the undercount but also to explain it and to discover the sources of underenumeration can become an important tool for census planners in their attempts to reduce the undercount in future censuses, by attaining a deeper understanding about the underlying mechanism of underenumeration.

As pointed out in the paper by Freedman and Navidi, the proponents of the New York adjustment procedure failed to provide sufficient justification for the model used. This failure was both with respect to the inclusion of variables and the resulting potential bias and with respect to the specification of the variance and of the error structure. A long list of additional potential variables is recommended for consideration. This list includes "geographical location" and interactions, so that the possibility of different regression models for geographical regions, not only with different constants but also with different regression coefficients, must be considered. If we add to this the various possibilities for error structure (model errors, sampling errors, and correlations between them), the number of different models to be considered and the number of their parameters becomes very large indeed. The choice of the correct model among these and the estimation of its parameters all on the basis of 66 observations becomes a formidable problem. To this are added the problems due to the fact that the observations are based on data from a complex sample design, rather than on simple random sampling, so that, for instance, the diagonality of the sampling variance matrix, K, is indeed difficult to justify.

However, in fact, the 66 estimates of undercounts are each based on many observations (the Post Enumeration Program sample size in each area) and this individual information for subunits might be utilized for more efficient model search and identification. For instance, some method of sample re-use or cross-validation based on sample-splitting as proposed by Pfeffermann and Nathan (1985) could be used. It is shown there that efficient cross-validation can overcome both the problem of overfitting and underestimation of error due to the search among a large number of alternatives and the problem of testing goodness of fit on the basis of data from complex samples.

The empirical results and simulation study of Section 6 illustrate clearly the faults of the proposed adjustment. However, it should be pointed out that the fact that replacement of the crime rate variable by an urbanization rate results in approximately the same quality of fit (as measured under the model assumptions) does not in itself invalidate either model for purposes of adjustment. Similarly, the lack of consistency in the choice of the best subset of three explanatory variables in the simulation study does not necessarily show inadequate adjustment. It is possible that more than a single choice of a set of explanatory variables can provide equally adequate estimates of undercount, although, of course, the explanation provided by the models is thereby limited. In any case, as pointed out, the estimates of standard errors used to judge the quality of these models are definitely deficient.

Finally, although the results of this paper show, without doubt, that the adjustment procedure proposed by New York is not "statistically defensible," this should, under no circumstances, be regarded as a demonstration that an adequate adjustment procedure cannot be found. The negative result should rather be interpreted as implying that an adequate procedure for adjustment of census counts has not yet been found, either for a specific aim or for an official, all purpose one. However, the methods proposed by Ericksen and Kadane (1985) are certainly worthy of further consideration and, above all, for further empirical testing. In particular, suitable methods for model choice and model identification for these circumstances should be developed and applied. The results obtained should be continually scrutinized and appraised by methods similar to those of the present paper.

ADDITIONAL REFERENCES

PFEFFERMANN, D. and NATHAN, G. (1981). Regression analysis of data from a cluster sample. *J. Amer. Statist. Assoc.* **76** 681–689.

PFEFFERMANN, D. and NATHAN, G. (1985). Problems in model identification based on data from complex sample surveys. *Bull. Int. Statist. Inst.* **51** 12.2.1–12.2.17.

SPENCER, B. D. (1982). A note on statistical defensibility. *Amer. Statistician* **36** 208–209.

Rejoinder

D. A. Freedman and W. C. Navidi

To begin with, we would like to thank Morrie DeGroot for his editorial support and the discussants for their careful work. We wish Jay Kadane weren't quite so angry with us, but then we are being very negative about some of his work. He and Gene Ericksen are good statisticians who believe in what they do;

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science*. www.jstor.org

but at a technical level we do not buy their story at all and have gone public with the reasons. In part, that is because we have other ideas for adjusting the 1990 Census—as will be explained at the end.

We will respond to some of the major points raised by Kadane, Dempster, Madansky, and Ericksen, and then draw conclusions about 1990. With respect to the friendly comments by Wolter, Felligi, Moses, and Nathan, we can only admire their insight and prose style.

