

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

*Technical report#:2006-010
December 2006*

Title: On the design of simulation studies in medical statistics

Author: Hakan Demirtas

Affiliation: University of Illinois at Chicago, Division of Epidemiology and Biostatistics.

On the design of simulation studies in medical statistics

December 22, 2006

Hakan Demirtas

Division of Epidemiology and Biostatistics (MC923)

University of Illinois at Chicago

School of Public Health

1603 West Taylor Street, Chicago, IL, 60612-4336, USA

Burton et al. ([1]) elegantly present key operational attributes of medical simulation studies in a fairly comprehensive fashion. In this paper, I address some inaccuracies, and discuss a few important underlying aspects of simulation designs in the context of complete data generation, model formation, inferential validity, assessment and evaluation on philosophical and practical levels.

As mentioned by the authors, creating simulated data sets that are generated around a real data set has been increasingly common in medical statistics, with the rationale being re-producing the real data trends with compatible distributional characteristics. Because there is usually no consensus among statisticians about which of the competing methods is best, many advocate sensitivity analyses that could be performed by trying a variety of methods, or varying the model parameters over a plausible range to see what happens. This approach is valuable, but limited. Instead, I suggest simulating the performance of a method when its assumptions are wrong by proposing a variety of populations that are capable of producing data like those actually seen, simulating behavior of various methods over repeated samples from each population, and subsequently identifying methods that

seem to perform well for most of the populations. To elaborate further, suppose we identify a family of models that, from a likelihood standpoint, fit the data equally well. If our basic conclusions about effects of interest do not change drastically over this family, then the scientific validity of these conclusions is enhanced. Conversely, if the answers do exhibit great variation, drawing firm conclusions seems unwise. Robustness of results over the domain of parameters is desirable and fortunate when it occurs. Yet there is another type of analysis which may lead us to prefer one model, M_1 , to another, M_2 , even when M_1 and M_2 achieve the same likelihood for the current data set. Suppose that we devise a variety of plausible population models, different in nature but all tending to produce samples that resemble the observed data. If, by simulation, we discover that M_1 performs better than M_2 across many of these populations, then we may be more inclined to trust M_1 than M_2 [2, 3].

The need for multivariate data generation is frequently encountered in practice. While I realize that the focus of Burton et al.'s paper is not on random number generation, a few additional points can be made. Considering the virtually unlimited range of data behavior, modeling the first two moments (mean and variance-covariance structure) via a multivariate normal distribution may be overly simplistic. Although real data rarely conform with normality, it has been regarded as a mathematical convenience for inferential purposes due to its nice distributional properties. Despite its critical importance, it only represents a single point in the skewness-elongation plane; and general classes of continuous distributions that span a broader spectrum in terms of symmetry and peakedness behavior ([4]) have received increased interest among statisticians. In view of this, multivariate data generation should be performed under more general classes of densities (e.g., Tukey's classes, the Burr family, the Johnson Family, the Pearson family, generalized Lambda and Beta families, Fleishman polynomials) that include many standard distributions as exact or approximate special cases [5-13], under the assumption that the researcher has necessary expertise. Furthermore, as far as categorical data generation is concerned, one common approach is generating latent continuous variables and converting them to binary/ordinal versions using pre-specified threshold points. This latent variable approach is generally

inappropriate as correlations between the derived categorical variables are not of a simple form or interpretation. In addition, after the conversion process the magnitudes of the correlations become almost certainly distorted. Given these limitations, categorical variables should be created directly, whenever feasible [14-18].

Accuracy appears to be defined in a manner which is inconsistent with widely established terminology. It is a measure of bias, and has nothing to do with precision. However, the authors denote MSE (mean square error) which is an integrated measure of accuracy and precision, as an accuracy measure. Of note, the authors' definition of MSE as an accuracy measure is in conflict with their term for accuracy in equation 1 (section 2.6), in which accuracy is solely a measure of bias. On a related note, the authors correctly mention two different bias quantities : the relative magnitude of the raw bias to the true value of the parameter, and to the overall uncertainty in the system (percentage and standardized biases, respectively). These accuracy measures have relative merits and pitfalls. Standardized bias depends on the total inherent variability which may be too small or too large, causing misleading interpretations; and percentage bias has the assumed true value of the estimand in the denominator which similarly may take extreme values. In our limited experience, it is advisable to consider both accuracy benchmarks simultaneously. The authors also present an interesting discussion on the trade-off between the amount of bias and variability. Narrower confidence intervals that smaller variability induces are good, but only subject to the correct coverage rate which is a direct function of the bias. When type I errors are not properly controlled as measured by the coverage rate, it means that the procedure is not working well. Do we err on the side of bias or variability? In the less-than-perfect world we live in, if one is forced to choose between the two of them, the "less bias" option is more attractive. The main reason is that variance reduction techniques are more general, better-developed and better-accepted while bias reduction techniques are typically problem-specific. Another reason is that large variability implies a higher chance of capturing the truth in the confidence intervals. In other words, a reasonable degree of inflation in the variation is acceptable if the reward is unbiasedness. If the estimates are biased, what sense does it make to have small variability? Needless to say, the "smaller

variation is better” cliché is not really valid here. Furthermore, a “larger than necessary” number of simulation replications may result in spurious biases, as correctly pointed out by the authors.

