Rejoinder

G. Alastair Young

I am grateful to the Editors for arranging such a perceptive and informative discussion of my article. Discussants range from the designer of the bootstrap, through some of the principal architects of bootstrap methodology, to those at the sharp end of statistical practice who can provide sound judgments on the usefulness of the bootstrap in action, and even to someone who is firmly antibootstrap. I should thank the discussants especially for providing a number of concrete examples, more incisive than those in my article, which both endorse and refute some of the arguments I put forward. The discussion also highlights a number of potential applications of bootstrap which were not described in my article.

There are a number of themes which recur throughout the discussion. I should like to make some brief remarks on these before, in random order, replying to other points raised by each of the discussants in turn.

THINK FIRST, THEN BOOTSTRAP

The bootstrap is no surrogate for careful thought on a statistical problem and, despite the name, it is foolish to think of it as such or to portray it as such. The bootstrap must be applied consciously, not blindly. Applied blindly, the bootstrap often cannot be trusted, and it is always necessary to formulate in precise terms the problem being tackled. Instead, what bootstrap allows is a means, usually via Monte Carlo and the computer, of incorporating the fruits of careful thought into an analysis in a way that is often impossible within the restrictions of standard, off-the-shelf, statistical procedures. It is desirable that implementation of a bootstrap analysis should be automatic, but this should not be confused with the idea that bootstrap should automatically be applied.

COMMUNICATION OF IDEAS

A number of the remarks made by discussants relate to education and communication, and there is strong agreement that something needs to be done here. Among practitioners, there is much dissatisfaction with the processes by which research findings are transmitted to potential users. There is too little exposition of bootstrap in the applications-oriented literature and too much bootstrap research is driven by abstract thinking, rather than by the particular needs of specific data analyses. There are exceptions, as many of Professor Efron's contributions to the literature demonstrate, but these are too few. As well as researchers on bootstrap methodology becoming more involved in applications, so that the really relevant research questions can be formulated, broadcasting the bootstrap message requires, as Professor Beran points out, a considerable update of statistical education. Also, communication must be a two-way process, with practitioners indicating more loudly what they would like from the bootstrap and researchers advertising their products more keenly.

THE COMPUTATIONAL PLATFORM

Another recurring theme of the discussion relates to the computational platform upon which applied statisticians do their work. Efron suggests that use of bootstrap and other resampling methods will develop rapidly with increased availability and use of interactive computing environments for data analysis such as S-PLUS. The support for this view from Drs. Meredith and Morel suggests that my assessment of the difficulties of packaging the bootstrap paradigm is unnecessarily pessimistic. The trend in applied statistics is toward the use of computing environments within which the bootstrap has a very natural place. In the meantime, like Professor Hinkley, I feel there is an urgent need for better software products specifically designed to implement bootstrap analyses.

DEVELOPMENT OF PROTOCOLS

A theme of much of the discussion relates to the rather bewildering array of potential bootstrap algorithms and of the need for stronger practical guidance of what to use when. My article focussed strongly on the complexities of the more sophisticated variants of bootstrap and suggested that such complexities inhibit adoption of bootstrap ideas by practitioners. The discussion suggests, however, that this may not be the main problem and that the primary need may be to provide practically useful guidelines on use of bootstrap in a number of quite specific settings. Hinkley suggests that simple protocols could be laid down for bootstrapping

in the regression setting. Meredith and Morel indicate an interest in understanding robustness of commonly used bootstrap procedures to routinely encountered departures from assumptions and also in understanding circumstances where it is better to use permutation procedures rather than bootstrap. Development of protocols ought surely to be made a focus of research in this area.

REPLY TO BERAN

Professor Beran takes me to task on a number of points. While I suspect his confidence in the abilities of skilled hands to produce practical, dependable versions of bootstrap for many kinds of complex statistical model, as well as for situations where simple bootstrapping fails, is well placed, that stage has certainly not been reached. Beran confuses the rather negative view I project of the current state of play with a pessimistic view of the future and the whole bootstrap industry. The overall tone of much of the discussion does little to refute my suggestion that much theoretical work has been focussed on problems which theoreticians find important or interesting, rather than on the more awkward versions of simpler statistical problems which practitioners consider important. My article offers a snapshot of the field as it appears today. While I plead as guilty as anyone, I believe most of the discussion supports my view that bootstrap research is tackling practical concerns only obliquely. Of course watertight guarantees can never be issued, but before submitting to the surgeon's knife the patient deserves to have a reasonable picture of its likely operative value for his or her condition.

