# DISCUSSION OF PAPERS BY HUDDLESTON & HOCKING AND PATRICK

Daryl Pregibon, Univ. of Toronto

Before proceeding directly with the discussion of the papers by Drs. Huddleston & Hocking and Patrick, a short history of my experience with missing data is in order. As part of the Quantitative Edit and Imputation Team at Statistics Canada in 1976, I was examining current methods of handling missing data and also trying to develop new methodology. New methodology was needed because the state of the art at that time was indeed constrained. In particular, we wanted an imputation procedure which could handle:

(1) survey data where the multivariate response space was dominated by zeroes, and

(2) a non-random "blank generating mechanism".

Current methods usually assumed multivariate normal data and that data were missing at random, i.e., without reference to the true but unreported data. This latter point is extremely important and I have always believed that blanks do not occur randomly. In particular, observed blanks can usually be classified as one of the following types:

(a) the "I missed the question" blank,

(b) the "I meant a zero" blank,

(c) the "failed edit" blank,

(d) the "I don´t know" blank, and

(e) the "I know but won´t tell you" blank.

Only blank types (a) and (d) may qualify as being random. Blank types (b), (c) and (e) are usually the ones encountered, but have never received the recognition they deserve. For my particular survey, blank type (b) dominated the field of possibilities. I resolved the problem to the best of my ability (Pregibon,1976), but was aware of the impractibility of the routine application of the method. That was two years ago and I haven´t done much since, except perhaps keeping up with the literature on missing data. In fact, in those two years, nothing much has come by which addressed the issues (1) and (2). Then I came to San Diego and everyone is talking about non-random missing data! See for example the papers given here by Rubin, Nordheim and McFarland. These papers have renewed my interest and hopefully, next year I´ll have more to offer than this discussion.

Returning then to the papers of this session, I find I have very little to say. Drs. Huddleston & Hocking have formulated a tight, specific problem, for which they have a very tight, specific solution. The novel feature of their method is the incorporation of auxiliary information (through total fields and the like) into the estimation procedure. This is indeed a step forward and was also referred to by Rubin(1978) as good policy. It also indicates that an important goal of survey and questionnaire design should be to minimize the loss of information due to missing data. On the other hand, their method does not address issues (1) and (2), which from my point of view is unfortunate. Their method is also constrained to handle only quantitative (continuous) data. This rules out the application to surveys which are of the mixed (categorical - continuous) type. A more subtle problem with the method is the lack of "commutativity" between estimation and imputation. That is, using their ESTMAT routine, we can estimate the relevant parameters and using these, fill in the missing data. If, however, we (or some other user of the data base) use the completed data set to estimate the same parameters, the results don´t coincide. This is not a criticism of their technique, but rather a warning to those who would use ESTMAT for both estimation and imputation.

Charles Patrick on the other hand, has a tight formulation of a general problem, for which he has a very tight, general solution. In particular, just about every imputation system can be cast in his decision theoretic framework by proper specification of priors , error distribution and utility functions! This generality is attractive, not from the point of view of actual implementation, but in that it gives us an objective basis by which to compare different imputation schemes --- something which we have never enjoyed. Hopefully, his next efforts will be along these lines, indicating the assumptions implicit in various procedures, including the hot-deck. Another relevant point concerning his approach is the incorporation of prior information into the estimation scheme. I see this as being a necessary requirement of any imputation system which attempts to address the issue of non-random missing data. This re-emphasizes the importance of close communication between survey methodologist and subject matter in order to determine the required subjective input (either implicit or explicit) into the system. My criticism of the method is that since it is estimation based, two different estimators of the same quantity can lead to two quite different imputations! It seems impossible to detect when this will be the case and would be embarrassing to say the least.

Finally, to do justice to the title of this session, I feel I must stress my opinion on the role of imputation in the editing and imputation framework. In particular, since I am of the Fellegi & Holt school (1976), I believe that they must not be separate sytems! The efficiency and reliability of an imputation system can be seriously affected by wholesale disregard of information and "flags" from the edit stage. I hope that researchers in this area will take this into account and design imputation schemes flexible enough to stand alone, but also with the capability of accepting as input, output from a pre-edit or edit processor. Whether this takes the form of prior distributions or specialized coding and/or weighting is not really important right now. What is important is that we start thinking of the two stages as one. Only then will we start to bridge the gap between statistical theory and practical analysis of complex survey data.

## REFERENCES

Fellegi,I.P. and Holt,D.(1976). A Systematic Approach to Automatic Edit and Imputation. JASA 71.

McFarland,B.H. and Fisher,L.(1978). Estimation and Testing in the Missing Data Problem. Presented at the Joint Statistical Meetings (JSM), San Diego.

Nordheim,R.(1978). Obtaining Information From Non-Randomly Missing Data. Presented at JSM, San Diego.

Pregibon,D.(1976). Incomplete Survey Data: Estimation and Imputation. Methodology Journal of Household Survey Division, Statistics Canada.

Rubin,D.(1978). Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Non-Response. Presented at JSM, San Diego.