Improving the Utility of Imputed Values in Survey Datasets

David R. Johnson and Rebekah Young
Department of Sociology
The Pennsylvania State University
University Park, PA 16802

1. Introduction

Missing data are found in almost all large survey datasets. Multiple imputation (Rubin 1976) has emerged as a general and widely used technique for analysis in the presence of missing data. The key idea of multiple imputation is that missing values are imputed with plausible values drawn from the conditional distribution of the missing data given the observed data under a specified model. This produces a series of "complete" datasets which can then be used for analysis. For a detailed review of multiple imputation see Rubin (1987) and Little and Rubin (2002).

Advances in statistical software packages that support multiple imputation (e.g. Stata, SAS, R, and SPSS) have produced considerable flexibility and ease of use for practical researchers. Many software packages assume joint multivariate normality of the variables being imputed, and this is widely viewed as a robust and unbiased approach (Schafer 1997; Schafer 1999). In practice, researchers often deal with variables that are not "normal" in a sense that they are required to be measured at an interval or categorical level for the data analysis. When non-normal variables are imputed under the normal assumption, the imputation produces an implausible value. For example, when imputing the variable "number of children" a value of 1.43 could be imputed under the normal model. This poses a practical problem for researchers who wish to utilize the benefits of multiple imputation but require their data to have discrete values.

When dealing with values imputed under the normal model, researchers have often been advised to round the imputed values to the nearest integer (Schafer 1997). If a value of 1.43 was imputed for a variable representing number of children, for example, the value would be rounded to a 1.0. Other research has shown that this naïve rounding method can cause more bias than the original "implausible" imputation values (Horton, Lipsitz and Parzen 2003). A number of alternative rounding methods have recently been proposed to deal with this limitation (Demirtas 2007; Yucel, He and Zaslavsky 2008).

A second option for multiple imputation in the presence of categorical data is a model based approach. When all the variables in the model are categorical, for example, a log-linear imputation model can be used (Schafer 1997). Most survey datasets are more diverse than this and there is a clear need for an imputation procedure than can handle a relatively complex data structure. Logical or consistency bounds have been proposed as one method for dealing with this (Heeringa, Little and Raghunathan 1997; Raghunathan et al. 2001). Another approach is to impute each variable conditional on all others, an iterative univariate imputation procedure proposed by Kennickell and McManus (1994). This procedure is known as regression switching, chained equations, sequential regressions, or variable by variable Gibbs sampling (van Buuren and Oudshoorn 1999). The basic idea behind this procedure is that since there are a variety of regression techniques for modeling non continuous data (e.g. logit, multinomial), each variable to be imputed can be fit with a tailored prediction equation, reducing the multivariate imputation task to a series of regression models. This procedure is implemented in the Stata ICE package (Royston 2004).

The primary advantage of multiple imputation is that it provides unbiased estimates of the means and covariances and accurately accounts for the standard errors. An important limitation when dealing

with categorical data is that it does not provide accurate estimates of the frequency distributions of the imputed values. Particularly when data are imputed under multivariate normality, analysis methods that use crosstabs or report frequencies may be in error, even with a planned missing design where the missing data are distributed completely at random (MCAR). This is of no small consequence for researchers who rely upon descriptive statistics.

In this paper we use ICE in Stata to multiply impute both planned missing and regular missing data from a large national telephone interview survey on social factors in infertility. We first demonstrate the problem of biased frequency distributions for the imputed data in a set of Likert-type items included on the survey instrument. We next apply to these imputations a method developed by Yucel, He and Zaslavsky (2008; hereafter referred to as YHZ) to correct the frequency distributions when the data are assumed to be missing completely at random. Finally, we develop and apply a modification of the YHZ approach to yield approximately correct distributions when the missing data follow a missing at random (MAR) pattern. The effect of these adjustments on the covariances is also empirically examined. Finally, we evaluate the different calibration approaches in a simulation.