In addition, there are a few minor issues that need to be addressed. First, the definition of “scenario” seems to be restricted to different sample sizes. In simulation studies, it has a broader meaning that encompasses different data generation schemes, statistical methodologies and competing parameterizations within them, nonresponse mechanisms in incomplete data setup, etc. as well as sample sizes. In conjunction with this, the “independent samples” argument only applies to different sample sizes. Replications of simulated data sets employing the same underlying truth inherently translate to some degree of dependence among samples. What the authors probably mean is that researchers should not take a subset of a larger data set in an attempt to evaluate a different scenario that is pertinent to a smaller sample size. Otherwise, we find the “independent-dependent samples” notion unconvincing. Second, the terminology is occasionally confusing. In the context of censoring, random/non-informative, and dependent/non-random/informative have been used interchangeably in the statistical literature. The use of these qualifiers together (page 4284) creates a false impression that they depict distinct meanings. Furthermore, “null hypothesis of *no effect*” (page 4287) should not have been used in the general discussion of power, type I and II errors, since the null hypothesis may correspond to another simplified view of the data generation process relative to the alternative hypothesis. Third, the number of simulations required (B) is based on the assumption that the test statistics follows an approximately normal distribution (page 4285). Simulation studies typically include many parameters that are estimated by test statistics of different assumed distributional forms. For this reason, it is recommended that B be calculated with respect to all parameters of interest in the system, and the highest value which would correspond to the worst-behaving parameter be chosen. Finally, in Section 2.7.4, a finite number of simulations obviously leads to a range of acceptable coverage rates. Under the assumption that $\alpha = 0.05$, a two-fold increase in nominal type I errors which translates to 90% coverage is considered reasonable by some authors [19-20] for a procedure to work properly, among other criteria.

In other words, while the authors' statement is not wrong, it is incomplete; acceptable lower limit of the coverage rate and undetectable differences in the coverage rates due to the finite replication size are two grossly different concepts.

Acknowledgment: I thank Donald Hedeker for helpful comments and suggestions.

References

- [1] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine* 2006; **25**: 4279-4292.
- [2] Demirtas H, Schafer JL. On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2003; **22**: 2553-2575.
- [3] Demirtas H. Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine* 2005; **24**: 2345-2363.
- [4] Genton MG. (Ed) *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton, FL: Chapman and Hall/CRC, 2004.
- [5] Morgenthaler S, Tukey JW. Fitting quantiles: Doubling, HR, HQ, and HHH distributions. *Journal of Computational and Graphical Statistics* 2000; **9**, 180–195.
- [6] Field C, Genton MG. The multivariate g-and-h distribution. *Technometrics* 2006; **48**, 104–111.
- [7] Burr IW. Cumulative frequency functions. *Annals of Mathematical Statistics* 1942; **13**, 215–232.
- [8] Johnson NL. Systems of frequency curves generated by methods of translation. *Biometrika* 1949; **36**, 149–176.
- [9] Parrish RS. Generating random deviates from multivariate Pearson distributions. *Computational Statistics and Data Analysis* 1990; **9**, 283–295.

- [10] Ramberg JS, Schmeiser BW. An approximate method for generating asymmetric random variables. *Communications of the ACM* 1974; **17**, 78–82.
- [11] Schmeiser BW, Deutch SJ. A versatile four parameter family of probability distributions suitable for simulation. *AIIE Transactions* 1977; **9**, 176–182.
- [12] McDonald JB, Xu YJ. A generalization of the beta distribution with applications. *Journal of Econometrics* 1995; **66**, 133–152.
- [13] Fleishman AI. A method for simulating non-normal distributions. *Psychometrika* 1978; **43**, 521–532.
- [14] Emrich JL, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. *The American Statistician* 1991; **45**, 302–304.
- [15] Lee AJ. Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician* 1993; **47**, 209–215.
- [16] Gange SJ. Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 1995; **49**, 134–138.
- [17] Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. *The American Statistician* 1996; **50**, 306–310.
- [18] Demirtas H. A method for multivariate ordinal data generation given marginal distributions and correlations. *Journal of Statistical Computation and Simulation* 2006; **76**, 1017–1025.
- [19] Collins LM, Schafer JL, Kam CH. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 2001; **6**, 330–351.
- [20] Demirtas, H. Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica* 2004; **58**: 466-482.