DISCUSSION

John C. Bailar III, National Cancer Institute

I believe it is fair to say that statistical theory is largely concerned with the study of variability, especially errors. Generally speaking, errors are of two types: random and non-random, or bias. The concepts of variation, determinism, random error, non-random error, and related matters I will consider later are all subject to much refinement, but for my purpose here the terms can be used in a very rough sense.

Speaking roughly then, we can say that statistical theory (including non-mathematical theory) has been concerned largely with random variation, while non-random variation was left for the subject matter experts to deal with however they could. There were exceptions, of course, and I will give only one fine example. The U.S. Bureau of the Census was much concerned about the effects of census enumerators on data. By treating interviewer effects as being themselves random, statistical models were developed, imaginative studies were designed and carried out, parameters for interviewer effects on various census items were estimated, and significant conclusions were drawn. Those conclusions were responsible, in part, for the progressive shift from complete reliance on trained enumerators to the widespread use of self-enumeration in the U.S. Decennial Censuses.

One can find other examples of the application of statistical theory and mathematical development to a problem previously considered to be one of bias rather than random variation, but it has only been in very recent years that statistical theorists have begun to pay much attention to these matters. In retrospect, this is surprising, since biases are often much more serious sources of uncertainty than sampling variation.

With this year's American Statistical Association program, we can say that the study of non-sampling errors has arrived in force. At least five full sessions, including the invited paper by Donald Rubin and a discussion of policy issues related to non-response, are devoted to just one aspect of non-sampling errors: non-response and imputation. I believe it is significant that the new Section on Survey Research Methods is sponsor or cosponsor for four of these five sessions.1/ Clearly, the newcomers know where the problems are and where the action ought to be. Further, all are morning sessions, a very practical schedule when one wants people's attention. I will comment on each of the papers at this session, and more briefly on those at two of the other sessions. I will focus on item non-response rather than total or complete non-response.

Item non-response is a common and difficult problem in most areas of applied statistics, and many different methods have been developed to deal with it, after reasonable follow-up efforts have failed:

--One can simply tabulate non-responses and let someone else worry about what to do with them - generally someone who knows less about the data than the tabulator does.

--One can delete item non-responders from tabulations of single items, or even expunge them from the whole data set.

--One can assign a value to each non-respondent based on the reported value for some similar population element.

--One can use related data files or other external sources to fill in the gaps.

--One can use the responses obtained on some items to estimate responses missing for other items; formulas for this purpose may be derived from the data set at hand or from external sources (regression methods).

--One can arbitrarily and randomly assign responses actually obtained to the sample elements not responding; this is generally done only within sample strata (hot-deck methods).

--One can alter the relative weights assigned to sample elements within strata ("weighting class procedures").

These methods overlap to some extent; and, of course, not all methods may be applicable to any given data set. There must also be other approaches I have not thought of or have never heard about.

With this rich array of techniques to handle missing data, how is a statistician or data analyst to choose one for use in a specific analysis? I believe the choice should be made on a rational basis by means of comparing their properties. None of these procedures is perfect, or even optimal, for all uses, and the choices involve some difficult trade-offs.

The properties that seem likely to be most important in most applications include the following:

- Simplicity
- Cost
- Availability of relevant statistical theory
- Effect on variance (sampling errors)
- Effect on bias (non-sampling errors)
- Dependence on auxiliary files
- Development of estimates for individual values
- Acceptability in applications (e.g., economic and social policy issues)
- Application to cross-tabulations and other multivariance manifestation.

The last two points need brief comment. In some applications, individual missing values must be estimated at least implicitly. For example, total sales of a retail store may be used to determine whether the store is in the sampling frame for a future special survey. If total sales are not reported, we may need some means to guess whether the store should be in the new study.

A much more common problem may be considered a variation on the need to estimate individual values: that of preparing cross-tabulations on several variables simultaneously. We may want to avoid compounding the non-response loss on one item with that on another or we may suspect that biases are more serious when more items are missing. I am deeply concerned about the effects of various imputation methods on cross-tabulations of two or more items.

