Increased Accuracy of Distribution Based Missing Value Imputation: An Alternative to Mean Imputation in Real World Environment Survey Research

Thomas Emerson Wasser[*]
Articles

Anh Thu Burks[2],
Kelly Bristol[3],
Lauren Walton[4]

Tags: bias reduction, survey methods, mean imputation, missing value imputation
Survey Practice
Vol. 7, Issue 3, 2014

Objective: Missing values within variables prevents data analysis on many levels including both univariate and multivariate analysis. This research presents distribution based missing value imputation, where the distribution of nonmissing values is simulated to create a set of values that are then randomly inserted into to the missing values in the actual data.
Methods: Distribution imputation was compared to mean imputation in 12 different simulation conditions based on four sample sizes (50, 100, 150, 200) and three different missing value percentages for each of the sample sizes (10 percent, 20 percent, and 30 percent). Each simulation created 1,000 test datasets within each condition for a total of 12,000 simulated datasets.
Results: Mean based imputation was biased and less accurate than distribution based imputation (DBI) in all 12 simulations combinations. DBI was more accurate in matching the number of rejected hypotheses as compared to the gold standard. Comparing the calculated p-values for bias where an unbiased estimator would demonstrate a 50/50 split being greater than and less than the gold standard, DBI was closer to the gold standard with at 48.7/51.3 split as compared to the 25.8/74.2 split of mean based imputation.
Conclusions: Distribution imputation was found to be more accurate and unbiased as compared to mean based imputation methods. As a result, when studies are small and do not contain a large number of variables or even in situations where more elaborate imputation methods cannot be done, DBI is an accurate and unbiased method.

INTRODUCTION

As clinical trials move into late phase, real-world research, patients are often surveyed for many years. In other types of sampling, patients are tracked in large insurance claims databases where they are later surveyed. In these studies, the issue of how to handle missing data becomes more important than single

* **Institution:** HealthCore, Inc.
2 **Institution:** Nielsen
3 **Institution:** Nielsen
4 **Institution:** Nielsen

surveys because a single missing value will remove a patient from the entire analysis. (Graham et al. 2003; Rubin 1987; Tsikriktsis 2005). Missing values complicate the research process at several levels. From the methodological perspective, missing data could represent bias (Seaman and White 2001; Shih 2002). Patients with missing data could be different than those with complete data for many different reasons including demographics, previous therapy, comorbidities, existing conditions and others (Tsikriktsis 2005).

From the research design perspective missing values in repeated measures (within subjects) designs could be handled differently than in group (between subjects) analysis where only a single observation has been made on a smaller number of variables (Graham et al. 2003; Shih 2002). Additionally, there has been much research from the analytical perspectives including questions regarding the type of imputation (Schlomer et al. 2010) to be made, determination of whether the data is missing at random (Rubin 1976; Schlomer et al. 2010), and what percentage of missing data within a variable makes imputation less desirable (Barzi and Woodward 2004; Rubin et al. 2007).

The literature demonstrates that mean based imputation is biased (Tsikriktsis 2005). This bias occurs because substituting missing values with a constant decreases the standard deviation. By extension the significance between the two groups will increase. This means the likelihood of incorrectly rejecting the null hypothesis (H_o) increases. This is the definition of Type I error. Similar to mean based imputation, last value carried forward imputation cannot be used in group based analysis as it is a method best suited for repeated measures, and studies have shown it is not accurate (Dinh 2013; Donders et al. 2006; Graham et al. 2003).

This research introduces distribution based imputation (DBI). DBI works by calculating the mean and standard deviation for a variable that is not missing, and uses those statistics a random distribution of data is simulated. Values from this simulated data are then sampled at random and inserted into the missing values for the real variable. This method is more flexible in databases with a small number of variables were multiple imputation is not possible or ill-advised.

While multiple imputation methods may be shown to be effective in research were a large number of variables with complete data on those variables are available; it would not be applicable in research where only a small number of study variables are used. This research tests DBI against mean based imputation and compares the results against the exact gold standard results that would have been obtained if there were no missing data.

METHODS

Data simulations were conducted using the R (sample executable code for performing a simulation is included as an Appendix). To create the simulation,

two large distributions (n=100,000) were constructed with the intention of representing the population of values for two variables. These two variables can be thought of as the outcome variable for two independent groups. Simulation instructions for 'R' required the X variable to have a mean of 100 and standard distribution (SD) of 15, the second Y variable was instructed to have a mean of 105 and SD also equal to 15. This simulation would then have an overall effect size of 0.33 or 33 percent of one SD.

