Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

DEVELOPMENT OF MISSING DATA ESTIMATION TECHNIQUES WITH STATISTICS METHODS

Orathai Chuacharoen*,
Department of Statistics, Faculty of Science, Ramkhamhaeng University, Thailand
Maniratt Jaroongdaechakul,
Department of Statistics, Faculty of Science, Ramkhamhaeng University, Thailand
Montree Piriyakul
Department of Statistics, Faculty of Science, Ramkhamhaeng University, Thailand
*Corresponding Author Email: orathai_c@rumail.ru.ac.th

Abstract

Development of missing data estimation techniques with statistics methods is an experimental research in order to study solution process for missing data and develop new data analysis techniques and statistics estimation methods by using data with uniform distribution (0,1), binomial distribution (50, 0.2), binomial distribution (50, 0.5), binomial distribution (50, 0.8), normal distribution (0.1), and 1,000 sets of real data with 10,000 replications. Research results are shown as follows:

1. Mean Imputation (MI) is suitable for normal distribution data.
2. K-Nearest Neighbor Imputation (KNN) is suitable for normal distribution and real data when the sample size is large.
3. Extreme Imputation (EI) is suitable for discrete random variables especially in uniform and binomial distribution, MSE and MMRE will be low when the sample size is small.
4. Side Imputation (SI) is suitable for binomial distribution data. Proportion of missing data is high and the sample size is large.
5. Side Mean Imputation (SMI) is suitable for normal distribution and real data when the sample size is large.
6. New Multiple Imputation (Multi-Im) is suitable for binomial distribution data with more than 10 sample sizes.

Concept of mean value application basis is used in study and development of missing data estimation technique. For further studies, concepts from other theories might be applied.

Keywords: Missing data, Missing data estimation

Introduction

Not only science data study and analysis but also of social science focuses on data integrity. Data, especially used in the analysis of research, experiment, business and economic analysis, organizational administration and development especially needs

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

to be integral in order to obtain correct and accurate analyzed results which can be used for solving problems, planning, making policies and consideration without error.

What usually found in data collection is that data are always missing. Researchers need to find proper solutions as this problem effects on data analysis efficiency. There are various alternatives for missing data management, however, choosing the one that is improper with data characteristics absolutely cause error of data analysis.

Missing data are observed values needed to be determined, still unknown. Such values should be known if proper method for data collection or efficiency measurement is used. Missing data is generally found in quantitative research even the survey or experiment is well controlled. (de Leeuw, Hox, & Huisman, 2003) Nowadays, researchers have been gradually more focusing on missing data and trying to solve this problem by collecting integral data. As missing data come from several causes, researchers need to consider proper guidelines for missing data management every time this problem is found. There are various alternatives for missing data management. Choosing the improper one will absolutely cause error of data analysis. Statistics methods are developed for integral data analysis. In the past, there were methods for fixing missing data before an analysis such as Listwise Deletion and Mean Substitution. These methods, however, cause bias of estimator especially in multivariate analysis. If data of only one variable are missed from unit of analysis, such unit will be cut off without considering whether there are other perfect variables or not. (Tsikrikstsis, 2005)

Cutting data is the beginning method of multivariate analysis in general statistics programs. The study and experiment of Kim and Curry (referred in Uchenna & Nduka, 2012) was found that if only 10% of each variable data are randomly missed, 59% of the unit of analysis will cut off. It can be seen that this is a very-high-rate loss. Analyzed results will be bias and incorrect. Information and expenses are lost. Therefore, proper missing data estimation techniques should be determined in advance because if it is used for replacing lists of missing data before data analysis, test and other statistics will have higher power of a test. Missing data estimation for replacing the data increases efficiency of the estimation and research conclusion. (Raymond, 1986)

As aforementioned, missing data problem is therefore so important that statisticians have always tried to find solutions. Therefore, this research studies solution process of missing data problems in order to compare and develop data analysis techniques when data are missed and so that the efficient estimator will be gained and able to replace the missing data. This will lead to more accurate and correct data analysis.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

Literature Review

Concept of Missing Data

According to Piyaporn and Sukon Prasitwattanaseri (2006), the study of missing data and its management scoping that missing data are observed values needed to be determined, still unknown. Such values should be known if proper method for data collection or efficiency measurement is used. Focusing on the importance of missing data on researches, effects of missing data on researches should be emphasized. The data might severely or might not cause any effects. However, the effects caused by missing data can be briefly classified as follows:

1. Missing data can cause losses of power of a test because the sample size is decreased by cutting the data from the study. When smaller sample size is analyzed, losses of confidence level and increase of variance in this study are absolutely affected.

