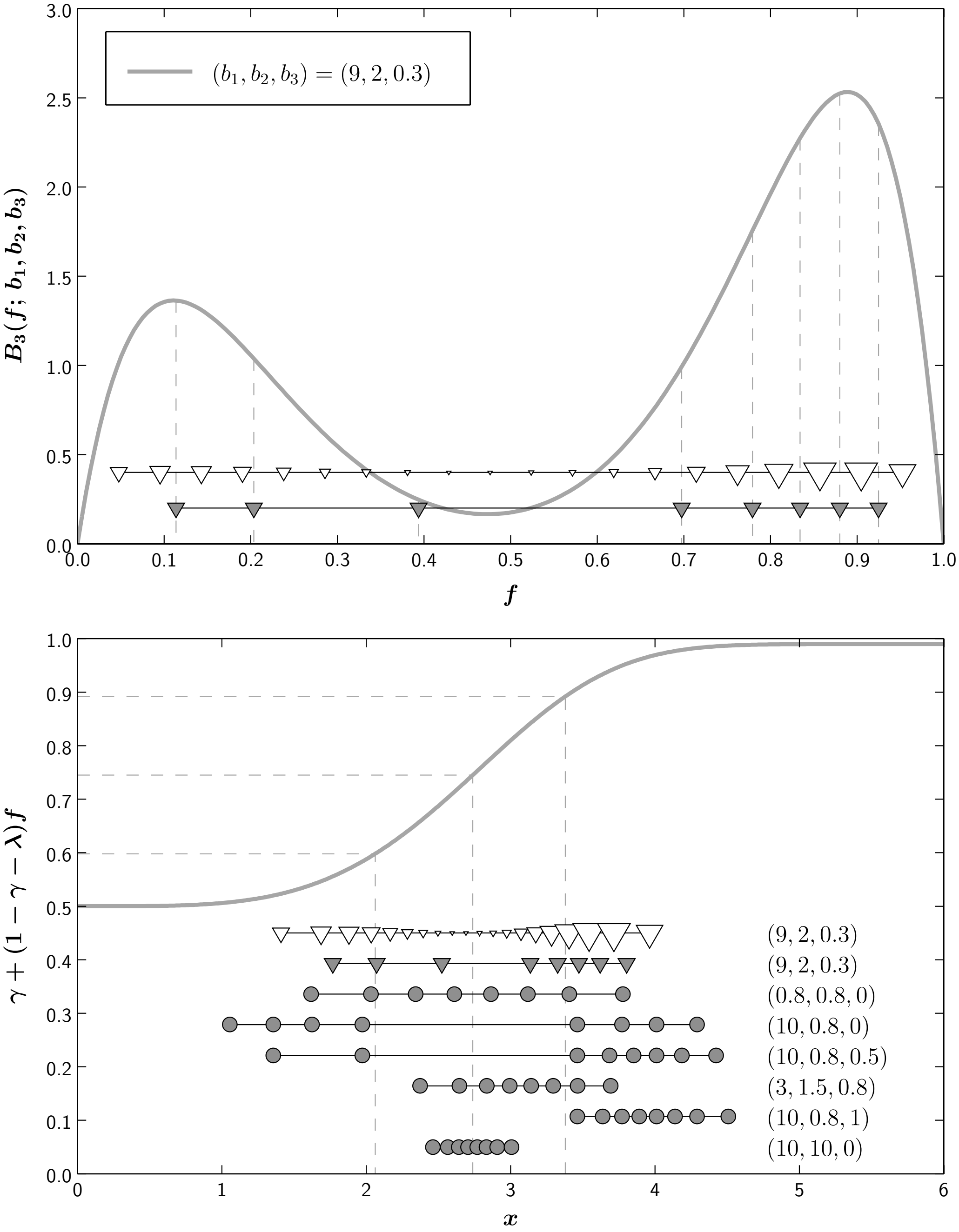


Fig. 5.11: Examples of sampling scheme generation using a modified beta distribution. See section 5.5.1 for details.

sampling schemes were tested at $\{k = 20, N \approx 240\}$ and at $\{k = 20, N \approx 480\}$. can be found in the results archive under:

- `simulations/optimal/2AFC/para/equal/l01/weibull/`
- `simulations/optimal/2AFC/para/unequal/l01/weibull/`

for the equal and unequal methods, respectively.

5.5.2 Non-parametric sampling scheme generation

While the parameterization of section 5.5.1 includes many sampling schemes, it does not quite succeed in approximating all sampling schemes of interest—for example, no set of parameters (b_1, b_2, b_3) comes very close to capturing the essential characteristics of scheme ⧫ from figure 1.2, which has many of its points clustered around the mid-point of the psychometric function and just one high point.

To explore sampling scheme characteristics still further, a random non-parametric generation method was developed. First, the sampling scheme was defined by the sequence $\boldsymbol{f} = (1, 2 \ldots k)/(k + 1)$. Such a set of values defines k points and $k + 1$ "gaps"—there is a gap between 0 and the first point, between each consecutive pair of points, and between the kth point and 1. The points were then "shuffled" 100 times: a single shuffle involved choosing one of the points at random, and moving it to the centre of a randomly chosen gap. All points had equal probability of selection, and all eligible gaps had equal probability of selection, but only gaps greater than $0.3/(k + 1)$ were eligible (thus, there was a lower bound of $0.15/(k + 1)$ on the spacing between successive f-values). Block sizes were all equal: $n_i = N/k$.

Examples of 8-point sampling schemes generated by the random method are shown in figure 5.12.

Sets of 1000 random sampling schemes were tested at a time. One set was run at each of the combinations of values of k and N (21 sets in all, for a total of 21,000 tests or 4.20×10^7 fits). For a given k, the same 1000 schemes were tested at each N. Results can be found in the archive under:

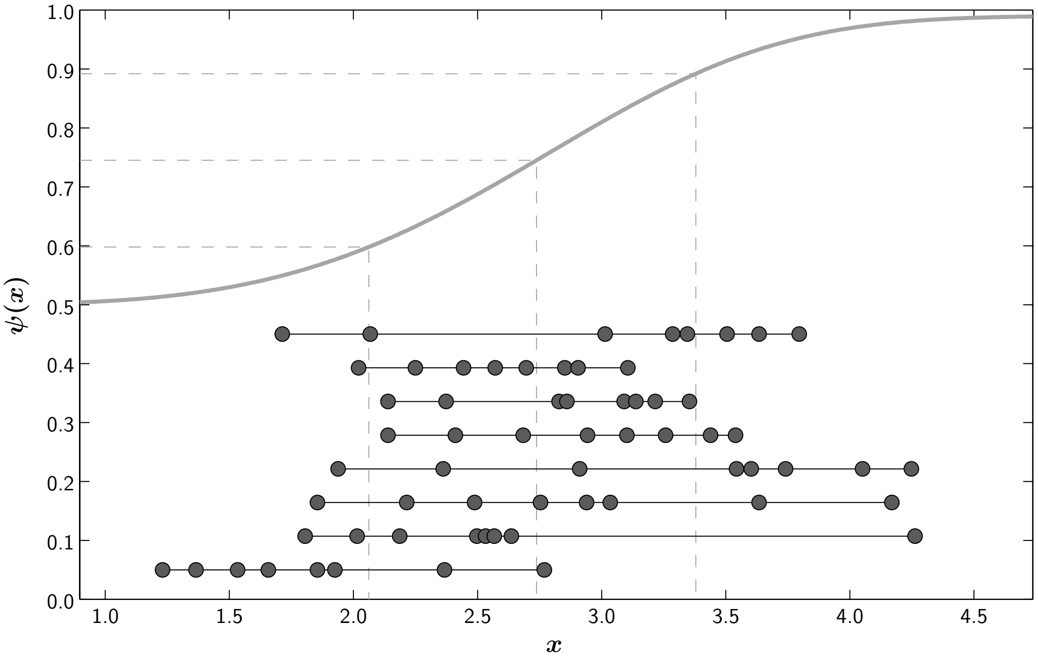


Fig. 5.12: Eight examples of 8-point sampling schemes generated by the random method of section 5.5.2. A 2-AFC Weibull function is shown for comparison, with lines marking the thresholds $t_{0.2}$, $t_{0.5}$ and $t_{0.8}$ (corresponding to performance levels of 0.598, 0.745 and 0.892, respectively).

- simulations/optimal/2AFC/nonpara/l01/weibull/

5.5.3 Representation of results

The parametric sampling scheme generation space (section 5.5.1) was sampled at regular intervals in each of its three dimensions, because it had been intended to present the results (bias or efficiency scores) as a function of the three dimensions without smoothing. Smoothing in three dimensions was considered undesirable because any failure of the parameterization to account for variation in the scores (which would show up as sharp apparently random changes) would then go unnoticed.

Three-dimensional representation of the results will not be necessary because the results from the parametric simulations were found to be generally smooth even when re-represented on a two-dimensional plot. A two-dimensional format will therefore be used for all the results: the dependent variable (bias or efficiency) is plotted as a function of the weighted standard deviation σ_p of the sampling scheme's performance levels (on the ordinate), and the weighted mean performance level $\bar{p}$ (on the abscissa). The weights used were simply equal to the number of observations used at each performance level,[†] so $\bar{p}$ and σ_p are equal to the first moment and second central moment, respectively, of the distribution of expected performance values corresponding to each individual trial location:

$$\bar{p} = \sum_{i=1}^{k} p_i n_i, \tag{5.5}$$

$$\text{and} \quad \sigma_p = \left[\sum_{i=1}^{k} n_i (p_i - \bar{p})^2 \right]^{\frac{1}{2}}. \tag{5.6}$$

Figure 5.13 shows the way in which sampling schemes map onto the space

[†] Other mapping methods were attempted, including the weighted mean and standard deviation of the stimulus values rather than the performance levels, and the weighted mean and standard deviation using the probit weights of equation (2.1). However, none of the alternatives produced noticeably smoother results, and $\bar{p}$ and σ_p as defined above were found to produce the most conveniently shaped and intuitively interpretable space.

defined by $\bar{p}$ and σ_p. The first column shows the layout of the parametrically generated 3-, 6- and 12-point sampling schemes using the 8000 parameter sets of section 5.5.1. In the first panel, the dots mark the positions corresponding to the sampling schemes explored by McKee *et al.*[5] at $k = 3$ (the positions corresponding to their 4- and 5-point schemes are almost identical to these). The second column shows the sets of 1000 3-, 6- and 12-point sampling schemes generated by the random method of section 5.5.2. The third column shows what would happen if the sampling schemes were generated by drawing each f-value independently from a uniform distribution on the interval $(0, 1)$. Note that when the parametric system is used, there is considerable overlap in the space covered by sampling schemes with different numbers of blocks—this facilitates examination of the effect of k. The area covered by the random method of section 5.5.2 is not quite so invariant with respect to k, but the method is better in that regard than the purely random independent selection of performance values.

Bias and efficiency scores will be denoted by colour. For bias scores w, a linear scale will be used, in which black denotes zero bias, "hot" colours denote $w > 0$ (underestimation of the measure of interest), and "cold" colours denote $w < 0$ (overestimation)—see, for example, figure 5.20. For interval width scores, an ordinal scale will be used, because inefficient sampling schemes (notably those with all sample points below the mid-point $t_{0.5}$) occasionally yielded extremely large numbers. The distributions of interval widths from a set of 8000 (or 1000) sampling scheme tests often had such long, thin upper tails that even a log transform did not allow the whole range to be represented while preserving sufficient precision at the low end. In figure 5.14 and others like it, the colour of each point is therefore determined by the rank of each interval width score within the set of 8000 or 1000 to which it belongs, dark blue denoting the lowest values, and dark red denoting the highest. The bar at the side of the figure shows how the colours relate to the actual interval width values, and also shows the distribution of widths in the form of a histogram. The text at the top of the bar indicates that (for example, in the lower panel of figure 5.14) the top 15.2% of values, up to a

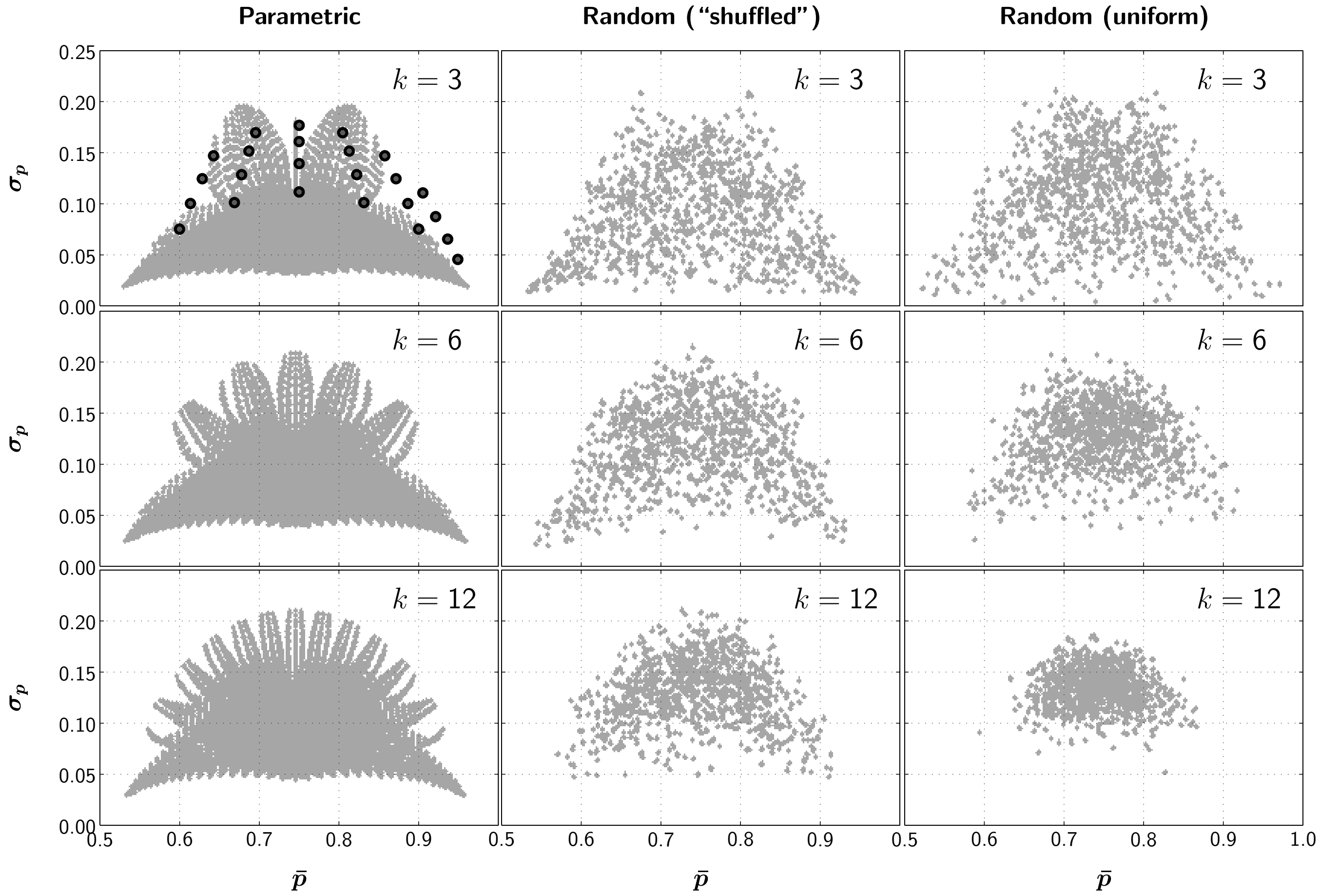


Fig. 5.13: Distribution in $(\bar{p}, \sigma_p)$ space of some of the test sets used in the simulations of section 5.5. From left to right, the columns show parametrically generated test sets (section 5.5.1), non-parametrically generated test sets (section 5.5.2), and (for comparison) sampling schemes with independent uniformly distributed performance values. Black dots in the top left panel show the schemes used by McKee *et al.*[5]

maximum interval width of 362, are not represented on the histogram.

To investigate the effect of k and N on bias and efficiency scores, two methods of analysis will be used. The first is to examine how the minimum moves with changes in k or N, the minimum being the location in $(\bar{p}, \sigma_p)$ space where the smallest absolute biases or smallest interval widths occur. To find the minimum, the best 1% of the schemes of a particular test set are identified (e.g. the 80 sampling schemes with the smallest interval widths out of a parametrically generated set of 8000) and the median $\bar{p}$ value, median σ_p value and median interval width from that group are reported. This method of finding the minimum was chosen in preference to smoothing the results and reporting the location of the minimum smoothed value, because interval widths may not always vary smoothly in $(\bar{p}, \sigma_p)$ space—if the best 80 sampling schemes happened to be intermingled with the worst 80, then a smoothing technique would not reveal the location of the minimum because the smoothed score in that region would reflect an average of the adjacent low and high scores. Effectively, the chosen method asks where the best schemes are located in $(\bar{p}, \sigma_p)$ space, rather than which location in $(\bar{p}, \sigma_p)$ space is best. An example of the results is figure 5.15.

The second test is to examine the effect of k and N on the magnitudes of non-optimal scores, i.e. the rest of the space, outside of the optimum region. To do this, some method had to be found of comparing, fairly, the distribution of score values in, a $k = 3$ test set with those in a $k = 12$ set. Simply to compare, for example, the median of the 1000 scores from a non-parametric test set with $k = 3$, with the median from the non-parametric set with $k = 12$, would not be fair with regard to the respective distributions of the two sets in $(\bar{p}, \sigma_p)$ space: the former contains a greater concentration of sampling schemes at low values of σ_p than the latter, which may affect the distribution of interval widths, confounding any direct effect of k. Therefore, for this test, a smoothing method *will* be used: first the space is divided into a grid with resolution 0.01 in the $\bar{p}$ dimension and 0.005 in the σ_p dimension. The smoothed score value at each grid location is then obtained by weighted local averaging, using a Gaussian kernel with a standard deviation of 0.02 in the $\bar{p}$

dimension and 0.01 in the σ_p dimension (the kernel is truncated after ± 1.67 standard deviations, in order to limit spatially the influence of very large values). Then, only grid locations within the rectangle defined by $0.65 \leq \bar{p} \leq 0.85$ and $0.05 \leq \sigma_p \leq 0.18$ are considered—judging by figure 5.13, this is roughly the area in which all the parametric and non-parametric test sets overlap. The rectangle contains 567 grid locations or "pixels", and quantiles of the distribution of these 567 pixel values are reported in, for example, figure 5.16.

5.5.4 Threshold results

Efficiency

Results from a typical parametric test set are shown in figure 5.14, using the graphical conventions of section 5.5.3. In this example, $k = 6$, $N = 240$, and the dependent measure is the width of bootstrap intervals for threshold $t_{0.5}$—the upper panel shows the results for 68.3% intervals, and the lower panel shows results for 95.4% intervals. In each panel, a triangle marks the median position of the most efficient 1% of the test set.

Three aspects of figure 5.14 are of particular interest. First, interval width scores vary smoothly with changes in $\bar{p}$ and σ_p. Second, the surfaces are only very slightly asymmetric, and only noticeably so for 95.4% intervals: in the lower panel, when σ_p is very low (i.e. when the performance values of the scheme are grouped close together) sampling schemes with a mean performance value below the mid-point produce larger interval widths than those with mean performance value that lies an equal distance above the mid-point (the mid-point is at $\bar{p} = 0.745$, which is the performance level corresponding to threshold $t_{0.5}$). Otherwise, the surfaces are generally symmetrical, and the smallest interval widths are to be found with means close to the mid-point ($\bar{p} \approx 0.745$). This is perhaps surprising given the asymmetry of expected binomial variability about the mid-point. Nevertheless it is consistent with the results of McKee *et al.*,[5] who also found stimulus ranges centred on the mid-point to be maximally efficient. The third thing to notice about figure 5.14

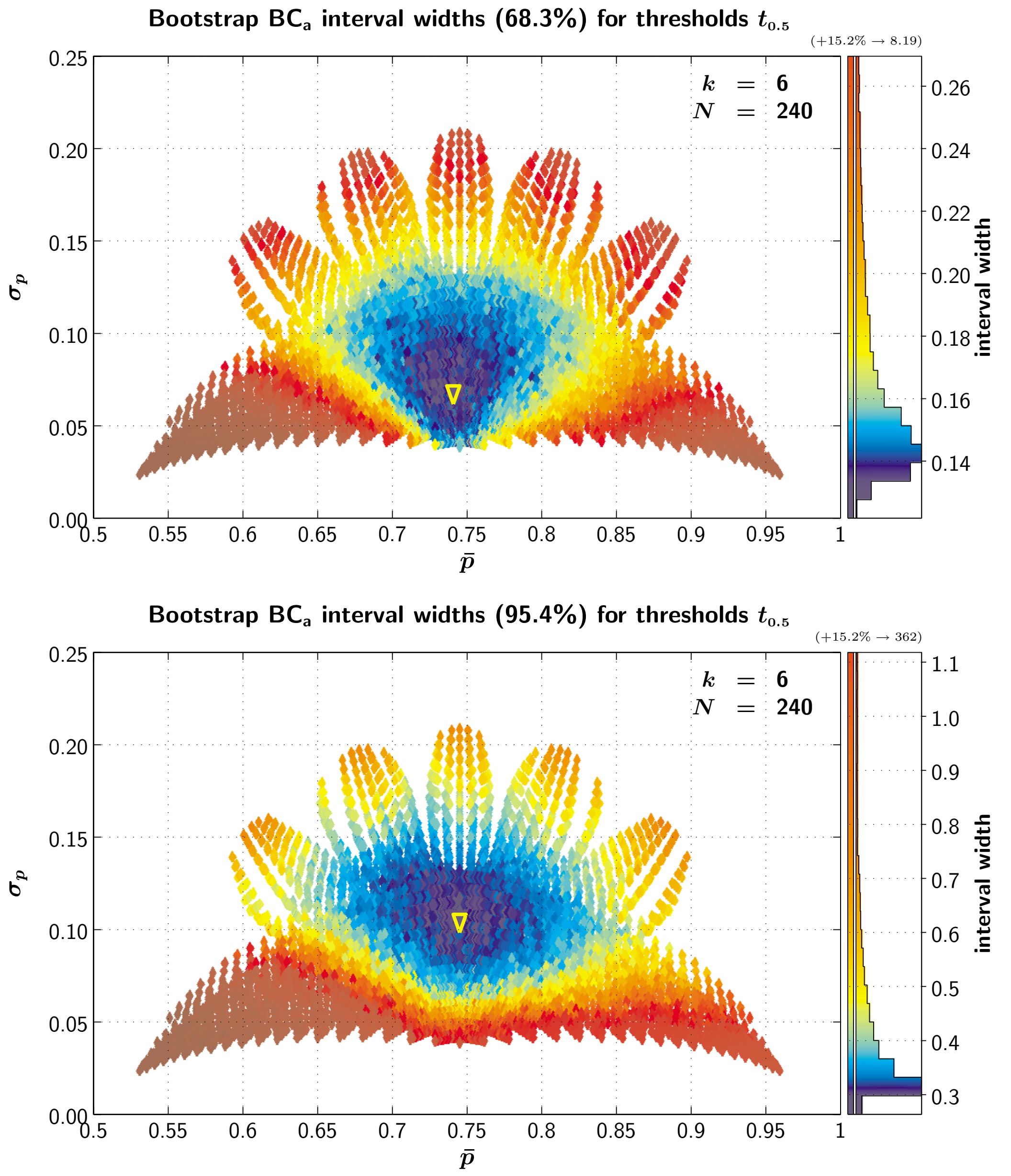


Fig. 5.14: Widths of 68.3% (upper panel) and 95.4% (lower panel) bootstrap BC$_a$ intervals for threshold $t_{0.5}$ are shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 8000 sampling schemes. Schemes were generated using the parametric method of section 5.5.1 with $k = 6$ and $n_i = 40$. The triangle marks the median position of the most efficient 1% of the test set. See sections 5.5.3 and 5.5.4 for further details.

is that the position of the minimum depends on the desired coverage of the interval. The optimal strategy is to place samples closer to the mean if one is interested in 68.3% intervals (optimum $\sigma_p \approx 0.07$), than at 95.4%, (optimum $\sigma_p \approx 0.11$). At 95.4%, the optimum strategy is to centre the stimulus values around the mid-point, but clearly *not* to target the mid-point as closely as possible: in the $k = 6$ test set, the narrowest grouping around the mid-point occurs at $(\bar{p} = 0.745, \sigma_p = 0.039)$, and this particular sampling scheme yields a bootstrap interval width of 1.69, which is a factor of about 6 larger than the minimum interval width, a ratio which corresponds to a relative efficiency of roughly 3%.

Although the above result may seem counter-intuitive when one considers that the Taylor-Creelman "ideal" is to concentrate all trials at the mid-point, it must be borne in mind that the Taylor-Creelman ideal is *only* ideal as $N \rightarrow \infty$ and the distribution of threshold values becomes perfectly normal. Thus, the following advice, from researchers over the past 30 years, should be treated as applying only in the limit, when N exceeds the values typically found in psychophysical experiments:

> "If one is interested in estimating [threshold level] X_p,... one should place observations as close to X_p as possible."
>
> *Levitt (1971)*[19]

> "Presentation on many trials far above or far below threshold is a waste of time, because the responses to such stimuli have little bearing on the question of the location of the threshold."
>
> *Emerson (1984)*[20]

> "Standard error is minimal for a normalized intensity of 0 corresponding to 50% probability of seeing [in a yes-no experiment]. This will be true for any value of n."
>
> *King-Smith and Rose (1997)*[21]

> "The optimal placement level should be equal to the mean, which is located at the 50% point for the yes/no paradigm and the 75%

> point for the 2-AFC paradigm."
>
> *Yuan (1999)*[22]

For *realistic* block sizes and numbers of blocks, the finding that a widely spread sampling scheme can produce lower threshold variability than a narrowly spaced scheme has been previously noted by O'Regan and Humbert[9] in their Monte Carlo simulation study. The benefit of sampling at very high or very low performance values was also reported by Hawley and Colburn,[23] who used an asymptotic confidence region method to plot bands similar to Finney's fiducial bands[17] around the psychometric function (see also Hawley, 1990,[24] Appendix B). The effect can readily be predicted from the probit equation for threshold limits, equation (2.5), as can the finding that the optimum spacing of samples increases as the desired coverage of the interval increases. The latter effect can be seen in table 8.1 from Finney (1971),[17] where both 95% and 99% intervals are considered for 2- and 3-point sampling schemes on a yes-no psychometric function. Indeed, in the current study, bootstrap BC$_\mathrm{a}$ interval widths are highly correlated with the probit interval lengths: their rank correlation coefficient is 0.91, indicating a high correspondence between the shapes of the surfaces defined by the two interval methods.[†] (Section 5.5.6 will examine the correspondence between interval methods in more detail, using the rank correlation coefficient measure.)

The results from Monte Carlo percentile intervals were almost identical to the bootstrap BC$_\mathrm{a}$ intervals, with a rank correlation coefficient of 0.99 for both 68.3% and 95.4% intervals. Therefore, they will not be shown separately.

Threshold results are highly consistent across variations in k. This can be seen in figure 5.15, which shows the location of the surface minimum as a function of k. Symbol size denotes N—the smallest size corresponds

[†] There were slight differences, however, which may reflect the inaccuracy of the probit method as revealed in chapter 3: the lowest probit scores were slightly larger than the corresponding BC$_\mathrm{a}$ scores, and the minimum of the surface defined by the probit scores was shifted noticeably to the left, to about $\bar{p} = 0.73$ (the former discrepancy would not be picked up by the rank correlation measure, but the latter probably accounts for most of the 9% shortfall in correlation). See section 5.5.6 for more correlation results.

to $N = 120$ and the largest to $N = 480$. Lighter symbols denote 68.3% bootstrap intervals, and darker symbols denote 95.4% intervals. As described in section 5.5.3, the minimum was located by finding the median $\bar{p}$ and median σ_p of the best 1% of the test set. The corresponding minimum interval width score, WCI_{min}, was taken to be the median of the best 1% of interval widths.

The left panel of figure 5.15 shows the $\bar{p}$ location of the minimum as a function of k, the central panel shows the σ_p location, and the right panel shows the Taylor-Creelman efficiency score computed from the WCI_{min}. Note that both efficiency and the location of the minimum are virtually independent of k. However, there are differences as N increases from 120 to 240 to 480: the optimal spread of performance values becomes narrower, as the threshold distributions become more normal and come closer to the Taylor-Creelman ideal. Accordingly, the Taylor-Creelman efficiency score increases. The effect can be seen at both the 68.3% level and the 95.4% level, although threshold distributions behave more normally (tighter sampling schemes are favoured more highly, and efficiency is higher) when only the central 68.3% is considered.

