A Contingency-Table Model for Imputing Data Satisfying Analytic Constraints

William E Winkler william.e.winkler@census.gov
http://www.census.gov/srd/www/byyear.html

UNECE Worksession on Statistical Data Editing
Madrid, Spain, October 20-22, 2003

FH Theorem. Implicit edits allow one to fill in an edit-failing record in one pass so that the record satisfies all edits.

Italian Labour Force Survey (product space of fields)
> 10^35 data points, > 10^25 edit patterns

Winkler (1997) – 100+ times as fast as IBM-ISTAT (1996) that used GKL algorithms (Garfinkel, Kunnathur, Liepins 1986), Chen (1998), Winkler and Chen (2002), Chen (2002)

Need probability distribution

Hot-Deck generally does not work because of collapsing, joint distributions destroyed.

Let $X = (x_{ij})$, $1 \leq i \leq n$, $1 \leq j \leq k$ be an array of data where the first $m < n$ rows have complete data

Lemma 1. (Little and Rubin 1987). Let X be an incomplete data array. Let I_1 be a set of interactions and let M_1 be a set of margins determined by the complete data records and the non-missing items in the incomplete records of X. Then the EM fitting procedure yields a complete data representation X_C that is a model for filling-in incomplete data records with probability distributions that are consistent with the observed data in X

X_C along with the interactions I_1 and margins M_1 is a parametric form of the model for the data.

Lemma 1 of Winkler (1990, *Ann. Pro*b.) and Theorem 1 of Winkler (1993) yield that the above Lemma 1 can be extended to Lemma 2 for the situation of edit restraints that are considered as structural zeros.

As long as the set of interactions and edit constraints are logically consistent, then the theorems of Winkler (1990, 1993) yield an iterative fitting procedure that monotonely increases the likelihood and yields a solution X_C.

Theorem. For each incomplete record *r* in X, we can match against X_C to fill in data in a manner that satisfies edit restraints and preserves the probability distribution of X_C.