From these two distributions, four different sample sizes were extracted at random for both the X and Y variables (50, 100, 150, and 200). Within each of the sample sizes, three different missing value percentages were used for each variable (10 percent, 20 percent and 30 percent). This resulted in 12 (4×3) distinct simulations being performed. Missing values for both the distribution and mean imputation methods were selected at random from the given sample size and were the same values within each X and Y array. Imputation methods were performed as follows:

- For mean based imputation, the mean was calculated for the nonmissing data. That mean value was then substituted into the randomly selected places in the array where the missing values had been extracted.
- For DBI, the mean and standard deviation were calculated on the nonmissing data. These values were then used to create a faux-distribution based on 1,000 observations. From this distribution, the number of missing values were extracted without replacement and placed into the array where the missing values had been extracted.

From each of these simulations, three independent t-tests were calculated for the sample sizes mentioned. The first served as the gold standard test and was the result of the X and Y variable with no missing data. The second was the t-test comparing X and Y variables where distribution based imputation had been used, and the third was the t-test comparing X and Y variables where mean based imputation had been used.

Within each sample size and missing value combination (12 simulations), this process was repeated 1,000 times, selecting new samples from the 100,000 X and Y variables at each pass. The means and standard deviations for each of the three conditions (gold standard, distribution imputation, and mean imputation as well as the *p*-value were written to an external file to facilitate further analysis. This process yielded a master database of 12,000 results specific to all of the conditions mentioned above. Calculation of the t-values and *p*-values for all of the t-test procedures were performed with the R program t-test procedure, and no test validation was needed.

DATA ANALYSIS

From the master database, the number of statistically significant results using

a *p*-value of 0.05 were counted for the gold standard and both distribution and mean imputation methods. The number of rejected ($p<0.05$) tests were then placed into a table for each of the sample size and missing value percent stratifications and expressed as a percent. The most efficient method is the one that yielded a closer count (percentage) of rejected hypothesis values to the gold standard test. The more unbiased imputation methods would be the one that was both accurate and had counts (percentage) of *p*-values both above and below the count of the gold standard.

An examination was performed to determine the bias of the imputation method, without regard to the arbitrary cutoff value of $p=0.05$. In order to examine bias, the calculated value of the *p*-value was used for both distribution and mean based imputation compared to the gold standard. An unbiased estimator would have approximately the same number of calculated *p*-values greater than and less than the gold standard *p*-value. In other words based on 12,000 simulations, an unbiased estimator would have approximately 6,000 calculated *p*-values both greater than and less than the gold standard *p*-value. In order for a statistical method to be conclusively determined to be more efficient than another, it must outperform the other method across a variety of different experimental conditions. In this simulation study, the more efficient method is the one that most closely approximates the gold standard values in terms of the absolute number of *p*-values that would be rejected as well as the percentage of time that the resulting *p*-value was an over estimate or underestimate of the gold standard *p*-value.

The last analysis was a descriptive based analysis calculating the average SD value for the X variable across all simulations. This analysis was performed to illustrate the suspicion that using mean based imputation would decrease the SD values, and this then leads to a greater count of statistically significant results and an increase in Type I error.

RESULTS

With regard to the accuracy of the simulation of the population based X and Y values, the results of the R-based population simulation were checked and validated for accuracy (X-variable mean=100.00, SD=14.99. Y-variable mean=105.00, SD=15.01).

Mean based imputation resulted in a higher percentage of rejected HO hypotheses than the gold standard tests in all 12 of the simulation scenarios (100 percent). The range of these differences was less than 1 percent in only two simulations and greater than 5 percent in five of the 12 simulations (Table 1). The average increase in Type I error was 4.57 percent with the smallest increase at 0.2 percent and the greatest increase at 11.6 percent. Five of the simulations yielded increases greater than five percent. Distribution based imputation was not biased as an increase in Type I error was seen in only three of 12 simulations (25 percent). The largest increase in Type I error was only 4.2 percent in a

condition where mean based imputation had a Type I error rate of 9.7 percent (n=50, missing 20 percent).

