Comment

Michael P. Meredith and Jorge G. Morel

We are very pleased for the opportunity to comment on this provocative paper by Professor Young. While we agree that the promise offered by the extensive research efforts on the bootstrap over the past decade has not been fully realized in practice, there is evidence to suggest that the use of bootstrap methods is making significant inroads into the applications literature of the biomedical sciences. A lag period between theoretical development and incorporation into practice is expected as newer methods trickle into the various areas of application. Young's expectations for the timing of this transfer from theory to practice may simply be unrealistic when viewed in historical context. As practicing biometricians in the biopharmaceutical sciences, we will address, at least in part, questions posed by Young as to why the bootstrap methods have been slow to catch on among practitioners. We will also make a few comments on areas that may serve to enhance the practical appeal of the bootstrap and thereby hasten its use in practice.

First, the bootstrap is making the slow journey into the mainstream of statistical curricula and out of the strictly research-oriented seminars and special-topics courses where students may never actually perform any bootstrap computations. Thus, the current cohort of statistics graduate students is probably the first to have such broad exposure to the basic theoretical and practical aspects of bootstrap techniques. This is evidenced by the recent proliferation of theory-oriented texts (e.g., Beran and Ducharme, 1991; Hall, 1992a; LePage and Billard, 1992; Mammen, 1992) and practical texts (e.g., Efron and Tibshirani, 1993; Noreen, 1989; Westfall and Young, 1993) to serve as a foundation on which faculty can base their courses.

Accepting that this transition to mainstream is indeed taking place, we then have a delay before some fraction of this newly educated cohort makes its way into other than academic research-oriented positions. This delay is not surprising, as exemplified by now-common methodology for the general linear model where theory was well established in the statistical literature and texts (e.g., Scheffé, 1957; Graybill, 1961; Searle, 1971) before broad evidence of practical application. Also note that there are many results in general linear model theory that have not been embraced in practice simply because they lack practical utility. The general linear model did not really become an integral part of the practicing statistical armamentarium until readily accepted and documented software such as SAS or BMDP were widely available in the 1970's and 1980's. Realistically, we should expect similar delays for the bootstrap, although one could argue that today's population of technical and scientific people are far more computer literate than those of a generation ago, thereby hastening the transition from theory to practice. It is absolutely imperative that the proper applications and limitations of bootstrap methods are clearly conveyed to students and currently practicing statisticians. This clear communication will facilitate a more rapid incorporation of bootstrap methods into routine statistical practice.

Researchers should also recognize that for the practicing statistician there are often significant roadblocks to the deployment of "new" methodology. Roadblocks can be a simple lack of knowledge by individuals who completed their statistical education years ago and have had few opportunities to further their professional development by sifting through the morass of current research literature (at best, these individuals may read a text, when available, or attend a continuing education short course on the topic). As people who "grew up" using the delta method to derive asymptotic standard errors for complicated functions of random variables and have occasionally jackknifed nonlinear least squares parameter estimates to remove first-order bias, we were very interested to follow the development of bootstrap methodology. "Tried-and-true" techniques, like the delta method, frequently serve one's needs quite well, provided that you exercise caution. Hence, there may appear to be little motivation to employ

Michael P. Meredith is Principal Statistical Scientist, Biometrics and Statistical Sciences Department, The Procter & Gamble Company, Cincinnati, Ohio 45241-2422, and Adjunct Associate Professor of Biological Statistics, Biometrics Unit, Cornell University, Ithaca, New York 14853-7801. Jorge G. Morel is Senior Statistical Scientist, Biometrics and Statistical Sciences Department, The Procter & Gamble Company, Cincinnati, Ohio 45241-2422.

newer, less well understood techniques (the practical law of inertia, you practice what you know!). Of course, numerous pathological examples of the delta method are widely documented in the bootstrap literature, although the delta method may demonstrate good performance relative to resampling methods in some special circumstances (e.g., Wu, 1986).

Another roadblock occurs when newer methods are utilized on a problem and the research client, a referee or a journal editor balks at their use because the techniques are not commonly recognized in the individual's existing literature. This can be a very difficult hurdle to overcome and requires publication of convincing articles in relevant subject matter area journals that juxtapose the new methods with standard (familiar) methods. Statistical research articles rarely attack a practical problem of substance and, by necessity, typically consider very much simplified versions of real problems or a revisitation of data that have been used in a long string of statistical research articles. The extra step must be taken to collaborate on publications in an applications-area journal in order to hasten acceptance.

The lack of suitable software for easily implementing an important subset of bootstrap methods will clearly continue to be problematic as pointed out by Young. Specific bootstrap applications can require quite distinct programming and, consequently, the methods may not lend themselves to inclusion in well-known statistical packages. In spite of these inconveniences, there are some indications that development of relevant statistical software is underway. For example, SAS software (release 6.07) has recently incorporated the MULTTEST procedure, where p-values are adjusted for high multiplicity using the bootstrap methods. The S language may continue to provide the simplest bootstrap programming in the future. S must continue its transition from academic and some industry research workstations to a broader base of practicing statisticians. The pace of this transition appears to be accelerating with the introduction of language extensions like S-PLUS and its transfer to PC platforms with operating systems other than UNIX. Certainly, the plummeting costs for computing power enable almost anyone to conduct rather sophisticated bootstrap computations and permutation tests on relatively inexpensive desktop computers.