Now that I have presented a list of techniques that may be used to deal with non-response and a list of properties to be considered in evaluating these techniques, the next step, obviously, is to cross them. It would be helpful, though impossibly difficult and time-consuming, to fill in each cell of such a table with some objective measure of how well each procedure meets each criterion. I will instead come now to the focus of this session and two others, and simply suggest how various papers and sessions fit into the cross-classification.

Tupek and Richardson, in a largely theoretical paper, discuss the use of ratio estimates to compensate for the bias associated with total non-response, using information available apart from and prior to the survey. There are obvious extensions of this procedure to item non-response. The primary concern of Tupek and Richardson is the effect of their procedure on non-response bias.

Robison and Richardson consider a variety of imputation methods used on separate items in a single survey. This work is focused again on control of non-response biases but applies to the whole list of techniques.

Next is the paper of Proctor, who compares two approaches to the problems of item non-response: deletion of a non-responder vs. substitution of the item mean. The latter method is not shown on my list of techniques, since for estimation of item means or totals, it is equivalent to deletion. Proctor's problem, however, is to estimate a function of several items, where deletion might have consequences quite different from substitution of the mean. He also considers hot-deck methods and touches on the use of data from external sources, all from the point of view of finding an optimum balance between variance and bias.

Scheiber studies a single data item for which some responses were temporarily deleted. He then compares three different methods of imputation (hot-deck procedure, an external record match, and a regression formula) with the known true value.

Cox and Folsom then discuss the hot-deck and weighting-class procedures with respect to variance and bias. These first five papers are directly or indirectly concerned with item non-response. The Thornberry-Massey paper describes a poststratified ratio adjustment to deal with undercoverage bias - a most important topic, but one that must usually be dealt with by techniques even cruder than those for item non-response, and where, accordingly, progress is even more important.

I will turn more briefly to the second session of papers on non-response and imputation. Bailar and Bailar have made a long delayed start on theoretical studies of the hot-deck procedure, including comparisons with other imputation methods linear in a single variable across a collection of sample elements. Ernst compares the hot-deck and weighting-class procedures in terms of their variances but notes that bias is likely to be more important than variance. Hill presents empirical evidence on the strengths and weaknesses of the hot-deck method in one very large data set. Huddleston and Hocking describe a multiple regression approach, with numeric examples and some data on computer time requirements. Patrick discusses conditional expected utility, and mentions some applications of the theory he develops to imputation.

In an invited paper, Rubin describes a Bayesian approach to non-response, in which sample elements are first sorted into groups with similar patterns of items missing. He

assumes that sample elements are independent and that a posterior distribution exists; then, in effect, he uses the more complete sample elements to fill the missing items in the less complete elements. This paper is highly theoretical and does not fit well into my classification, but I believe it is the only one given at this meeting that combines some of the benefits of the hot-deck and regression approaches. I cannot yet tell whether it combines some of their defects, too, but Rubin has clearly taken a large step forward in conceptual approaches to problems of item non-response.

Likewise, it is difficult to determine at this time how the panel discussion on policy issues related to imputation may fit into my classification, but I suspect that it will concentrate on policy issues arising when missing data require that some means be adopted. How much that panel discussion will focus on item non-response is not clear; to the extent that it does, it may emphasize the last row of my classification.

It is good to see such strong interest, at last, in problems of non-response. The papers I have mentioned deal with a wide variety of topics, but not one cell of my cross-tabulation of methods and properties has yet been explored in depth. I believe we will see even more interest in this general topic at future statistical meetings. We should. Statisticians have not really been preempted in this field as much as they have abandoned it. Application of the full range of analytic tools could have the same profound effect on our understanding of non-random errors, particularly non-response problems, as it has already had on our understanding of random variation.

I will close by congratulating the authors of all six of this morning's papers on the success they have had in tackling a wide variety of topics in this difficult and important field.

FOOTNOTE

1/ The sessions sponsored directly by the Section on Survey Research Methods were the invited address by Donald Rubin and two topic contributed sessions on the imputation and editing of survey data. Another invited session on missing data analysis, with papers by Frane and Nordheim, was cosponsored with the Section on Statistical Computing. The panel discussion referred to was a General Methodology session, entitled "Adjustments for Missing and Faulty Data in Surveys and Censuses: Methods and Policy Issues."