Note that in each panel of figure 5.15, the last column of results is marked "unequal"—this shows the results of the unequal block distribution method (see section 5.5.1) for which results were taken at $N = 240$ and $N = 480$ with k equal to 20. The locations of the minima are very similar to those for the equal block distribution method.

Not only does k have little effect on WCI_{min} and on the location of the minimum, it also has little effect on the distribution of interval widths as a whole. Efficiency values corresponding to quantiles $\{0.025, 0.159, 0.5, 0.841, 0.975\}$ of the distribution of bootstrap interval widths (obtained by the smoothing method described in section 5.5.3) are plotted in figure 5.16 as a function of k. From left to right, the three panels show the results for $N = 120$, $N = 240$ and $N = 480$. Again, lighter symbols denote 68.3% intervals and darker symbols denote 95.4% intervals. Circles denote the median values, upward triangles denote the higher efficiencies that correspond to quantiles

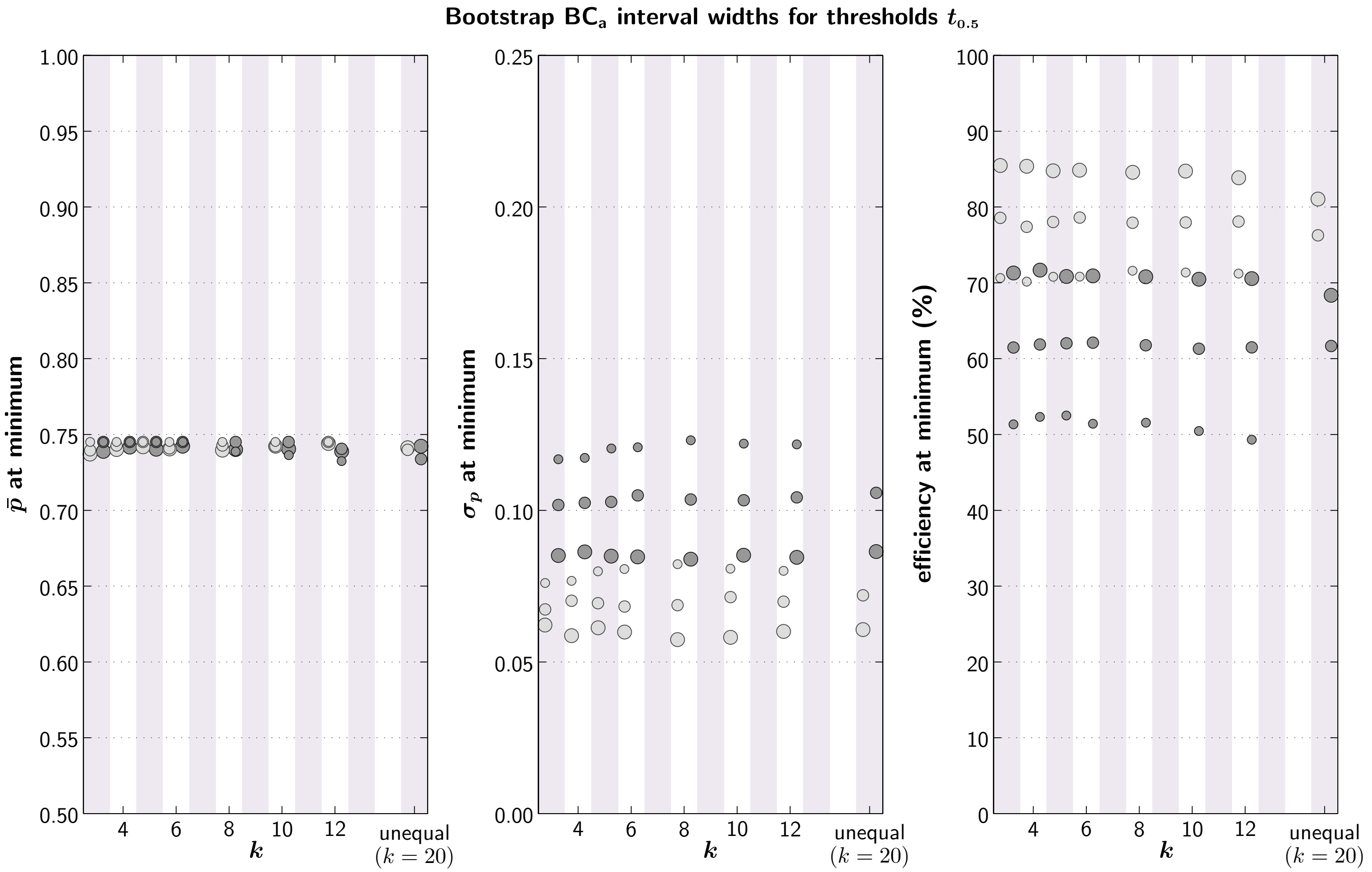


Fig. 5.15: For bootstrap BC$_\mathrm{a}$ threshold intervals in sets of parametrically generated sampling schemes (section 5.5.1), $\bar{p}$ (left panel), σ_p (central panel) and Taylor-Creelman efficiency (right panel) corresponding to WCI$_\mathrm{min}$ are shown. The number of blocks k is on the abscissa, and symbol size relates to the total number of trials N. Lighter symbols denote 68.3% intervals, and darker symbols denote 95.4% intervals. See section 5.5.4 for details.

0.025 and 0.159 of the distribution of interval widths, and downward triangles denote the lower efficiencies corresponding to quantiles 0.841 and 0.975 of the interval width distribution. Note that, as was also the case in section 5.4.1, efficiency generally improves as N increases. There are few trends to be discerned as k varies however. One possible exception is that, when $N = 480$, the worst sampling schemes (lowest chain of symbols) are more efficient at higher values of k.

What do the optimally efficient sampling schemes from the parametric test sets look like? Figure 5.17 shows the sampling schemes with the smallest 95.4% bootstrap threshold intervals in the parametric test sets—the upper, middle and lower panels shows the best schemes from the sets in which $N = 120$, 240 and 480, respectively. In all three cases, the best sampling schemes from the parametric test sets are generally evenly spaced. However, simulations from the non-parametric test sets show that the parametric system does not fully capture the possible range of sampling schemes.

The results of the non-parametric test set at $(k = 6, N = 240)$ are plotted in figure 5.18. Again, the upper panel shows the results for 68.3% $\mathrm{BC_a}$ bootstrap intervals in thresholds, and the lower panel shows the results for 95.4% intervals. The general shape of the surface is very similar to that of the parametric results in figure 5.14. However, there are a few exceptions. For example, in the lower panel of figure 5.18, the point at $(0.798, 0.131)$, coloured red, corresponds to a sampling scheme with performance values $\{0.52, 0.77, 0.84, 0.86, 0.88, 0.91\}$ which yields an interval width of 1.52. Yet this inefficient red point is surrounded by more efficient blue: for example, very close to it is the point at $(0.796, 0.134)$ which represents a very different sampling scheme: $\{0.59, 0.68, 0.75, 0.89, 0.92, 0.96\}$ yielding an interval width of 0.33. The schemes are very different despite having almost identical values of $\bar{p}$ and σ_p.

Such exceptional cases are rare. In general, the surface is nearly as smooth as that of the parametric results. Nevertheless, the exceptional cases can sometimes be among the most efficient sampling schemes. Figure 5.19 shows the best sampling schemes from the non-parametric test sets, in the same

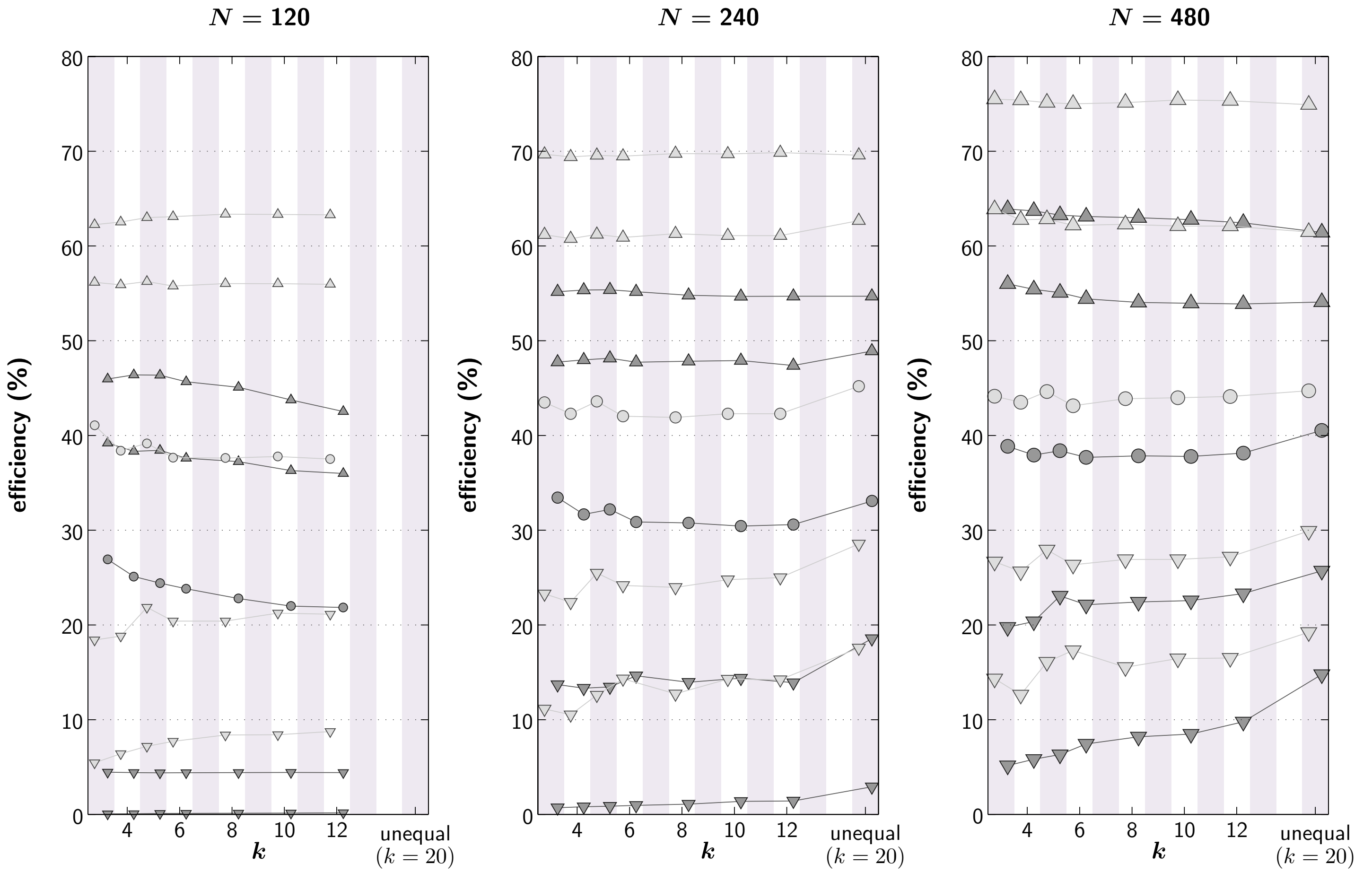


Fig. 5.16: Efficiency scores corresponding to certain quantiles of the distribution of widths of bootstrap $\mathrm{BC_a}$ intervals on thresholds $t_{0.5}$ are shown as a function of k. See sections 5.5.3 and 5.5.4 for details.

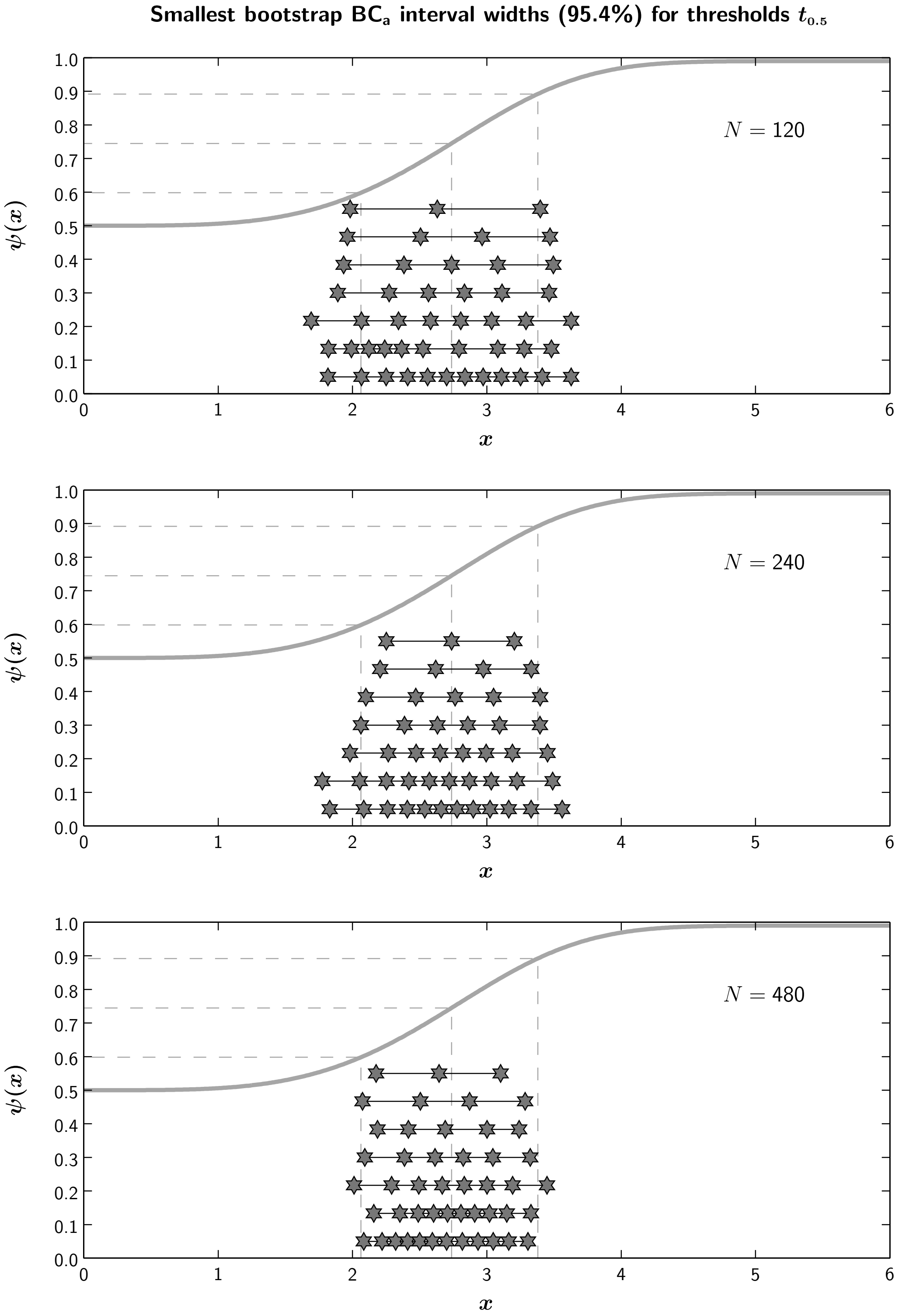


Fig. 5.17: For each set of 8000 sampling schemes generated by the parametric method (section 5.5.1), the sampling scheme with the smallest 95.4% threshold interval, obtained by the bootstrap BC$_a$ method, is shown. Results from the test sets in which $N = 120$, $N = 240$, and $N = 480$ are shown in the upper, middle and lower panels, respectively.

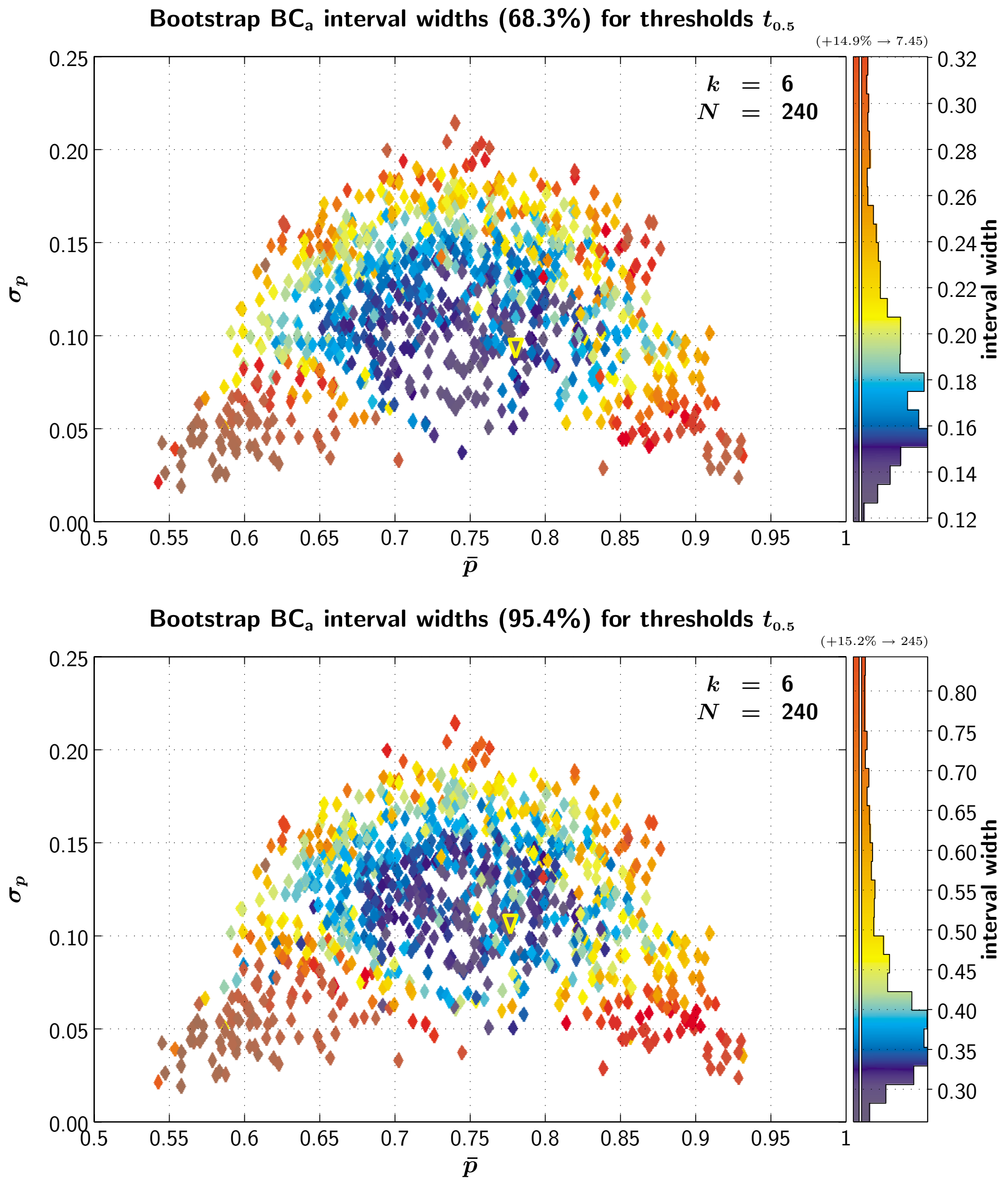


Fig. 5.18: Widths of 68.3% (upper panel) and 95.4% (lower panel) bootstrap BC$_a$ intervals for threshold $t_{0.5}$ are shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 1000 sampling schemes. Schemes were generated using the non-parametric method of section 5.5.2 with $k = 6$ and $n_i = 40$. The triangle marks the median position of the most efficient 1% of the test set. See sections 5.5.3 and 5.5.4 for further details.

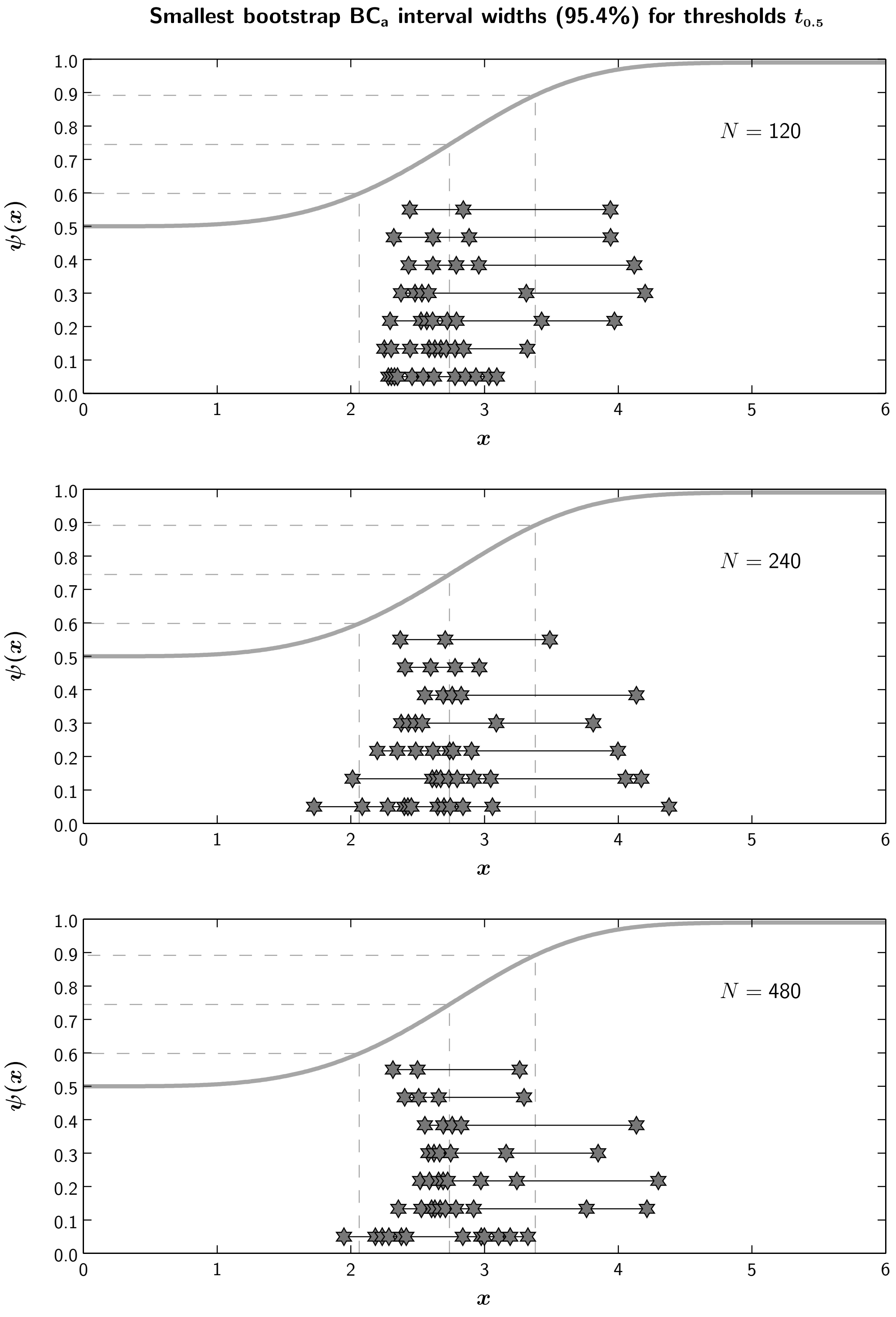


Fig. 5.19: For each set of 1000 sampling schemes generated by the non-parametric method (section 5.5.2), the sampling scheme with the smallest 95.4% threshold interval, obtained by the bootstrap BC_a method, is shown. Results from the test sets in which $N = 120$, $N = 240$, and $N = 480$ are shown in the upper, middle and lower panels, respectively.

manner as figure 5.17. The similarity in the sampling schemes with different numbers of blocks is striking: there is a tendency for the best schemes to be characterized by a single very high-performance block (or perhaps two blocks, depending on the total number), with the rest very close to each other in the region of the threshold. They are very reminiscent of Wichmann's sampling scheme **s6** (⧫ on figure 1.2). However, there is a surprising tendency to group the majority of blocks *below* the threshold, which has the effect of balancing the high point somewhat and bringing the mean performance level back towards the mid-point $\bar{p} = 0.745$, whereas Wichmann's ⧫ has a slightly higher mean value, $\bar{p} = 0.78$. The coordinates of these winning sampling schemes, along with their bootstrap interval widths, are given in columns 6–8 of table 5.1 (the preceding three columns contain the same information for the winners of the parametric test sets shown in figure 5.17). The last column of the table gives the efficiency of the best randomly-generated sampling scheme, relative to the best parametrically generated sampling scheme. Note that the best random schemes are, in most cases, a little better than the best parametric ones, particularly at $N = 120$, where they bring anything up to a 30% increase in efficiency. Even this is slight, however,[†] and the difference disappears by the time N reaches 480.

Bias

Figure 5.20 shows bias w in the estimation of threshold $t_{0.5}$, as a function of $\bar{p}$ and σ_p, for the parametric test set (upper panel) and the non-parametric test set (lower panel) in which $k = 6$ and $N = 240$.

For both parametric and non-parametric test sets, there is a prevalence of positive rather than negative bias (hence, thresholds are generally underestimated), but the magnitude of the bias was generally low. In particular, it is important to note that the location of the most efficient sampling schemes (see the lower panels of figures 5.14 and 5.18) is an area of low bias ($w < 0.2$).

[†] A relative efficiency of 130% corresponds to a 12% decrease in interval width, and the interval widths of roughly 0.4 are already very small (the widths quoted are standardized as described in section 5.3.1, so they must be considered in relation to a psychometric function with a slope of 1).

		parametric			non-parametric			relative
N	k	$\bar{p}$	σ_p	WCI$_{\mathrm{min}}$	$\bar{p}$	σ_p	WCI$_{\mathrm{min}}$	efficiency
120	3	0.73	0.13	0.431	0.80	0.12	0.393	120.7 %
	4	0.74	0.12	0.431	0.78	0.12	0.384	125.5 %
	5	0.75	0.12	0.433	0.78	0.11	0.378	131.7 %
	6	0.74	0.11	0.426	0.77	0.12	0.387	121.5 %
	8	0.74	0.13	0.422	0.76	0.10	0.394	115.0 %
	10	0.71	0.12	0.436	0.73	0.07	0.389	125.6 %
	12	0.75	0.12	0.445	0.72	0.07	0.401	123.2 %
240	3	0.74	0.09	0.280	0.77	0.11	0.282	99.1 %
	4	0.76	0.10	0.279	0.73	0.05	0.274	103.5 %
	5	0.75	0.10	0.281	0.79	0.10	0.255	121.0 %
	6	0.75	0.10	0.281	0.75	0.11	0.271	107.5 %
	8	0.74	0.10	0.282	0.74	0.10	0.266	112.5 %
	10	0.73	0.11	0.285	0.78	0.11	0.268	112.8 %
	12	0.74	0.11	0.276	0.72	0.11	0.275	101.0 %
480	3	0.72	0.09	0.186	0.73	0.10	0.186	100.9 %
	4	0.74	0.10	0.188	0.74	0.08	0.185	102.8 %
	5	0.74	0.09	0.189	0.79	0.10	0.177	114.1 %
	6	0.74	0.09	0.187	0.78	0.09	0.182	105.1 %
	8	0.75	0.10	0.185	0.78	0.09	0.186	99.0 %
	10	0.75	0.08	0.185	0.77	0.10	0.185	99.8 %
	12	0.73	0.09	0.187	0.73	0.10	0.193	94.1 %

Table 5.1: The coordinates $(\bar{p}, \sigma_p)$ and the interval width score are given for the sampling scheme with the smallest 95.4% bootstrap threshold interval in each set of 8000 sampling schemes generated by the parametric method of section 5.5.1 (columns 3–5), and in each set of 1000 sampling schemes generated by the non-parametric method of section 5.5.2 (columns 6–8). The last column gives the efficiency of the best non-parametric scheme relative to the best corresponding parametric scheme.