2. Empirical example of bias in distributions

When data are imputed under a fully normal model, as many modern imputation programs implement (e.g., SAS MI, MI in SPSS and Stata, NORM), this poses problems when conducting frequency distributions. We illustrate this by examining the distributions of four-category Likert-type items in the National Survey of Fertility Barriers (NSFB). The first item we examine is a social support item which asks the respondent to indicate how often "Someone gives you give advice in a crisis". There are four ordinal response categories (often, occasionally, sometimes, and never). To reduce survey length and participant burden, this item was a part of a set of questions given to only two-thirds of randomly selected respondents. This planned missing design results in missing values for about one-third of the respondents that are missing completely at random (MCAR). The distribution of the missing values would be expected to have roughly the same distribution as the fully observed values. The missing values were imputed using the normal model in ICE. Table 1 compares the distributions of the observed and imputed values. The imputed values were rounded to the nearest whole valid response (1-4). It is clear that the distribution of the imputed values is biased, with particularly large differences in the first two categories. The reason for this is that the imputed values are generated to fit a normal distribution. In this case the observed data do not conform to a normal distribution so the imputed data are biased when naively rounded.

Table 1. Observed and imputed distributions when the missing data are MCAR

	Observed Data		"Nieve" Rounded Imputed Data		YHZ Calibration of Imputed Data	
	N	%	N	%	N	%
1. Often	2,656	78	531	63	640	76
2. Occasionally	413	12	226	27	100	12
3. Seldom	220	6	68	8	53	6
4. Never	124	4	20	2	52	6
Total	2,656	100	845	100	845	100

Figure 1 illustrates this difference from another item on the survey—the respondents rating of the importance of having children. Red bars (integers 1–4) give the density of the observed values. Blue bars (non-integer values that follow an approximately normal distribution) show imputed values which were not rounded. The distributions are quite distinct and it is clear that rounding the imputed values would not reproduce the observed distribution into the four ordinal categories.

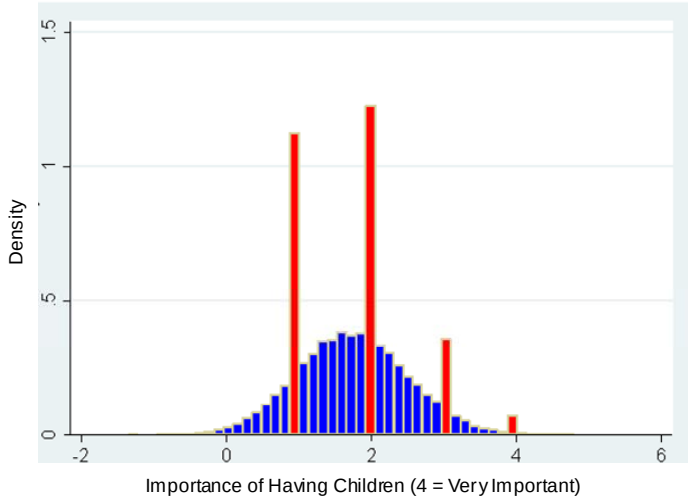

Figure 1. Comparison of distributions of observed and imputed values

3. Calibration approaches

Because rounding the imputed values to discrete categories does not adequately approximate the observed distribution, Yucel, He, and Zaslavsky (2008) developed a method to calibrate thresholds for assigning the imputed data to the observed categories. The Yucel-He-Zaslavsky(YHZ) method forces the distribution of the imputed data to match the distribution of the observed data. The method generates a centile score for each imputed value and assigns the score to the category based on thresholds computed from the non-missing data. The distribution of the recoded imputed data now closely matches to the distribution of the observed data and is much more accurate depiction of the distribution than was found for the "naive" rounding method. Column three in Table 1 shows the YHZ calibrated distribution for this example with MCAR data.

3.1 Limitations of the YHZ Method

When the missing data are distributed completely at random (MCAR) such as is the case with planned missing designs, the YHZ method works quite well in reproducing the distribution in the categories. When data are missing at random (MAR), however, the expectation of the true distribution assuming no missing data is not equal to the observed distribution when missing data are present. In this case, calibrating the imputed data to match the observed distribution can yield a biased estimate of the true distribution.

We illustrate this limitation with MAR data from the NSFB. In this survey targeting women (primary respondents), interviews with husbands (secondary respondents) were also sought if the primary respondent was married. An interview was successfully completed with the husband in less than half of the cases, producing substantial missing data. Because husband's response was correlated with a number of characteristics of his wife, the data were not missing completely at random and the MAR assumption is more plausible. We use the husband's response to the social support item: "Someone to give you information to help you understand a situation" which has four ordered response categories (often, occasionally, seldom, and never). The results of both naïve rounding and the YHZ calibration method are shown in Table 2. In this case there is a substantial difference between the distribution of the rounded imputed data and the observed data and the YHZ calibrations. We suspect that because the missing data are MAR that the naïve rounding may indeed be a closer approximation to the distribution that would have been found if these values had been observed.

