one PEP series (2/9) with another (5/9) as the dependent variable in regression had only a slight effect on the adjustments.

Freedman and Navidi, who have raised some good points, have properly drawn attention to two important issues: the need to incorporate all sources of error into our measure of uncertainty and problems of extrapolating from the set of areas on which the regression equation is calculated to the set of areas where estimates are needed. These points modify, but do not obviate, the use of our method for adjusting the census. They also fail to demonstrate that our adjustments do not improve upon the census-estimated population distribution for 1980.

An ideal composite estimate would incorporate information from demographic analysis, make allowances for other independent variables that could have been included in the regression equation, and give some weight to alternative series of PEP estimates. Use of the additional sources of information would improve the estimates while increasing our measures of uncertainty. This uncertainty would not increase to the point where we would consider the adjusted population for New York City to be less accurate than the census count. Moreover, the uncertainty associated with our adjustment would be less than the uncertainty with which we must currently view the count.

ADDITIONAL REFERENCES

ERICKSEN, E. P. (1980). Affidavit submitted to U.S. District Court, Southern District of New York, in *Carey v. Baldrige,* 80 Civ., 4550 (HFW).

ERICKSEN, E. P. (1974). A regression method for estimating population changes of local areas. *J. Amer. Statist. Assoc.* **69** 867–875.

ERICKSEN, E. P. and KADANE, J. B. (1985). The robustness of local undercount estimates in the 1980 U.S. Census. Presented at the *International Symposium on Small Area Statistics, Ottawa, Canada, May 24, 1985.*

FAY, R. E., III, and HERRIOT, R. A. (1979). Estimates of income for small places—an application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.* **74** 269–277.

UNITED STATES BUREAU OF THE CENSUS (1982). Coverage of the national population in the 1980 Census by age, sex, and race. *Curr. Population Rep.* Series P-23, No. 115.

Comment

A. P. Dempster

In their provocative article, Freedman and Navidi argue vigorously against the use of "statistical models" for adjustment of 1980 census counts for both large and small regions of the U.S., "even compared to nothing," but indicate that they might allow exceptions if the assumptions were "made explicit" and were "shown to be appropriate." I agree with the authors that explicitness of assumptions is a virtue, but I question whether anyone actually assumes models in so true versus false a form as Freedman and Navidi appear to suggest. Hence, the concept of what is appropriate is considerably more subtle than they allow.

I will discuss below the aspects of modeling which I believe are most critical for regression adjustment of undercount rates derived from the Post Enumeration Program (PEP). I will also take a brief look at the logical foundation of the argument of Freedman and Navidi and I will argue that they have fallen into traps of their own choosing. I agree with them that the frequentist concept of modeling the production of data as "random draws from a box" is only marginally relevant to the applied problem, not only because the methodology is questionable in the specific circumstances, but also and more fundamentally because my attempts to find or construct a satisfying and explicit general account of frequentist logic have all failed. Freedman and Navidi apparently recommend doing "nothing," which I take to be a recommendation to report raw census counts and no more. I prefer a more cheerful outlook. Statistical logic does have merit, and we do have formal tools capable of addressing problems which most professions relegate to guesswork by acknowledged experts. I suggest pushing ahead with a more satisfactory logic. Finally, my comments will conclude with a brief review of the technical development of Freedman and Navidi.

In their zeal to attack certain formal assumptions, Freedman and Navidi risk demolishing statistical principles which lie at the root of our profession's claim to make a scientific contribution to uncertainty assessment. I wish to elaborate on two of these: the principle of randomization and the principle of regression to the mean.

The PEP program does rely on data from formally randomized surveys. The advantage of randomization does not lie primarily in providing a basis for mean square error computations or for randomization tests or confidence intervals, although these may sometimes

A. P. Dempster is Professor of Theoretical Statistics, Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138.

be useful. The primary advantage is the statistician's knowledge that certain aspects of his information were acquired in accord with carefully controlled standards, which provide in turn a measure of confidence, not otherwise attainable in constructed probability models.

