

**FRACTIONAL IMPUTATION, VARIANCE ESTIMATION, AND TESTING FOR
INCOMPLETE TWO-WAY CONTINGENCY TABLES**

K. J. Koehler, S. S. Kang, and M. D. Larsen
Iowa State University

April 27, 2006

ABSTRACT

Partially observed multivariate categorical data can be handled in at least three ways. One choice is by restricting analyses to cases that are fully observed. This reduces sample size and potentially leads to bias. A second choice is modeling the data and performing maximum likelihood estimation. This requires the analyst to have a capability of implementing non-standard analyses and programs. A third choice is imputation, or filling-in the missing values to create a complete data set. Options for imputing categorical data include assigning missing observations into the most common category conditional on observed information or randomly allocating missing observations based on observed conditional proportions. Pretending all the data were actually observed leads to understatement of variability. Imputation procedures such as fully efficient fractional imputation (FEFI) or multiple imputation (MI) create multiple versions of the missing observations, thereby reflecting uncertainty about their true values. Multiple imputation generates a finite set of imputations through a posterior predictive distribution. Fractional imputation assigns weights to the observed data. The focus of this article is the development of FEFI for partially classified two-way contingency tables. Variances of FEFI estimators of population proportions are derived. Simulation results, when data are missing completely at random, show that FEFI is comparable in performance to maximum likelihood estimation and multiple imputation and superior to simple stochastic imputation and complete case analysis. Wald and deviance tests of independence under FEFI are proposed. Simulations are used to compare type I error rates and Power of the various test/imputation procedures. The partially observed marginal information is useful for estimating the joint distribution of cell probabilities, but is not useful for testing association.