Comment [RY1]: I think this might be sort-of confusing to people who are not familiar with the survey

Table 2. Rounding and Calibration Distributions with MAR data

	Observed Data		Naïve Rounded Imputed Data		YHZ Calibrated Imputed Data		Johnson-Young Calibrated Imputed Data	
	N	%	N	%	N	%	N	%
1 Often	550	66	995	45	1448	66	1322	60
2 Occasionally	170	20	886	40	447	20	443	20
3 Seldom	78	9	283	13	205	9	254	12
4 Never	34	4	26	1	90	4	172	8
Total	832	100	2190	100	2190	100	2190	100

3.2 A revised calibration approach for MAR data

Because the YHZ method yields biased estimates under MAR conditions and naïve rounding has also been shown to be biased, a revised approach is needed. We next develop a modified method (Johnson-Young) that is designed to work with MAR data. Rather than developing a single set of thresholds based on the observed distribution to allocate the missing values, we instead develop a set of thresholds (5 are used here) based on the observed distributions of the variable for different levels of predicted values of the variable. We estimate the distribution of the missing values from the distribution of the observed scores based on the predicted score from a regression including all variables used in the imputation. For example, assume we have four variables Y, X1, X2, X3. All may have missing values and are all used in the imputation model. In the first step we regress each variable on the other three variables and generate a predicted score for each observed case on the variables. We then divide the predicted scores into quintiles. Within each of the quintiles we generate the distribution of the observed scores. These distributions are used to develop thresholds for assigning the case to the categories of the variable. The final step is to calibrate the imputed values with the thresholds from the quintile in which the predicted score of the imputed value falls. This is illustrated in Figure 2. The vertical lines divide the predicted social support scores into five quintiles. Within each quintile the distribution of the observed scores falling is calculated. The scores for the missing values predicted from the variables in the model which have observed values are used to select the thresholds for calibrating the imputed values.

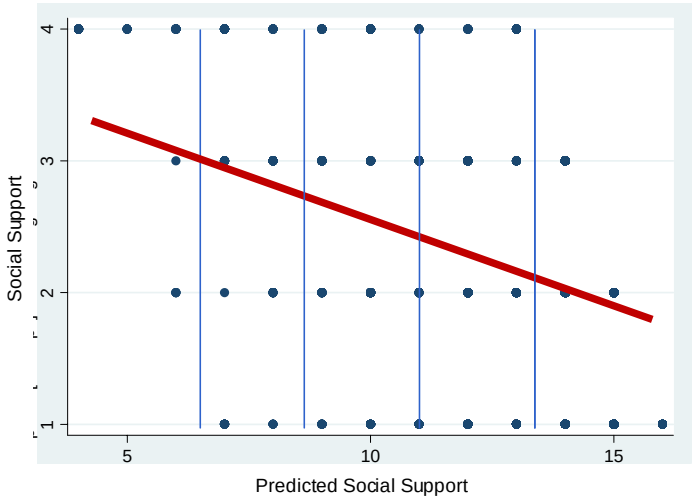

Figure 2. Illustration of the Johnson-Young Calibration Approach

Table 2 shows the distribution obtained when the Johnson-Young calibration method is used with MAR data. Although the YHZ method and the Johnson-Young method yield similar distributions, with the largest difference in the highest and lowest scores, both are substantially different from those obtained under naïve rounding. Although the distributions are likely less biased using these calibration approaches it is possible that because this method changes the imputed values that the estimates of the correlations, covariances, and regression coefficients may be biased by applying the procedure. To assess this we conduct a simulation study.

4. Testing the New Calibration Method with a Simulation Study

To assess how the different methods of calibrating the imputed values affect the estimates obtained when analyzing the imputed data we conducted a simulation using Stata with 1,000 replications under the following conditions.