Despite the substantial disagreements among the discussants, there is also a remarkable degree of agreement on the technical facts. The debate is largely about what conclusions to draw; on balance, the discussion does not change our opinions. The one serious technical dispute has to do with our comparison of the Post Enumeration Program (PEP) 2/9 and 10/8 series, and we take that up first.

1. PEP 2/9 AND 10/8

Several discussants comment on the superiority of PEP 2/9 over 10/8. In a way, this is a red herring. If the assumptions of the model apply to both series, 10/8 is preferred because it gives smaller standard errors (see our Table 2). The real question is why the assumptions apply much better to 2/9 than to 10/8, and to that we see no answer. How can anyone show, for example, that the combined effect of mover bias and correlation bias is an order of magnitude smaller in PEP 2/9 than in 10/8?

However, for sentimental reasons at least, we are willing to defend 10/8 in a quality comparison with 2/9. As we read the trial transcript, Ericksen and Kadane favored 2/9 for two reasons: (i) they liked the imputation rules used to create 2/9, and (ii) at the aggregate national level 2/9 agreed closely with the demographic estimates for undercount by race (see Madansky's second table).

With respect to reason (i), PEP 2/9 and 10/8 use exactly the same imputation rules. With respect to reason (ii), Jeff Passel of the Census Bureau gave testimony showing that while 2/9 did agree with demographic analysis at the aggregate level, the agreement evaporated when the totals were broken down by age and sex; indeed, none of the PEP series were in reasonable agreement with the demographics. (At the national level, the samples are so large that sampling error is irrelevant to this argument.)

By contrast with these two central points, differences in handling post office information, previous interviews, and movers seem to us to be distinctly minor. For example, New York was willing to use some PEP series based on the August Current Population Survey (CPS). Nor do we grant that the differences all favor 2/9, as explained in our paper.

Some discussants also object to our near-random choice of 10/8 as the foil. One hints darkly at cheating, one says all the series must be studied before conclusions can be drawn, and one points to a series agreeing with 2/9. The Bureau, perhaps at our suggestion, did run the Ericksen-Kadane model on all 12 main PEP series and used the results to adjust all 3000 counties in the United States. There were substantial differences across the different series at the level of the 66 study areas, and even more at the county level.

As a statistical bureaucracy, if it adjusted at all the Census Bureau would probably adjust all the way, down to the 39,000 minor civil divisions or even the tract level: in government tables, detail adds to total. The choice of which PEP series to use really would have serious implications for the distribution of tax moneys, especially for local governments.

In our view, there is no reasonable way to choose one PEP series over another. The desirability of an imputation rule depends on unmeasured characteristics of various kinds of nonrespondents. This is hard to settle a priori, although New York did its best; for a devastating critique of that effort, see the affidavit filed in the case by Ken Wachter.

2. KADANE

The precise issue is stated by Kadane as follows: according to the defense, "there is at present no feasible method for ... reliably adjusting the official census counts to reflect more accurately the true population distribution of the United States," and that is certainly our view. He continues, "The latter proposition is very difficult to support, other than by repeating it many times." We view Ericksen and Kadane as having taken their best shot in court at adjusting the census to make it more accurate using their model. But in our opinion, that adjustment just amounts to taking the census numbers and adding a component of noise with an unknown error structure. Our paper sets out the reasoning, we hope without too much repetition. We think similar arguments apply to any of the other, less well developed, model-based adjustments that have been proposed.

Kadane defends his model as being "simple and tractable," and asserts that "only a very naive user would believe in the literal truth of the assumptions"; for him, the only issue is the robustness of the model, and the sensible way to validate a model is by embedding it in a more general model and then testing to see whether the additional complexity is necessary.