2. Missing data might cause bias of estimates, for example, sensitive questions; drug abusive or sexual behaviors or even general questions like annual income, might not be desired to be answered. Therefore, some collected data cannot completely be representatives of the population.

3. Missing data might lead to difficulties in examination of effects of variables as they are lack of some population attributes that results in mistaken conclusion.

It can be seen that missing data effect on researches, both in analysis and interpretive conclusion. Severity level of such effects depends on several components such as size and type of missing data, importance of variables that have missing data in the research, and missing data management techniques.

Type of Missing Data

Consideration of type of missing data is an important step because if characteristics of missing data are recognized, guidelines for incomplete data management will be properly considered. Missing data are generally classified into 3 types as follows (Piyaporn and Sukon Prasitwattanaseri, 2006)

1. Missing completely at random (MCAR) are randomly missing data from entire observed values, i.e., the missing data are independent from variables. There are various reasons causing missing data. It can be broken tool, defecting equipment, bad weather, sick studied target group, or incorrect input. This type of missing data is considered the least problematic cause as it does not relate to data results. Therefore, complete data can be chosen for analysis.
2. Missing at random (MAR) are the missing data which are nonrandom from entire observed values, but from some internal parts or groups of them, i.e., values of the missing data depend on other variables in database.
3. Not missing at random (NMAR) are nonrandom missing data. Values of the missing data depend on complete data values in the same or other variables. In

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

some cases, values of missing data might not rely on any variables in database, but other variables that are not collected for that study. Characteristics of this type of missing data are considered seriously effecting on data analysis.

Practically, characteristics of MCAR are not commonly found, but of MAR vice versa. Therefore, developed statistics methods for solving missing data problems are often carried out under MAR conditions.

Methods of Handling Missing Data

There are several methods for handing missing data. Considering choosing which of them depends on the characteristics of missing data. If improper method is chosen, error might be increased and expected results can be failed. Methods of handling missing data that are often chosen are as follows:
(Sukchareon W., 2015)

1st Group Method Collect further data: This means to cut incomplete data out of the entire data and collecting further data for replacing the cut number of samples. However, further data collection absolutely needs cost and time.

2nd Group Method Delete no-data case: This means to delete data in case they are incomplete. It can be classified into 2 types: 1) Listwise deletion is to cut entire set of missing data by analyzing only complete data. Advantages of this method are that it is easy to understand and each type can be analyzed and compared while disadvantages are that chances of having missing data is higher and leads to lower power of an analysis. Moreover, expenses are spent with the missing data and error might be found and 2) Pairwise deletion an analysis of relationships between pairwise variables by analyzing data with 2 complete variables and in every case. Advantages of this method are that all collected data are used and the highest number of data is provided. However, disadvantages are that in some topics, data cannot be comparatively analyzed as the number of sample groups is unequal or data of each variable provided by the sample groups might be in different sets of data.

3rd Group Method Guess values of missing data by using the same value, for example, Mean Substitution is a method of replacing missing data by mean of known data. It is supposed that characteristics of similar sample units should have similar interesting values. Advantage of this method is that replacing data can be all found while disadvantage is the missing data might not be totally the same. The provided values are not corresponding to the fact even they decrease variance.