As every applied statistician knows, randomized samples are beset with many problems such as inaccurate sampling frames, nonresponse, and response error. Since these problems exist and are documented and studied in most serious surveys (c.f., the case studies in Madow, Nisselson, and Olkin, 1983), it follows that evidence, arguments, and judgments about their extent and effects have always been an essential part of intelligent analyses of survey results. Ericksen and Kadane (1985) plus the accompanying sets of prepared comments provide a rich and detailed set of perspectives on the PEP surveys in particular, including many valuable suggestions for addressing nonsampling errors. By contrast, Freedman and Navidi stress the mathematics-like test of needing "to show the assumptions were true." The world of applied statistics requires more.

The report of Freedman and Navidi troubles me because it makes no attempt to separate what is good from what is bad about the PEP surveys, but instead suggests that the bad taints the good so that all must be thrown out as a piece. An advantage of randomization deserving more recognition is that it facilitates a thought experiment which makes relatively clear how to separate strong components from weak. Imagine the data which would have been available had the PEP surveys been based on 100% samples of their respective sampling frames. The results would still be flawed due to inadequate frames, difficulties with matching, and so on. But one can logically separate the problem into two parts, first the problem of inference from the observed data to the imagined 100% sample data, and second from the imagined data to the real undercount/overcount problem. The former is susceptible to statistical inference procedures which are relatively uncontroversial, while the latter is totally dependent on evidence from outside the PEP surveys. An assessment of the value of the PEP surveys must in the end depend on quantitative assessments of the effects of both sources of error, where the two assessments are largely separate exercises. Freedman and Navidi appear to assume that the statistician's task becomes impossible when the existence of serious biases is established. I maintain that the quantitative assessment of nonsampling errors is a professional responsibility.

Ericksen and Kadane suggest that the raw PEP estimates of 1980 undercount rates for 66 areas should be adjusted by partially pulling toward a hyperplane in 66-dimensional space representing exogenous factors judged on the basis of various sorts of evidence to be associated with undercount. I believe that the gradual introduction of such shrinkage estimators into statistical practice is one of a few major advances in applied statistics over the past 20 years. I have remarked elsewhere (Dempster, 1980) that such estimation procedures which combine two sources of knowledge should be understood as a variant of Galton's principle of regression to the mean. My preference is for a Bayesian interpretation of such adjustments (Dempster, 1983), but respected non-Bayesian statisticians such as Morris (1983) adopt sampling models like those of Freedman and Navidi. A crucial point is that techniques quite similar to the method of adjustment of Ericksen and Kadane are appropriate under wider conditions than the narrow technical assumptions criticized by Freedman and Navidi. For example, Tukey (1983) in an appendix to his long and thoughtful affidavit on behalf of New York argues for a regression adjustment "that is not based on restrictive assumptions but which is likely to lead to improved estimates over a broad range of assumptions."

Procedures of the kind discussed by Ericksen and Kadane, or by Tukey, are essentially compromises between extremes, namely, the extreme of no regression adjustment, and the extreme of 100% adjustment back to a regression hyperplane. Once the two extremes have been identified it becomes necessary either to choose between them or fashion a compromise. The issue is not directly one of models, although the factors bearing on the choice are often represented in terms of models. Rather, the issue is one of knowledge of the relative sizes and shapes of two kinds of variation. It seems a bit illogical for Freedman and Navidi to argue that the raw PEP estimates of net undercounts have too much error in them to use, but then decline to smooth them back in a way which would curtail the error.

One aspect of most discussions of the undercount problem leaves me decidedly uncheerful. There were, for example, true but unknown counts $T_1, T_2, \cdots, T_{66}$ for the 66 areas on Census Day in 1980, modulo a very small margin of fuzziness of definition. Many statisticians, including official statisticians and their critics, are willing to put forward numerical surrogates for $T_1, T_2, \cdots, T_{66}$, or at least schemes for obtaining such estimates, and other statisticians come forward to analyze and support or criticize the proposals. Despite voluminous writing, it is very rare to find a clear statement of what it means for one estimate to be better than another. Phrases like closer to the true value on the average are used, but averages over what are not explained.