We reject this as an operating philosophy for reasons given in the papers cited by Madansky. The great models of science are validated by simple, harsh tests: the investigators use the models to make predictions,

and their peers get to see whether these are true or close enough. Now the Ericksen-Kadane model makes seven quite strong and implausible assumptions about the world with no supporting evidence. And even on Kadane's own robustness test, the model fails because the conclusions are so sensitive to the assumptions. Our paper documents this in detail.

Coming now to more technical issues, Kadane asserts that we devote only one sentence to showing bias in PEP; and if correlation bias exists, it can only lead to an underestimate of the undercount. Why is the burden on us: does Kadane get to make an estimator unbiased by fiat? In any event, there is a lot of empirical evidence demonstrating the bias in Section 6 of our paper, and Wolter reports still more. The point on correlation bias is also wrong; the Census Bureau argued that people interviewed in the April CPS were less likely to cooperate with the census, a correlation bias in the opposite direction. (On this too, see Wachter's affidavit.) Finally, the real issue is differences in the various biases across areas, and these can go in either direction.

Kadane is seriously offended by our simulation study: "When it comes to making bold assertions from unsupported assumptions, Freedman and Navidi, by the end of their paper, show themselves to be in a class by themselves." Let's take this more slowly. If the Ericksen-Kadane assumptions hold, so do their conclusions. We wanted to see what happened if some of the assumptions failed a little bit. In this connection, a simulation study is not a brand new idea, and a simulation study is what we did.

We started by granting the first assumption in the model, that PEP = truth + random error. We then had to choose, for our computer microcosm, values for "truth" and for the variances of the "random errors." We elected to use the PEP 10/8 numbers. Contrary to what Kadane says, this hardly commits us to the view that either the equation or PEP 10/8 is true of the United States in 1980. Our addition of "random error" to "truth" also comes in for negative comment: we thought this just granted one of the two main assumptions in the model.

We did run the simulation starting with PEP 2/9 rather than 10/8; in some ways this came out better for Kadane, and in other ways, worse. It was a close call, but in the end we decided to stay with our foil. At a minimum, we think our study shows that in a simulation world rather like the Ericksen-Kadane model, their data processing will give quite misleading results. Now it's their turn: Why should anybody rely on their computations as applied to the 1980 Census?

We close by protesting Kadane's attack on our scholarship. We did our best to give a fair summary of the position taken by New York's three statistical experts in rebuttal as reflected in the trial transcript (pp 2140–2495) and in the Ericksen-Kadane paper (Franklin Fisher was the third expert). After reviewing the criticisms and the documents, we think we did a pretty good job. For example, Kadane says, "This invention of a position to attack is most egregious in the matter of estimation of the population of subareas." Well, here is Gene Ericksen under examination by New York's own lawyers (pp 2426–2427 of the trial transcript):

> Q. Can the regression equation be used for areas within those areas (i.e., subareas of the 66 study areas)?
> A. Yes, it can.
> Q. Could you describe how that would be done?
> A. Well, just to give an example, if you wanted to have regression estimates for central cities which were not included in the regression equation, like let's say Newark, New Jersey, one would take the crime rate for Newark, percentage minority for Newark [and percentage conventionally enumerated for Newark], plug it into the regression equation, and you could have a regression estimate for Newark, and that would give you a pretty good idea of what the undercount was in Newark, New Jersey.

The statistician who thinks that New York's lawyers were surprised or displeased by these responses does not follow major league litigation. And just to show that we are not seizing on an isolated lapse, the same position was sketched by Franklin Fisher (pp 2222–2231). It seems virtually the same as the position we impute to "New York's experts." Likewise for the standard errors, which are evidently another sore point. Fisher, who introduced New York's exhibits reporting the standard errors, interpreted them exactly as we indicate (pp 2207–2209). Kadane testified after Fisher and did nothing to correct that interpretation.

3. DEMPSTER

Given an animal which is either a horse or a donkey, some Bayesian statisticians (and probably some frequentists too) will evidently take the average and declare it a mule. On the farm, nobody makes such blunders. But in the world of model-based census adjustments, the empirical consequences of assumptions and methods are much harder to grasp.