Table 1 Comparisons of simulation and mean based imputation against population results with no missing data. Analysis based on 1,000 simulations. Effect size is 0.33 of a standard deviation.*

Sample size for t-test (per group)	Percent of missing data	Percent of 1,000 rejected H_o hypothesis				
		Gold standard with no missing data	DBI	Percent from gold standard	Mean based imputation	Percent from gold standard
50	10	39.0	39.3	0.3	42.4	3.4
	20	36.3	40.5	4.2	46.0	9.7
	30	36.8	38.0	1.2	48.4	11.6
100	10	64.3	63.0	−1.3	67.5	3.2
	20	66.2	65.3	−0.9	72.6	6.4
	30	67.2	64.5	−2.7	74.9	7.7
150	10	83.2	80.5	−2.7	84.1	0.9
	20	84.2	80.4	−3.8	85.6	1.4
	30	80.4	79.6	−0.8	85.9	5.5
200	10	92.2	89.2	−3.0	92.4	0.2
	20	91.4	89.4	−2.0	93.4	2.0
	30	92.2	90.7	−1.7	95.0	2.8

*Most accurate estimator is shaded.

In order to examine the bias of distribution and mean imputation without regard to the alpha level of a statistical test, Table 2 presents the data on the imputation methods based on the calculated value compared to the gold standard. For all of the individual simulations a higher percentage of the mean based imputation, *p*-values were smaller than the *p*-values calculated on the gold standard. Across all 12,000 simulations, the *p*-value for Mean based imputation was greater than the gold standard 3,091 (25.8 percent) of the time and smaller than the gold standard 8,909 (74.2 percent) times. In four of 12 simulations, the *p*-values were smaller than the gold standard more than 80 percent of the time. This indicates a high level of bias.

For distribution based imputation, there was an even split 50/50 split on one of the 12 simulations which indicates a perfect non-bias condition (n=150, missing 10 percent). The maximum departure from the desired 50 percent value was only 54.6 percent (n=50, missing 20 percent) as compared to 68 percent in the Mean based imputation method. Across all 12,000 simulations Distribution based imputation demonstrated results much closer to the 50/50 results with *p*-value greater than the gold standard 5,845 (48.7 percent) and *p*-value less than the gold standard 6,155 (51.3 percent) The difference between distribution and mean based imputation across all 12,000 simulations can be seen graphically displayed in Figure 1.

Table 2 Bias analysis of distribution imputation versus mean imputation. (There are 1,000 simulations for each condition.)*

Sample size	Percent missing	DBI				Mean imputation			
		$P_{Random}>P_{Gold}$		$P_{Random}<P_{Gold}$		$P_{Mean}>P_{Gold}$		$P_{Mean}<P_{Gold}$	
		Count	Percent	Count	Percent	Count	Percent	Count	Percent
50	10	509	50.9	491	49.1	390	39.0	610	61.0
	20	454	45.4	546	54.6	320	32.0	680	68.0
	30	482	48.2	518	51.8	279	27.9	721	72.1
100	10	483	48.3	517	51.7	353	35.3	647	64.7
	20	471	47.1	529	52.9	271	27.1	729	72.9
	30	471	47.1	529	52.9	188	18.8	812	81.2
150	10	500	50.0	500	50.0	310	31.0	690	69.0
	20	502	50.2	498	49.8	230	23.0	770	77.0
	30	488	48.8	512	51.2	161	16.1	839	83.9
200	10	487	48.7	513	51.3	281	28.1	719	71.9
	20	495	49.5	505	50.5	175	17.5	825	82.5
	30	503	50.3	497	49.7	133	13.3	867	86.7
Total		5,845	48.7	6,155	51.3	3,091	25.8	8,909	74.2

*Most accurate estimator is shaded.

