

**University of Illinois at Chicago
School of Public Health
Division of Epidemiology and Biostatistics**

*Technical report#:2006-008
September 2006*

Title: On Tukey's gh distribution for multiple imputation

Authors: Hakan Demirtas and Donald Hedeker

Affiliation(s): University of Illinois at Chicago, Division of Epidemiology and Biostatistics

On Tukey's gh distribution for multiple imputation

September 12, 2006

He and Raghunathan (2006) elegantly present a connection between multiple imputation and one general class of distributions (Tukey's gh) that includes a variety of standard distributions (e.g. lognormal, Weibull) as exact or approximate special cases and covers a broader area in the skewness-elongation plane. Creating multiply imputed data sets under general families that allow for distributional features such as heavy tails, skewness and non-zero peakedness (kurtosis), which are clearly not accommodated by a Gaussian model, had not yet been fully explored (Liu, 1995), partly owing to the complexity of forming a Bayesian predictive distribution of the missing data given the observed data under these settings. Instead of adopting a Bayesian approach (Schafer, 1997), He and Raghunathan (2006) account for the parameter uncertainty by obtaining nonparametric bootstrap samples that anchor the estimation procedure for the parameters of Tukey's gh distribution. For univariate data, sampling with replacement is equivalent to using the inverse CDF (cumulative distribution function) method on the empirical CDF . Although it is a reasonable approach for moderate and large sample sizes, it may lead to unacceptable random samples in the small-sample case where the number of distinct data values is limited, which in turn may generate an unduly large degree of variability among parameter estimates in the subsequent step. In order to circumvent this potential complication, two simple ideas from nonparametric density estimation can be employed (Silverman, 1986) based on using a smoothed variant of the empirical CDF : 1) Binning the data and forming a frequency polygon, and using the inverse CDF approach on the resulting distribution function which is a piecewise quadratic polynomial; 2) Connecting the jump points of the empirical CDF with line segments to form a piecewise linear function (Gentle, 2003).

In addition, a few comments regarding estimation of the g and h parameters can be made. Since finding the maximum likelihood or method of moments estimates is a computationally daunting task, it is natural to resort to the empirical quantile technique, as was done in He and Raghunathan (2006). In this context, there are some issues that need to be addressed. 1) While it may be obvious, some readers may

not realize that p should be less than 0.5 in the estimation of g , the parameter that governs the symmetry behavior. 2) Different values of p correspond to different estimates of g , conveying information on the fluctuations in the data skewness. Although not explicitly stated, the authors probably used the median value of g . While the median is an intuitively sensible measure and it might be the right quantity on average, it is unsettling that one can get wildly varying values of g . This, in turn, may distort estimation of the kurtosis parameter h which is estimated conditional upon the value of \hat{g} . 3) Somewhat related to 2, based on empirical evaluations, we have observed that it is better to choose p 's in a non-linear fashion (e.g. geometrically increasing) rather than using an evenly spaced (linearly increasing) set of numbers. 4) When the data exhibit different tail behavior for the two halves of the data, it may be constructive to estimate h for the lower and upper halves separately.

References:

- Gentle, J.E. (2003), *Random Number Generation and Monte Carlo Methods*. New York City: Springer-Verlag New York, Inc.
- He, Y. and Raghunathan, T.E. (2006), "Tukey's gh distribution for multiple imputation" *The American Statistician*, 60, 251–256.
- Liu, C. (1995), "Missing data imputation using the multivariate t distribution," *Journal of Multivariate Analysis*, 53, 139–158.
- Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

Hakan Demirtas and Donald Hedeker

University of Illinois at Chicago