Dempster argues for the PEP estimates because they are based on probability samples. This is indeed

a strength, but in our view there are also fatal flaws: dependence between the census and the Current Population Survey; too much missing data. That is, the PEP estimates superimpose unreasonable statistical models (capture-recapture, imputation rules) on a good probability sample. Dempster's solution is to call for "quantitative assessment of nonsampling errors." This is fine as far as it goes, but we don't know how to do it for PEP; if he knows, he should tell.

He favors the Ericksen-Kadane model because he likes Bayesian methods, regression to the mean, and shrinkage estimators. We like his taste, but don't see that his argument forces his conclusion. He goes on to say, "A crucial point is that techniques quite similar to Ericksen and Kadane's method of adjustment are appropriate under wider conditions than the narrow technical assumptions criticized by Freedman and Navidi." What techniques and what conditions? He cites Tukey's affidavit. Well, we've read that affidavit more than once; it is clearly the product of a brilliant mind; but we still do not understand either the techniques or the conditions, never mind the argument.

"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." Our view, which we tried to document in our paper, is that the regression hyperplanes proposed by Ericksen and Kadane have no real connection with the problem; while the choices of exactly which hyperplane to use and of which PEP series to start from have profound impacts on the results. The problem is more complicated than choosing a point on the line connecting Dempster's two "extremes."

Like Kadane, Dempster is quite critical of our simulation study, but for somewhat different reasons: "points (i) and (ii) are as expected." In other words, he agrees that the variables in the equation can't be identified or σ^2 estimated from the data. He then makes a compensating bow to Ericksen and Kadane, pointing out that in the t test, σ^2 is poorly estimated too. This gesture, although well meant, is peculiar. Student's great achievement was to incorporate the uncertainty about σ^2 into his procedure; by contrast, Ericksen and Kadane pretend this uncertainty is absent.

The consequences are shown in our study, which shows the Ericksen-Kadane formulas to be seriously underestimating the standard errors of their proposed adjustments. Dempster says, "this is suggestive, but again not followed up." We thought the follow-up was pretty obvious. The standard errors are what "show" the Ericksen-Kadane model to be an improvement on PEP. If Dempster can't trust the standard errors, then in logic he shouldn't conclude that the averaging and shrinking have done any good. After all, Charles Stein had to compute mean square errors to prove that shrinkage estimators work. Or is there some higher truth available to real statisticians?

Dempster says "Statistical logic does have merit, and we do have formal tools capable of addressing problems which most professions relegate to guesswork by acknowledged experts." He is quite right, and that brings with it a real responsibility: not to push the techniques beyond their limits. One hallmark of science is to respect the boundary between the possible and the impossible.

4. MADANSKY

Madansky begins by showing that if the assumptions underlying the capture-recapture model hold, capture-recapture estimates are useful. What makes him think the assumptions hold? He also argues that absent complete dependence, independence is an adequate approximation. Why?

More interesting are his "comments on Freedman's general quest." Basically, he suggests that if statisticians were to pay serious attention to the assumptions underlying their methods, they would all be out of jobs. This seems to us to be unduly alarmist. There are a lot of good applied statistical studies, including a famous one by Madansky on how to find the best pastrami sandwich in New York. Of course, and this may be the real point, there are also a lot of bad studies, which share the flaws of the Ericksen-Kadane model under discussion here. Well, a hundred wrongs don't make a right.