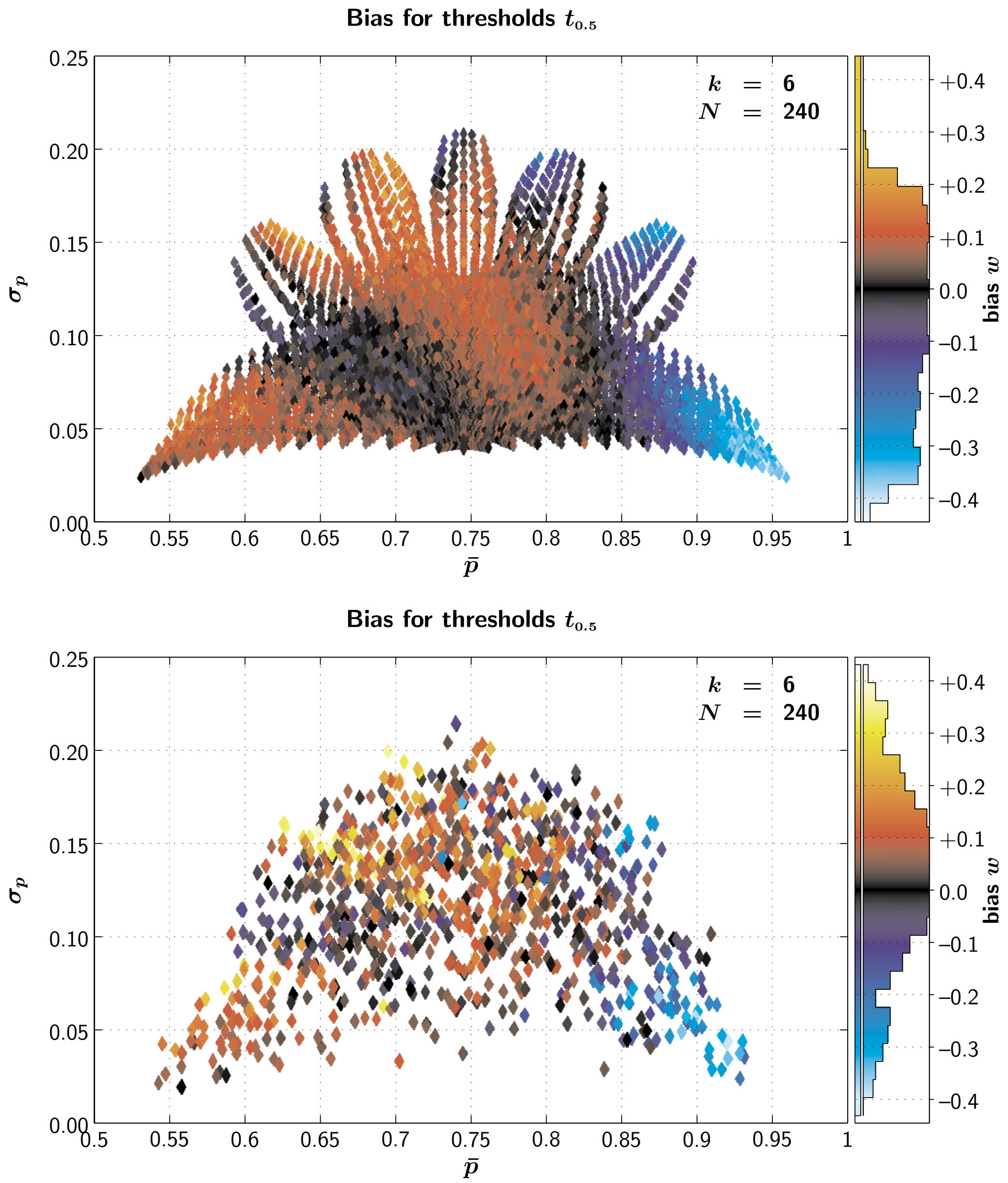


Fig. 5.20: Bias in the estimation of threshold $t_{0.5}$ is shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 8000 parametrically generated sampling schemes (upper panel) and 1000 non-parametrically generated sampling schemes. All schemes had $k = 6$ and $n_i = 40$. See sections 5.5.3 and 5.5.4 for further details.

In the parametric test sets, no sampling schemes in this region had a value of w exceeding 0.25, although there are a few schemes with large biases (yellow points) in the non-parametric sets—these are all schemes with $\bar{p} < 0.745$, so it is clearly better to place sample points a little too high, rather than a little too low, to avoid bias.

Figure 5.21 shows, as a function of k, quantiles $\{0.025, 0.159, 0.5, 0.841, 0.975\}$ of the distribution of bias values obtained by the smoothing method described in section 5.5.3. The magnitude of bias shows little effect of k or of N. (When $N = 120$, there is a slight tendency for the distribution to be more skewed towards negative bias scores as k increases, but in any case the absolute values of the bias scores involved, even at the extreme ends of the distribution, are small).

5.5.5 Slope results

Efficiency

Bootstrap $\mathrm{BC_a}$ interval widths for slopes are plotted in figure 5.22 for the parametric test set with $k = 6$ and $N = 240$, and in figure 5.23 for the non-parametric test set with the same values of k and N. As before, the upper panels show 68.3% intervals and the lower panels show 95.4% intervals.

The parametric results are smooth, as they were for thresholds. This time, however, the surfaces are distinctly asymmetric, with the most efficient sampling schemes generally being those with higher mean performance levels. The narrower the spread of the sampling scheme σ_p, the higher the optimum mean level $\bar{p}$, hence the diagonal boundary between efficient and inefficient schemes. For 68.3% intervals, there is no ceiling, among the sampling schemes studied, on the desirable spread of performance values. For 95.4% intervals, on the other hand, the minimum occurs at a slightly lower value of σ_p, larger values being less efficient.

The non-parametric results of figure 5.23 show a similar general trend, with the most efficient sampling schemes lying in the same general area of the space, but there is a great deal of “noise” corresponding to those sam-

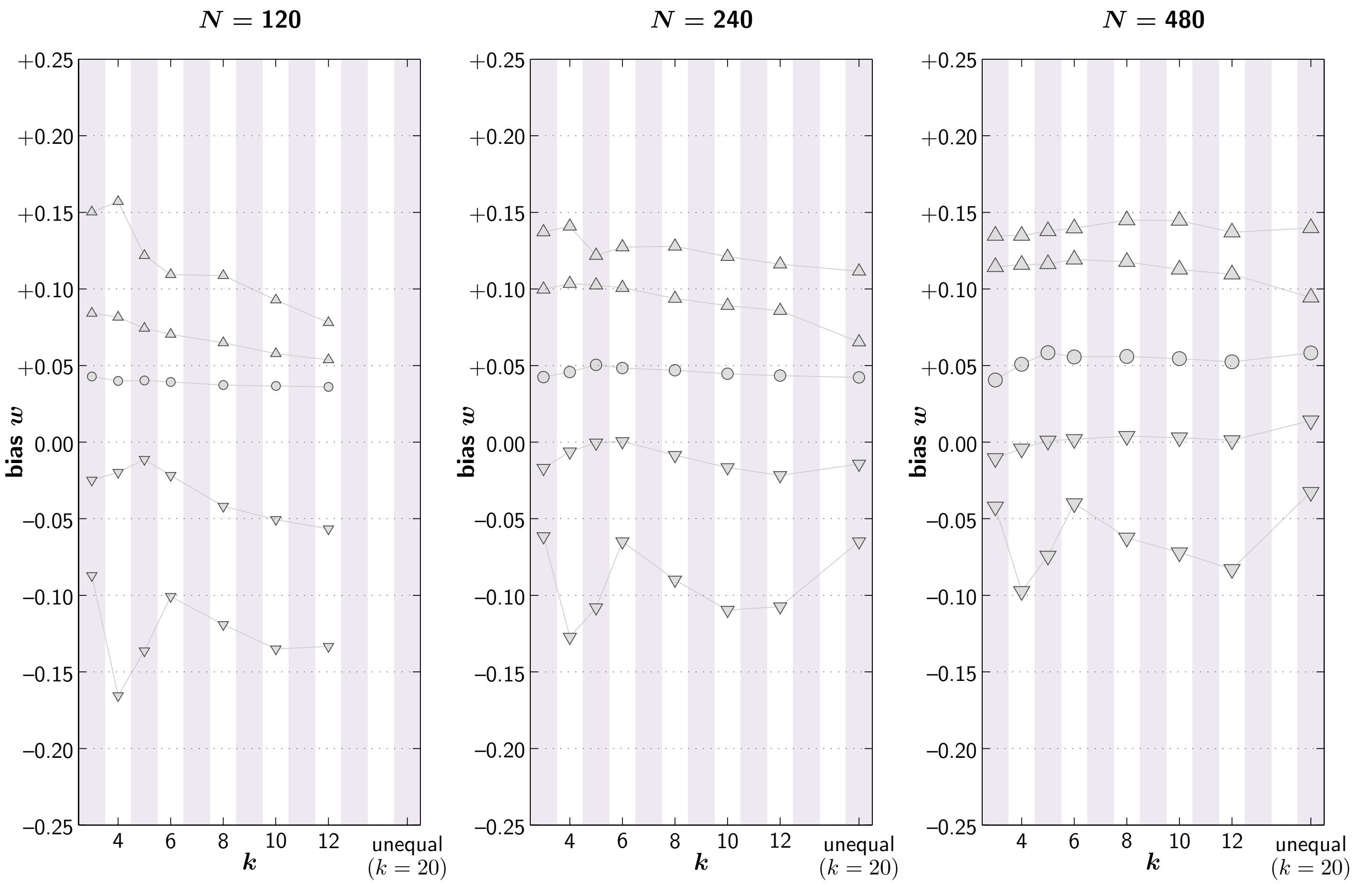


Fig. 5.21: Certain quantiles of the distributions of bias in the estimation of threshold $t_{0.5}$ are shown as a function of k. See sections 5.5.3 and 5.5.4 for details.

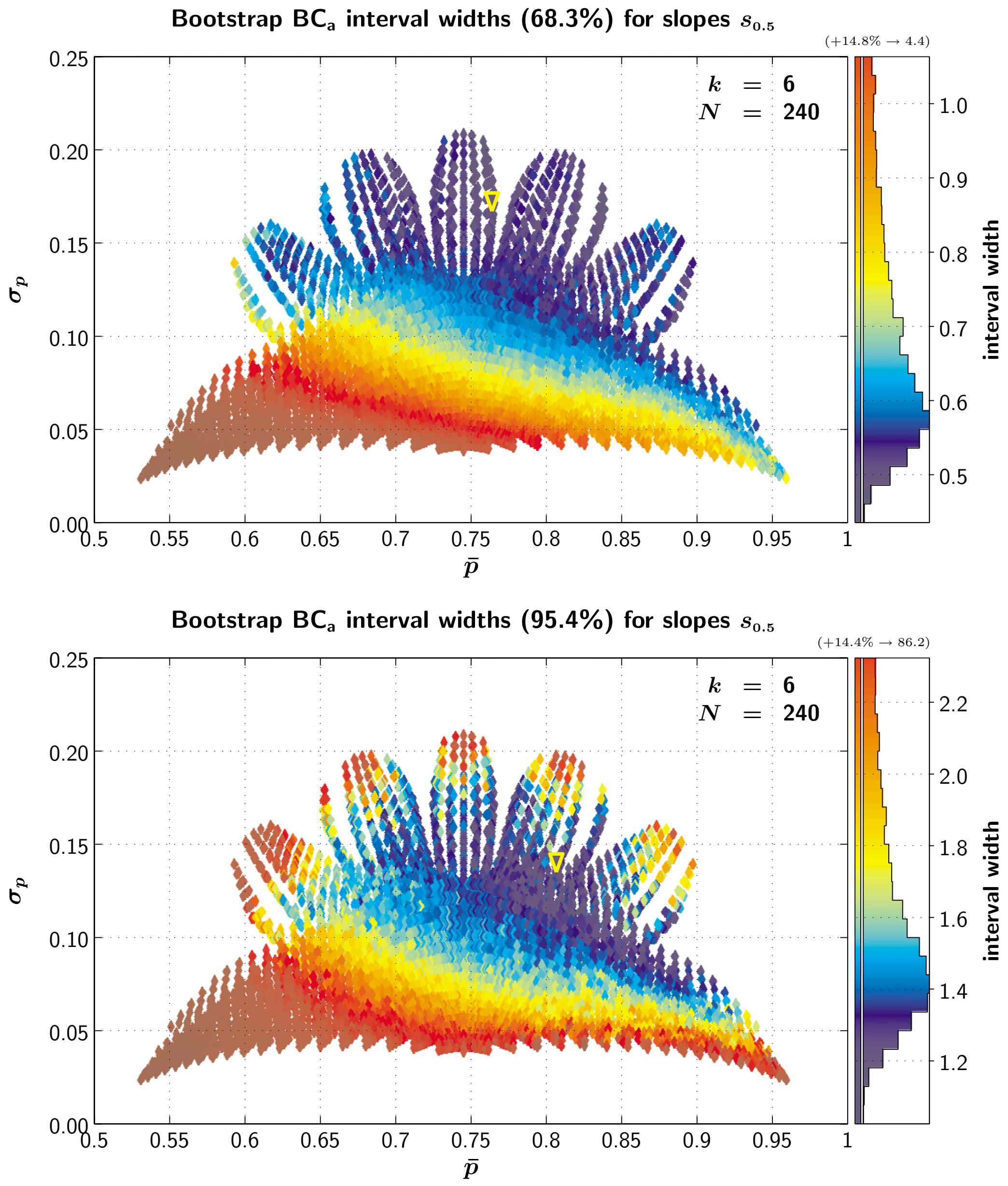


Fig. 5.22: Widths of 68.3% (upper panel) and 95.4% (lower panel) bootstrap BC$_a$ intervals for slope $s_{0.5}$ are shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 8000 sampling schemes. Schemes were generated using the parametric method of section 5.5.1 with $k = 6$ and $n_i = 40$. The triangle marks the median position of the most efficient 1% of the test set. See sections 5.5.3 and 5.5.5 for further details.

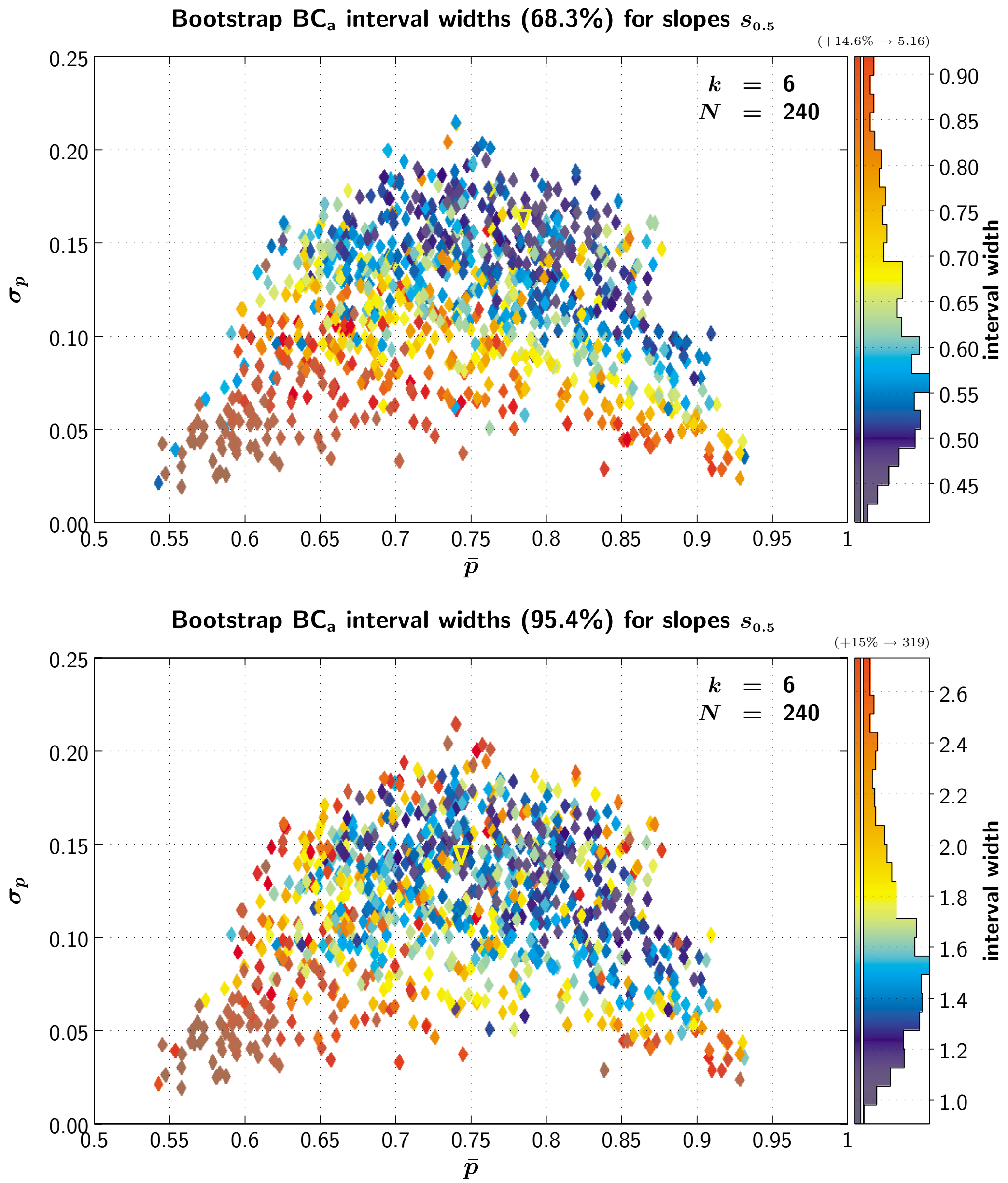


Fig. 5.23: Widths of 68.3% (upper panel) and 95.4% (lower panel) bootstrap BC$_a$ intervals for slope $s_{0.5}$ are shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 1000 sampling schemes. Schemes were generated using the non-parametric method of section 5.5.2 with $k = 6$ and $n_i = 40$. The triangle marks the median position of the most efficient 1% of the test set. See sections 5.5.3 and 5.5.5 for further details.

pling schemes that the parametric system does not adequately capture. The mapping of sampling schemes using the coordinates $(\bar{p}, \sigma_p)$ is therefore not ideal for representing the efficiency of slope estimation—a more sophisticated system is required. Unfortunately the parametric system of section 5.5.1 does not fill this requirement, for it only includes schemes that make up a smooth surface in $(\bar{p}, \sigma_p)$ coordinates.

Some of the differences between the parametric and non-parametric systems can also be seen by comparing figures 5.24 and 5.25, which show the most efficient parametrically and non-parametrically generated sampling schemes, respectively. As was also the case for thresholds, the best parametric schemes are more evenly spaced than the best non-parametric schemes, although there is perhaps more correspondence between the results from the two generation methods than there was for thresholds. The non-parametric schemes have a tendency to divide their sample points between two clusters, at performance levels of around 0.65 and 0.92, but there is also a tendency to add a single block at a higher performance level, around 0.98 or 0.99 (N.B. as we saw in section 5.4.2, and will see again in section 5.5.5, schemes containing very high performance values are particularly prone to bias).

Table 5.2 gives the coordinates $(\bar{p}, \sigma_p)$ and the interval widths $\textsc{wci}_{\min}$, for the schemes shown in figures 5.24 and 5.25, as well as the efficiency of the best non-parametric scheme relative to the best corresponding parametric scheme. The difference in efficiency between the two sets is generally larger than it was for thresholds, and it does not disappear as N increases. However, it still represents only a small benefit, of around 10–15% in terms of absolute interval width.

Note that there is considerably more apparently random variation in $\bar{p}$ and σ_p than there was in the threshold results of table 5.1. This can also be seen in the locations of the minima found by taking the median positions of the best 1%: figure 5.26 shows the relationship between k and the location of the minimum to be somewhat noisy. Both the location of the minimum and the efficiency of $\textsc{wci}_{\min}$, seem to be more stable with respect to variation in N when k is large than when k is small. Most such instability is confined to the

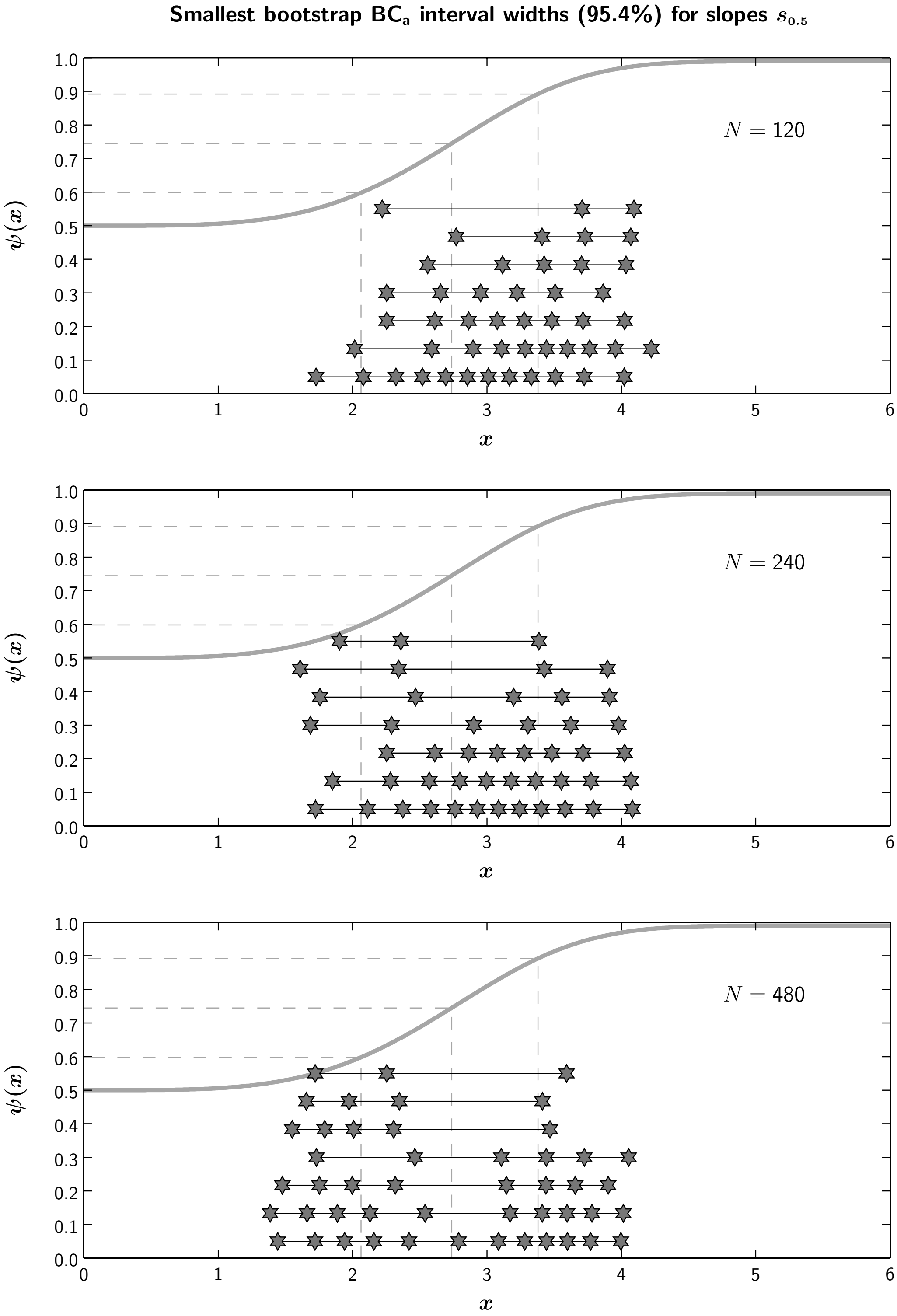


Fig. 5.24: For each set of 8000 sampling schemes generated by the parametric method (section 5.5.1), the sampling scheme with the smallest 95.4% slope interval, obtained by the bootstrap BC$_a$ method, is shown. Results from the test sets in which $N = 120$, $N = 240$, and $N = 480$ are shown in the upper, middle and lower panels, respectively.

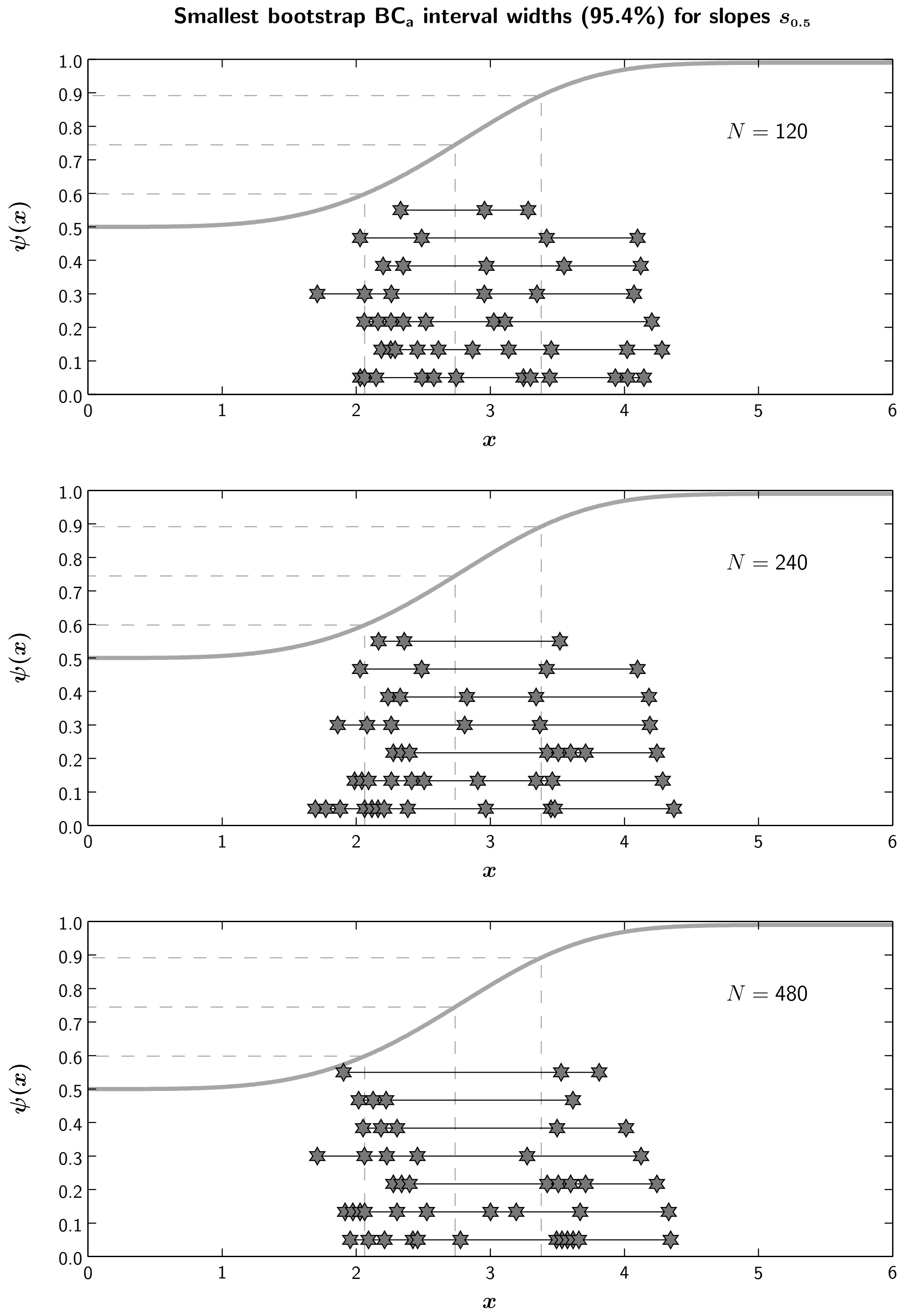


Fig. 5.25: For each set of 1000 sampling schemes generated by the non-parametric method (section 5.5.2), the sampling scheme with the smallest 95.4% slope interval, obtained by the bootstrap $\mathrm{BC_a}$ method, is shown. Results from the test sets in which $N = 120$, $N = 240$, and $N = 480$ are shown in the upper, middle and lower panels, respectively.

		parametric			non-parametric			relative
N	k	$\bar{p}$	σ_p	$\mathrm{WCI}_{\mathrm{min}}$	$\bar{p}$	σ_p	$\mathrm{WCI}_{\mathrm{min}}$	efficiency
120	3	0.85	0.16	1.603	0.77	0.09	1.347	141.6 %
	4	0.89	0.08	1.495	0.79	0.16	1.320	128.1 %
	5	0.87	0.10	1.512	0.80	0.14	1.276	140.5 %
	6	0.82	0.11	1.548	0.74	0.16	1.392	123.6 %
	8	0.83	0.11	1.552	0.73	0.13	1.337	134.8 %
	10	0.85	0.12	1.627	0.78	0.13	1.387	137.8 %
	12	0.78	0.13	1.610	0.79	0.14	1.368	138.5 %
240	3	0.71	0.14	1.010	0.73	0.13	0.917	121.4 %
	4	0.76	0.17	1.074	0.79	0.16	0.965	124.0 %
	5	0.80	0.15	1.084	0.78	0.13	0.960	127.6 %
	6	0.79	0.15	1.050	0.74	0.15	0.944	123.7 %
	8	0.83	0.11	1.034	0.83	0.14	0.891	134.6 %
	10	0.81	0.13	1.061	0.73	0.14	0.957	123.0 %
	12	0.79	0.13	1.042	0.70	0.15	0.967	116.0 %
480	3	0.70	0.16	0.628	0.82	0.17	0.587	114.1 %
	4	0.67	0.14	0.712	0.69	0.14	0.614	134.4 %
	5	0.65	0.14	0.754	0.75	0.16	0.685	121.1 %
	6	0.81	0.15	0.752	0.72	0.15	0.698	116.0 %
	8	0.74	0.17	0.761	0.83	0.14	0.654	135.1 %
	10	0.75	0.17	0.754	0.73	0.15	0.687	120.4 %
	12	0.76	0.16	0.758	0.79	0.15	0.672	127.0 %

Table 5.2: The coordinates $(\bar{p}, \sigma_p)$ and the interval width score are given for the sampling scheme with the smallest 95.4% bootstrap slope interval in each set of 8000 sampling schemes generated by the parametric method of section 5.5.1 (columns 3–5), and in each set of 1000 sampling schemes generated by the non-parametric method of section 5.5.2 (columns 6–8). The last column gives the efficiency of the best non-parametric scheme relative to the best corresponding parametric scheme.

region where $k \leq 5$. Of particular interest is the interaction between k and N with regard to optimum efficiency: when $N = 120$, there is an advantage to sampling with 6 blocks or more whereas, at $N = 240$ and $N = 480$, 3- or 4-point sampling is more efficient (this suggests that the best strategy would be to use a large number of blocks in the early stages of the experiment and then "home in" on a small number of points once a certain number of trials, perhaps 200, have been taken). There is little relative advantage between larger values of k (in the region where $k \geq 6$), and results from the unequal block distribution method are similar to those for the equal block distribution method with large k.