A point revealed by Beran needs restating. The bootstrap research literature has tended to be somewhat inward looking, pursuing accurate statistical procedures through the bootstrap alone. Instead, bootstrap should be viewed as complementary to other methods, in providing a means of calibrating, and hence improving, other procedures. Beran gives the example of bootstrap calibration of empirical likelihood confidence intervals and, indeed, for that purpose bootstrap has much to offer. Hall and La Scala (1990) show how coverage error comparable to that obtained from a double bootstrap confidence interval can be obtained by a bootstrap calibration, using a single level of resampling, of empirical likelihood.

REPLY TO EFRON

Let me make again the point I make in the paper concerning practical use of the bootstrap. I do not dispute that the bootstrap is used quite a bit in practice. My concerns relate to the type of use. Efron's example typifies the main use of bootstrap to be found in the applied literature: simple, classical, ideas of error assessment being applied to clumsy, awkward and difficult problems which do not fit easily into the simple mathematical formulations of classical statistical analysis which have been the predominant focus of theoretical research. Efron's analysis is persuasive of the utility of bootstrap, but, especially when set against the contribution of Grambsch, Cowles and Louis, one wonders whether in the hands of someone less expert—and one unaware of the subtleties behind use of the bootstrap in such settings—it would produce answers of such practical value.

REPLY TO GRAMBSCH, COWLES AND LOUIS

The underlying message of my article is basically a warning that bootstrap is not all that it is sometimes cracked up to be. As Professors Grambsch, Cowles and Louis note, the apparent general applicability of bootstrap is both its strength and its weakness. Many examples of use of the bootstrap are given by discussants of my article. To my mind they validate the claim that while theoreticians have spent their time primarily in sharpening the bootstrap, to obtain refined answers in simple settings or to provide simple answers in dependent data settings, the primary appeal to the user is likely to be for simple error assessment, such as bias and variance estimation, in more non-standard settings. Grambsch, Cowles and Louis provide the example of error assessment for a loess curve, while Efron describes use of the bootstrap for error assessment with the related lowess scatterplot smoother. The contribution of Grambsch, Cowles and Louis demonstrates vividly that there are subtleties in use of bootstrap in such problems which we ignore at our peril. Again we see a reminder of why theoreticians need to spend more time listening to the interests of the applied statisticians and to focus more of their efforts into addressing issues relating to use of the bootstrap in settings that practitioners deem valuable. The apparent universality of bootstrap and the successes of theoreticians in producing workable versions of bootstrap for complicated data structures has tended to encourage an unjustified belief in bootstrap for simple, but non-standard, data structures.

REPLY TO HINKLEY

Much research effort has focussed on answering very precise questions on performance of the bootstrap: does the bootstrap produce asymptotically correct inference in some narrowly defined sense? Hinkley makes an appeal for a change of tack which

echoes my own concerns, and those of Navidi, that more must be done to examine what bootstrap does in small samples and whether the answers it produces in that setting are acceptable or not. His point that it is important to compare the results of bootstrap analyses with more traditional parametric analyses is an important one, which has not been taken on board much by bootstrap researchers. Hinkley provides two examples which shed light on what bootstrap actually does. These illustrate that bootstrap results will tend to mimic parametric model results under a best-fitting parametric model rather than the simplest model which fits the data. As bootstrap was launched as a means of relaxing assumptions of simple, tractable models, this is surely what should be wanted.

Professor Hinkley mentions, too, a number of practical concerns relating to implementation of bootstrap theory. He highlights the idea of bootstrap diagnostic plots as a means of investigating the effect of changing model assumptions or the bootstrap procedure. In the light of remarks of other discussants, the ideas he presents surely represent an important future direction for research. I should like to add two further references on procedures which enable fast execution of the double bootstrap. Lee and Young (1993a) illustrate how approximate versions of the iterated bootstrap may sometimes be implemented without any simulation. It will be of interest to compare the performance of approximate versions of these sophisticated bootstrap procedures with approximate versions, such as the ABC method, of simpler bootstrap procedures. Also, recent work by Lee and Young (1993b) has shown how sequential sampling ideas may be used to reduce computational expense.

This contribution mentions too a number of potential applications of bootstrap given little attention in my article, for which I am grateful. Finally, I should like to remark briefly on Hinkley's question on the status of the parametric bootstrap. In the same way that it is sensible to consider standard likelihood analyses in parallel to bootstrap analyses, it is surely sensible to consider parametric and nonparametric analyses in parallel. Indeed, there can be strong benefits in allowing the data analysis to choose adaptively between parametric and nonparametric bootstraps, and the answers are sensible (see Lee, 1994).