- Two normally distributed random variables were generated, W and X, that were set to be correlated (r ~ .3).
- Two sample sizes were used with N of 250 and 2,000.
- Missing data was generated for W so that approximately one-third of the values would be missing. The missing were assigned in two ways. In the first case, the missing were assigned completely at random (MCAR) and in the second the missing data were assigned to be correlated with X (MAR).
- W was recoded into 4 ordinal values (1, 2, 3, 4) where the cases were distributed in one of 4 ways:
- The X was either kept as a normally distributed continuous variable or was recoded into 4 ordinal uniformly distributed values
    - uniformly distributed (.25, .25, .25, .25)
    - normally distributed (.20, .30, .30, .20)
    - triangularly distributed (.05, .15, .30, .50)
    - bimodally distributed (.40, .10, .10, .40)

The missing data were imputed using ICE in Stata under two different models. The first was the widely used normal model similar to the approach used in SAS MI, NORM, and the MI procedure in SPSS. The second approach used the multinomial regression method to impute the missing data in W. In both approaches we used 20 imputations with 200 burn-in and 100 between dataset iterations. Because of space limitations we only present the results with 250 cases in the dataset, when the missing was MAR and the X variable was recoded into 4 uniformly distributed categories. The results of the other models are available from the authors on request.

4.1 Imputation and Calibration Methods Compared

For each of the simulation conditions we compared statistics derived in the true score model which contained no missing data in W with the statistics derived from six different approaches. The first approach involved no calibration. Here, the imputed values were not rounded or transformed in any way. The second approach was naïve rounding, where the imputed values were rounded into the four integer categories. The third approach was the use of the multinomial option in ICE. The W variable was treated as a categorical variable and multinomial regression models were used to assign the missing value imputations to one of the four categories. The YHZ and the Johnson-Young calibration methods were the fourth and fifth methods. Finally, we also reported the results from a complete cases analysis where the missing cases were excluded from the model.

For each approach we report five outcome measures. These are all expressed as differences from the "true score" model with no missing data. These are the mean of W, the correlation of W and X, the b coefficient from a regression model where X was the dependent variable and W was the independent variable, and the standard error of this b coefficient. For the situations in which we use discrete values for the imputed values of W (all models except for the default ICE model in which the imputed values were not recoded) we also report the absolute value of the difference in the proportions in the categories in the true model compared to the imputed model. The results from these models are summarized in Table 3.

4.2 Summary of the results of the simulations.

When comparing how well the different models conformed to the true estimates, we begin with certain expectations about the pattern that we expect to emerge. The multiple imputation normal model is designed to provide unbiased estimates of the means and covariances when computed using the imputed values assigned by the procedure without rounding or assignment to the observed range. Therefore we would expect the no calibration model to provide the least biased estimates of the mean, correlation, and the regression coefficient. Because this same normal model makes a normality assumption about the distribution of the variables, we would expect it to perform more poorly when the distribution departs most from normal. The multinomial imputation model makes no distributional assumptions, so we would expect it to perform well in reproducing the true score distribution. Finally, because the data in the simulation we report are MAR, we would expect the YHZ method to not perform well. We also report the complete cases analysis, primarily for comparison purposes, but expect it to be biased in some of these measures with MAR data.

Comparing the estimates of bias in the means, no calibration, multinomial and Johnson-Young methods consistently showed the least bias across the distributions. Complete case analysis and YHZ were similar, as would be expected. The YHZ approach calibrates the distribution to that of the non-missing cases, but substantially underestimate the mean. Naive rounding works relatively well for the normal condition, but not as well when the distribution deviated from the normal. When examining the bias in the correlation coefficient the no calibration and Johnson-Young were very similar and showed the least bias. The estimates were quite good for the multinomial approach. As the distribution departed

further from normality, the naïve rounding method showed increasing bias. The most bias was found in the YHZ calibration method and the complete cases.

Because the imputed values in the no calibration approach have decimal values, we could not compare the differences from the true distributions to the categories for this method. We would expect the multinomial and the Johnson-Young method to perform best here, which indeed they did. The other three methods were all more biased, with naïve rounding the most biased as the distribution departed from normality. We expected the no calibration approach to yield the least biased estimates of the b coefficients, but instead found less biased estimates for all distributions for the Johnson-Young approach. The Multinomial estimates were also better than the uncalibrated estimates and similar to those for Johnson-Young. Naïve rounding also performed well except when the distribution was triangular. The YHZ and complete cases methods showed the most bias for the b coefficient. Finally, when examining the difference from the true standard error of the b coefficient, complete cases analysis came closest to the true estimate with no clear pattern of differences for the other methods which tended to have similar differences.