4th Group Method is an analysis by using statistics methods which are described below:

1. Dummy Variable Adjustment: For example, forecasting analysis whose disadvantage is to cause error.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

2. Regression Model or Regression Imputation whose disadvantage is that every happening data will be on the regression model. This situation is so called “Overfit”.
3. Maximum Likelihood Estimation is to calculate missing data from the existing ones by Log-Likelihood technique or maximum-likelihood values. The disadvantage is that this technique is not supported by current programs.
4. Multiple Imputation (MI) is the combination of Expectation Maximization (EM) and Raw Maximum Likelihood Methods together with capability of hot deck qualifications in order to simulate a set of data model that the missing data are substituted by several sets of replacing values. Then, the sets of data are analyzed and adjusted. After that, analyzed results are gradually recorded until the least error is provided.
5. Expectation Maximization (EM) is dependent on 2-step repetition process. The first process is called “Expectation (E) step” which will estimate the expected values from likelihood function under the condition of complete data. The second process is called “Maximization (M) step” where expected missing data will be substituted by E-step values and estimated from Likelihood function. In case of no missing data, the 2 steps will be repeated until convergence or small-changed values are obtained. The missing data are substituted by such values.
6. Raw Maximum Likelihood methods depend on using complete data for setting maximum likelihood values under proper statistics model like structural equation model, regression model, ANOVA and ANCOVA models etc.

Research Methodology

Development of missing data estimation techniques with statistics methods is the experimental research with the following steps:

1st Step Determine characteristics of data distribution, sample size, missing data proportion, and repetition internal by identifying the characteristics of data distribution for variety of attributes of data and real data whose distribution is still unknown as they are from at least 1,000 sets of questionnaire. Then, determine sample size of each distribution; n = 10, 30, 50, 100 and 400 and missing data proportion for every distribution at 5%, 10%, 20% and 50%. Complete data are used as based data for comparison. Repetition interval is determined at r = 10,000.

2nd Step Simulate all to-be-studied sample values by Monte Carlo method from Excel program according to the determined distribution. In each distribution, all 4,000,000 values are simulated with 10,000 columns and 400 rows.

3rd Step Sample data examples according to specified sizes, scopes and sets of missing values

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

4th Step Calculate estimates by using statistics estimation methods to substitute the missing data with the specified repetition interval. Then, consider accuracy of the estimation in each case by using MRE and MMRE and compare cases for the best estimation method.

Statistics missing data estimation used in this research consists of

1. Mean Imputation (MI) is carried out by using mean. It was firstly proposed by Wilks in 1932. This method estimates dependent variables by using mean of existing data of dependent variables as follows:

$$\hat{x}_m = \bar{x}^* = \frac{\sum_{i=1}^{n^*} x_i}{n^*} \quad ; i= 1 , 2 , \dots , n^*$$

When $\hat{x}_m$ = missing data estimator by using mean of remaining data

$\bar{x}^*$ = estimates of missing data by using mean of remaining data

n* = the number of remaining data

$\hat{x}_m$ is used as a representative of missing data

2. K-Nearest Neighbor Imputation (KNN)

(1) Determine groups of *n* data and consider the group data. Find distance of the most approximate values by determining $\delta(x_i, x_j)$ is distance of data and calculating $\delta(x_i, x_j)$ with the given equation.

$$\delta(x_i, x_j) = \sqrt{\sum (x_i - x_j)^2}$$

1) Determine the most repeated min $\delta(x_i, x_j)$
2) Calculate estimates for substituting the missing data by using

$\hat{x}_{knn}$ = $\frac{x_i + x_j}{2}$; $\hat{x}_{knn}$ คือ missing data estimator by using K-Nearest Neighbor Imputation

3. Extreme Imputation (EI) $\hat{x}_{ei} = \frac{Max + Min}{2}$

when $\hat{x}_{ei}$ = missing data estimator by using extreme imputation

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

Max = maximum value from remaining data

Min = minimum value from remaining data

4. Side Imputation (SI) $\hat{x}_{si} = \frac{Over + Upper}{2}$

when $\hat{x}_{si}$ = missing data estimator by using side imputation

Over = data value found before missing data location

Upper = data value found after missing data location

5. Side Mean Imputation (SMI) is the combination of Side Imputation (SI) and Regime Switching Imputation (RSM). This can be used in case of both dependent and independent data by dividing data considering intervals until each missing data is found. Data found between each missing data are used for determining data representatives. In this research, the representatives will be determined from data mean values and therefore periodically provided. Mean of side representative values are calculated to be used as the representative of missing data again.