REPLY TO MEREDITH AND MOREL

Meredith and Morel share their own experiences of the difficulties inherent in the transfer of new statistical ideas from the research environment into practice. I think their point that current levels of computer literacy might be expected to expedite transition of bootstrap ideas, which at the practical level are essentially computational, endorses very firmly the views of other discussants that the fundamental problem is one of education. This is something that is difficult to see from the ivory tower.

Meredith and Morel also make a powerful point which workers in the statistical research environment, where one of the key criteria for publishability is precisely novelty of method, do well to note. In the applications literature novel statistical ideas are naturally viewed with suspicion. Greater strides must be taken to provide convincing cases of the utility of bootstrap, not just as an abstract device, but in specific applications areas.

This contribution provides concrete evidence of the extent to which bootstrap ideas are filtering into the biomedical literature, but not of the forms of bootstrap procedure found most useful in that area. It provides, too, evidence of the importance of supplemental application of bootstrap. Too much discussion of bootstrap has focussed on its use to replace standard statistical methods and not on its confirmatory use. Meredith and Morel also provide examples of problems which researchers have not yet tackled, but to which it would be sensible to turn. Once again the feeling is that research effort is not always focussed in the correct direction.

REPLY TO NAVIDI

Professor Navidi argues that bootstrap must be tailored to the dependence structure of the data and that, as such, the degree to which universal methods can be developed may be limited. While agreeing with his point, I would reiterate a key point made in my article. Procedures for bootstrapping in the dependent data setting not only lack the automatic, universal nature of Efron's original bootstrap for the independent, identically distributed setting, but also somewhat disappointingly they do not appear to enjoy the same level of immediate success as in that setting.

Navidi endorses my appeal for more work to be done on small-sample behaviour of the bootstrap and stresses the important point that uncertainties over basic assumptions, such as independence assumptions and model specification, cast into doubt the relevance of higher-order asymptotic accuracy considerations in many settings. Are ideas such as iterated bootstrap actually needed for the kinds of purpose to which they have primarily been applied, such as reducing coverage error of confidence intervals even further? Or are they going to be revealed, as both Beran and Hinkley suggest, as most valuable for fixing up deficiencies of bootstrap, in providing em-

pirical diagnostics and methods for practical implementation of the sophisticated procedures required in many settings? I feel Navidi's powerful point suggests the latter use is the one that will matter in the long run.

I am grateful to Navidi for providing a nice example on bootstrapping in the regression setting with independent errors. His example involves setting the bootstrap to work on the problem of assessing the accuracy of model selection procedures, where standard existing procedures are not very helpful. This example is revealing in that it shows that bootstrap, at least in its simplest form, does not work well. It characterizes a number of problems which make the following point. Bootstrap seems to work best in circumstances where standard alternatives exist and work reasonably, but bootstrap is most appealing in circumstances where alternatives are scarce or known to perform badly and it is precisely here where the basic bootstrap is often least successful and most in need of refinement. Similar comments apply to the problem of bootstrapping the variance of a sample quantile discussed by Hinkley.

REPLY TO SCHERVISH

Professor Schervish takes a firm stance against the bootstrap. His elegant analysis of the problem of estimating the sampling distribution of $T(X, F) = n(\theta - X_{(n)})/\theta$ illustrates well how much more may be achieved by careful dissection of a problem than by naive application of the bootstrap. However, no one pretends that this example is anything other than pathological, and I would remind him of Efron's point that real applications tend to be less pathological than clumsy and awkward. I find myself wondering how the true sampling distribution of $T(X, F)$ compares with the asymptotic distribution for the sample size $n = 50$ considered in the example. Further, as my article discusses, there is a way of bootstrapping in this problem, involving use of bootstrap samples of size m rather than n, which works and is simpler than Schervish's analysis. A fair comparison would consider too such methods, but the example does make again the important point that the bootstrap may discourage careful thinking about underlying assumptions necessary for meaningful analysis. There are echoes here of Hinkley's appeal for bootstrappers to make a precise theoretical definition of the problem they are studying. Development of protocols must surely help to encourage the practitioner toward more careful thought about the background to bootstrap analyses. Hybrid procedures of the kind discussed by Lee (1994), which explicitly combine in an adaptive way nonparametric bootstrap analyses with carefully considered parametric analysis, also help in this direction.

ADDITIONAL REFERENCES

BANKS, D. L. (1988). Histospline smoothing the Bayesian bootstrap. *Biometrika* **75** 673–684.