Much of the prior discussion highlights the difficulties faced in moving new results from theory to practice. Many of these difficulties, however, are being rapidly overcome. Therefore, we must disagree with Young's thesis that bootstrap methodology is making slow inroads among practicing statisticians. Our impression is that bootstrap techniques seem to be finding their way into the solution of problems in many areas of the applied biological and health sciences. To assess this impression we searched MEDLINE (a computerized database of over 7 million references from the biomedical literature spanning 1966–1993) and found over 130 references using statistical bootstrap methods. The majority of these references are from subject matter oriented journals in areas including physiology, dentistry, diagnostic test performance, chronobiology, neuroscience, immunology, behavior, psychology, cardiology, genetics, epidemiology, toxicology, pharmacokinetics, risk analysis and bioequivalence (see also Efron and Tibshirani, 1993, Chapter 25). These references are predominantly from 1988 to 1993. Not surprisingly, there are clear examples of inappropriate applications of bootstrap methodology, but these transgressions are certainly no worse than the common misuse of standard statistical methods to be found in the applications literature. We suspect that other areas (e.g., engineering, economics or chemistry) are experiencing equally rapid introductions of bootstrap methods to their relevant literature, and it would seem just a matter of time before the more useful techniques that exist in the theoretical statistical literature are weeded out from those techniques that provide less promise for the practitioner. This whole process of bootstrap technology transfer would, of course, be more rapidly facilitated by close collaboration of statistical researchers with subject matter researchers. Expanding sabbatical leaves to industry and government may assure more rapid assimilation of new methods among practicing statisticians and their clients. In addition, these alliances could provide a synergistic focusing for further academic research efforts and could possibly avoid efforts to solve poorly motivated problems that may look good from one's desk.

We do agree with Young that the bootstrap is often used to supplement standard statistical analyses along with investigating complex statistics. The supplemental application of bootstrap methods is rather important in biopharmaceutical research, as data will frequently be scrutinized by both nonstatisticians and statisticians within a regulatory agency such as the FDA. Consequently, although the bootstrap results may not be submitted for review, the methods may have been employed to confirm results derived from more standard analyses and to provide some reassurance that others' scrutiny will not uncover surprises.

In addition to their supplemental use described above, bootstrap methods are frequently useful in biopharmaceutical and clinical trial research. This is partly due to the nature of randomized clinical trials, especially smaller trials used in earlier phases of drug development. These trials typically study a nonrandom sample of patient volunteers meeting an

exhaustive list of inclusion and exclusion criteria to become a trial participant. In the strictest sense, valid inference is based solely upon the act of randomization to treatment, and inferences extend little beyond the specific study population. Use of permutation tests and nonparametric bootstrap methods has clear advantages over more standard parametric analyses for such data.

Relevant resampling techniques not discussed by Professor Young are permutation tests, where sampling is done without replacement. With a powerful, inexpensive computer on your desk, it is very attractive to derive the "exact" distribution of a statistical test under the null hypothesis. Even though it is generally not feasible to derive the entire permutation distribution computationally, the technique remains attractive and permutation-based inferences can be made as "exact" as the user desires.

This resampling technique is gaining momentum in statistical practice, as evidenced by the proliferation of computer programs now available such as RT (Manly, 1992) and StatXact and LogXact (Cytel Software Corp., 1992). Permutation tests (and their resampling without replacement) may be confused with bootstrap methods by those new to the bootstrap. Distinctions between methods, and the advantages and disadvantages of the methods should be clearly stated to aid those new to the bootstrap. Romano (1989) has shown the asymptotic equivalence of both techniques in hypothesis tests for several applications. Whether to use permutation or bootstrap techniques remains an area where more research would prove helpful. Bootstrap methods would likely receive added attention in statistical practice if their advantages were made clear relative to permutation tests.

Bootstrap techniques for dependent data emphasized in Young's paper (e.g., moving blocks) rarely find application in our work as we do not often encounter single time series with large numbers of observations. However, we routinely design studies with longitudinal data on experimental units where the units have been randomized to one of several treatment regimens. This scenario occurs frequently in many of the applied biological sciences and has received considerable attention as repeated measures or longitudinal data analysis. The number of repeated measures on each unit is typically small, say, six or fewer, and measurements may not be equally spaced in time or there may be unequal numbers of measurements for each unit due to missed visits, dropout or withdrawals. A practical approach to these situations where we have dependence within experimental units and independence between experimental units is to analyze functions of within experimental unit observations that are meaningful to the researcher. These functions may themselves then be bootstrapped.

Further bootstrap research focus, from a practicing statistician's viewpoint, would be extremely useful if directed toward areas that are frequently encountered in our applications. For example, the dependent data problem described above is far more common than standard time series. Areas of practical importance include the following, to name a few: heteroscedastic linear models and generalized linear models; multivariate response; measurement error; the influence of outlying observations; binary response with extra variation; and bootstrapping studies with complex restrictions on randomization.

Difficulties with nonexchangeable random variables, as in heteroscedastic regression, were punctuated in a study by Wu (1986), where he shows that the naive application of bootstrap methods intended for i.i.d. data can lead to a variance estimator that is biased and inconsistent. These results emphasize the need for systematic exploration of the robustness of commonly used bootstrap methods versus other procedures in routinely encountered departures from underlying assumptions. Unfortunately, the name "bootstrap" alone seems to imply immunity from the usual statistical ills.

Finally, it should be mentioned that the practical use of resampling techniques for variance estimation has a much longer history than indicated by the date associated with the coinage of the term "bootstrap." Its origin can be traced back to Mahalanobis (1944) and Deming (1956), who have supported the use of interpenetrating or replicated samples. McCarthy (1969) developed the idea of using repeated replications based on half-samples. Recent works of Rao and Wu (1988) and Rao and Yue (1992) show that the bootstrap provides a resampling method for inference with complex survey data.

In conclusion we must applaud the efforts of Professor Young to bring the focus of research on bootstrap techniques and their properties back to an arena that will ultimately serve practical needs.