For 95.4% intervals, the optimal spread of performance values, σ_p, increases as N increases (this is the reverse of the trend observed for thresholds). For 68.3% intervals, the optimal value for σ_p is more or less constant at around 0.17 or 0.18. This suggests that the increasing trend in σ_p for 95.4% intervals may asymptote at this value, because as N increases, we can expect increasing similarity between the results for 68.3% and those for 95.4% as the likelihood distribution of slopes becomes more normal. For thresholds, the optimal values for σ_p for 68.3% and 95.4% intervals converge at 0 as $N \to \infty$, corresponding to the Taylor-Creelman ideal of sampling at a single point. For slopes they appear, from the middle panel of figure 5.26 to converge at around 0.17 or 0.18, a figure which is consistent with the "ideal" 2-point sampling scheme found in section 5.3.1: for the ideal pair $\boldsymbol{p} = \{0.61, 0.965\}$, $\bar{p} = 0.788$ and $\sigma_p = 0.178$.

The pattern of figure 5.26 is echoed in figure 5.27, which shows quantiles $\{0.025, 0.159, 0.5, 0.841, 0.975\}$ of the distribution of bootstrap slope intervals, obtained using the smoothing method described in section 5.5.3, as a function of k. There is little change in the middle and lower parts of the distributions as k varies, but the more efficient sampling schemes, like the optimum of figure 5.26 show an advantage for low k. Note also the general increase in efficiency as N increases (left → centre → right panel): this is noticeable for 95.4% intervals, but negligible for 68.3% intervals, which are already, for $N \geq 120$, behaving in the normal-ideal manner.

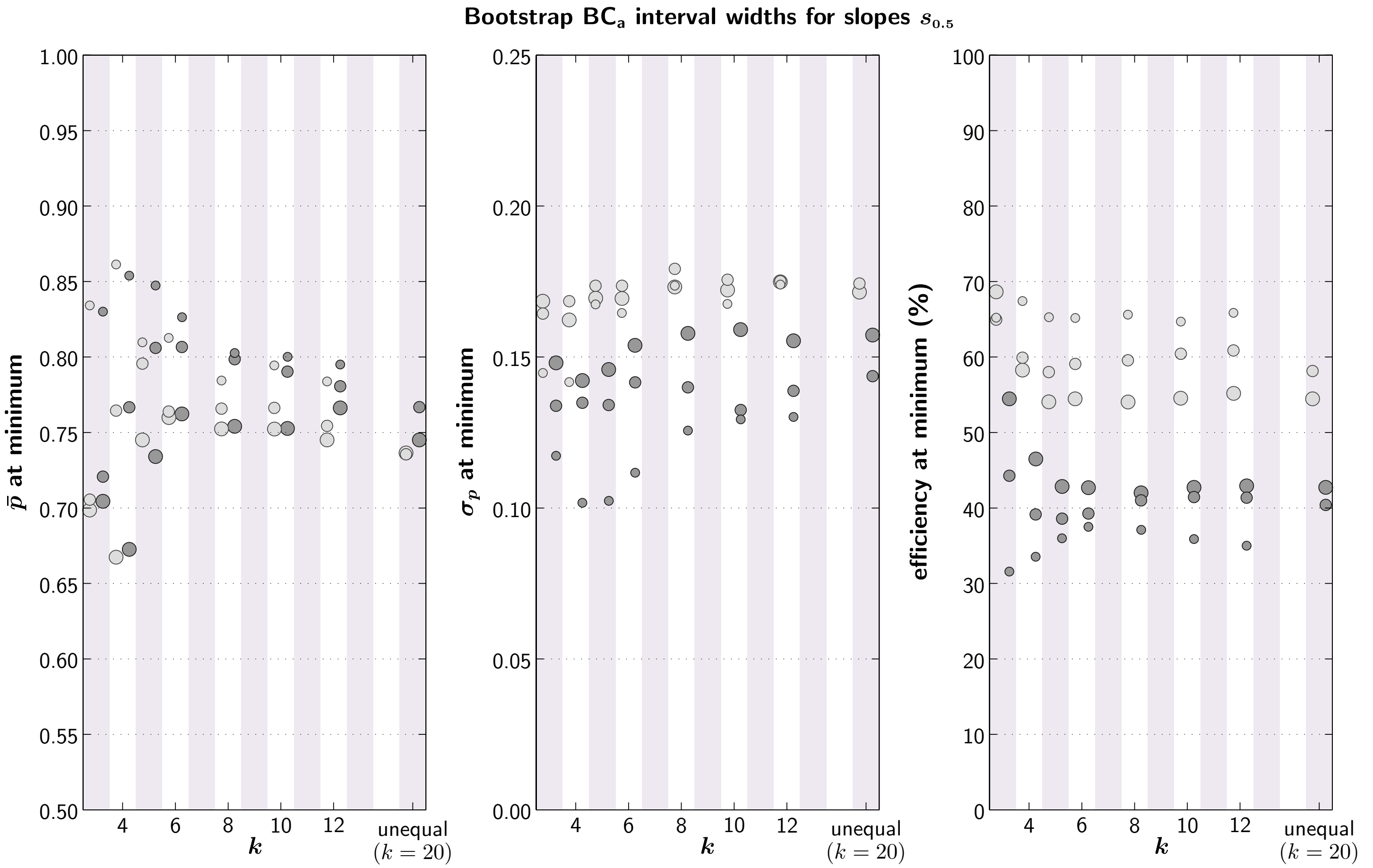


Fig. 5.26: For bootstrap BC$_a$ slope intervals in sets of parametrically generated sampling schemes (section 5.5.1), $\bar{p}$ (left panel), σ_p (central panel) and Taylor-Creelman efficiency (right panel) corresponding to WCI$_{\text{min}}$ are shown. The number of blocks k is on the abscissa, and symbol size relates to the total number of trials N. Lighter symbols denote 68.3% intervals, and darker symbols denote 95.4% intervals. See section 5.5.5 for details.

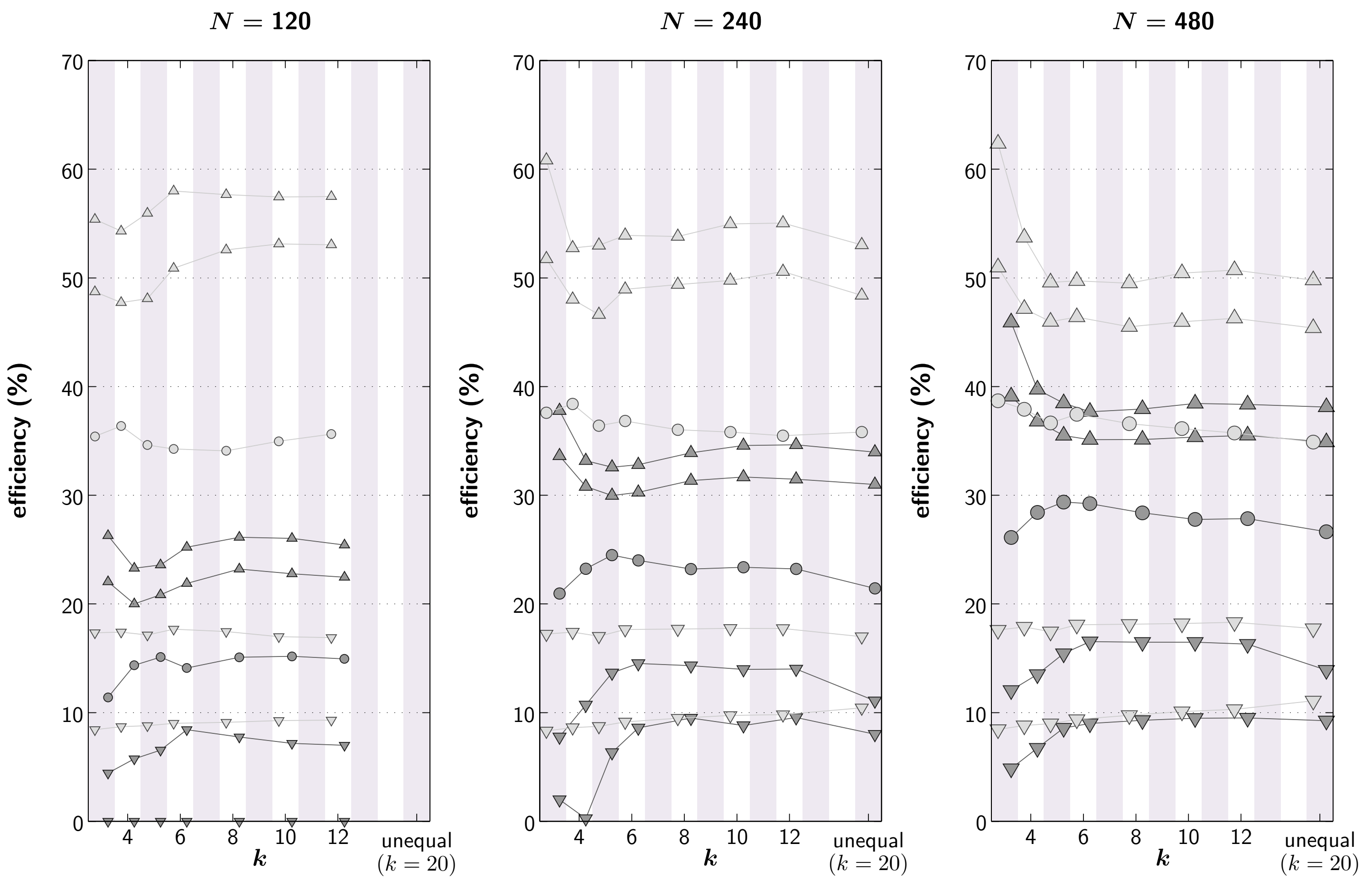


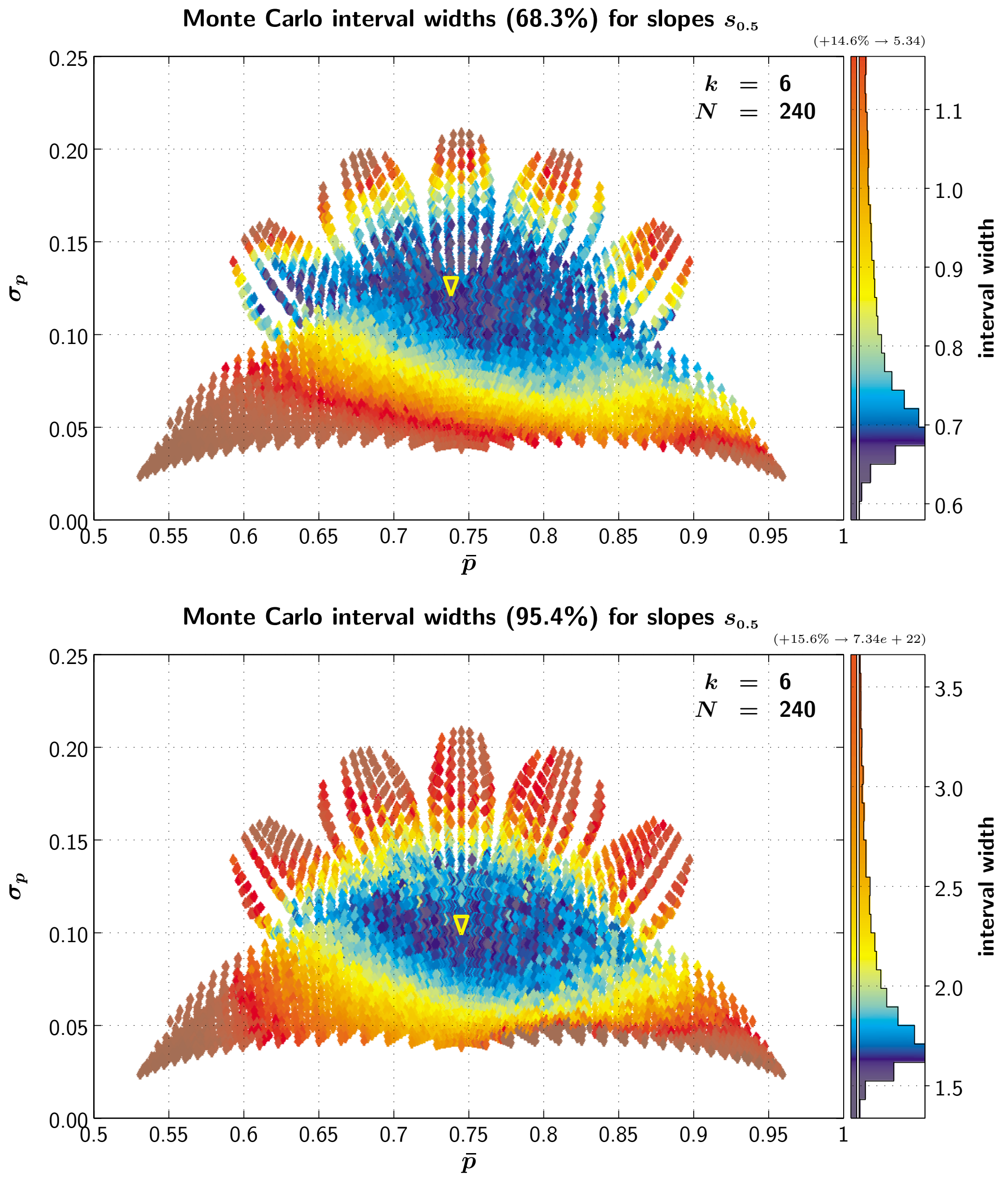
Fig. 5.27: Efficiency scores corresponding to certain quantiles of the distribution of widths of bootstrap $\mathrm{BC_a}$ intervals on slopes $s_{0.5}$ are shown as a function of k. See sections 5.5.3 and 5.5.5 for details.

One further thing to note about the slope interval widths is that the general tendency towards negatively biased slope estimates (see section 5.5.5) means that the bias-corrected ($\mathrm{BC_a}$) intervals on slopes are different from the Monte Carlo percentile intervals, and in fact they give different information about the relative efficiency of sampling schemes at different points in the space. Figure 5.28 shows 68.3% and 95.4% Monte Carlo intervals for the example test set ($k = 6$, $N = 240$), in the upper and lower panels, respectively. Comparison with figure 5.22 shows that the optimum sampling scheme has rather different characteristics, depending on whether one takes Monte Carlo interval width or bootstrap interval width as the criterion. This is reflected in the low rank correlation coefficient between interval widths computed by the two methods: in this set, for 95.4% intervals, the coefficient is 0.54 (from table 5.4). Note also, in table 5.4, that intervals from the probit standard error method are better correlated with the bootstrap intervals than with the Monte Carlo intervals. The bootstrap intervals should be taken as the more reliable guide to efficiency when the underlying psychometric function is not known, given that both the $\mathrm{BC_a}$ method and the probit standard error method were found to have better coverage properties than unadjusted percentile intervals (compare performance of the $\mathrm{BC_a}$ and probit methods of figure 3.22 with that of the bootstrap percentile method in figure 3.8).

Bias

Figure 5.29 shows bias w in the estimation of slope $s_{0.5}$, as a function of $\bar{p}$ and σ_p, for the parametric test set (upper panel) and the non-parametric test set (lower panel) in which $k = 6$ and $N = 240$.

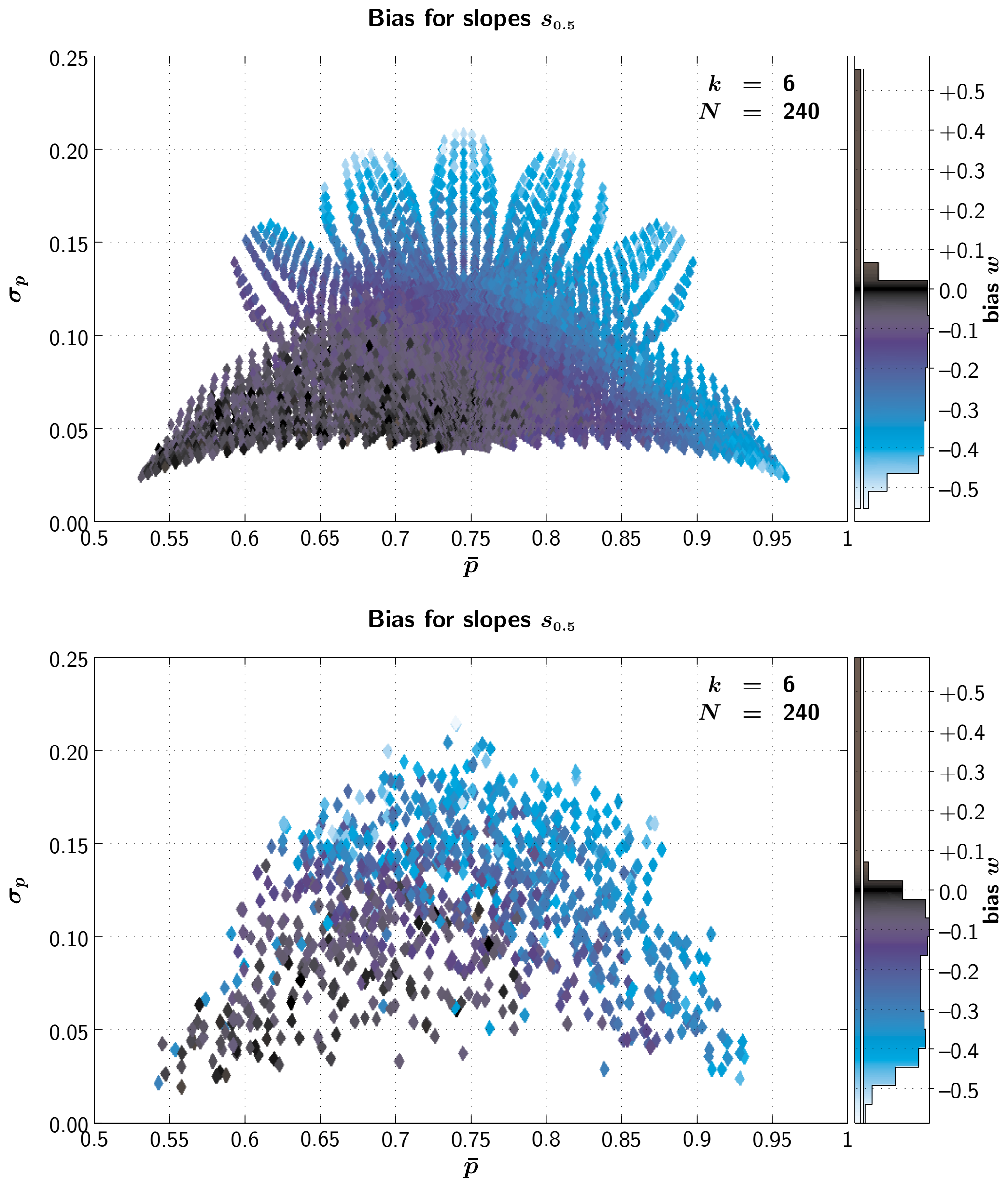
Slope is consistently overestimated by almost all the sampling schemes tested, leading to negative w values at nearly all points on both the parametric and non-parametric surfaces. This is likely to be the same effect as that previously discussed in section 5.4.2, and can be attributed to the difficulty in obtaining accurate estimates of λ when $\lambda_{\mathrm{gen}} = 0.01$. The least biased schemes are those with a low mean performance value and very closely spaced sampling, similar to ◀, which is the least biased scheme in figure 5.9. As



Monte Carlo interval widths (68.3%) for slopes $s_{0.5}$

Monte Carlo interval widths (95.4%) for slopes $s_{0.5}$

Fig. 5.28: Widths of 68.3% (upper panel) and 95.4% (lower panel) Monte Carlo percentile intervals for slope $s_{0.5}$ are shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 8000 sampling schemes. Schemes were generated using the parametric method of section 5.5.1 with $k = 6$ and $n_i = 40$. The triangle marks the median position of the most efficient 1% of the test set. See sections 5.5.3 and 5.5.5 for further details.


Bias for slopes $s_{0.5}$

Bias for slopes $s_{0.5}$

Fig. 5.29: Bias in the estimation of slope $s_{0.5}$ is shown as a function of the weighted mean $\bar{p}$ and weight standard deviation σ_p of the performance values comprising each of 8000 parametrically generated sampling schemes (upper panel) and 1000 non-parametrically generated sampling schemes. All schemes had $k = 6$ and $n_i = 40$. See sections 5.5.3 and 5.5.5 for further details.

before, the least biased schemes are the ones with the largest interval widths with regard to slopes.

Comparing the upper panel of figure 5.29 and the lower panel of figure 5.14, it appears that area of optimum sampling with regard to 95.4% threshold intervals is *just* within the area in which slope bias is acceptable ($w > -0.25$). If accurate slope estimation is important, it may be advisable to err on the side of too-narrow rather than too-wide spacing. Comparison with the lower panel of figure 5.14 reveals the unfortunate fact that optimally efficient sampling, with regard to 95.4% slope intervals, cannot be unbiased, at least when $\lambda_{\text{gen}} = 0.01$.

Additional simulations were conducted with the logistic function ($\alpha_{\text{gen}} = 0, \beta_{\text{gen}} = 1$), using the non-parametric generation method with $k = 6$ and $N = 240$, under three conditions: $\lambda_{\text{gen}} = 0.01$, $\lambda_{\text{gen}} = 0.025$ and an "idealized" condition in which $\lambda_{\text{gen}} = 0$ and λ was fixed at 0 during fitting. Results from the first set produced results that were indistinguishable, on any measure of bias or efficiency, from the corresponding Weibull set. Results from the second and third conditions produced almost identical patterns of slope bias to the first, but with reduced magnitude: when λ_{gen} was equal to 0.025, slope bias was reduced by a factor of about 2. When λ_{gen} was known exactly, bias was reduced by a factor of about 3 (although it is interesting to note that, even in the idealized condition when the bias values were very low, they were still nearly all negative). These results are consistent with the idea that it is mis-estimation of λ, which is particularly problematic when the unknown underlying value λ_{gen} takes a very low value such as 0.01, that is the root cause of the slope bias problem. Results from the three logistic conditions will not be shown here, but can be found in the results archive under:

- `simulations/optimal/2AFC/nonpara/l01/logistic/`
- `simulations/optimal/2AFC/nonpara/l025/logistic/`
- `simulations/optimal/2AFC/nonpara/l0f0/logistic/`

Figure 5.30 shows the effect of k and N on the distribution of smoothed bias values, obtained by the method described in section 5.5.3. There is some improvement as N increases, with each distribution contracting towards 0. There is also a noticeable effect of k: the median the distribution moves towards 0 as k decreases, although the distribution retains its long negative tail. Given that efficient sampling of either threshold or slope requires the use of a non-optimal sampling scheme with regard to slope bias, this indicates that it is sensible to divide the total number of observations into a smaller rather than a larger number of blocks (fewer than, say, 6) when an unbiased estimate of slope is important.

5.5.6 Comparison with probit methods

In addition to the bootstrap $\mathrm{BC_a}$ and Monte Carlo percentile intervals discussed above, 95.4% probit interval widths were also calculated for each of the parametric and non-parametric test sets described above, using equation (2.5) for thresholds and equation (2.4) for slopes.

Previous studies that have tested the validity of probit threshold intervals using Monte Carlo simulation[5,25] have generally them to be fairly accurate, even in 2-AFC,[5] once N exceeds about 100. Nevertheless, a considerable improvement was found in the current study, between $N = 120$ and $N = 240$.

The measure used was not the correspondence between the actual interval widths produced by the probit method and those produced by simulation, but rather the correspondence between the way in which the probit and bootstrap methods rank sampling schemes *relative to each other*. First, sampling schemes were discarded wherever they produced invalid results on either of the two interval width scores to be compared—for example in the $(k = 6,\ N = 240)$ parametric test set, 1295 of the 8000 results were removed because of undefined probit limits, and all $\mathrm{BC_a}$ results were valid, so no additional results had to be removed. Then, the rank correlation coefficient was computed between the two sets of interval widths from the remaining results, indicating the degree of correspondence between the surfaces defined by the two methods.

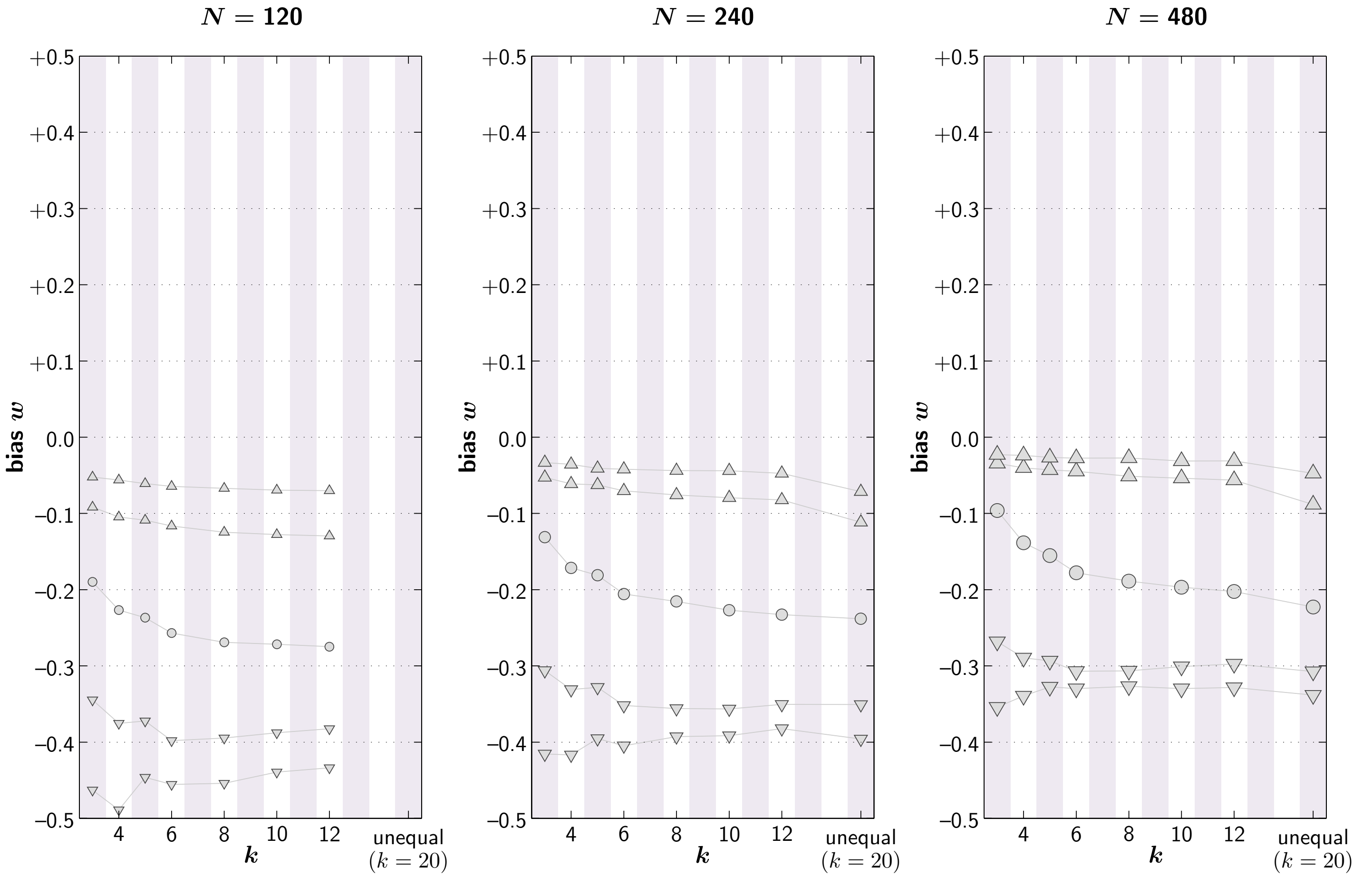


Fig. 5.30: Certain quantiles of the distributions of bias in the estimation of slope $s_{0.5}$ are shown as a function of k. See sections 5.5.3 and 5.5.5 for details.