Steps of Side Mean Imputation are as follows:

1) Remaining data x_1 , x_2 , … , x_{n-m} (Given m = the number of missing data) is considered by groups which are divided according to the number of all found missing data locations + 1. Therefore, total data group will be m+1 groups which are $\bar{x}_{s_1}, \bar{x}_{s_2}, \bar{x}_{s_2}, \ldots, \bar{x}_{s_s}, \bar{x}_{s_{m+1}}$

2) Data from side sets, like $\bar{x}_{s_1}, \bar{x}_{s_2}$ are calculated for mean values again , for example $\hat{x}_{SMI_{s1}} = \frac{\bar{x}_{s_1} + \bar{x}_{s_2}}{2}$

when $\hat{x}_{SMI_{s1}}$ = missing data estimator by using side mean imputation

$\bar{x}_{s_1}$ = mean of data groups found before missing data location

$\bar{x}_{s_2}$ = mean of data groups found after missing data location

3) Missing data estimators by using side mean imputation are obtained $\hat{x}_{SMI_{s1}}, \hat{x}_{SMI_{sm2}}, \hat{x}_{SMI_{s3}}, \ldots, \hat{x}_{SMI_{ss}}$ to substitute *m* missing data.

6. New Multiple Imputation (Multi-Im) is the combination of estimation methods applying Mean Imputation (MI), theory concept of Maximum Likelihood, sampling theory, and possibility theory with the following steps

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

1) Consider sets of missing data and determine data sampling with ratios of sets of remaining data

(p). Sample the specified remaining data (p) proportionally without focusing on the order. Proportioned data of the remaining data (p) are then individually calculated for mean and repetition intervals.

2) When sampling according to the number of proportions of the remaining data (p) and repeating

until the number of sample space in experimental set is reached (remaining data group), resample data for ${}^{n-m}C_{p(n-m)}$ intervals without having repeated data. ${}^{n-m}C_{p(n-m)}$ mean values are given.

when $${}^{n-m}C_{p(n-m)} = \frac{(n-m)!}{((n-m)-p(n-m))!(p(n-m))!}$$

when n = the number of all data

m = the number of missing data

p = data sampling ratio of remaining from missing data

3) After that, ${}^{-m}C_{p(n-m)}$ mean values are recalculated for mean again. The new mean will be used

for substituting missing data.

5th Step Collect real data from 1,000 sets of questionnaire and sets of missing data in order to apply the best statistics missing data estimation method with the real data.

6th Step Find estimates for substituting missing data by using 6 techniques of statistics missing data estimation and apply them.

7th Step Calculate mean, variance, Mean-MRE (MMRE) and MSE from complete and substituting data (estimates)

The lowest MSE ($\overline{X}$) sampling method should be considered because it implies that such estimation provides the most approximate mean compared to the mean of complete data.