BARRON, A. R. (1986). Discussion of "On the consistency of Bayes estimates," by P. Diaconis and D. A. Freedman. *Ann. Statist.* **14** 26–30.

CHAMBERS, J. M. and HASTIE, T. G. (1992). *Statistical Models in S*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

CLEVELAND, W. S. and DEVLIN, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

COX, D. R. and SNELL, E. J. (1981). *Applied Statistics. Principles and Examples.* Chapman and Hall, London.

CYTEL SOFTWARE CORP. (1992). StatXact and LogXact. Cytel Software Corporation, 675 Massachusetts Avenue, Cambridge, MA 02139.

DEMING, W. E. (1956). On simplifications of sampling design through replication with equal probabilities and without stages. *J. Amer. Statist. Assoc.* **51** 24–53.

DIACONIS, P. and FREEDMAN, D. A. (1986). On the consistency of Bayes estimates (with discussion). *Ann. Statist.* **14** 1–67.

EFRON, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** 316–331.

EFRON, B. and FELDMAN, D. (1991). Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.* **86** 9–26.

FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230.

FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615–629.

FREEDMAN, D. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

FREEDMAN, D., NAVIDI, W. and PETERS, S. (1988). On the impact of variable selection in fitting regression equations. In *On Model Uncertainty and Its Statistical Implications* (T. K. Dijkstra, ed.) 1–16. Springer, Berlin.

GONG, G. (1986). Cross-validation, the jackknife, and the bootstrap: excess error estimation in forward logistic regression. *J. Amer. Statist. Assoc.* **81** 108–113.

GRAYBILL, F. A. (1961). *An Introduction to Linear Statistical Models* **1**. McGraw-Hill, New York.

HALL, P. and LA SCALA, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Rev.* **58** 109–127.

HÄRDLE, W. and BOWMAN, A. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

JANAS, D. (1993). *Bootstrap Procedures for Time Series.* Shaker, Aachen.

JÖCKEL, K.-H., ROTHE, G. and SENDLER, W., eds. (1992). *Bootstrapping and Related Techniques. Lecture Notes in Econom. and Math. Systems* **367**. Springer, Berlin.

KIPNIS, V. (1992). Bootstrap assessment of prediction in exploratory regression analysis. In *Exploring the Limits of Bootstrap* (R. LePage and L. Billard, eds.) 363–387. Wiley, New York.

LAVINE, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Ann. Statist.* **20** 1222–1235.

LEE, S. M.-S. (1994). Optimal choice between parametric and nonparametric bootstrap estimates. *Math. Proc. Cambridge Philos. Soc.* **115** 335–363.

LEE, S.M.-S. and YOUNG, G. A. (1993a). Asymptotic iterated bootstrap confidence intervals. Research Report 93-18, Statistical Laboratory, Univ. Cambridge.

LEE, S. M.-S. and YOUNG, G. A. (1993b). Sequential iterated bootstrap confidence intervals. Research Report 93-17, Statistical Laboratory, Univ. Cambridge. (To appear in *J. Roy. Statist. Soc. Ser. B.*)

LO, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Ann. Statist.* **15** 360–375.

MAHALANOBIS, P. C. (1944). On large-scale sample surveys. *Philos. Trans. Roy. Soc. London Ser. B* **231** 329–451.

MANLY, B. F. J. (1992). *A Program for Randomization Testing*. WEST, Inc., Cheyenne, WY.

MAULDIN, R. D., SUDDERTH, W. D. and WILLIAMS, S. C. (1992). Polya trees and random distributions. *Ann. Statist.* **20** 1203–1221.

MCCARTHY, P. J. (1969). Pseudo-replication: half-samples. *Internat. Statist. Rev.* **37** 239–264.

PRESS, W. H., FLANNERY, B. P., TEUKOLSKY, S. A. and VETTERLING, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing*, 1st ed. Cambridge Univ. Press.

RAO, J. N. K. and WU, C. F. J. (1988). Resampling inference with complex survey data. *J. Amer. Statist. Assoc.* **83** 231–241.

RAO, J. N. K. and YUE, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18** 209–217.

ROMANO, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17** 141–159.

RUBIN, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9** 130–134.

SCHEFFÉ, H. (1957). *The Analysis of Variance*. Wiley, New York.

SEARLE, S. R. (1971). *Linear Models*. Wiley, New York.

THERNEAU, T. M. (1993). How many stratification factors are "too many" to use in a randomization plan? *Controlled Clinical Trials* **14** 98–108.