Table 5.3 shows the coefficients relating to 95.4% intervals on threshold $t_{0.5}$. B ↔ P denotes the correlation between bootstrap and probit intervals, M ↔ P denotes the correlation between Monte Carlo and probit intervals, and B ↔ M denotes the correlation between bootstrap and Monte Carlo intervals. As was mentioned in section 5.5.4, there is almost perfect correspondence between the Monte Carlo and bootstrap $\mathrm{BC_a}$ results. Probit calculations correspond quite well with the $\mathrm{BC_a}$ results, with a coefficient of at about 0.7 or more. There is a considerable increase in correlation, of about 10–15%, between $N = 120$ and $N = 240$.

Table 5.4 shows the coefficients for 95.4% slope intervals. As discussed in section 5.5.5, the correlation between the Monte Carlo and bootstrap $\mathrm{BC_a}$ results is poor, although there is some improvement as N increases. At $N = 120$ and $N = 240$, there is very little correlation between the probit standard errors and Monte Carlo percentile intervals, but probit results are fairly well correlated with the bootstrap $\mathrm{BC_a}$ results. This is consistent with the results of chapter 3, in which it was found that both the probit standard error method and the $\mathrm{BC_a}$ method were better, for slopes, than the unadjusted percentile method.

5.5.7 Summary

With regard to thresholds, the optimally efficient sampling strategy was found to be to centre sample points on the threshold, despite the asymmetry of expected variability in a 2-AFC context. However, for a realistic total number of trials, the best strategy is not to try to group sample points as closely as possible around the threshold. Rather, the optimal spacing increases as the desired coverage of the interval increases, or as N decreases, both of which effects are predicted by the probit formulae. The strategy of placing a single block at a high performance level, and the rest of the blocks grouped close together slightly below threshold, was found to be the most efficient of the schemes studied. However, the gain in efficiency, relative to a (more conventional) evenly spaced sampling scheme of optimal spread, was found to be small.

		parametric			non-parametric		
N	k	B ↔ P	M ↔ P	B ↔ M	B ↔ P	M ↔ P	B ↔ M
120	3	0.68	0.78	0.99	0.66	0.74	0.98
	4	0.73	0.81	0.99	0.70	0.76	0.97
	5	0.76	0.82	0.99	0.66	0.69	0.97
	6	0.78	0.83	0.99	0.66	0.69	0.96
	8	0.81	0.85	0.98	0.68	0.66	0.95
	10	0.82	0.85	0.98	0.70	0.66	0.93
	12	0.83	0.85	0.97	0.67	0.58	0.92
240	3	0.90	0.93	0.99	0.85	0.87	0.99
	4	0.90	0.93	0.99	0.83	0.85	0.99
	5	0.91	0.93	0.99	0.83	0.84	0.98
	6	0.91	0.93	0.99	0.82	0.83	0.97
	8	0.91	0.93	0.99	0.83	0.82	0.98
	10	0.92	0.93	0.99	0.86	0.84	0.97
	12	0.92	0.93	0.99	0.82	0.78	0.97
480	3	0.96	0.96	0.99	0.92	0.93	0.99
	4	0.95	0.96	0.99	0.91	0.91	0.99
	5	0.95	0.95	0.99	0.91	0.90	0.99
	6	0.94	0.95	0.99	0.90	0.89	0.99
	8	0.94	0.95	0.99	0.92	0.91	0.99
	10	0.94	0.94	0.99	0.92	0.90	0.98
	12	0.95	0.94	0.99	0.90	0.87	0.98

Table 5.3: For parametrically generated (columns 3–5) and non-parametrically generated schemes (column 6–8), rank correlation coefficients are given between the 95.4% threshold interval widths computed using the $\mathrm{BC_a}$, probit and Monte Carlo interval methods. See section 5.5.6 for details.

		parametric			non-parametric		
N	k	B ↔ P	M ↔ P	B ↔ M	B ↔ P	M ↔ P	B ↔ M
120	3	0.68	0.54	0.74	0.69	0.52	0.75
	4	0.68	0.37	0.62	0.62	0.27	0.58
	5	0.67	0.24	0.49	0.59	0.09	0.49
	6	0.65	0.15	0.44	0.50	0.01	0.44
	8	0.69	0.08	0.41	0.50	-0.14	0.43
	10	0.68	0.03	0.41	0.44	-0.19	0.40
	12	0.66	-0.01	0.43	0.33	-0.28	0.40
240	3	0.80	0.60	0.86	0.72	0.55	0.87
	4	0.82	0.44	0.74	0.66	0.32	0.77
	5	0.82	0.32	0.62	0.65	0.16	0.67
	6	0.81	0.22	0.54	0.58	0.08	0.63
	8	0.81	0.14	0.49	0.61	0.03	0.56
	10	0.81	0.11	0.48	0.58	-0.02	0.52
	12	0.79	0.09	0.49	0.56	-0.02	0.53
480	3	0.88	0.74	0.92	0.75	0.62	0.93
	4	0.91	0.67	0.86	0.67	0.42	0.88
	5	0.91	0.60	0.81	0.68	0.34	0.83
	6	0.90	0.55	0.77	0.62	0.25	0.79
	8	0.89	0.48	0.72	0.69	0.25	0.73
	10	0.89	0.46	0.71	0.66	0.23	0.73
	12	0.89	0.45	0.71	0.65	0.27	0.75

Table 5.4: For parametrically generated (columns 3–5) and non-parametrically generated schemes (column 6–8), rank correlation coefficients are given between the 95.4% slope interval widths computed using the BC$_{\mathrm{a}}$, probit and Monte Carlo interval methods. See section 5.5.6 for details.

For slopes, the most efficient sampling schemes generally had a wide spacing, which was wider for larger numbers of observations, and wider for 68.3% intervals than for 95.4%. The most efficient schemes also tended to place more sample points at high expected performance levels than at low levels, but this is unfortunately associated with a considerable level of bias in slope estimation. Slope bias is linked to bias in the estimation of λ: when λ itself can be estimated without bias, as is the case when $\lambda_{\mathrm{gen}} = 0.025$, or when λ is known exactly, slope estimation bias need not be large enough to cause concern, even for fairly efficient schemes. However, a λ_{gen} value of around 0.01 is particularly problematic in this regard.

For thresholds, the optimum value of the mean performance level $\bar{p}$, the optimum spread of performance levels σ_p, and optimum efficiency (corresponding to the smallest interval width $\textsc{wci}_{\mathrm{min}}$) all show little or no dependence on the number of blocks k. Furthermore, where results were taken for the unequal block distribution method, they are very similar to those of the equal block distribution method. Taken together these two findings indicate that, if the mean and standard deviation of the performance values that make up a sampling scheme remain constant, the number of blocks into which the total number of trials is divided does not matter. Slopes are a slightly different matter: the advantages of lower bias and increased efficiency are to be gained by concentrating trials into fewer blocks, particularly when N is high.[†] For both thresholds and slopes, the effect of k on the distributions of non-optimal bias and efficiency scores follows a similar pattern to the effect on the optimal sampling scheme.

Despite the inaccurate coverage of probit threshold intervals, probit anal-

[†] This is probably a consequence of the fact that the sampling schemes explored in both the parametric and non-parametric simulations had a certain minimum spacing between stimulus values. If the optimal sampling strategy with regard to efficient slope estimation is to concentrate all observations at two particular points, as has been suggested,[19,21] the lower limit on inter-stimulus spacing prevents all but two of the blocks from occupying the two optimal positions. This constraint is of course artificial, and in a real experiment there is no restriction on running more than one block at the same stimulus level. Therefore the set of possible sampling schemes at, for example, $k = 6$ should really be considered to be a super-set of the possible sampling schemes at $k = 3$.

ysis can be used to make fairly accurate predictions of the relative efficiency of different sampling schemes, with regard to thresholds. Accuracy improves as N increases, and there is still some substantial improvement at values of N above 120. For slopes, the probit predictions are less accurate, but they are more accurate predictors of the bootstrap interval widths that an experimenter would actually report, than are Monte Carlo percentile intervals based on variation from a known psychometric function.

The fact that the most efficient sampling schemes are sometimes unevenly spaced means that the parameterization proposed by McKee *et al.*,[5] which uses the mid-point and extent of the range of the stimulus values used, is inadequate to represent fully the possible variations in sampling schemes. The representation of sampling schemes by the pair of coordinates $(\bar{p}, \sigma_p)$, which are the first moment and second central moment of the distribution of expected performance values of the individual trials, proved to be useful in that it is easily interpretable, and captures most of the variation in threshold efficiency, threshold bias and slope bias smoothly. With regard to threshold efficiency, there was just a small number of exceptions that did not fit into a smooth pattern, as figure 5.18 showed. For measures of slope efficiency, however, the two-dimensional representation was poor, as can be seen in figure 5.23. A more sophisticated representation was attempted, using three parameters (section 5.5.1). While the three-parameter method provided a useful, smoothly varying set of bias and efficiency scores from which to draw conclusions, the parameterization did not wholly succeed, in the sense that it captured little more of the possible variation in slope estimation efficiency than did the two parameters $(\bar{p}, \sigma_p)$.

References for chapter 5

[1] Lam, C. F, Mills, J. H. & Dubno, J. R. (1996). Placement of observations for the efficient estimation of a psychometric function. *Journal of the Acoustical Society of America*, **99**(6): 3689–3693.

[2] Lam, C. F, Dubno, J. R, Ahlstrom, J. B, He, N. J. & Mills, J. H. (1997). Estimating parameters for psychometric functions using the four-point sampling method. *Journal of the Acoustical Society of America*, **102**(6): 3697–3703.

[3] Lam, C. F, Dubno, J. R. & Mills, J. H. (1999). Determination of optimal data placement for psychometric function estimation: a computer simulation. *Journal of the Acoustical Society of America*, **106**(4, pt. 1): 1969–1976.

[4] Teller, D. Y. (1985). Psychophysics of infant vision: Definitions and limitations. In Gottlieb, G. & Krasnegor, N (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life: a Methodological Overview.* Norwood, NJ: Ablex.

[5] 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.

[6] 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` .

[7] Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, **47**(1): 87–91.

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

[9] 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.

[10] Leek, M. R, Hanna, T. E. & Marshall, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception and Psychophysics*, **51**(3): 247–256.

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

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

[13] Taylor, M. M. & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**(4): 782–787. An erratum is given in issue 42, number 5, page 1097.

[14] Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 505–508.

[15] Robbins, H. & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**: 400–407.

[16] Wetherill, G. B. (1963). Sequential estimation of quantal reponse curves. *Journal of the Royal Statistical Society, Series B*, **25**(1): 1–48.

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

[18] Dai, H. P. (1995). On measuring psychometric functions: a comparison of the constant-stimulus and adaptive up-down methods. *Journal of the Acoustical Society of America*, **98**(6): 3135–3139.

[19] Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 467–477.

[20] Emerson, P. L. (1984). Observations on a maximum likelihood method of sequential testing and a simplified approximation. *Perception and Psychophysics*, **36**: 199–203.

[21] King-Smith, P. E. & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, **37**(12): 1595–1604.

[22] Yuan, H. (1999). Comparative study of adaptive psychophysical procedures.(threshold estimation, maximum likelihood). *Dissertation Abstracts International: Section B: The Sciences and Engineering*, **60**(4-b): 1910.

[23] Hawley, M. L. & Colburn, H. S. (1995). Application of confidence intervals and joint confidence regions to the estimation of psychometric functions. *Journal of the Acoustical Society of America*, **97**: 3277.

[24] Hawley, M. L. (1990). Comparison of adaptive procedures for obtaining psychophysical thresholds using computer simulation. Master's thesis, Boston University.

[25] 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.

6. A note on sequential estimation

6.1 Introduction

The simulations of chapter 5 indicate that, for a constant total number of trials N, the number of blocks k into which the trials are divided makes little difference from the point of view of bias or efficiency. For thresholds, the optimal mean location and spread of the trial positions did not change as k increased, and neither the optimum efficiency nor the distribution of non-optimal efficiencies was affected. When k was larger than about 4, the same was also true of slope estimation.

In any real experiment, however, one clear advantage to the division of trials into a larger number of blocks is that there are more opportunities to stop, assess the data gathered so far, and estimate the best stimulus value for the next block of trials. The logical extreme is to re-assess the data after every individual *trial*, positioning each separate observation at the current estimate of the optimal location.[†] The advantage of updating after every trial is that, in theory, the number of non-optimally placed stimuli is minimized. The concomitant drawback is that early estimates of the optimal stimulus location are based on small numbers of trials, and are accordingly

[†] Such a strategy is an example of a *single presentation design*, in which every individual trial plays a rôle, either in deciding whether to update the stimulus value yet (and if so, in which direction), or in contributing to the estimate of the optimal stimulus value for the next trial. By contrast, experiments in which a predetermined number of consecutive trials ($n_i > 1$) is presented at each stimulus level will be referred to as *block designs* (the term "constant stimuli" is deliberately avoided here, for reasons to be discussed). The usual connotation of the term "adaptive procedure" in psychophysics is that a single presentation design is in use, but there is no reason why the sequential selection of fixed-size blocks should not also be described as adaptive, or why one of the many available adaptive procedures[1] should not be applied to the problem.

prone to error. This may be a particular problem if, for example, "lapses" occur at a high stimulus level early in the adaptive run. Some adaptive procedures are better than others at recovering from such a situation: Taylor, Forbes and Creelman[2] argue that PEST[3] is more robust in this respect than some maximum-likelihood adaptive procedures,[4–6] and simulations by Madigan and Williams[7] confirm that Taylor and Creelman's PEST is indeed less seriously affected by lapses than Pentland's Best PEST,[4] although Watson and Pelli's QUEST[6] was in fact superior. King-Smith *et al.*[8] give an elegant graphical illustration of the way in which their minimum-variance adaptation of QUEST is even better than the original at recovery from early lapses.

Nevertheless, there are other considerations which may affect the choice of k, and there are situations in which a block design may be preferable, even to one of the more robust single-presentation adaptive procedures. The performance of real subjects can be non-stationary, either due to learning or to fatigue in the course of a block of trials. Such effects depend on the psychophysical phenomenon under study, as subjects learn more quickly in some tasks than in others, and seem to have greater stamina in some tasks than in others. Measurable decreases[9,10] and increases[11] in a subject's threshold have both been reported over a relatively small number of trials. Running a small number of practice trials before each block, in order to allow the observer to acclimatize to the stimulus level to be measured and thus stabilize performance, may be advisable in the former circumstance, particularly if the observer uses a different cue at low stimulus levels from those available at high levels—in such a case (such as that reported by Henning, Millar and Hill[12]) an adaptive procedure that changes the stimulus level rapidly on successive trials might be counter-productive. An experimenter must decide on the optimal block size for an experiment, depending on the task and on the subject.

When a block design is used, it is sometimes referred to as the "method of constant stimuli", a name which applies to a subset of block designs but which is often somewhat imprecisely applied to all. This can lead to misunderstanding. In 1988 an article by Simpson[13] entitled *The Method of Con-*

stant Stimuli is Efficient demonstrated by Monte Carlo simulation that the efficiency of a block-design experiment compared favourably with that of a single-presentation maximum-likelihood adaptive procedure, adding weight to the earlier statement of McKee, Klein and Teller[14] that the variability of estimates from adaptive methods "can never be less than those from the method of constant stimuli selected for the optimal deployment of trials." A subsequent study by Watson and Fitzhugh,[15] entitled *The Method of Constant Stimuli is Inefficient*, used Monte Carlo simulation to demonstrate the converse finding. Watson and Fitzhugh took the literal, classical interpretation of the phrase "constant stimuli" in which all the stimulus positions are predetermined at the beginning of each simulated experiment, their positions presumably based on pilot data, and are not adjusted during the course of the experiment. The imprecision of the initial guess therefore means that the psychometric function is often poorly sampled, which leads to large variance in the estimated threshold, as Watson and Fithugh's simulations showed. By contrast, the adaptive procedure QUEST, which does not commit itself in advance to particular (imprecisely chosen) stimulus values, naturally performed more efficiently under the same initial conditions of uncertainty.

In their conclusion, referring to the statement of McKee *et al.*, Watson and Fitzhugh write:

> "We must presume that 'optimal deployment' means 'optimal based on the true location of the threshold.' Of course, if this location is known, there is no need to run the experiment. In the real world, threshold is never known exactly, even after the experiment is completed."

It is unlikely, however, that McKee *et al.* were simply being naïve about the problem of estimating an unknown quantity. More probably, they considered optimal or close-to-optimal sampling schemes worthy of study because they were quite accustomed to producing something roughly similar to the "optimal deployment of trials" in block-design experiments of their own. In practice it is not difficult to do so, if the experimenter stops after every block or small number of blocks to assess the best stimulus value for the next. Often this is done "by eye", but nonetheless it should be considered a kind of adaptive procedure, because it is an algorithm (albeit an algorithm which cannot be written down, and which varies from experimenter to experimenter) for computing the next stimulus level based on the data gathered so far. If this is truly what McKee *et al.* meant (as seems probable given their use of the word "selected") then their mistake was not in the assumption of prescience on the part of the experimenter, but rather in ambiguous application of the term "constant stimuli."[†]

Section 5.5 took the approach of McKee *et al.*[14] in asking what the optimal block-design sampling patterns look like, and how accurate and precise they are, without specifying how the experimenter is supposed to achieve such a sampling pattern. In effect, the results provide a target to aim for, without specifying how best to hit that target. By contrast, the adaptive approach provides an array of (more or less accurate) methods of successive approximation to a target, which generally aim for a *non*-optimal target: when attempting to find a threshold, most adaptive procedures follow the advice of Levitt[16] in aiming directly for the threshold performance level. Such a strategy is unfortunately only asymptotically optimal (see section 5.5.4). Some adaptive procedures are exceptions to this, and take a broader view of optimal sampling. They include:

- APE by Watt and Andrews,[17] which updates after every four blocks, based on a sliding estimate of the psychometric function parameters—a description is also provided by Treutwein;[1]

- the Minimum Variance Method of King-Smith and colleagues,[8,18] a Bayesian method which minimizes the predicted variance of the posterior distribution of thresholds at each step, and which is shown[8] to be more efficient at threshold estimation than three variations of QUEST (the method could also be adapted to optimize slope estimation[19,20]);

[†] Simpson, incidentally, explicitly took the same classical definition of "method of constant stimuli" as Watson and Fitzhugh, so the latter authors were probably right to point out that he had failed to take account of its vulnerability to initial uncertainty.

- the Ψ Method of Kontsevich and Tyler,[21] another Bayesian method, whose stimulus selection algorithm is based on the "entropy" measure of Pelli's Ideal Psychometric Procedure[22]—this cost function minimizes the expected variance of prediction for the whole psychometric function, and thus provides a way of measuring the "optimal" sampling position with respect to thresholds and slopes simultaneously.

Any of the above procedures might be adapted in order to select the position of *blocks* rather than individual trials (in fact, APE was formulated specifically with block designs in mind). A fourth option is the "by eye" method, which works according to unknown, variable, and probably non-optimal principles, but which nevertheless should not be underestimated in its ability to produce accurate and efficient sampling patterns.

Watson and Fitzhugh[15] observe that, in simulation and in the laboratory, the method of constant stimuli "is inefficient largely because threshold may lie outside the testing interval." Yet it is highly unusual to see a published set of psychophysical data from a block design in which the threshold of interest lies outside the range of stimulus values at which trials were performed. Indeed, to conclude one's study in such a situation would be viewed as bad experimental procedure. It is therefore probable that the majority of practitioners of block-design experiments respond to their data in some way while gathering it: either they update the stimulus intensity after every block (or perhaps after every two or three blocks) taking into account the data so far, or they genuinely determine the stimulus levels in advance according to the method of constant stimuli, and then perform extra observations afterwards if the threshold of interest does not lie within the stimulus range. The latter strategy is inefficient, to be sure, but it is still, in an important sense, adaptive because the stimulus intensities for the new trials are determined by the old data. It is therefore quite common for the experimenter to play an adaptive rôle in stimulus selection.

The simulations of chapter 3, in common with previous Monte Carlo studies by Foster and Bischof,[23,24] Swanepoel and Frangos,[25] and Lee,[26] do not take into account the possibility of adaptive sequential stimulus selection.

With the exception of Lee,[26] all the above studies assume that the stimuli occur at fixed intensities $\boldsymbol{x}_{\text{gen}}$, relative to the true curve $\psi(x;\, \boldsymbol{\theta}_{\text{gen}})$. All the different $\boldsymbol{x}_{\text{gen}}$ examined (with the possible exception of the scheme ◀ from figure 1.2) were fairly reasonable distributions of stimulus levels for estimating the threshold and/or slope of the psychometric function, *assuming that the true psychometric function were already known.* The inaccuracy of such an assumption meant that repeated Monte Carlo simulation of the experiment would sometimes produce situations that would never happen in practice. Due to random variation in the simulated data, all the measured performance levels might, for example, happen to fall below the generating curve on one particular replication. The initial fit $\hat{\boldsymbol{\theta}}_0$ would therefore indicate a higher threshold than the true threshold from $\boldsymbol{\theta}_{\text{gen}}$, and consequently the stimulus values $\boldsymbol{x}_{\text{gen}}$ would appear to the experimenter to lie on a lower part of the curve than they in fact do. On some such occasions, the threshold level of interest might even appear to lie outside the stimulus range, and it is in counting such a case that the Monte Carlo simulation with fixed $\boldsymbol{x}_{\text{gen}}$ is unrealistic. In reality the experimenter would not be satisfied: he or she would either perform additional observations to remedy the situation, or would have selected the stimulus values carefully in an adaptive sequence in order to prevent the situation from occurring in the first place.

To summarize: in a real experiment the experimenter usually takes steps to ensure that the psychometric function appears well sampled—in other words, that the sampling scheme looks reasonable[†] relative to the estimated curve $\psi(x;\, \hat{\boldsymbol{\theta}}_0)$. There will be some error in $\hat{\boldsymbol{\theta}}_0$, which means that the sampling scheme might not always appear so reasonable if it were possible to

[†] "Reasonable" here is a loose criterion that means simply "however the experimenter would ideally like the pattern of stimulus values to appear, to within some subjective tolerance." Depending on the demands of the experimental context, a "reasonable" sampling scheme might be one that is approximately optimal for the estimation of a particular threshold level, or for the estimation of slope, or for simultaneous threshold and slope estimation according to some compromise measure of efficiency. Alternatively, a reasonable sampling scheme might simply be one that is roughly evenly spaced and which covers a wide range of performance levels. The important point is not what the experimenter's criterion for reasonableness is, but rather that such a criterion exists and is applied.

plot it against the true psychometric function $\psi(x;\, \boldsymbol{\theta}_{\mathrm{gen}})$. However, the fixed-stimulus Monte Carlo approach simulates the converse set of circumstances: the stimulus locations may be reasonable relative to the true psychometric function but, during simulation, they do not necessarily constitute a sampling pattern of the sort than an experimenter would accept, relative to the estimated curve.

Lee[26] takes a different approach, in which the generating stimulus levels are not fixed, but are instead drawn independently from a uniform distribution on each replication of the experiment. This does not, however, solve the problem, because the sampling schemes are not selected for their reasonableness relative to the true psychometric function *or* the estimated psychometric function.[†]

What are the implications for hypothesis testing? Sequential selection may present a problem for the application of bootstrap methods because, in the form applied here and by others[23,24,28–32] to psychometric functions, bootstrap confidence intervals themselves are based on simulations that assume fixed stimulus levels. In a real experiment it is more likely that each stimulus value x_i is chosen by an algorithm that uses the previous observed performance levels $y_1 \ldots y_{i-1}$, which are themselves variable, but the bootstrap simulations do not take into account the extra variability that this entails. Therefore, the bootstrap may underestimate the variability of the estimated parameters. Another way of viewing the same effect is to say that the apparent "reasonableness" of the distribution of stimulus values in each observed data set, ensured each time by the experimenter, means that the bootstrap variability around the estimate $\hat{\boldsymbol{\theta}}_0$ is less than the variability of estimates around the true parameter set $\boldsymbol{\theta}_{\mathrm{gen}}$, because the pattern of stimulus values tends to be *less* reasonable relative to the true curve. Either argument leads to the same prediction, namely that coverage is too low.

[†] This was not, in fact, Lee's intention. His approach was adopted from Gong,[27] who introduced random regressors in a logistic regression problem in order to represent naturally occurring phenomena (clinical markers that might explain deaths from chronic hepatitis) whose values are beyond the control of the investigator. In such a context, random independent selection of $\boldsymbol{x}_{\mathrm{gen}}$ is more appropriate.

The same arguments may even be applied in the case of the strict method of constant stimuli—even though the stimulus levels are chosen in advance and left unchanged throughout the experiment, the assumption that they are fixed during bootstrap simulation may lead to underestimation of variability because the stimulus levels were in fact usually chosen on the basis of pilot data, for which the process of collection is ill-defined but certainly prone to random variation, and presumably in some sense adaptive.

6.2 Simulations

6.2.1 Method

In order to investigate the possible effects of sequential stimulus selection on confidence interval coverage, four further sets of coverage tests were run. In each, a cumulative normal function with $\alpha_{\mathrm{gen}} = 0$ and $\beta_{\mathrm{gen}} = 1$ was taken to be the true psychometric function. The four conditions were identical to the four general cases used in chapter 3:

- **Idealized yes-no:** $\gamma_{\mathrm{gen}} = 0$, $\lambda_{\mathrm{gen}} = 0$, and both γ and λ were fixed at 0 during fitting.

- **Realistic yes-no:** $\gamma_{\mathrm{gen}} = 0.02$, $\lambda_{\mathrm{gen}} = 0.01$, and both γ and λ were free to vary in the range $[0, 0.05]$ during fitting.

- **Idealized 2-AFC:** $\gamma_{\mathrm{gen}} = 0.5$, $\lambda_{\mathrm{gen}} = 0$ and λ was fixed at 0 during fitting.

- **Realistic 2-AFC:** $\gamma_{\mathrm{gen}} = 0.5$, $\lambda_{\mathrm{gen}} = 0.01$, and λ was free to vary in the range $[0, 0.05]$ during fitting.

Results from the four test sets are to be found in the result archive under:

- `simulations/sequential/yesno/g0f0l0f0/cumnorm/`

- `simulations/sequential/yesno/g02l01/cumnorm/`

- simulations/sequential/2AFC/l0f0/cumnorm/
- simulations/sequential/2AFC/l01/cumnorm/

The general method is the same as that described in section 3.1.1, except that the data set in each simulated experiment is generated by an adaptive stimulus selection method. The first requirement, however, is to generate randomly varying starting conditions, in other words to simulate conditions under which an experimenter might *start* using a sequential stimulus selection method. To do this, it is assumed that, through some pilot testing method, the experimenter has succeeded in running two blocks of trials in which substantially different performance levels were measured. Two f-values are drawn independently from a uniform random distribution on the interval $(0, 1)$, the lower of the two being designated as f_1 and the higher as f_2. These are the detection levels at which the experimenter has happened to strike the unknown psychometric function, corresponding to stimulus values x_1 and x_2. At each stimulus level, a simulated block of n_i trials is run, yielding observed performance levels y_1 and y_2, each y_i being drawn from the binomial distribution $\mathrm{Bi}\left[n_i, \psi(x_i; \boldsymbol{\theta}_{\mathrm{gen}})\right]$. The apparent detection levels $\hat{f}_1$ and $\hat{f}_2$ are equal to y_1 and y_2 in the yes-no design, and equal to $y_1/(1-\gamma)$ and $y_2/(1-\gamma)$, clipped in the range $[0, 1]$, in a forced-choice design. If $\hat{f}_2 - \hat{f}_1 \geq 0.5$, the pair is taken to represent apparently successful pilot data, i.e. the sort of data that an experimenter might then take as a basis for further exploration using an adaptive block-by-block stimulus selection method. If not, the process is repeated until a successful pair is generated.