5. ERICKSEN

Ericksen makes a thoughtful contribution to the discussion, the issues he raises are not simple ones, and readers will have to make their own decisions. Here are some comments.

i. In many areas, the PEP standard errors are quite a bit less than the estimated undercounts, but in other areas they are substantially more. The problem is, of course, most acute in the smaller areas. And in this regard too much depends on which series is considered. Furthermore, the standard errors do not address the issue of bias or differences in estimated undercounts across the different series. Ericksen probably thinks this bias is small, and the interseries differences do not matter. We disagree for reasons already given.

ii. We focused on the regression adjustment, since (as Ericksen says) it was presented as the best of the model-based procedures. Do the regression adjustments bring the numbers closer to the mark? We tried

to show that this question is, as a technical matter, almost unanswerable. In our opinion, there is on the whole no improvement; in Ericksen's, there is. But it is really a question of opinion rather than technical statistical calculation, because the calculation rides on unverified and perhaps unverifiable assumptions.

iii. Ericksen thinks that it does not make much difference whether one starts from 2/9 or 10/8, and whether crime or urbanization goes into the equation. We think Tables 2 and 3 show such choices have serious resource implications, especially at the local level. And the Bureau made a much more thorough argument to this effect, with respect to the PEP series and the explanatory variables, as noted by Wolter.

Often, there is no objective way to make critical choices in modeling. For example, as we understood Ericksen's testimony, he did not have urbanization or rate of population change on his original list of variables; he did not exclude them by minimizing σ^2 but by judgment. These subjective elements are almost unavoidable when statisticians develop regression models in the absence of strong theory. How can you foresee what variables someone else will find relevant in this kind of problem?

When decisions about variables or functional form matter and there is no sound technical way to justify them, the idea of using a model to assist in decision making becomes on reflection a peculiar one. Under such circumstances, which are the rule not the exception in policy modeling, important subjective elements are pushed into the back of the computer rather than being made available for informed judgment.

These issues have been canvassed at least since the Keynes-Tinbergen exchange in the *Economic Journal* (Vol. 49 (1939), 558–572, and Vol. 50 (1940), 141–156); the questions are as relevant now as they were then.

6. WHAT TO DO IN 1990?

For 1980, we prefer the census counts (warts and all) to the PEP adjustments, and the PEP adjustments to Ericksen-Kadane. The Bureau of the Census has been in business for a long time, they're quite good at what they do, and they even provide clear descriptions of their activities. Each step in the adjustment process makes the census more complicated and less tangible, with no demonstrable gain in accuracy.

Some of our discussants suggest that the 1990 Census can be successfully adjusted using more sophisticated models along the lines proposed by Ericksen and Kadane. However, this does not seem practical to us, unless a lot of additional empirical knowledge can be developed. For example, if the problem really is the illegals, then much more has to be learned about their characteristics; textbook statistical models may help with that endeavor, but are not by themselves adequate substitutes.

Successful model development cannot take place in a vacuum; at some point, predictions have to be tested against facts, and this could be a valuable by-product of our proposal for adjusting the 1990 Census. The idea is hardly original, but may be quite useful. It is to take a probability sample of small areas (for example, blocks), and count them very carefully in a "super census." The ratio between the super census counts and the regular census counts can then be used to adjust small areas not in the sample. It is possible to extrapolate from a probability sample to a population without introducing a series of unverifiable assumptions.

In this context, regression models might be useful for smoothing, provided they are developed and calibrated using appropriate techniques of cross-validation. The stochastic assumptions would be derived from the sampling procedure, rather than being assumed ad hoc as in Ericksen-Kadane. Too, the super census differs from Wolter's proposal in that capture-recapture models would not be needed, eliminating one layer of dubious assumptions. However, as indicated above, the super census would provide an arena in which model-based adjustment procedures (like capture-recapture or Ericksen-Kadane) could really be tested. Indeed, model predictions could be compared to the results of the super census on the sample areas or on aggregates of those areas.

There are two basic presumptions behind the super census idea: (i) that sample areas can be counted more accurately than in the regular census, and (ii) that this counting can be done without changing the way the people in the sample areas would respond to the regular census. Point (i) is almost bound to be right. Point (ii) is neither obviously right nor obviously wrong. It is worth investigating.

7. CONCLUSION

We are often accused of being "against statistics." However, like all academics, we insist on making a distinction. We are only against certain kinds of statistics. Our serious point is that from time to time modelers should look to make sure they are getting the right answers, by testing their models against reality. Otherwise, they may only be poking at the entrails of electronic chickens.