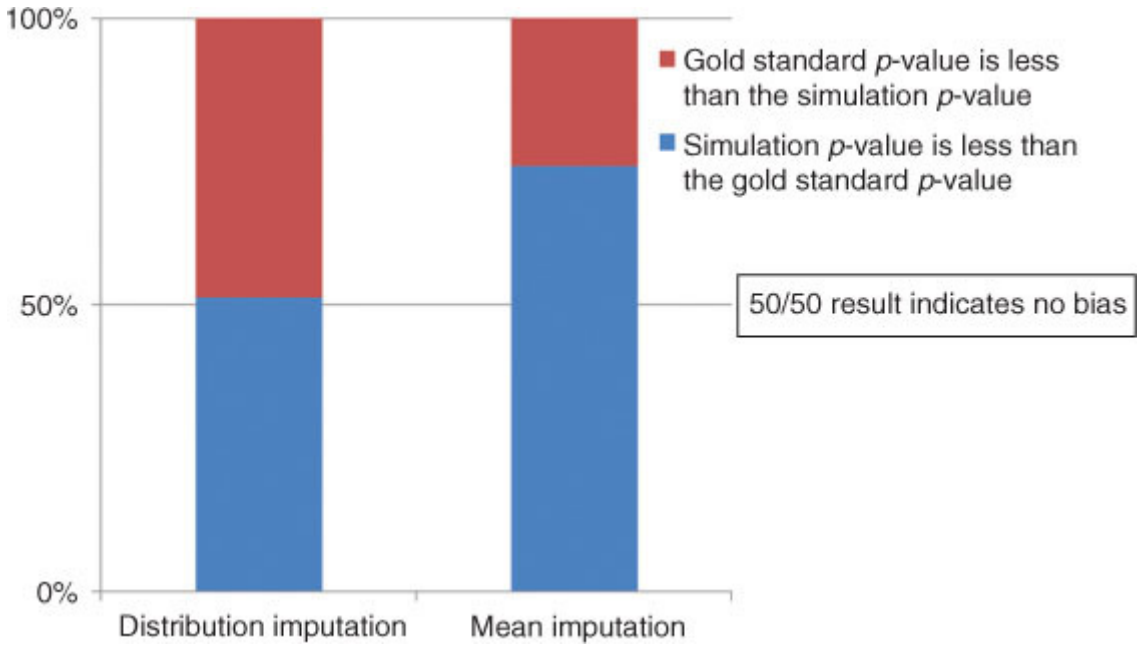

Figure 1 Total simulation bias measurement. (Each bar is based on 12,000 simulated t-tests, 1,000 in each experimental condition [Sample size and percent missing]).

Desired outcome for evidence of no bias is 50/50 split in randomization.

Table 3 displays data regarding the accuracy of the standard deviations for the simulated data on the X variable. Recall that the simulation specification was for a SD=15.00, as can be seen on all occasions the distribution based method of imputation yielded SD values closer to the gold standard and in all occasions the mean based imputation method had SD values that were smaller than the gold standard.

Table 3 Illustration of the accuracy of the standard deviations of distribution and mean imputation methods compared to the gold standard value.*

Sample size	Percent missing	Imputation method				
		Gold standard	DBI		Mean	
			Avg SD	Δ	Avg SD	Δ
50	10	14.92	14.90	−0.02	14.17	−0.75
	20	15.06	15.05	−0.01	13.46	−1.60
	30	14.83	14.74	−0.09	12.34	−2.49
100	10	14.96	14.93	−0.03	14.18	−0.78
	20	14.93	14.89	−0.04	13.33	−1.60
	30	14.91	14.84	−0.07	12.44	−2.47
150	10	14.94	14.92	−0.02	14.15	−0.79
	20	14.91	14.90	−0.01	13.34	−1.57
	30	14.97	14.96	−0.01	12.49	−2.48
200	10	14.96	14.97	0.01	14.20	−0.77
	20	15.00	15.00	0.00	13.42	−1.58
	30	15.00	15.00	0.00	12.55	−2.45

*Most accurate estimator is shaded.

DISCUSSION

In order for a statistical method to be conclusively determined to be more efficient than another, it must outperform the other method across a variety of different experimental conditions. In order to determine bias of an imputation method, the method cannot systematically over or underestimate the gold standard. In this study, DBI was both more efficient and had less bias than mean based imputation.

The fact that mean based imputation had a higher percentage of rejected Ho hypothesis all of the simulation conditions demonstrates a high level of bias for the method. Not only were these percentages higher but in five of the 12 conditions the percentage was more than five percent greater. These percentages are Type I errors and the increase in this error averaged 4.57 percent. This provides clear evidence that mean based imputation is biased. DBI was not biased as the percent of rejections were both higher and lower than the gold standard and averaged only a −1.1 percent which indicates no increase in Type I error when averaged across all simulation conditions.

Similar results were seen when examining the exact, calculated value of the *p*-value. An unbiased estimate of the gold standard can be determined by and even split of imputated *p*-values both greater than and less than the gold standard. In this application with 12,000 simulations, a perfect unbiased method would result in 6,000 values both greater and less than the gold standard for a 50/50 split. Distribution based imputation was remarkably close (48.7/51.3) as compared to mean based imputation which was highly biased (25.8/74.2).