Any of the adaptive methods mentioned in section 6.1 (APE, the Minimum Variance Method or the Ψ Method) might have been adapted for the purpose of selecting subsequent blocks. However, Bayesian methods were avoided because the requirement for many additional assumptions (specifically the shape of the prior probability distribution of each parameter, representing prior experimenter knowledge about the psychometric function) would introduce many new variables into the simulation, a thorough investigation of which is beyond the scope of the current research.[†]

Instead, a rather simpler, non-optimally efficient system is used. To choose each stimulus value, a psychometric function is first fitted to the k data points collected so far, to obtain estimated parameters $\hat{\boldsymbol{\theta}}_0$. The estimated expected detection levels $\hat{\boldsymbol{f}}$ are computed using $\hat{f}_i = \psi(x_i; \hat{\boldsymbol{\theta}}_0)$. The values are sorted into ascending order, and then 0 is appended to the beginning and 1 to the end to form a series consisting of $(0, \hat{f}_1 \ldots \hat{f}_k, 1)$. The f-value corresponding to the next stimulus, $\hat{f}_{k+1}$, is chosen to bisect the largest interval between consecutive members of the series. The next stimulus value x_{k+1} is then equal to $F^{-1}(\hat{f}_{k+1}; \hat{\boldsymbol{\theta}}_0)$. The method was found to be fairly robust, and tended to produce widely spread sampling schemes whose predicted performance levels were quite evenly spaced. At $k \geq 5$, the stimulus values nearly always covered most of the estimated psychometric function from $F(x) = 0.2$ to $F(x) = 0.8$.

In each simulated experiment, after the initial randomly generated pair of blocks, 10 further blocks were simulated, the 10 stimulus values being chosen in sequence using the above method. After each block, confidence intervals were computed using the bootstrap standard error, basic bootstrap, bootstrap percentile, $\mathrm{BC_a}$, probit standard error and probit fiducial methods. Block size n_i was taken to be 40. Thus, coverage results were obtained for 10 conditions of sampling density ranging from $k = 3$ $(N = 120)$ to $k = 12$ $(N = 480)$. As in chapter 3, $R = 1999$ bootstrap simulations were performed on each of $C = 500$ experimental replications. The Freeman-

[†] Some pilot simulations were run using a method which selected each stimulus value to yield minimum predicted confidence interval widths from probit analysis, based on a psychometric function fitted to the data so far. However, the procedure occasionally demonstrated a lack of robustness: after the first two blocks, it was sometimes the case that the method placed the third block at a stimulus level very close to one of the other two. This could lead to extreme over-estimation of slope, from which it was difficult for the method to recover (subsequent blocks were then also placed very close together, because their position was computed relative to the very steep estimated curve). A practical solution to this in a real experiment might be to use a Bayesian prior to constrain the fitted slope. Again, the Bayesian approach was rejected because it compromised the generality of the simulations—the realism of any particular prior can only satisfactorily be judged in the context of a specific psychophysical application.

Tukey transformation described in section 3.1.2 and the graphical conventions of section 3.1.3 will be used to display the results.

6.2.2 Results

Idealized yes-no

Figures 6.1 and 6.2 show coverage results for thresholds and slopes respectively, at a target coverage probability of 95.4%, in the idealized yes-no condition. They may be contrasted with the results of the fixed-stimulus Monte Carlo approach shown in figures 3.1–3.2 (bootstrap methods), and 3.17–3.18 (probit methods).

In each group of symbols, the leftmost symbol represents the results for $k = 3$, progressing to $k = 12$ on the right. The size of the symbols also increases from left to right, as N increases from 120 to 480.

For thresholds, all the confidence interval methods tested perform very similarly (as was also the case for the fixed-stimulus simulations) and are highly accurate. The fact that symmetrical methods (the bootstrap and probit standard error methods) perform as well as the other methods indicates that the distributions of thresholds are close to normal. Note the slight increasing trend in the intervals' coverage as the number of blocks (and hence the total number of observations) increases. Coverage is slightly too low when fewer blocks have been taken—this is as we might expect, because early in the experiment there is greater variability in the adaptively selected stimulus positions, which leads to lower coverage by the argument of section 6.1.

For slopes, on the other hand, the situation has changed: all the procedures except the basic bootstrap have a distinct positive imbalance that was not present in the fixed-stimulus results. Somewhat surprisingly, the basic bootstrap is now very accurate and well-balanced.

Realistic yes-no

Figures 6.3 and 6.4 show the threshold and slope results for the realistic yes-no condition. They may be contrasted with the fixed-stimulus Monte

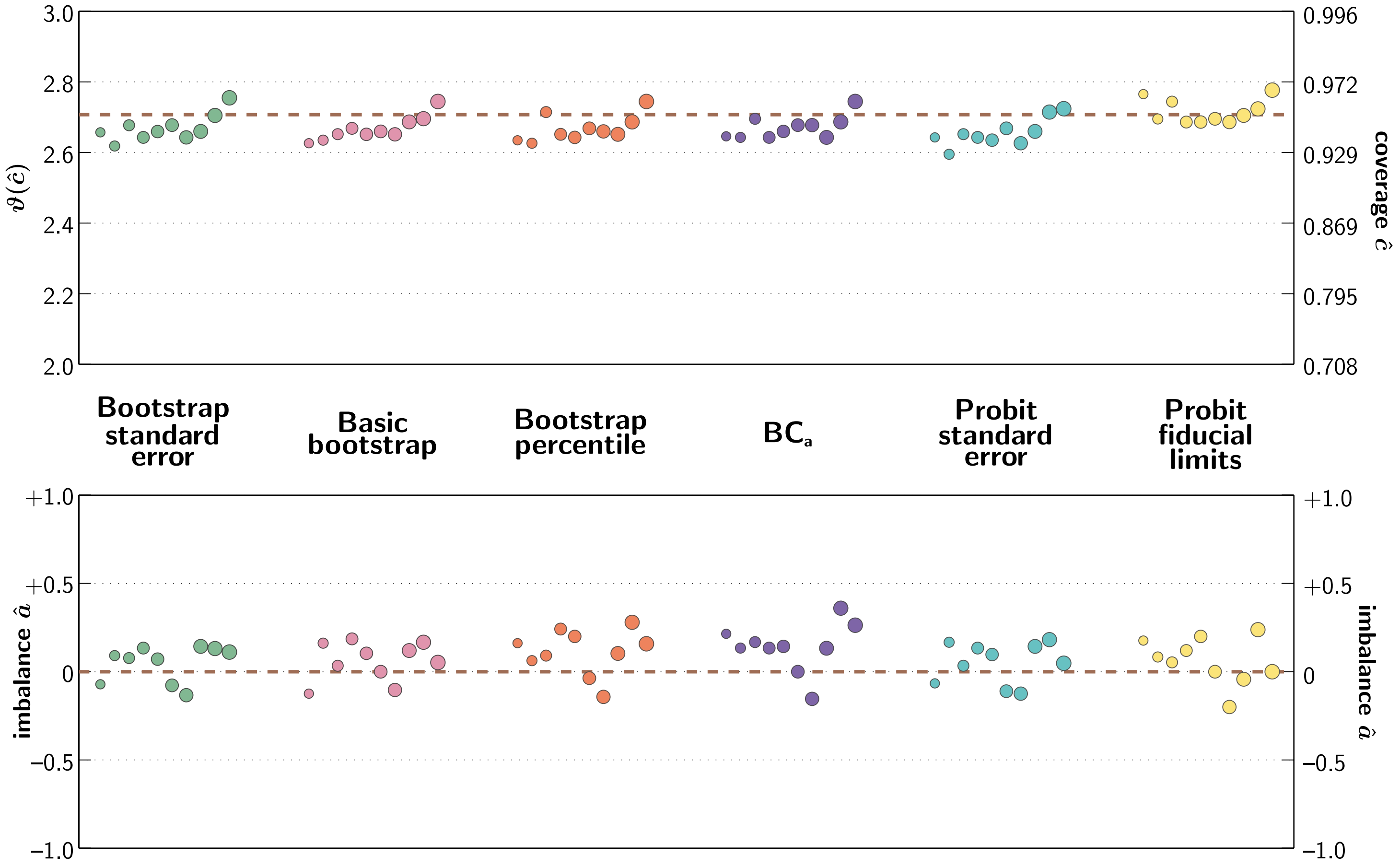


Fig. 6.1: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in the idealized yes-no case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

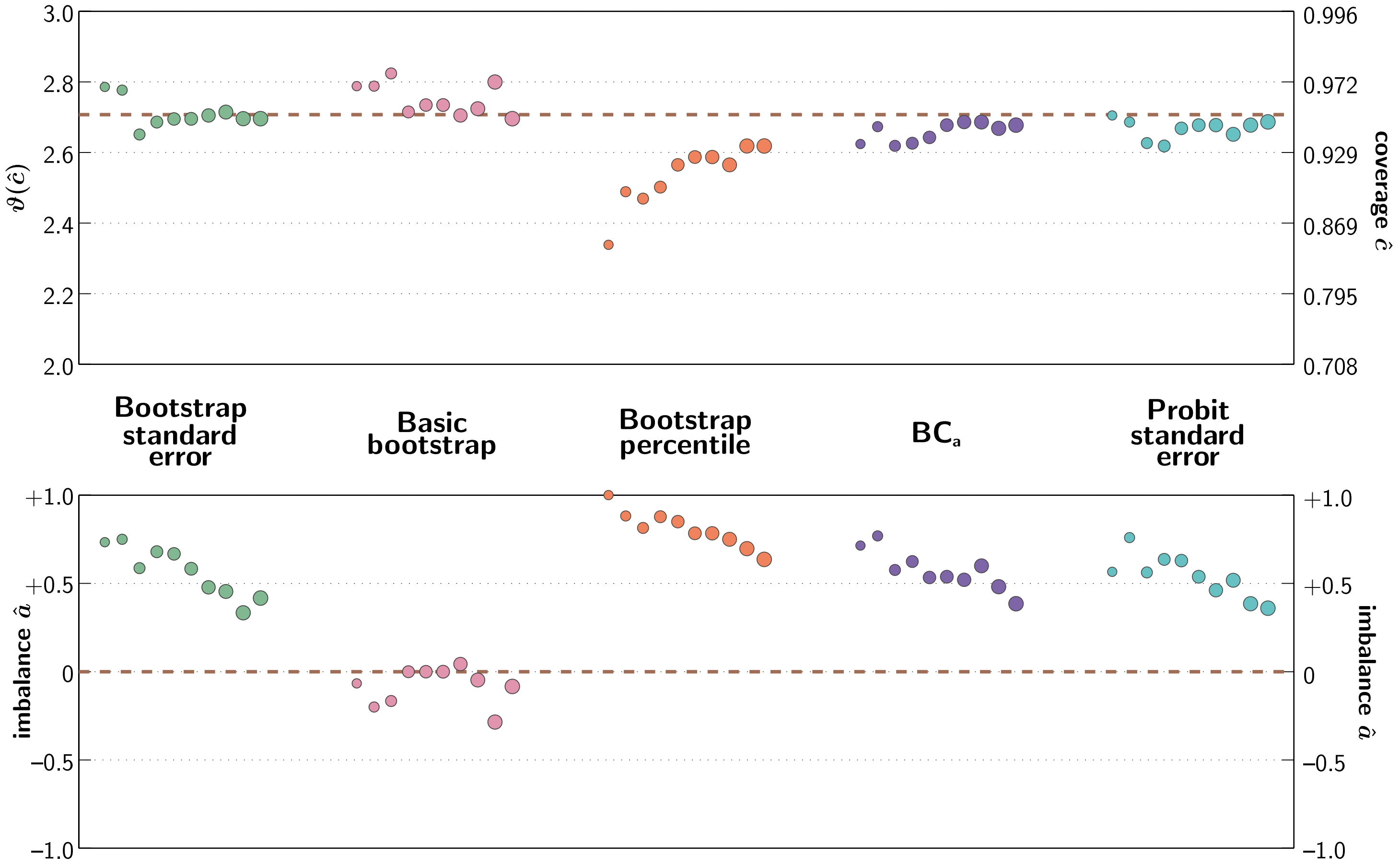


Fig. 6.2: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in the idealized yes-no case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

Carlo results of shown in figures 3.3–3.4 (bootstrap methods), and 3.17–3.18 (probit methods).

There is now a tendency for coverage of the true threshold value to be too low, and to be unbalanced towards the negative side. The pattern of threshold results from one confidence interval method to the next is very similar to that observed in the fixed-stimulus simulations. Note, however, the trend towards increasing imbalance as k increases: the addition of nuisance parameters γ and λ interacts with the effect of sequential stimulus selection, becoming more pronounced as the experiment proceeds. For slopes, the results are qualitatively similar to those of the idealized yes-no condition, above, but the positive imbalance of most of the methods is greater. The basic bootstrap method is still very accurate.

Idealized 2-AFC

Figures 6.5 and 6.6 show the idealized 2-AFC simulation results. The corresponding fixed-stimulus results are found in figures 3.5–3.6 (bootstrap methods), and figures 3.20 and 3.22 (probit methods).

For thresholds, there is more of a difference between confidence interval methods than there was in the idealized yes-no case. This was also observed in the fixed-stimulus simulations of chapter 3. The pattern of imbalance values between confidence interval methods is also similar to that observed in chapter 3. However, for nearly all the methods, coverage is too low at lower values of k (and consequently of N), and rises towards the target as k and N increase—this trend was *not* observed with increasing N in the fixed-stimulus simulations.

For slopes, the sequential selection has again increased the imbalance of the methods. In 2-AFC unlike yes-no (above), the basic bootstrap is now negatively unbalanced. The $\mathrm{BC_a}$ method, which was well-balanced in the fixed-stimulus simulations, is now highly unbalanced in the positive direction. The probit method, which was negatively unbalanced under the assumption of fixed stimuli, is now somewhat positively unbalanced, but it is the most accurate of the methods studied.

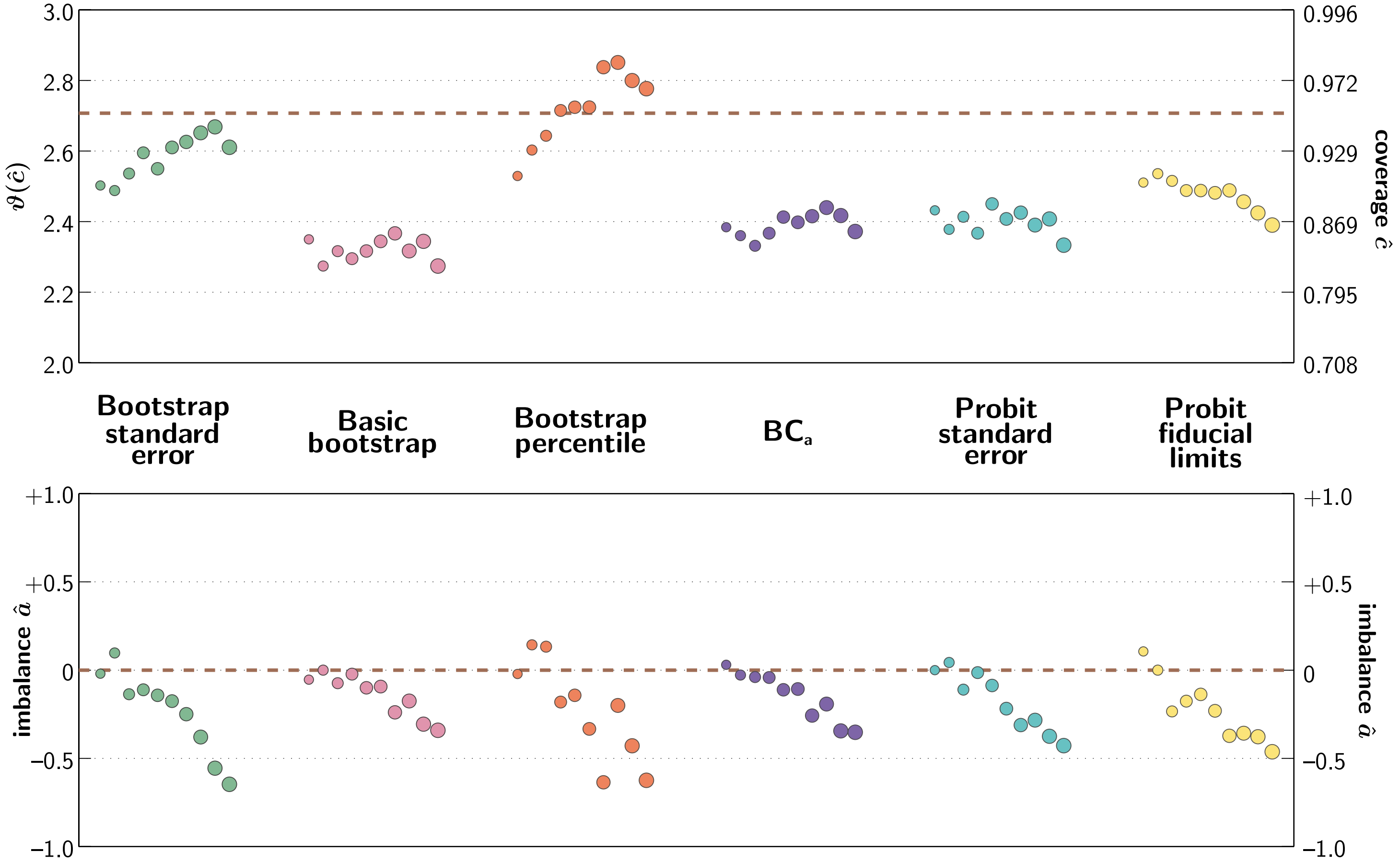


Fig. 6.3: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in the realistic yes-no case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

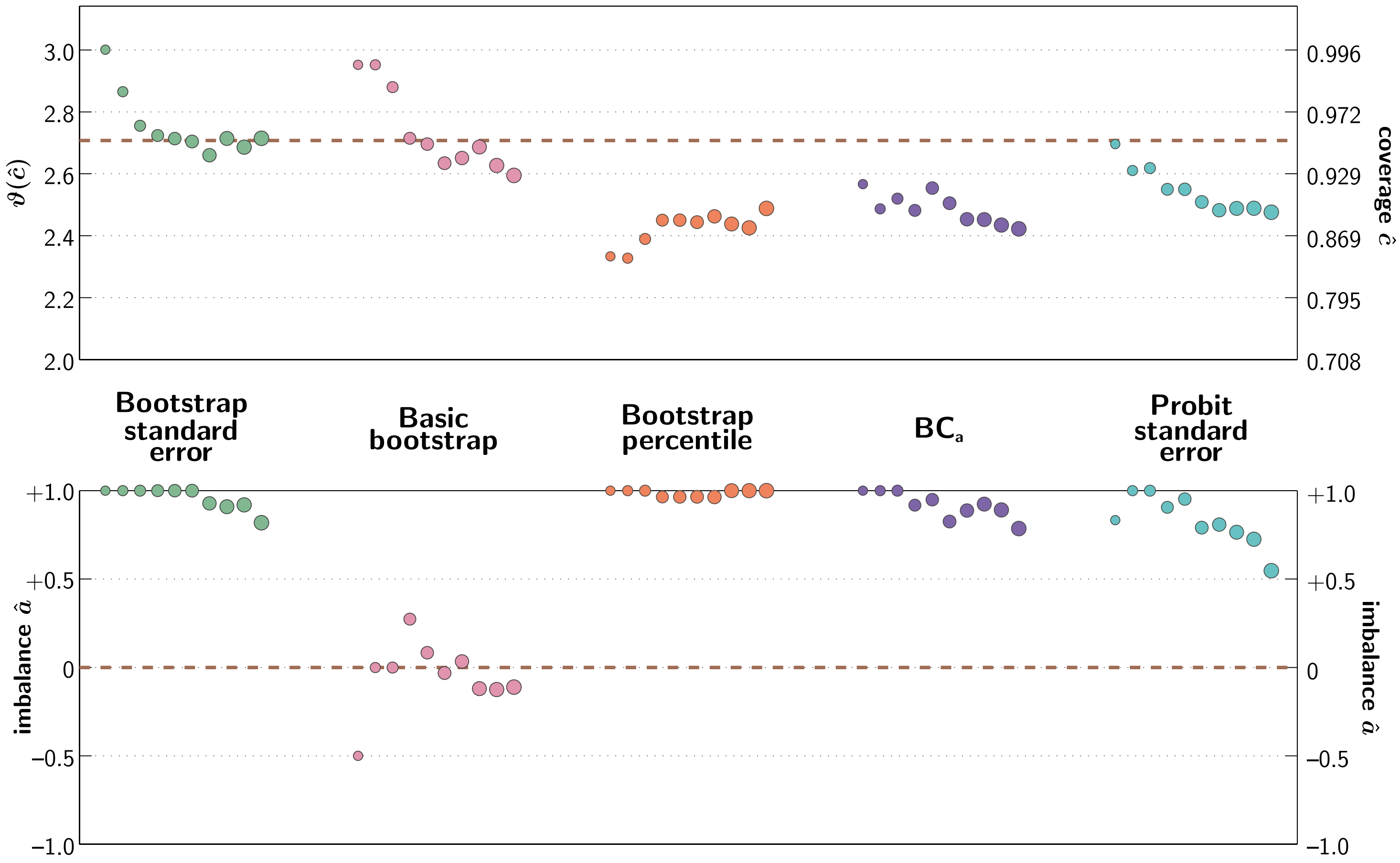


Fig. 6.4: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in the realistic yes-no case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

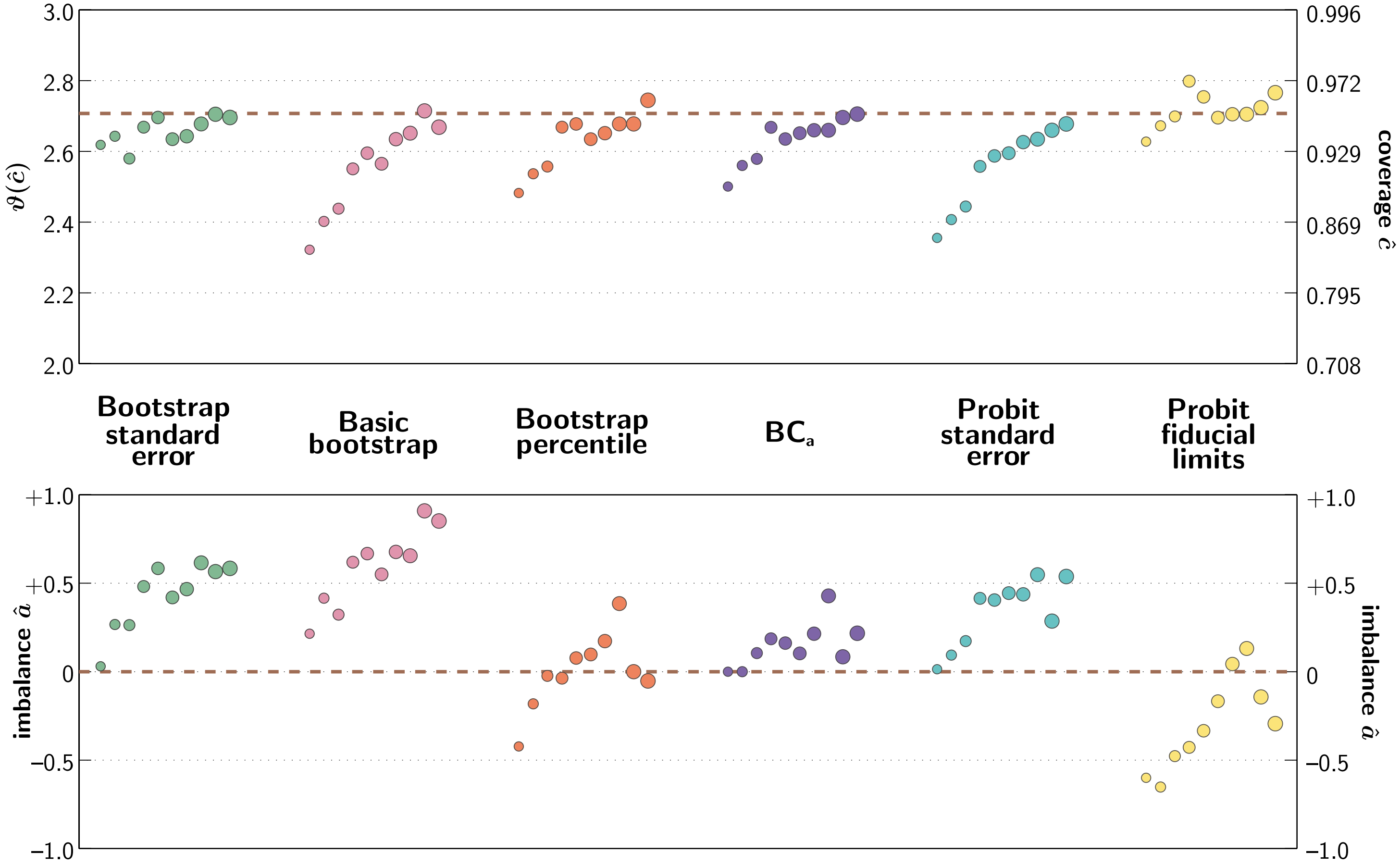


Fig. 6.5: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in the idealized 2-AFC case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

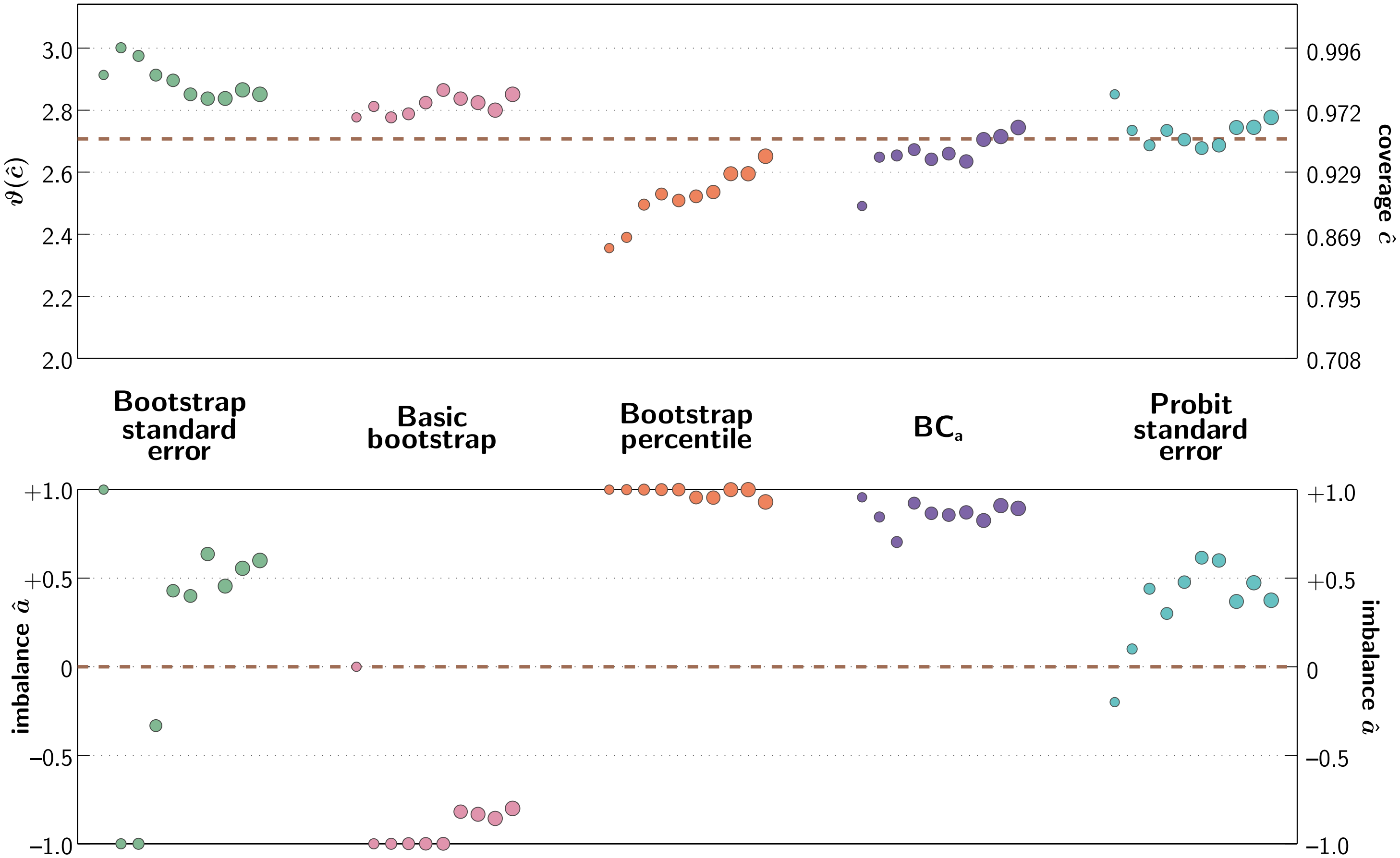


Fig. 6.6: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in the idealized 2-AFC case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

Realistic 2-AFC

Finally, figures 6.7 and 6.8 show the realistic 2-AFC simulation results, which may be compared with the fixed-stimulus results of figures 3.7–3.8 (bootstrap methods), and figures 3.20 and 3.22 (probit methods).

The threshold results show a drop in coverage for most of the confidence interval methods, relative to that observed in the fixed-stimulus simulations. The only interval method to reach target coverage is the bootstrap percentile method, which climbs towards the target as k and N increase. All the methods are somewhat negatively unbalanced, except for the bootstrap standard error and basic bootstrap methods. The bootstrap standard error method is the surprising winner, with very good balance and the highest coverage—still too low, but invariant with respect to k and N over the range studied.