Suppose I state that the population of Jonesville was 10,132 on Census Day, and you say that it was 10,858. Absent the true value T, what meaning can attach to a claim that my estimate is better than yours? I submit that the answer, whereby I assert that I have a better rule that you, has not worked and will not work, except in practically rare and ideal circumstances, so should be abandoned as a modus operandi. Instead we should agree that the goal of analysis is to produce a prospectively meaningful distribution for T derived from evidence and arguments which are specified as explicitly as possible, whence point estimates are conventionally agreed upon summaries of the posterior distribution of T. The task of statistical science is to develop prescriptions and practices for approaching the goal in a manner that commands respect.

Freedman and Navidi would argue that T is "fixed and unknown," while my estimate 10,132 is "fixed and known," which is true. Echoing Neyman, they proceed to argue that, since $T - 10{,}132$ is therefore fixed, it should not be regarded as "stochastic" or a "random variable." I agree that terms like stochastic and random are associated with examples where fixing a value is an important event taking place in space and time. In such examples, it is natural before the event occurs to regard the probability as a prospective measure of uncertainty. Having established the existence of prospectively valid probabilities in limited circumstances, it is reasonable to seek to extend to wider circumstances, as distinguished mathematical scientists began to do by the 17th century. For example, it is surely sensible to apply the game of chance model after the draw has taken place but before the outcome is known. In this way, the barrier associated with the label "fixed" is breached, and there should be no legislated limit on extending the tool of prospective probability assessment to any and all well defined unknowns, subject only to the requirement that the case be argued in a way that commands credibility and respect from scientific peers.

My advocacy of this position is based in part on the view that the present confusion about dealing with uncertainty in practical problems is harming the profession, and in part on the perception that most practical problems requiring analysis of statistical data *require* bringing together sources of information to be combined with the statistical data. That is, scientifically or operationally meaningful inferences simply cannot be obtained from the data alone. We need to abandon the mindset that the statistician's task is to deal with his data, and only his data, which has always been a myth anyway. For example, we should accept the task of assessing nonsampling errors as professionally symmetric to the task of assessing sampling errors, equally demanding of probabilistic treatment, only harder and less familiar, therefore demanding more research and more resources. The unifying thread should be the goal of probability assessment.

Finally, turning to the technical issues raised by Freedman and Navidi, their correctness need not be doubted, but the motive seems to be to sow doubts rather than to resolve issues. Table 1 is incomplete and mostly unsurprising. The anomaly in that table involving North Dakota, South Dakota, and Wyoming does require an explanation, but without the explanation conclusions are hard to draw. The selection of PEP 10/8 for study alongside PEP 2/9 "because it was listed next to 2/9" is odd because it does nothing at all "to prevent any accusation" of picking an extreme series. A preferable method would have been to report variations across the dozen or so PEP estimates deemed worthy of consideration. The existence of biases related to X was surely never in doubt, but how large are they? Regarding conclusions from the simulation study, points i and ii are as expected. I prefer to avoid variable selection methods, on the grounds of their non-Bayesian ad hockery. Also, statisticians do need to be reminded that σ^2 is often poorly estimated, but the implications are not always bad. For example, Student's t on 10 or so degrees of freedom can be a trustworthy tool even though variance estimates on 10 degrees of freedom are poor. Point iii is suggestive, but again not followed up. It would be important to see some standard diagnostics for the regression analyses before accepting a Gaussian model. If the linear dependence of Y on X appears acceptable, while the residuals appear non-Gaussian, then it is necessary to face the concerns raised by Tukey as quoted above.

ACKNOWLEDGMENT

This work was faciliated in part by National Science Foundation Grant DMS-82-01820.

ADDITIONAL REFERENCES

Dempster, A. P. (1980). Comment on "Using empirical Bayes techniques in the law school validity studies" by Rubin. *J. Amer. Statist. Assoc.* **75** 817.

Dempster, A. P. (1983). Comment on "Parametric empirical Bayes inference theory and applications" by Morris. *J. Amer. Statist. Assoc.* **78** 57.

Madow, G., Nisselson, H. and Olkin, I. (1983). *Incomplete Data in Sample Surveys.* Academic Press, New York, Vol. 1.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.* **78** 47–55.

Tukey, J. W. (1983). Affidavit of plaintiff's expert, *Cuomo vs. Baldrige*, 80 Civ. 4550 (JES), Oct. 12, 1983.