The last analysis highlighted in Table 3 displays the reason why mean based imputation leads to more rejected Ho hypothesis. As can be seen in the table the SDs for the mean based imputation are smaller in each of the 12 simulations across all conditions of sample size and missing value percentage. The DBI yields nearly exact SD values as compared to the gold standard. Therefore, the suspicion that the mean based imputation method is biased by the SD being much smaller and leading to a higher or increased Type I error rate is satisfied using the methodology of this study.

LIMITATIONS

This study did not address the data missing at random issue. This is more of a research design based issue and would rely on different methods of data analysis or design interpretation. Second, only a limited number of experimental simulations were used in this study and clearly different sample sizes, effect sizes and percentages of missing data could be comprehensively examined. While it is expected that the findings of this study would extend beyond this simulation, it cannot conclusively be determined.

While this study confirms what has been suspected regarding the relative inaccuracy of mean based imputation, it also presents a highly accurate method of imputation that is not biased with regard to that calculated value of the *p*-value but also highly efficient. The technique of DBI is easy to perform in R and is able to be performed on large datasets regardless of the sample size and percent of missing values within. Whether to use multiple imputation methods or DBI or even no imputation depends on the nature of the experiment and the hypotheses being tested. However, in studies with a smaller number of variables, where other imputation methods cannot be done, DBI is an accurate and unbiased method.

REFERENCES

Barzi, F. and M. Woodward. 2004. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology 160*(1): 34–45.

Dinh, P. 2013. Using multiple imputation to estimate cumulative distribution functions in longitudinal data analysis with data missing at random. *Pharmaceutical Statistics 12*(5): 260–267.

Donders, A., G. Heijden, T. Stijnen, and Moons, K. 2006. Review: a gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology 59*: 1087–1091.

Graham, J.W., P.E. Cumsille and E. Elek-Fisk. 2003. Methods for handling missing data. In: (J.A. Schinka and W.F. Velicer, eds.) *Research methods in psychology* Volume 2. of Handbook of Psychology (I. B. Weiner, Editor-in-Chief). John Wiley & Sons, New York.

Rubin, D. 1976. Inference and missing data. *Biometricka 63*(3): 581–592.

Rubin, D. 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York.

Rubin, L., K. Witkiewitz, J. Andre, and Reilly, S. 2007. Methods for handling missing data in the behavioral neurosciences: don't throw the baby rat out with the bath water. *The Journal of Undergraduate Neuroscience Education 5*(2): A71–A77.

Schlomer, G., S. Bauman and N. Card. 2010. Best practices for missing data management in counseling psychology. *Journal of Counseling Psychology 57*(1): 1–10.

Seaman, S. and I. White. 2001. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research 22*(3): 278–295.

Shih, W. 2002. Problems in dealing with missing data and informative censoring in clinical trials. *Current Controlled Trials in Cardiovascular Medicine 4*(3): 1–7.

Tsikriktsis, N. 2005. A review of techniques for treating missing data in OM survey research. *Journal of Operations Management 24*: 53–62.

APPENDIX

Sample code for R that was used to calculate the mean and standard deviation from simulation data and then substitute's distributional data and substitutes values back to the data. This code can easily be modified to be used on any continuous variable. For detailed instructions, readers may contact the author.

#Create the populational distribution with fixed parameters such as an IQ variable.

```
xpop<-rnorm(n=100,000, mean=100.0, sd=15)
```

#Create memory variables for use 'Pull' is the number of missing values needed to be pulled from the #simulated data.

```
n=200

pull=60
```

#Create the sample one value at a time from the population of 100,000.

```
for(i in 1:1,000)

x <- sample(xpop,n,replace=F)

xrand <-x
```

#storing the values to be used in the false distribution to be compared back to simulation for accuracy.

meanx<-mean(x)

sdx<-sd(x)

#selects the numbers that will go into the random replacement.

xselect<-rnorm(1,000,meanx,sdx)

xinsert<-sample(xselect,pull)

#Performs the substitution of the values (60 in this case) that will need to change for each N change.

usexrand<-replace(xrand, c[6:65], xinsert)

meanxrand<-mean(usexrand)

sdxrand<-sd(usexrand)

usexcon<-replace(xconstant, c[6:65], meanx)

meanxcon<-mean(usexcon)

sdxcon<-sd(usexcon)

}

#The next two lines prints the two variables and the reader can see values 6 through 65 have been replaced by the new random variable substitution.

xrand

usexrand