For slopes, the imbalance previously observed in the fixed-stimulus simulations is now exaggerated, the basic bootstrap and bootstrap percentile methods in particular being completely unbalanced in many cases. The overall coverage of the bootstrap percentile method has also dropped. The probit method, previously negatively unbalanced in the fixed-stimulus simulations, is now slightly positively unbalanced. Its overall coverage is a little low at higher values of k, but is still the closest to target of any of the methods studied.

6.3 Summary and discussion

Sequential stimulus selection is presumed to occur to some extent in any psychophysical experiment, even in many experiments that claim to use the "method of constant stimuli". The experimenter must always decide where to place stimuli, and the decision will nearly always be based on previous measurements, so stimulus placement is itself a stochastically determined process. In the strict classical method of constant stimuli, the dependency is between pilot data and the data actually reported. In most cases, there will be additional stochastic sequential dependency between samples in the experiment proper, as the experimenter ensures good *apparent* sampling of

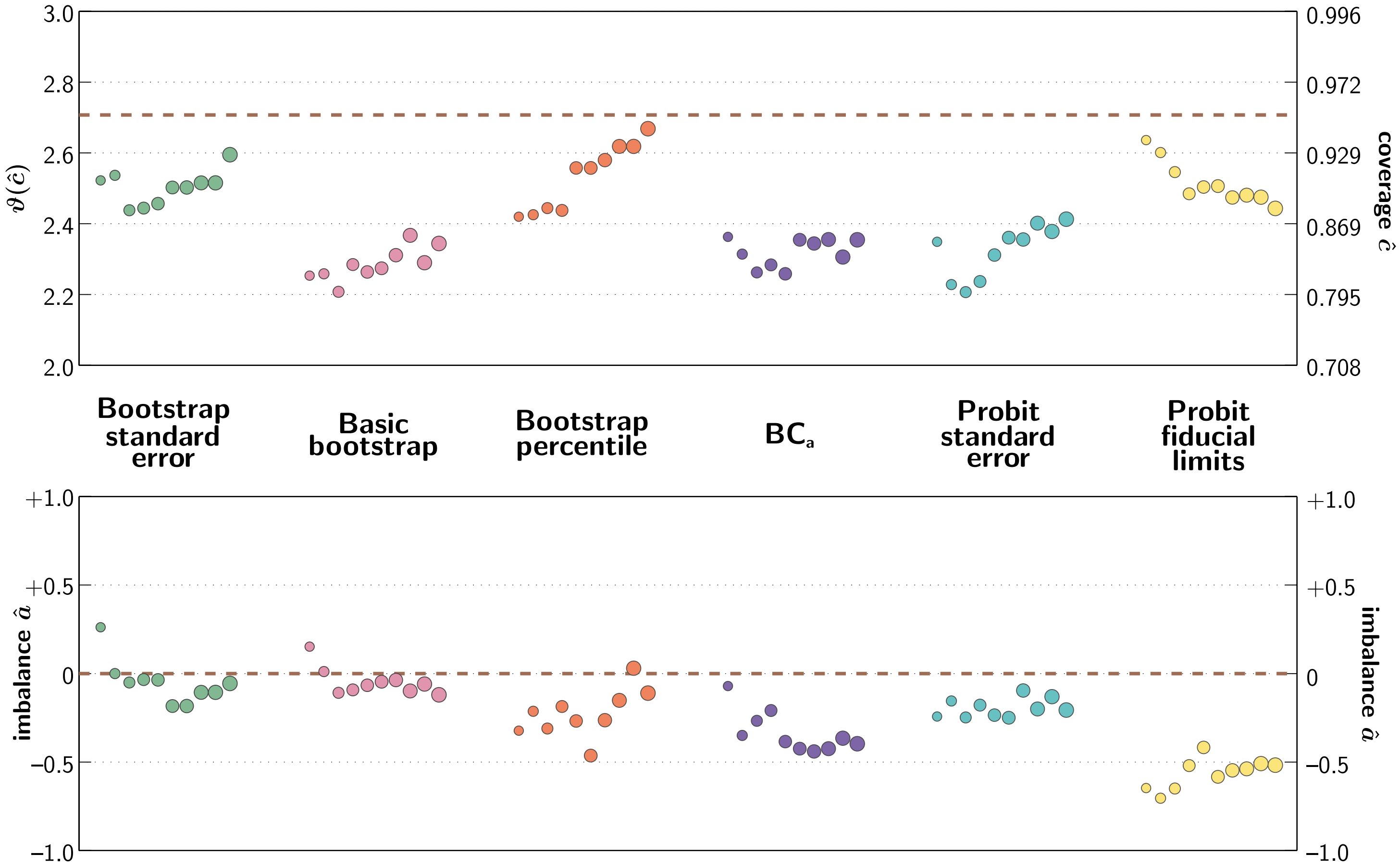


Fig. 6.7: Results of Monte Carlo coverage tests for 95.4% threshold confidence intervals obtained from the cumulative normal function in the realistic 2-AFC case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

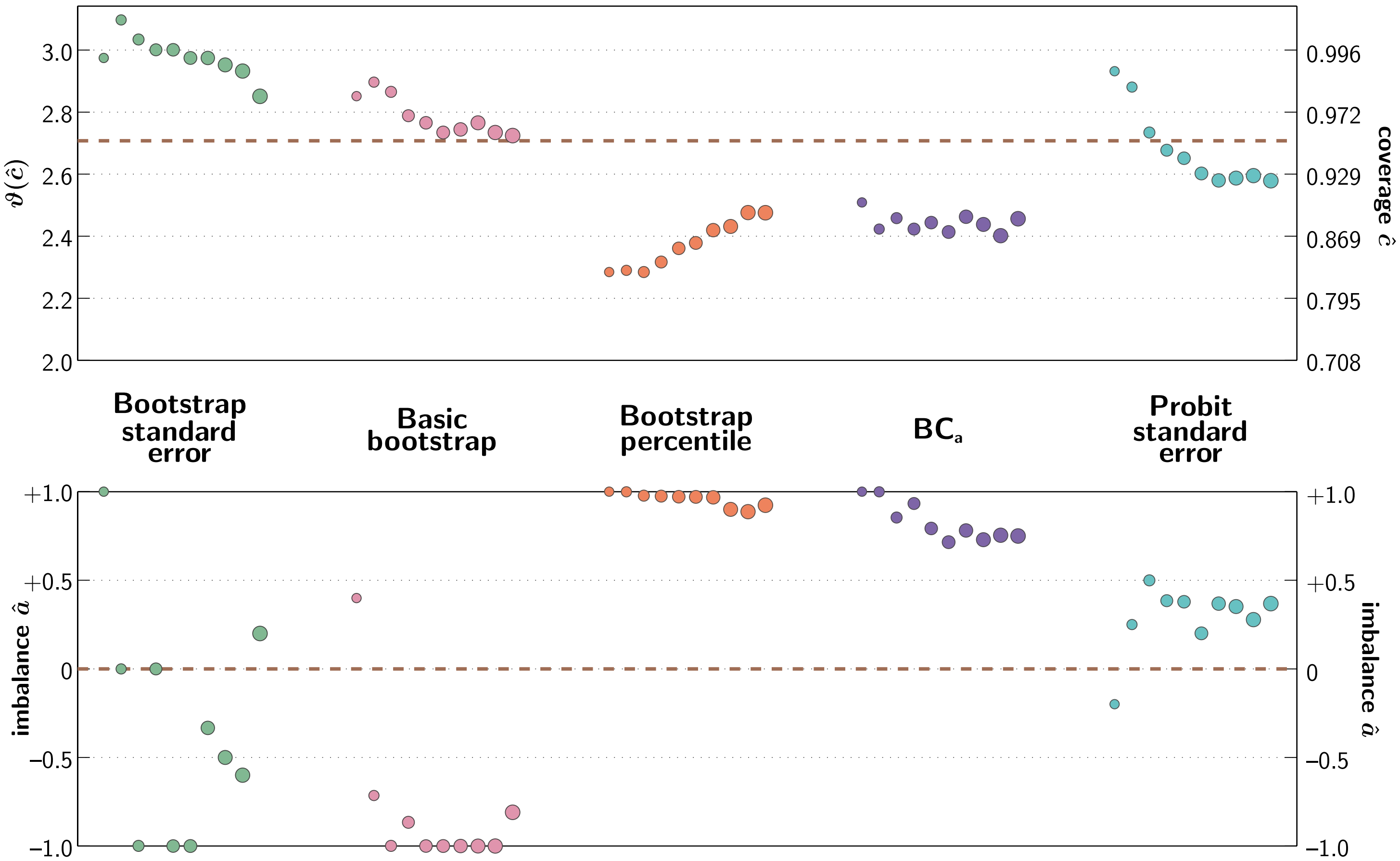


Fig. 6.8: Results of Monte Carlo coverage tests for 95.4% slope confidence intervals obtained from the cumulative normal function in the realistic 2-AFC case. Each group of symbols shows, from left to right, the effect of increasing k from 3 to 12, stimulus values being chosen sequentially by the method of section 6.2.1. See section 6.2.2 for details.

the psychometric function.

However, in the form in which they have been applied to psychophysics, bootstrap methods generally assume fixed stimuli.[23,24,28,32] Such assumption may only be appropriate under (non-psychophysical) conditions in which the explanatory variable is outside the investigator's control.[26,27] The simulations of the current chapter demonstrate that an adaptive algorithm for the sequential selection of stimulus values can yield somewhat different results from those that have been previously assumed to be appropriate in Monte Carlo studies.[†] The effect on threshold confidence interval methods is to lower their coverage, and to introduce an increasing relationship between coverage and the number of observations taken. The latter effect can be understood by considering that as more blocks are taken, so the variability of the stimulus positions of *subsequent* blocks decreases, and so the fixed-stimulus approximation becomes more accurate. For slopes, intervals tend to become more positively unbalanced. This last effect suggests that the overestimation of slope previously noted by some authors[28,33] may be greater in sequential selection methods than under the assumption of fixed stimuli. The effects are more pronounced for slopes than for thresholds, occur to a greater extent in the 2-AFC design than in the yes-no design, and are exacerbated by the presence of nuisance parameters in the model.

For both thresholds and slopes, the changes may lead to different conclusions about which confidence interval methods are most appropriate (the basic bootstrap, for example, appears to perform much better for the estimation of slopes than it did under the assumption of fixed stimuli). However, any such effects may depend on the precise details of the adaptive algorithm

[†] Note that the simulations of chapter 3 show a somewhat different aspect of confidence intervals' coverage properties than those of the current chapter. The former still have value: they highlight the fact that some confidence interval methods (such as the bootstrap standard error method) are more sensitive to variations in sampling scheme than others (such as the $\mathrm{BC_a}$ method). The current results report coverage probability estimates each of which is based on a probabilistic combination of 500 *different* sampling schemes, but do not show the differences that exist between confidence interval methods in terms of the reliability of inferences that may be made given any one *particular* observed pattern of performance values.

that the experimenter employs. The solution to the problem is likely to lie in the recreation of the adaptive algorithm within the bootstrap simulations themselves, i.e. to base bootstrap confidence intervals on simulated repetition of the entire experiment, *including* the process of stimulus selection. Clearly this cannot be done accurately if stimuli are chosen sequentially “by eye”. A valuable direction for future research would therefore be to examine the application of bootstrap simulation methods to adaptive procedures of all kinds, including those that can be applied to produce efficient block-based sampling schemes, and assess the coverage accuracy of confidence intervals that are based on the repeated simulation of the adaptive procedures themselves. Bootstrap methods involving sequentially dependent data are at a relatively early stage of research (see Davison and Hinkley,[34] chapter 8), and adaptation of their theory to account for the complexity of the sequential dependencies introduced by various adaptive procedures will require considerable development.

References for chapter 6

[1] Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**(17): 2503–2522.

[2] Taylor, M. M, Forbes, S. M. & Creelman, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America*, **74**(5): 1367–74.

[3] Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 505–508.

[4] Pentland, A. (1980). Maximum likelihood estimation: the best PEST. *Perception and Psychophysics*, **28**(4): 377–379.

[5] Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**(6): 1763–9.

[6] Watson, A. B. & Pelli, D. G. (1983). QUEST: a bayesian adaptive psychometric method. *Perception and Psychophysics*, **33**(2): 113–120.

[7] Madigan, R. & Williams, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception and Psychophysics*, **42**(3): 240–249.

[8] King-Smith, P. E, Grigsby, S. S, Vingrys, A. J, Benes, S. C. & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, **34**(7): 885–912.

[9] Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, **73**: 663–667.

[10] Wetherill, G. B. & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, **18**(1): 1–10.

[11] Lam, C. F, Dubno, J. R, Ahlstrom, J. B, He, N. J. & Mills, J. H. (1997). Estimating parameters for psychometric functions using the four-point sampling method. *Journal of the Acoustical Society of America*, **102**(6): 3697–3703.

[12] Henning, G. B, Millar, R. W. & Hill, N. J. (2000). Detection of incremental and decremental bars at different locations across mach bands and related stimuli. *Journal of the Optical Society of America A*, **17**(7): 1147–1159.

[13] Simpson, W. A. (1988). The method of constant stimuli is efficient. *Perception and Psychophysics*, **44**(5): 433–436.

[14] 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.

[15] Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, **47**(1): 87–91.

[16] Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 467–477.

[17] Watt, R. J. & Andrews, D. P. (1981). APE: adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, **1**(2): 205–213.

[18] King-Smith, P. E. (1984). Efficient threshold estimates from yes-no procedures using few (about 10) trials. *American Journal of Optometry and Physiological Optics*, **61**: 119P.

[19] King-Smith, P. E. & Pierce, G. E. (1994). Unbiased estimates of the slope of the psychometric function. *Investigative Ophthalmology and Visual Science (Supplement)*, **35**: 1295.

[20] King-Smith, P. E. & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, **37**(12): 1595–1604.

[21] KONTSEVICH, L. L. & TYLER, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, **39**(16): 2729–2737.

[22] PELLI, D. G. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Supplement)*, **28**: 366.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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 .

[33] 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.

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

7. Conclusions

Section 7.1 summarizes the principal findings of the current research, and is reproduced in the Extended Abstract, page 16ff.

Section 7.2 (page 257ff.) summarizes recommendations for the future development of hypothesis testing methods applied to psychometric functions.

7.1 Summary of findings

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 $\mathrm{BC_a}$ 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[4] 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,[5–7] 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[8–10] 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,[8,11,12] 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[13] and illustrated by Treutwein and Strasburger.[9] 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.[3,12,14,15] 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.

7.2 Recommendations

In their current form, bootstrap confidence intervals for the threshold and slope of a psychometric function should be interpreted with a conservative eye. Under many circumstances, the observed coverage probability of a confidence interval were found to fall below the target confidence level. In some cases, even when overall coverage was found to be accurate, the interval was not well balanced, so that coverage was too low in one of the two tails of the interval. A heuristic such as the expanded bootstrap method (suggested by Wichmann and Hill[15] and developed in section 2.2.6) may offer a way of ensuring certain minimum levels of coverage. However, the current version of the expanded method is not ideal because, notwithstanding its conservatism, it is somewhat unstable: its coverage may vary widely above the target level, and may be more or less unbalanced, depending on the distribution of stimulus values relative to the true curve.

The bootstrap-t and $\mathrm{BC_a}$ methods can offer improvements in stability for both threshold and slope intervals relative to simpler bootstrap methods. However, they can also suffer from imbalance and low coverage, which may be due to their reliance on the parametric approximation to the parameter covariance matrix obtained by inverting the expected Fisher information matrix. Reliance on the Fisher matrix may be particularly sensitive to parameter mis-estimation: certainly the imbalance of $\mathrm{BC_a}$ threshold intervals in the realistic 2-AFC case was linked to inaccuracy in the estimation of λ (see section 3.3.4). Future simulations might examine variations on the bootstrap-t and $\mathrm{BC_a}$ methods that use non-parametric alternatives to the Fisher matrix.

Some such alternatives are based on various forms of the jackknife technique, the application of which is outlined in the general case by Efron and Tibshirani[16] and by Davison and Hinkley.[17] Three different jackknife methods have been applied to the computation of bootstrap-t confidence intervals in the context of logistic regression by Swanepoel and Frangos.[4]

Confidence *region* methods offer potentially more powerful hypothesis tests than one-dimensional interval methods. One method in particular, the bootstrap deviance method, was also found to be fairly well balanced and very stable with regard to variations of sampling scheme, and tended to exceed its overall target coverage level. The method requires further development however, to adapt it fully for use in a situation in which nuisance parameters must be estimated. A contour-drawing algorithm might be used, for example, to compute region boundaries in three or four dimensions, which are then "flattened" into the two dimensions of interest. Thus, on any given bearing ϕ from the initial estimate $(\hat{\alpha}_0, \hat{\beta}_0)$, the radial distance d_ϕ from the estimate to the region boundary would be equal to the maximum d_ϕ encountered in a series of two-dimensional sections of the three- or four-dimensional region, each section being indexed by a different set of nuisance parameter values. Further testing would then be required in order to ensure that the modified method retained the desirable properties of balance, stability and conservatism.

All the confidence region methods studied (including, to a relatively minor extent, the bootstrap deviance method) showed signs of positive imbalance in the slope dimension, which is consistent with the tendency for slopes to be over-estimated. Another example of the relationship between the accuracy of the initial estimate and the accuracy of the confidence interval boundaries is the link between bias in the estimation of λ (and hence in the estimation of slope, which co-varies with λ) and imbalance in $\mathrm{BC_a}$ threshold intervals. Clearly, it is desirable that fitting methods be developed to address the problem of estimation bias, not only for its own sake, but also in order to improve the accuracy of confidence intervals and confidence regions. Potentially significant improvements might be made by attempting to reduce the influence of the nuisance parameters λ and γ. One approach would be to modify the fitting procedure to use a loss function that is less sensitive to extreme outliers than the likelihood metric. As Finney[1] notes, "...there is no *a priori* reason why minimum χ^2 should not be superior to maximum likelihood in small samples, or why some third method should not be superior to either" (page 52). The use of a more robust "third method" might reduce the influence of λ and γ, or perhaps even eliminate the need for them altogether. Naturally any such revision to the core fitting process would require extensive testing to ensure it achieved reductions in bias and improvements in confidence interval coverage over a wide range of conditions, while keeping any concomitant loss of efficiency within acceptable levels.

Finally, as suggested in chapter 6, another valuable line of development might be to incorporate simulations of sequential stimulus selection algorithms into the bootstrap, in order to address the problem of low coverage that may occur when stimuli are assumed to be fixed. The application of improvements such as the bootstrap-t or $\mathrm{BC_a}$ methods to such a framework might also be valuable, but is likely to require considerable theoretical development.

References for chapter 7

[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] 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.

[5] 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.

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

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

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

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

[10] 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 .

[11] 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.

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

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

[14] 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.

[15] 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 .

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

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

List of references

Beran, R. (1988). Balanced simultaneous confidence sets. *Journal of the American Statistical Association*, **83**(403): 679–686.

Beran, R. (1990). Refining bootstrap simultaneous confidence sets. *Journal of the American Statistical Association*, **85**(410): 417–426.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, **71**: 353–360.

Chang, Y. C. I. & Martinsek, A. T. (1992). Fixed size confidence regions for parameters of a logistic regression model. *Annals of Statistics*, **20**(4): 1953–1969.

Cox, D. R. & Snell, E. J. (1989). *Analysis of Binary Data.* London: Chapman and Hall., second edition.

Dai, H. P. (1995). On measuring psychometric functions: a comparison of the constant-stimulus and adaptive up-down methods. *Journal of the Acoustical Society of America*, **98**(6): 3135–3139.

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

Edwards, A. W. F. (1992). *Likelihood.* Baltimore: Johns Hopkins University Press. Expanded Edition.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, **82**: 171–200.

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

Emerson, P. L. (1984). Observations on a maximum likelihood method of sequential testing and a simplified approximation. *Perception and Psychophysics*, **36**: 199–203.

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

Foster, D. H. (1986). Estimating the variance of a critical stimulus level from sensory performance data. *Biological Cybernetics*, **53**: 189–194. An erratum is given on page 412 of the same volume.

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.

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.

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.

Freeman, M. F. & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, **21**(4): 607–611.

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.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley.

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

Hall, J. L. (1981). Hybrid adaptive procedure for estimation of psychometric functions. *Journal of the Acoustical Society of America*, **69**(6): 1763–9.

Hall, J. L. (1983). A procedure for detecting variability of psychophysical thresholds. *Journal of the Acoustical Society of America*, **73**: 663–667.

Hall, P. (1987). On the bootstrap and likelihood-based confidence regions. *Biometrika*, **74**(3): 481–493.

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

Hawley, M. L. (1990). Comparison of adaptive procedures for obtaining psychophysical thresholds using computer simulation. Master's thesis, Boston University.

Hawley, M. L. & Colburn, H. S. (1995). Application of confidence intervals and joint confidence regions to the estimation of psychometric functions. *Journal of the Acoustical Society of America*, **97**: 3277.

Henning, G. B, Millar, R. W. & Hill, N. J. (2000). Detection of incremental and decremental bars at different locations across mach bands and related stimuli. *Journal of the Optical Society of America A*, **17**(7): 1147–1159.

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.

Hinkley, D. V. (1978). Likelihood inference about location and scale parameters. *Biometrika*, **65**(2): 253–261.

Jennings, D. E. (1986). Judging inference adequacy in logistic regression. *Journal of the American Statistical Association*, **81**(394): 471–476.

Kendall, M. K. & Stuart, A. (1979). *The Advanced Theory of Statistics, Volume 2: Inference and Relationship.* New York: Macmillan, fourth edition.

King-Smith, P. E. (1984). Efficient threshold estimates from yes-no procedures using few (about 10) trials. *American Journal of Optometry and Physiological Optics*, **61**: 119P.

King-Smith, P. E, Grigsby, S. S, Vingrys, A. J, Benes, S. C. & Supowit, A. (1994). Efficient and unbiased modifications of the QUEST threshold method: Theory, simulations, experimental evaluation and practical implementation. *Vision Research*, **34**(7): 885–912.

King-Smith, P. E. & Pierce, G. E. (1994). Unbiased estimates of the slope of the psychometric function. *Investigative Ophthalmology and Visual Science (Supplement)*, **35**: 1295.

King-Smith, P. E. & Rose, D. (1997). Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, **37**(12): 1595–1604.

KONTSEVICH, L. L. & TYLER, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, **39**(16): 2729–2737.

KONTSEVICH, L. L. & TYLER, C. W. (1999). Distraction of attention and the slope of the psychometric function. *Journal of the Optical Society of America*, **A16**: 217–222.

LAM, C. F, DUBNO, J. R, AHLSTROM, J. B, HE, N. J. & MILLS, J. H. (1997). Estimating parameters for psychometric functions using the four-point sampling method. *Journal of the Acoustical Society of America*, **102**(6): 3697–3703.

LAM, C. F, DUBNO, J. R. & MILLS, J. H. (1999). Determination of optimal data placement for psychometric function estimation: a computer simulation. *Journal of the Acoustical Society of America*, **106**(4, pt. 1): 1969–1976.

LAM, C. F, MILLS, J. H. & DUBNO, J. R. (1996). Placement of observations for the efficient estimation of a psychometric function. *Journal of the Acoustical Society of America*, **99**(6): 3689–3693.

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

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

LEEK, M. R, HANNA, T. E. & MARSHALL, L. (1992). Estimation of psychometric functions from adaptive tracking procedures. *Perception and Psychophysics*, **51**(3): 247–256.

LEVITT, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 467–477.

MADIGAN, R. & WILLIAMS, D. (1987). Maximum-likelihood psychometric procedures in two-alternative forced-choice: Evaluation and recommendations. *Perception and Psychophysics*, **42**(3): 240–249.

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

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall, second edition.

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.

Nachmias, J. (1981). On the psychometric function for contrast detection. *Vision Research*, **21**(2): 215–223.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, **7**(4): 308–313.

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.

Patterson, V. H, Foster, D. H. & Heron, J. R. (1980). Variability of visual threshold in Multiple Sclerosis. *Brain*, **103**: 139–147.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *Journal of the Optical Society of America*, **2**: 1508–1532.

Pelli, D. G. (1987). The ideal psychometric procedure. *Investigative Ophthalmology and Visual Science (Supplement)*, **28**: 366.

Pentland, A. (1980). Maximum likelihood estimation: the best PEST. *Perception and Psychophysics*, **28**(4): 377–379.

Robbins, H. & Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, **22**: 400–407.

Simpson, W. A. (1988). The method of constant stimuli is efficient. *Perception and Psychophysics*, **44**(5): 433–436.

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.

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

Taylor, M. M. (1971). On the efficiency of psychophysical measurement. *Journal of the Acoustical Society of America*, **49**(2, pt. 2): 505–508.

Taylor, M. M. & Creelman, C. D. (1967). PEST: Efficient estimates on probability functions. *Journal of the Acoustical Society of America*, **41**(4): 782–787. An erratum is given in issue 42, number 5, page 1097.

Taylor, M. M, Forbes, S. M. & Creelman, C. D. (1983). PEST reduces bias in forced choice psychophysics. *Journal of the Acoustical Society of America*, **74**(5): 1367–74.

Teller, D. Y. (1985). Psychophysics of infant vision: Definitions and limitations. In Gottlieb, G. & Krasnegor, N (Eds.), *Measurement of Audition and Vision in the First Year of Postnatal Life: a Methodological Overview.* Norwood, NJ: Ablex.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, **35**(17): 2503–2522.

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.

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

Tyler, C. W. (1997). Why we need to pay attention to psychometric function slopes. *Vision Science and its Applications, Technical Digest*, **1**: 240–243.

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

Watson, A. B. & Fitzhugh, A. (1990). The method of constant stimuli is inefficient. *Perception and Psychophysics*, **47**(1): 87–91.

Watson, A. B. & Pelli, D. G. (1983). QUEST: a bayesian adaptive psychometric method. *Perception and Psychophysics*, **33**(2): 113–120.

Watt, R. J. & Andrews, D. P. (1981). APE: adaptive probit estimation of psychometric functions. *Current Psychological Reviews*, **1**(2): 205–213.

Wetherill, G. B. (1963). Sequential estimation of quantal reponse curves. *Journal of the Royal Statistical Society, Series B*, **25**(1): 1–48.

Wetherill, G. B. & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, **18**(1): 1–10.

Wichmann, F. A. (1999). *Some Aspects of Modelling Human Spatial Vision: Contrast Discrimination*. PhD thesis, University of Oxford, UK.

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 .

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 .

Yuan, H. (1999). Comparative study of adaptive psychophysical procedures.(threshold estimation, maximum likelihood). *Dissertation Abstracts International: Section B: The Sciences and Engineering*, **60**(4-b): 1910.

Appendix A

Notation

Lower-case Greek, and lower or upper case italic letters generally denote scalar values (c, k, R, α, σ). Vectors are denoted by lower-case bold symbols ($\boldsymbol{x}$, $\boldsymbol{n}$, $\boldsymbol{\theta}$, $\boldsymbol{\delta}$) and matrices by upper-case bold symbols ($\boldsymbol{I}$, $\boldsymbol{V}$). Estimates are be denoted by a "hat" ($\hat{\boldsymbol{\theta}}$, $\hat{a}$, $\widehat{\mathrm{se}}$).

An asterisk (*) denotes bootstrap or Monte Carlo replications of a certain quantity, and a subscript may also denote the simulation number, so that $\hat{u}_i^*$ is the i $^{\mathrm{th}}$ replication of the estimate $\hat{u}$. (N.B. the * is never used for footnotes—the symbols † and ‡ are used instead.) Without the subscript, the starred symbol refers to the entire set of simulated values but, even though the set contains multiple items, the symbol is not made bold unless the item that is being replicated was itself a vector. Thus, $\hat{u}^*$ is a set of simulated scalars $\hat{u}_i^*$, and $\hat{\boldsymbol{\theta}}^*$ is a set of simulated vectors $\hat{\boldsymbol{\theta}}_i^*$. A subscript in parentheses, as in $u_{(\varepsilon)}^*$ denotes an estimated quantile of a distribution (see page 55).

$\Phi(\cdot)$ denotes the cumulative of the standard normal distribution, $\Pr(\cdot)$ probability, $\exp(\cdot)$ the exponential and $\mathrm{Bi}(\cdot)$ the binomial distribution.

A list of symbols and their specific meanings follows.

2-AFC Two-alternative forced choice.

a Imbalance of coverage between the tails of a two-tailed interval (section 3.1.1).

$\boldsymbol{b}$ Vector of two or three parameters of $B(\cdot)$ or $B_3(\cdot)$. See section 5.5.1.

$B(\cdot)$ Beta distribution (section 5.5.1).