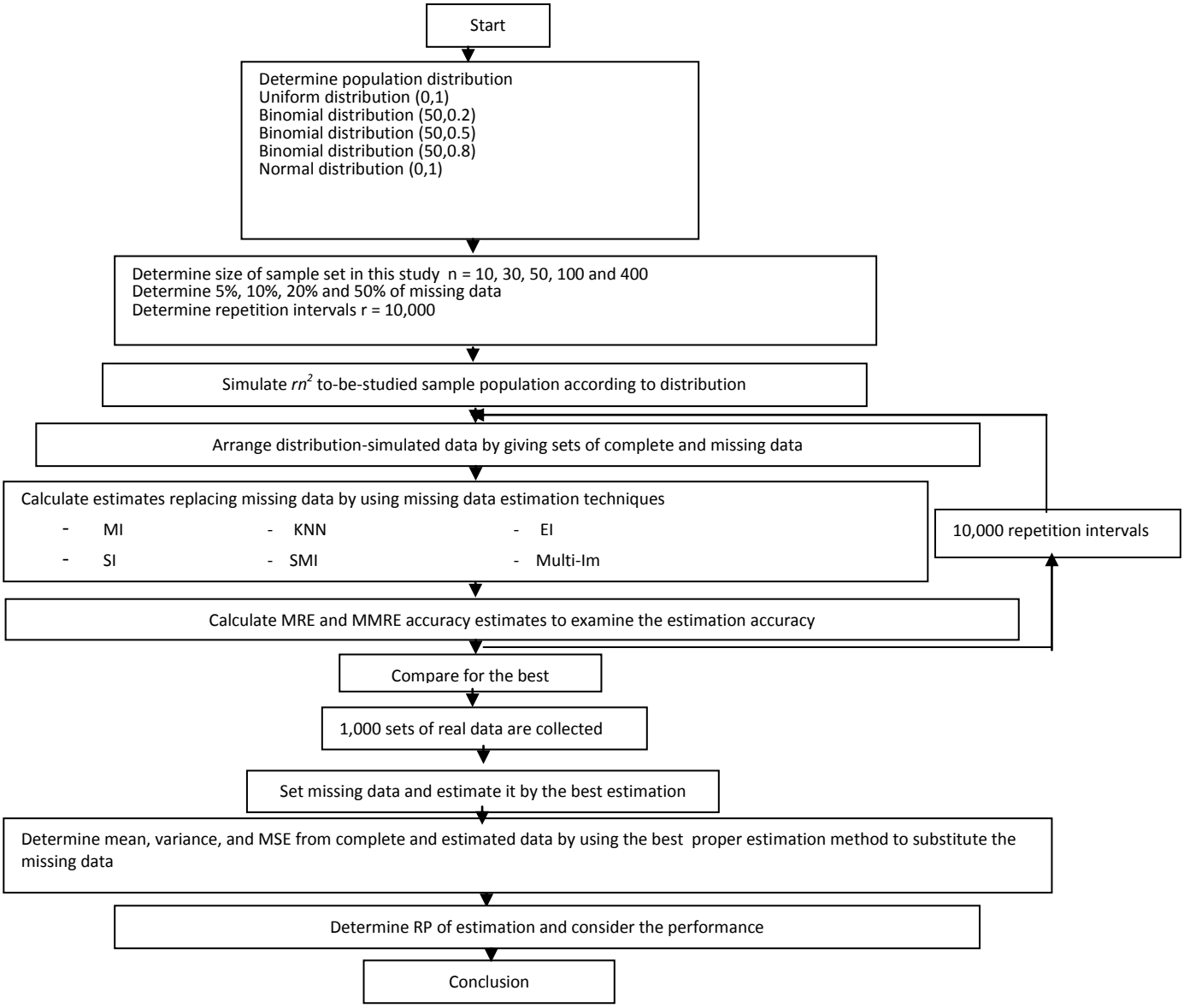
8th Step Calculate Relative PRecision (RP) of missing data estimation and consider comparing the performance of each studied missing data estimation method.

Comparison of the performance is carried out by considering Relative Precision (RP) of the estimation as follows: $$RP(\hat{\theta}_1, \hat{\theta}_2) = \frac{MSE(\hat{\theta}_1)}{MSE(\hat{\theta}_2)}$$

when $\hat{\theta}_1, \hat{\theta}_2$ = estimators of 1st and 2nd estimation methods, respectively

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

Consideration of sampling methods, having RP values more than 1, shows that the 2nd sampling method has better performance than the first one. Meanwhile, if the RP is less than 1, it shows that the 1st sampling method is better than the second one. The sampling method having the highest RP shows that the sampling method has the most correct performance in population mean estimation, compared to the first one.



Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

This above diagram shows methodology used in this research.

Research Results

Development of missing data estimation techniques with statistics methods aim to study solution process of missing data and develop new data analysis techniques and statistics estimation methods in order to solve missing data problems. Research results are as follows:

Table 1 shows the conclusion of proper missing data estimation techniques for various cases