Although not reported here, the findings from the simulations with a larger sample size (2,000) and when the X variable was a normally distributed continuous variable are very similar to the findings in this simulation model discussed here.

5.0 Discussion and Conclusions

Researchers who have a need for the distributions of the imputed missing values to accurately reflect the distribution that would have been found if no missing data were present have few tools available to meet this requirement. Moreover, the missing data literature has cautioned against rounding or otherwise recoding the imputed values because this is likely to lead to biased estimates of the covariances and regression coefficients computed from the data. Our findings here suggest that several approaches are available to calibrate the distribution even when the pattern of missing data are MAR and then when calibrated do not appear to distort the means, covariances, and regression coefficients. Two methods appear to work well at producing accurate distributions as well as means and covariances. These are the use of a multinomial imputation model, such as the one in ICE, and the other is by calibration using the Johnson-Young approach. The Johnson-Young method performed slightly better when the distribution departed most from normality. The differences, however, were small. Perhaps the most surprising finding is that recoding the imputed values follow the Johnson-Young approach did not distort the covariance-based estimates. The expectation was that when the imputed values were calibrated to yield more accurate distributions it would be at the cost of distorting the other coefficients. Because we did not find evidence of this in our simulation, it suggests that researchers can calibrate the values without distorting the models.

Comment [rly2]: ?

The multinomial model also worked well and may be easier for the researcher to use than the Johnson-Young calibration approach. When a large number of variables are being imputed, many of which have a relatively small number of ordinal categories such as those found in many survey studies, then conducting the imputation may be problematic, take too long to complete, or fail to converge. Currently, only ICE in Stata and the impute package in IVEware implement an imputation approach which tailors the imputation model to the level of measurement of the variables. Until this ability becomes more widely available in other statistical packages, the use of the Johnson-Young calibration may be the best choice.

Additional research is needed to confirm some of our findings. Simulations are needed in which we vary the proportion of data missing, the degree to which the missingess is correlated with the variables, and which include more independent variables with missing data. We also only tested the

models when there were four categories in the ordinal variable. Variables with 2 to 5 or more categories should also be tested. Finally, a generalized computer program, such as a Stata ado, applying the Johnson-Young method is needed that can work with several variables and with varying number of categories is needed before this procedure is accessible to other researchers.

Acknowledgements

This research was supported in part by a grant "Infertility: Pathways and psychosocial outcomes" funded by NICHD (P.I. David R. Johnson). Additional support was provided by the Social Science Research Institute at The Pennsylvania State University

References

Demeritas, H., (2007), "Rounding Strategies for Multiply Imputed Binary Data," Technical report#: 2007-005: University of Illinois.

Graham, J., (2009), "Missing Data: Making it Work in the Real World," *Annual Review of Psychology*, 60, 549 – 576.

He, Y. and Raghunathan, T.E., (2006), "Tukey's gh Distribution for Multiple Imputation," *The American Statistician*, 60, 251 – 256.

Heeringa, S.G., Little, R.J., and Raghunathan, T.E., (1997), *Imputation of Multivariate Data on Household Net Worth*, Ann Arbor: University of Michigan.

Horton, N.J., Lipsitz, S.R., and Parzen, M. (2003), "A Potential for Bias When Rounding in Multiple Imputation," *The American Statistician*, 57, 229 – 232.

Kennickell, A.B., and McManus, D.A., (1994), "Multiple Imputation of the 1983 and 1989 waves of the SCF,' Paper Presented at the 1994 Annual Meetings of the American Statistical Association, Toronto, Ont.

Kenward, M.G., and Carpenter, J., (2007), "Multiple Imputation: Current Perspectives," *Statistical Methods in Medical Research*, 16, 199 – 218.

Little, R.J., and Rubin, D.B., (2002), *Statistical Analysis with Missing Data*, Second ed., New York: J. Wiley & Sons.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P., (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models," 27, 85 – 95.

Royston, P., (2005), "Multiple imputation of missing values." *Stata Journal*, 4, 227 – 241.

Rubin, D. B., (1976), "Missing Data and Inference," *Biometrica*, 63, 581 – 592.

Rubin, D.B., (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Schafer, J.L., (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall: London.

Schafer, J.L. (1999), "Multiple Imputation: A Primer", *Statistical Methods in Medical Research*, 8, 3 – 15.