$B_3(\cdot)$ Modified (three-parameter) beta distribution (section 5.5.1).

$\mathrm{Bi}(n, p)$ Binomial distribution, with probability of success p in n trials.

c Coverage probability of a confidence interval or confidence region.

C Number of simulated repetitions of an experiment, each including determination of confidence limits, in a Monte Carlo coverage test.

CPE Cumulative probability estimate. See equation (2.6), page 55.

d_i Deviance residual (section 1.2.3).

D Deviance or log likelihood ratio, $D = \sum_{i=1}^{k} d_i^2$ (section 1.2.3).

$\mathrm{E}\{\cdot\}$ Expected value.

f Detection level $f = F(x;\, \alpha, \beta)$.

$F(x)$ The underlying two-parameter "detection probability" function that defines the shape of the psychometric function. The location and spread of $F(x)$ are determined by α and β, the first two parameters from the parameter vector $\boldsymbol{\theta}$. Its range is from 0 to 1, inclusive or exclusive, and its shape is usually some sort of sigmoid: examples include the cumulative normal, logistic and Weibull functions (section 1.2.1).

g Used in the computation of fiducial limits on thresholds in probit analysis (see section 2.1).

$_{\mathrm{gen}}$ Subscript denoting the true or "generating" value of a quantity (usually unknown to the experimenter).

h Smoothing parameter in density estimation (section 4.2).

$\mathrm{I}\{\cdot\}$ The indicator function: 1 when its argument is true, 0 otherwise.

$\boldsymbol{I}$ The Fisher information matrix of the parameters $\boldsymbol{\theta}$ (see section B.2).

k Number of blocks in single set of psychophysical data.

K Sweat factor (section 5.1.2).

$L(\boldsymbol{\theta})$ Likelihood (section B.2).

$\ell(\boldsymbol{\theta})$ Log-likelihood (section B.2).

m Number of alternatives in a forced-choice task, in which chance performance $\gamma = 1/m$.

$M(\cdot)$ Any model that is applied to psychophysical data, to fit and predict the proportion of positive or correct responses in response to stimulus x given additional explanatory variables $\boldsymbol{\rho}$.

MLE Maximum-likelihood estimate.

n Number of observations per block. Vector $\boldsymbol{n}$ of length k represents the block sizes in a single data set.

N Total number of observations in a data set. $N = \sum_{i=1}^{k} n_i$.

p Expected proportion of correct or positive responses. Vector $\boldsymbol{p}$ of length k represents the expected performance values from a fit to a single data set. $p_i = \psi(x_i; \boldsymbol{\theta})$.

$\bar{p}$ Mean of expected performance values p_i, weighted by n_i. Used to characterize sampling schemes—see section 5.5.3.

$\Pr(\cdot)$ Probability.

r Number of correct or positive responses in a block. Vector $\boldsymbol{r}$ of length k represents the performance values in a single data set.

R Number of simulated repetitions of an experiment, in a single bootstrap or Monte Carlo hypothesis test.

se Standard error.

s_f Psychometric function slope $\mathrm{d}F/\mathrm{d}x$ evaluated at $x = t_f$.

t_f Psychophysical threshold, i.e. the value of x such that $F(x; \alpha, \beta) = f$.

$^{\mathrm{T}}$ Transpose operator for vectors and matrices.

u Generic label for a measure of interest, such as a threshold or slope, that is computed from parameters $\boldsymbol{\theta}$.

$\dot{\boldsymbol{u}}$ Vector of derivatives of a measure of interest u with respect to each of the parameters $\boldsymbol{\theta}$ (see section B.1).

v
$\mathrm{var}\{\cdot\}$ Variance.

$\boldsymbol{V}$ The covariance matrix of the parameters $\boldsymbol{\theta}$ (see section B.2).

w Bias-correction term (section 2.2.5), and similarly estimation bias (section 5.1.1).

W_i Probit regression weighting coefficient for block i (section 2.1).

$\textsc{wci}_{\#\#}$ Width of confidence interval. ($\textsc{wci}_{68}$ = the width of a 68.3% confidence interval, and $\textsc{wci}_{95}$ = the width of a 95.4% confidence interval).

$\textsc{wci}_{\mathrm{min}}$ Width of the smallest confidence interval, corresponding to the most efficient sampling scheme in a set of simulations (section 5.5).

$\textsc{wnpi}_{\#\#}$ Width of non-parametric percentile interval. The $\textsc{wnpi}_{68}$ of a distribution X is equal to $X_{(0.841)} - X_{(0.159)}$. The $\textsc{wnpi}_{95}$ of X is equal to $X_{(0.977)} - X_{(0.023)}$. Division by $\frac{1}{2}\textsc{wnpi}_{68}$ is sometimes used as a robust method of standardization (for perfectly normal distributions, it is equivalent to division by the standard deviation).

x Stimulus intensity. Vector $\boldsymbol{x}$ of length k represents the stimulus values in a single data set.

y Observed proportion of correct or positive responses. Vector $\boldsymbol{y}$ of length k represents the performance values in a single data set, $y_i = r_i/n_i$.

α The first of the four parameters $\boldsymbol{\theta}$ of the psychometric function $\psi(x;\, \boldsymbol{\theta})$, α is used in the underlying function $F(x)$. Its precise rôle depends on the form of F (see section B.1), although it is usually related to the location of the psychometric function relative to the x-axis.

α_{pr} Location parameter in probit regression (section 2.1).

β The second of the four parameters $\boldsymbol{\theta}$ of the psychometric function $\psi(x;\, \boldsymbol{\theta})$, β is used in the underlying function $F(x)$. Its precise rôle depends on the form of F (see section B.1), although it is usually related to the spread or slope of the psychometric function along the x-axis.

β_{pr} Scale parameter in probit regression (section 2.1).

γ The third of the four parameters $\boldsymbol{\theta}$ of the psychometric function $\psi(x;\, \boldsymbol{\theta})$, γ represents the lower asymptote of the curve. In m-AFC designs, γ is fixed at $1/m$. In yes-no designs, it might be fixed at 0 under "idealized" assumptions, or it might need to be estimated, in which case it generally takes a small non-zero value that reflects the observer's "guess rate".

$\boldsymbol{\delta}_u$ Least-favourable unit direction vector for inference about a measure of interest u. Used in the estimation of the skewness correction factor ξ in the $\mathrm{BC_a}$ bootstrap method (section 2.2.5).

ε Confidence level corresponding to a confidence limit. For a two-tailed confidence interval of coverage $1 - 2\eta$, $\varepsilon_{\mathrm{LO}} = \eta$ and $\varepsilon_{\mathrm{UP}} = 1 - \eta$.

$\zeta(\cdot)$ Unknown normalizing transformation (section 2.2.4).

η Significance level associated with a confidence interval tail. In chapter 2 the target coverage probability of a two-tailed interval is $1 - 2\eta$.

$\boldsymbol{\theta}$ Vector of psychometric function parameters—effectively a shorthand notation for $(\alpha, \beta, \gamma, \lambda)^{\mathrm{T}}$.

$\vartheta(\cdot)$ Freeman-Tukey variance-stabilizing transformation for binomial probabilities (section 3.1.2).

λ The fourth of the four parameters $\boldsymbol{\theta}$ of the psychometric function $\psi(x;\, \boldsymbol{\theta})$, λ represents the offset between the upper asymptote of the curve and 1. It might be fixed at 0 under "idealized" assumptions, or it might need to be estimated, in which case it generally takes a small value (less than, say 0.05 for a trained adult observer) that reflects the observer's rate of stimulus-independent errors or "lapses".

ξ Skewness correction factor or "acceleration" (section 2.2.5).

$\boldsymbol{\rho}$ Vector denoting additional explanatory variables, besides x, in the general model $p = M(x, \boldsymbol{\rho}; \boldsymbol{\theta})$.

$\boldsymbol{\varrho}$ Confidence limit (sections 2.2.2 and 4.1) or likelihood contour value (section 4.2).

σ_p Second central moment of expected performance values p_i, weighted by n_i. Used to characterize sampling schemes—see section 5.5.3.

$\Upsilon(\cdot)$ Link function in probit regression (section 2.1).

$\Phi(\cdot)$ The cumulative of the standard normal distribution.

φ Angle corresponding to joint error in threshold and slope, after standardization (section 4.3.2).

$\psi(x)$ The psychometric function, predicting the probability of a correct or positive response to stimulus x. $\psi(x)$ is equal to the underlying function $F(x)$, scaled between lower and upper bounds γ and $1 - \lambda$.

$\Omega(\cdot)$ Bayesian prior (section B.2.1).

Appendix B

Formulae

B.1 The psychometric function and its derivatives

The psychometric function is written as

$$\psi\left(x;\,\alpha,\beta,\gamma,\lambda\right) \;=\; \gamma+\left(1-\gamma-\lambda\right)F\left(x;\,\alpha,\beta\right) \tag{B.1}$$

where $F\left(x;\,\alpha,\beta\right)$ is a monotonic two-parameter function with range 0 to 1 (inclusive or exclusive), such as the cumulative normal (section B.1.1), logistic (section B.1.2) or Weibull (section B.1.3).

In order to compute the likelihood derivatives of section B.2, the first and second derivatives of ψ with respect to the four parameters are required. The first derivatives of ψ are as follows:

$$\frac{\partial\psi}{\partial\alpha} \;=\; \left(1-\gamma-\lambda\right)\frac{\partial F}{\partial\alpha}, \tag{B.2}$$

$$\frac{\partial\psi}{\partial\beta} \;=\; \left(1-\gamma-\lambda\right)\frac{\partial F}{\partial\beta}, \tag{B.3}$$

$$\frac{\partial\psi}{\partial\gamma} \;=\; 1-F\left(x;\,\alpha,\beta\right), \tag{B.4}$$

$$\frac{\partial \psi}{\partial \lambda} \;=\; -F\left(x;\,\alpha,\beta\right), \tag{B.5}$$

and its second derivatives are given in table B.1:

$\partial^2\psi$	$\partial\alpha$	$\partial\beta$	$\partial\gamma$	$\partial\lambda$
$\partial\alpha$	$(1{-}\gamma{-}\lambda)\frac{\partial^2 F}{\partial\alpha^2}$	$(1{-}\gamma{-}\lambda)\frac{\partial^2 F}{\partial\alpha\,\partial\beta}$	$-\frac{\partial F}{\partial\alpha}$	$-\frac{\partial F}{\partial\alpha}$
$\partial\beta$	$(1{-}\gamma{-}\lambda)\frac{\partial^2 F}{\partial\beta\,\partial\alpha}$	$(1{-}\gamma{-}\lambda)\frac{\partial^2 F}{\partial\beta^2}$	$-\frac{\partial F}{\partial\beta}$	$-\frac{\partial F}{\partial\beta}$
$\partial\gamma$	$-\frac{\partial F}{\partial\alpha}$	$-\frac{\partial F}{\partial\beta}$	0	0
$\partial\lambda$	$-\frac{\partial F}{\partial\alpha}$	$-\frac{\partial F}{\partial\beta}$	0	0

Table B.1: Second derivatives of ψ with respect to its four parameters.

First and second derivatives of F are given in the relevant sections below. Also provided are the formulae for the threshold t_f and slope s_f of F at any arbitrary detection level f, along with their first derivatives with respect to α and β. The derivatives of t_f and s_f are used to construct the parametric influence vector:

$$\dot{\boldsymbol{u}} = \left(\frac{\partial u}{\partial\alpha}\,,\;\frac{\partial u}{\partial\beta}\,,\;\frac{\partial u}{\partial\gamma}\,,\;\frac{\partial u}{\partial\lambda}\right)^{\mathrm{T}},$$

where u is a generic term for a measure of interest, such as a threshold or slope. Since threshold and slope are defined purely in terms of $F(x)$, they do not involve the nuisance parameters γ or λ, so the last two elements of $\dot{\boldsymbol{u}}$ are 0. The derivatives with respect to α and β depend on the shape of F, and are given in the following equations:

Cumulative normal: (B.13)–(B.14) and (B.16)–(B.17), page 278.

Logistic: (B.25)–(B.26) and (B.28)–(B.29), page 279.

Weibull: (B.37)–(B.38) and (B.40)–(B.41), page 281.

The parametric influence vector is used in the BC$_\mathrm{a}$ bootstrap method to compute the "least-favourable" direction vector for u (section 2.2.5). When the i $^\mathrm{th}$ parameter is fixed, the i $^\mathrm{th}$ element of $\dot{\boldsymbol{u}}$ is set to 0.

B.1.1 The cumulative normal function

The cumulative normal function is given by

$$F\left(x;\, \alpha, \beta\right) \;=\; \Phi\!\left(\frac{x-\alpha}{\beta}\right) \;=\; \frac{1}{\beta\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{(\tau-\alpha)^2}{2\beta^2}}\,\mathrm{d}\tau,$$

which is equivalent to

$$F\left(x;\, \alpha, \beta\right) \;=\; \frac{1}{2}\,\mathrm{erf}\left(\frac{x-\alpha}{\beta\sqrt{2}}\right) + \frac{1}{2}\,. \tag{B.6}$$

Its first and second derivatives with respect to α and β are:

$$\frac{\partial F}{\partial \alpha} \;=\; \frac{-1}{\beta\sqrt{2\pi}}\;\mathrm{e}^{-\frac{(x-\alpha)^2}{2\beta^2}}, \tag{B.7}$$

$$\frac{\partial F}{\partial \beta} \;=\; \frac{\alpha - x}{\beta^2\sqrt{2\pi}}\;\mathrm{e}^{-\frac{(x-\alpha)^2}{2\beta^2}}, \tag{B.8}$$

$$\frac{\partial^2 F}{\partial \alpha^2} \;=\; \frac{\alpha - x}{\beta^3\sqrt{2\pi}}\;\mathrm{e}^{-\frac{(x-\alpha)^2}{2\beta^2}}, \tag{B.9}$$

$$\frac{\partial^2 F}{\partial \beta^2} \;=\; \frac{\left[2 - (x-\alpha)^2/\beta^2\right](x-\alpha)}{\beta^5\sqrt{2\pi}}\;\mathrm{e}^{-\frac{(x-\alpha)^2}{2\beta^2}}, \tag{B.10}$$

$$\frac{\partial^2 F}{\partial \alpha\,\partial \beta} \;=\; \frac{\partial^2 F}{\partial \beta\,\partial \alpha} \;=\; \frac{(\beta + x - \alpha)(\beta - x + \alpha)}{\beta^4\sqrt{2\pi}}\;\mathrm{e}^{-\frac{(x-\alpha)^2}{2\beta^2}}. \tag{B.11}$$

Threshold t_f is given by

$$t_f = \alpha + \beta\sqrt{2}\,\mathrm{erf}^{-1}(2f-1), \tag{B.12}$$

and its derivatives with respect to α and β are

$$\frac{\partial t_f}{\partial \alpha} = 1, \tag{B.13}$$

$$\frac{\partial t_f}{\partial \beta} = \sqrt{2}\,\mathrm{erf}^{-1}(2f-1). \tag{B.14}$$

Slope s_f is given by

$$s_f = \frac{1}{\beta\sqrt{2\pi}}\,\mathrm{e}^{-\left[\mathrm{erf}^{-1}(2f-1)\right]^2}, \tag{B.15}$$

and its derivatives with respect to α and β are

$$\frac{\partial s_f}{\partial \alpha} = 0, \tag{B.16}$$

$$\frac{\partial s_f}{\partial \beta} = \frac{-1}{\beta^2\sqrt{2\pi}}\,\mathrm{e}^{-\left[\mathrm{erf}^{-1}(2f-1)\right]^2}. \tag{B.17}$$

B.1.2 The logistic function

The logistic function is given by

$$F\,(x;\,\alpha,\beta) = \frac{1}{1+\mathrm{e}^{-(x-\alpha)/\beta}}\ \ , \tag{B.18}$$

and its first and second derivatives with respect to α and β are:

$$\frac{\partial F}{\partial \alpha} = \frac{-\mathrm{e}^{-(x-\alpha)/\beta}}{\beta \left[1+\mathrm{e}^{-(x-\alpha)/\beta}\right]^2} \;, \tag{B.19}$$

$$\frac{\partial F}{\partial \beta} = \frac{(\alpha - x)\,\mathrm{e}^{-(x-\alpha)/\beta}}{\beta^2 \left[1+\mathrm{e}^{-(x-\alpha)/\beta}\right]^2} \;, \tag{B.20}$$

$$\frac{\partial^2 F}{\partial \alpha^2} = \frac{\left(\mathrm{e}^{\alpha/\beta} - \mathrm{e}^{x/\beta}\right)\,\mathrm{e}^{\frac{x+\alpha}{\beta}}}{\beta^2 \left(\mathrm{e}^{\alpha/\beta} + \mathrm{e}^{x/\beta}\right)^3} \;, \tag{B.21}$$

$$\frac{\partial^2 F}{\partial \beta^2} = \frac{(x-\alpha)\,\mathrm{e}^{\frac{x+\alpha}{\beta}}\left[(x-\alpha+2\beta)\,\mathrm{e}^{\frac{\alpha}{\beta}} + (\alpha - x + 2\beta)\,\mathrm{e}^{\frac{x}{\beta}}\right]}{\beta^4 \left(\mathrm{e}^{\alpha/\beta} + \mathrm{e}^{x/\beta}\right)^3} , \tag{B.22}$$

$$\frac{\partial^2 F}{\partial \alpha\,\partial \beta} = \frac{\partial^2 F}{\partial \beta\,\partial \alpha} = \frac{\mathrm{e}^{\frac{x+\alpha}{\beta}}\left[(x-\alpha+\beta)\,\mathrm{e}^{\frac{\alpha}{\beta}} + (\alpha - x + \beta)\,\mathrm{e}^{\frac{x}{\beta}}\right]}{\beta^3 \left(\mathrm{e}^{\alpha/\beta} + \mathrm{e}^{x/\beta}\right)^3} . \tag{B.23}$$

Threshold t_f is given by

$$t_f = \alpha - \beta \log\left(\frac{1}{f} - 1\right), \tag{B.24}$$

and its derivatives with respect to α and β are

$$\frac{\partial t_f}{\partial \alpha} = 1, \tag{B.25}$$

$$\frac{\partial t_f}{\partial \beta} = -\log\left(\frac{1}{f} - 1\right). \tag{B.26}$$

Slope s_f is given by

$$s_f = \frac{f\,(1-f)}{\beta} \;, \tag{B.27}$$

and its derivatives with respect to α and β are

$$\frac{\partial s_f}{\partial \alpha} = 0, \tag{B.28}$$

$$\frac{\partial s_f}{\partial \beta} = \frac{f\,(f-1)}{\beta^2}. \tag{B.29}$$

B.1.3 The Weibull function

The Weibull function is given by

$$F\,(x;\,\alpha,\beta) = 1 - \mathrm{e}^{-(x/\alpha)^\beta}\;, \tag{B.30}$$

and is defined only when $x \geq 0$, $\alpha > 0$ and $\beta > 0$ (flat discontinuous Bayesian priors were used to constrain the parameter search to values of α and β for which the function was defined—see section B.2). The first and second derivatives of F with respect to α and β are:

$$\frac{\partial F}{\partial \alpha} = \frac{-\beta\,(x/\alpha)^\beta}{\alpha\,\mathrm{e}^{(x/\alpha)^\beta}}\;, \tag{B.31}$$

$$\frac{\partial F}{\partial \beta} = \frac{(x/\alpha)^\beta\;\log\,(x/\alpha)}{\mathrm{e}^{(x/\alpha)^\beta}}\;, \tag{B.32}$$

$$\frac{\partial^2 F}{\partial \alpha^2} = \frac{\beta\;(x/\alpha)^\beta\;\left\{1+\beta\;\left[1-(x/\alpha)^\beta\right]\right\}}{\alpha^2\;\mathrm{e}^{(x/\alpha)^\beta}}\;, \tag{B.33}$$

$$\frac{\partial^2 F}{\partial \beta^2} = \frac{(x/\alpha)^\beta\;\left[1-(x/\alpha)^\beta\right]\;\left[\log\,(x/\alpha)\right]^2}{\mathrm{e}^{(x/\alpha)^\beta}}\;, \tag{B.34}$$

$$\frac{\partial^2 F}{\partial \alpha\,\partial \beta} = \frac{\partial^2 F}{\partial \beta\,\partial \alpha} = \frac{\left(\frac{x}{\alpha}\right)^\beta\;\left\{\beta\;\log\left(\frac{x}{\alpha}\right)\;\left[\left(\frac{x}{\alpha}\right)^\beta-1\right]-1\right\}}{\alpha\;\mathrm{e}^{(x/\alpha)^\beta}}\;. \tag{B.35}$$

Threshold t_f is given by

$$t_f = \alpha \left[-\log(1-f)\right]^{\frac{1}{\beta}} \quad , \tag{B.36}$$

and its derivatives with respect to α and β are

$$\frac{\partial t_f}{\partial \alpha} = \left[-\log(1-f)\right]^{\frac{1}{\beta}} \quad , \tag{B.37}$$

$$\frac{\partial t_f}{\partial \beta} = -\frac{\alpha}{\beta^2} \left[-\log(1-f)\right]^{\frac{1}{\beta}} \log\left[-\log(1-f)\right] \quad . \tag{B.38}$$

Slope s_f is given by

$$s_f = \frac{\beta}{\alpha} (1-f) \left[-\log(1-f)\right]^{\left(1-\frac{1}{\beta}\right)} \quad , \tag{B.39}$$

and its derivatives with respect to α and β are

$$\frac{\partial s_f}{\partial \alpha} = \frac{\beta}{\alpha^2} (f-1) \left[-\log(1-f)\right]^{\left(1-\frac{1}{\beta}\right)} \quad , \tag{B.40}$$

$$\frac{\partial s_f}{\partial \beta} = \frac{1-f}{\alpha\, \beta} \left[-\log(1-f)\right]^{\left(1-\frac{1}{\beta}\right)} \left\{\beta + \log\left[-\log(1-f)\right]\right\} . \tag{B.41}$$

B.2 Log-likelihood and its derivatives

The likelihood $L(\boldsymbol{\theta})$ of a parameter set $\boldsymbol{\theta}$ given a set of responses $\boldsymbol{r}$ is equal to the probability of obtaining responses $\boldsymbol{r}$ given $\boldsymbol{\theta}$, assuming that each r_i is binomially distributed with probability of success p_i:

$$L(\boldsymbol{\theta}|\boldsymbol{r}) = \Pr(\boldsymbol{r}|\boldsymbol{\theta}) = \prod_{i=1}^{k} \left(\begin{array}{c} n_i \\ r_i \end{array} \right) {p_i}^{r_i} (1-p_i)^{n_i - r_i} \ , \tag{B.42}$$

where $p_i = \psi(x_i; \boldsymbol{\theta})$. The log-likelihood $\ell(\boldsymbol{\theta})$ is therefore given by

$$\begin{aligned} \ell(\boldsymbol{\theta}|\boldsymbol{r}) \ = \ \log L(\boldsymbol{\theta}|\boldsymbol{r}) \ &= \ \sum_{i=1}^{k} [\log n_i! - \log r_i! - \log{(n_i - r_i)!}] \\ &+ \ \sum_{i=1}^{k} [r_i \log p_i + (n_i - r_i) \log(1 - p_i)] \, . \end{aligned} \tag{B.43}$$

Note that the first term of (B.43) is independent of p_i and therefore need not be evaluated in the maximum-likelihood parameter search—likelihood is maximized simply by maximizing the second sum. Note also that, in (B.42) the probability of obtaining $r_i = 0$ when $p_i = 0$ is 1, as is the probability of obtaining or $r_i = n_i$ when $p_i = 1$. In these cases, terms in (B.43) that take the apparently undefined value $0 \log 0$ should therefore be evaluated as 0 (note also that $\lim_{b \to 0} \{a \log b\} = 0$ when $a = 0$).

The partial first derivative of $\ell(\boldsymbol{\theta})$ with respect to one of the parameters θ_i is

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{z=1}^{k} \left(\frac{r_z}{p_z} - \frac{n_z - r_z}{1 - p_z} \right) \left. \frac{\partial \psi}{\partial \theta_i} \right|_{x_z} \ , \tag{B.44}$$

and the partial cross-derivative with respect to two parameters θ_i and θ_j is

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \theta_i \, \partial \theta_j} = \sum_{z=1}^{k} \ &\left\{ \left(\frac{r_z}{p_z} - \frac{n_z - r_z}{1 - p_z} \right) \left. \frac{\partial^2 \psi}{\partial \theta_i \, \partial \theta_j} \right|_{x_z} \right. \\ &\left. - \left[\frac{r_z}{p_z^2} + \frac{n_z - r_z}{(1 - p_z)^2} \right] \left. \frac{\partial \psi}{\partial \theta_i} \right|_{x_z} \left. \frac{\partial \psi}{\partial \theta_j} \right|_{x_z} \right\} , \end{aligned} \tag{B.45}$$

where the relevant first and second derivatives of ψ are given in section B.1.

Equation (B.44) is used to compute the vector $\frac{\partial \ell}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \left(\frac{\partial \ell}{\partial \alpha}, \frac{\partial \ell}{\partial \beta}, \frac{\partial \ell}{\partial \gamma}, \frac{\partial \ell}{\partial \lambda} \right)$ which is used to estimate the skewness correction factor in the $\mathrm{BC_a}$ bootstrap method (section 2.2.5).

Equation (B.45) is used to compute the observed Fisher information matrix,

$$-\left.\frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\hat{\boldsymbol{\theta}}} ,$$

or the expected Fisher information matrix

$$\hat{\boldsymbol{I}} = -\mathrm{E}\left.\left\{ \frac{\partial^2 \ell}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathrm{T}}} \right\}\right|_{\hat{\boldsymbol{\theta}}} . \qquad \text{(B.46)}$$

The latter is obtained by making the substitution $r_i = p_i n_i$ in equations(B.44) and (B.45), and is used in the current study to provide an approximation to the parameter covariance matrix $\hat{\boldsymbol{V}} = \hat{\boldsymbol{I}}^{-1}$. When the i th parameter is fixed (for example, the third parameter, γ, is fixed in forced-choice designs), $\hat{I}_{ii}$ is set to 1 and the rest of the i th row and i th column are set to 0. The same pattern then appears in the corresponding row and column of $\hat{\boldsymbol{V}}$.

B.2.1 Bayesian priors

If a Bayesian prior is used to reflect the experimenter's prior knowledge about the psychometric function, it takes the form of a probability weighting function $\Omega(\boldsymbol{\theta})$ which acts as a multiplier for $L(\boldsymbol{\theta})$. Therefore, $\log\ \Omega(\boldsymbol{\theta})$ is added to $\ell(\boldsymbol{\theta})$. In the current study, only flat priors were used, but they could be discontinuous. The four parameters of the psychometric function were treated independently, so

$$\Omega(\boldsymbol{\theta}) = \Omega_\alpha(\alpha)\ \Omega_\beta(\beta)\ \Omega_\gamma(\gamma)\ \Omega_\lambda(\lambda).$$

Flat discontinuous priors were used to constrain λ (and γ in yes-no designs) within the interval $[0, 0.5]$ so, for example,

$$\begin{aligned} \Omega_\lambda(\lambda) &= 1\,; \quad 0 \leq \lambda \leq 0.05, \\ \Omega_\lambda(\lambda) &= 0\,; \quad \text{otherwise.} \end{aligned}$$

This forces the maximum-likelihood value of λ to lie within the desired interval, because $-\infty$ is added to $\ell(\boldsymbol{\theta})$ when λ is outside.

Discontinuity in the likelihood surface will cause problems for many search algorithms unless special care is taken. Many search procedures use the approach of gradient descent, in which a sudden jump to $\ell = -\infty$ provides no information and causes the parameter estimate to be undefined. The current study used the simplex search method, however, which automatically copes with a sheer drop in the log-likelihood surface: whenever log-likelihood drops (whether by a finite or infinite amount) it simply withdraws its last step and tries a smaller step. If the MLE lies at or beyond the edge, the simplex will converge on a point arbitrarily close to the edge, depending on the maximum number of iterations allowed and the minimum tolerance for fractional improvement.

The application of the prior $\Omega(\boldsymbol{\theta})$ adds

$$\frac{1}{\Omega(\boldsymbol{\theta})} \left. \frac{\partial \Omega}{\partial \theta_i} \right|_{\boldsymbol{\theta}}$$

to the log-likelihood derivative (B.44), and

$$\frac{1}{[\Omega(\boldsymbol{\theta})]^2} \left. \frac{\partial \Omega}{\partial \theta_i} \right|_{\boldsymbol{\theta}} \left. \frac{\partial \Omega}{\partial \theta_j} \right|_{\boldsymbol{\theta}} + \frac{1}{\Omega(\boldsymbol{\theta})} \left. \frac{\partial^2 \Omega}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta}}$$

to the cross-derivative (B.45). For the current application, using flat priors, the additions are 0 for all practical purposes.