DISCUSSION

Benjamin J. Tepping, Westat, Inc.

The papers by Kirkendall and Kelley contain much interesting material, with some of which I must take issue.

The Fellegi-Sunter model, on which these papers are based, recognizes that there are three possible outcomes, but (it seems to me) uses the wrong utility function. To simply minimize the probability of subjecting a case to clerical review conditional on bounds on the probabilities of erroneous matches and erroneous nonmatches ignores important facts:

(a) the value of an erroneous match is, in many (or perhaps most) applications, quite different from the value of an erroneous nonmatch;

(b) the cost and the probability of misclassification associated with the clerical review should be taken into consideration.

We do not necessarily want to minimize the number of clerical reviews. We do want to maximize the value of the record linkage operation. This implies that one must not only determine the costs of the various components of the operation, but must also set values on the possible outcomes. An illustration of this approach is the application of a theoretical model of record linkage to the Chandrasekar-Deming technique for estimating the number of vital events on the basis of data from two different sources. This was published in the Bureau of the Census Technical Notes No. 4, in 1971 [1].

It appears that neither author is aware of my paper [2] in JASA in 1968 in which is presented a model for the optimum linkage of records.

The authors treat the problem as an exercise in the testing of hypotheses. I think it is preferable to regard it as a problem of decision making, subject to a utility function which depends upon the state of nature. In these applications, the three possible decisions are to call the pair of records being compared a match or a nonmatch, or to make some kind of further investigation before deciding on a classification. That investigation may consist simply of subjecting the records to personal scrutiny or may involve seeking additional data. The utility function would specify a gain or loss for each of the possible decisions, conditional on whether the pair is in fact a match or a nonmatch.

Kirkendall's examples also ignore the problem of fixing the values of the probabilities of errors of the first and second kinds. Those probabilities should not be arbitrary. Any solution of the problem should depend upon evaluation of the loss or gain of alternative decisions as well as on the cost of non-decisions--e.g., resort to other means of arriving at a decision.

Kirkendall's first illustration assumes independence, both under H_0 and under H_1. In the real world, this assumption may be far from true. For example, under either of the hypotheses H_0 or H_1, an agreement on first name would increase the probability of an agreement on the item sex--two records both giving the first name as "Nancy" are not likely to indicate different sexes. Presumably the lack of independence could be treated as in her example of cancer patients, essentially by dividing the First Name item into two items: one for cases in which both records show the sex as male and one for cases in which both records show the sex as female. This comment also applies to Kelley's numerical example, in which independence of these components is assumed.

As is pointed out by Kelley, the literature that gives advice on the choice of blocking schemes is not extensive. Yet practical problems make blocking of the files being compared essential, and Kelley's work should contribute to the improvement of blocking designs. He does take account of costs, by considering both the decrease in operational costs, because blocking reduces the number of comparison pairs, and the increase in the probability of an erroneous nonmatch as a result of blocking. (I note, however, that he does not use the fact that the probability of an erroneous match decreases as a result of the blocking.) His numerical examples illustrate that the choice among competing admissible blocking schemes involves the implicit assignment of relative values to an increase in the probability of erroneous nonmatches and a decrease in the number of comparisons. In practice, no doubt, a similar implicit assignment of values to an erroneous match, an erroneous nonmatch and a case referred to personal review is made in order to fix the values of the parameters λ and μ of the Fellegi-Sunter model.

I think there is difficulty with the application of Kelley's Lemma 2 to the determination of a suitable blocking scheme even after dealing with the lack of independence of the components of the comparison vector. It seems that a choice must depend, among other things, on a knowledge of the probability, given that the pair is a match (or a nonmatch), that there is agreement between the units of the pair on specified components of the comparison vector. Estimates of such probabilities must ultimately depend upon extensive empirical investigations, although such estimates seem often to be made on the basis of assumed models.

REFERENCES

[1] Tepping, B.J., "The application of a linkage model to the Chandrasekar-Deming technique for estimating vital events," U.S. Bureau of the Census, Technical Notes No. 4, Washington, D.C., 1971, pp. 11-16.

[2] Tepping, B. J., "A model for optimum linkage of records," Journal of the American Statistical Association, 63, 1968, pp. 1321-1332.