Van Buuren, S. and Oudshoorn, K., (1999). "Flexible Multivariate Imputation by MICE," Leiden: TNO Preventie en Gezondheid.

Yu, L. M., Burton, A., and Rivero-Arias, O., (2007), "Evaluation of Software for Multiple Imputation of Semi-Continuous Data," *Statistical Methods in Medical Research*, 16, 243 – 258.

Yucel, R.M., and Demirtas, H., (2009), "Impact of Non-Normal Random Effects on Inference by Multiple Imputation: A Simulation Assessment," *Computational Statistics and Data Analysis*, Article in Press.

Yucel, R.M., He Y., and Zaslavsky, A. M., (2008), "Using Calibration to Improve Rounding in Imputation," *The American Statistician*, 62, 125 – 129.

Table 3. Differences from the true score for each method under each distribution

	No Calibration		Naive Rounding		Multinomial		YHZ		Johnson-Young		Complete Cases	
Difference	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Mean												
Normal	-0.0124	0.0466	-0.0191	0.0438	-0.0141	0.0451	-0.0956	0.0425	-0.0119	0.0459	-0.0970	0.0428
Uniform	-0.0124	0.0505	-0.0222	0.0465	-0.0148	0.0486	-0.0975	0.0458	-0.0119	0.0498	-0.1040	0.0464
Triangular	-0.0047	0.0370	-0.0435	0.0342	-0.0153	0.0353	-0.0747	0.0357	-0.0067	0.0357	-0.0757	0.0358
Bimodal	-0.0139	0.0630	-0.0312	0.0557	-0.0167	0.0607	-0.0747	0.0557	-0.0131	0.0620	-0.1179	0.0576
Correlation												
Normal	-0.0092	0.0434	-0.0144	0.0423	-0.0104	0.0431	-0.0571	0.0410	-0.0093	0.0469	-0.0215	0.0417
Uniform	-0.0088	0.0437	-0.0143	0.0424	-0.0103	0.0437	-0.0541	0.0415	-0.0088	0.0472	-0.0211	0.0419
Triangular	-0.0083	0.0394	-0.0321	0.0381	-0.0141	0.0397	-0.0527	0.0376	-0.0045	0.0434	-0.0200	0.0376
Bimodal	-0.0092	0.0454	-0.0157	0.0433	-0.0103	0.0451	-0.0379	0.0423	-0.0084	0.0489	-0.0208	0.0434
Proportion												
Normal	--	--	0.0154	0.0067	0.0159	0.0071	0.0233	0.0081	0.0161	0.0072	0.0235	0.0081
Uniform	--	--	0.0213	0.0079	0.0161	0.0071	0.0234	0.0084	0.0162	0.0072	0.0242	0.0083
Triangular	--	--	0.0224	0.0088	0.0132	0.0067	0.0220	0.0089	0.0129	0.0067	0.0221	0.0089
Bimodal	--	--	0.0505	0.0086	0.0149	0.0069	0.0274	0.0079	0.0148	0.0069	0.0247	0.0094
b - coefficient												
Normal	-0.0105	0.0452	-0.0075	0.0458	-0.0096	0.0458	-0.0590	0.0440	-0.0083	0.0501	-0.0371	0.0444
Uniform	-0.0095	0.0420	-0.0029	0.0428	-0.0086	0.0430	-0.0502	0.0413	-0.0073	0.0465	-0.0337	0.0410
Triangular	-0.0252	0.0494	-0.0346	0.0513	-0.0189	0.0529	-0.0745	0.0488	-0.0048	0.0581	-0.0535	0.0488
Bimodal	-0.0086	0.0362	0.0026	0.0372	-0.0073	0.0370	-0.0245	0.0354	-0.0063	0.0401	-0.0274	0.0349
SE of b												
Normal	0.0139	0.0060	0.0150	0.0056	0.0149	0.0059	0.0177	0.0059	0.0139	0.0059	0.0106	0.0029
Uniform	0.0127	0.0054	0.0140	0.0050	0.0139	0.0053	0.0160	0.0053	0.0129	0.0053	0.0097	0.0025
Triangular	0.0134	0.0063	0.0176	0.0061	0.0197	0.0095	0.0194	0.0066	0.0147	0.0068	0.0092	0.0030
Bimodal	0.0107	0.0043	0.0120	0.0037	0.0119	0.0042	0.0125	0.0039	0.0106	0.0040	0.0078	0.0019