Missing proportion	Sample Size	Uniform (0,1)	Binomial (50,0.2)	Binomial (50,0.5)	Binomial (50,0.8)	Normal (0,1)	Real Data
5%	n=10	EI	EI	EI	EI	EI	EI
	n=30	EI	Multi-Im	Multi-Im	Multi-Im	SMI	Multi-Im
	n=50	EI	Multi-Im	Multi-Im	Multi-Im	SMI	Multi-Im
	n=100	EI	Multi-Im	Multi-Im	Multi-Im	MI	Multi-Im
	n=400	EI	Multi-Im	Multi-Im	Multi-Im	SMI	KNN
10%	n=10	EI	EI	EI	EI	EI	EI
	n=30	EI	Multi-Im	Multi-Im	Multi-Im	SMI	Multi-Im
	n=50	EI	Multi-Im	Multi-Im	Multi-Im	SMI	Multi-Im
	n=100	EI	Multi-Im	Multi-Im	Multi-Im	SMI	Multi-Im
	n=400	EI	Multi-Im	Multi-Im	Multi-Im	SMI	KNN
20%	n=10	EI	EI	EI	EI	EI	EI
	n=30	EI	Multi-Im	EI	Multi-Im	SMI	Multi-Im
	n=50	EI	SI	Multi-Im	Multi-Im	MI	Multi-Im
	n=100	EI	SI	Multi-Im	Multi-Im	Multi-Im	Multi-Im
	n=400	EI	SI	SI	Multi-Im	SMI	SI
50%	n=10	EI	EI	EI	EI	SI	EI
	n=30	EI	SI	EI	Multi-Im	SI	SI
	n=50	EI	SI	Multi-Im	Multi-Im	SI	SI
	n=100	EI	SI	Multi-Im	Multi-Im	SI	SI
	n=400	EI	SI	SI	SI	SI	SI

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

Using Mean Imputation (MI) shows that Mean-MRE (MMRE) and MSE are low in case of normal-distribution data.
Using K-Nearest Neighbor Imputation (KNN) shows that Mean-MRE (MMRE) and MSE are low in case of normal-distribution data and large size of real data.

Extreme Imputation (EI) shows that Mean-MRE (MMRE) and MSE are low in case of data for discrete random variable. MSE and MMRE are low especially uniform and binomial distribution when sample size is small.
Side Imputation (SI) shows that Mean-MRE (MMRE) and MSE are low in case of binomial-distribution data, high ratio of missing data, and large data size.
Side Mean Imputation (SMI) shows that Mean-MRE (MMRE) and MSE are low in case of normal-distribution data, large size of real data.
New Multiple Imputation (Multi-Im) shows that Mean-MRE (MMRE) and MSE are low in case of binomial-distribution data, more than 10 sample sizes.

Research Discussion

The study of development of missing data estimation techniques with statistics methods can be discussed according to 6 types of studied statistics missing data estimation techniques as described below:

1. Mean Imputation (MI) is not remarkable when being compared with other types of estimation, however, it can be used with uniform distribution (0,1) and normal distribution (0,1) as it provides low MSE and is used in real data situation; 5% loss in 50 sample sizes and 20% loss in 100 sample sizes. This result implies that MI is good for larger sample sizes and in accordance with the study of Schmidt et.al. (2015) who found that MI can be fairly used if sample size is small. Supported by the research of Nilpattarachat P. (2016) who found that MI will be efficient when regression coefficient of missing independent variables is high and the sample size is large. Moreover, this is corresponding to the research of Sudsila P, Thongtheerapap A., and Chomtee B. (2017) who found that MI is the most efficient when sample sizes are 60, 100, or 300 (large sample sizes).

2. K-Nearest Neighbor Imputation (KNN) is not remarkable when being compared with other types of estimation, however, it provides low MSE when being used with uniform distribution (0,1), and normal distribution (0,1). Real data situations for binomial distribution b (50, 0.5), b (50, 0.8) will provide high MSE when data size is small (n = 10). This is in accordance with the study of Kaewrattanapat N., Kasemtheekarun P., and Manochayakorn C. (2012) who found that missing data are well substituted by continuous data using KNN and of Jonsson and Wohlin (2004) who found that KNN is suitable for Likert data which is continuous type.

3. Extreme Imputation (EI) is an estimation method suitable for uniform distribution (0,1) more than other methods used in this research. In addition, it is also found that EI is not the best method in case the data is binomial distribution b (50,

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

0.2), b (50, 0.5), b (50, 0.8) and normal distribution (0,1) including real data situation and provides low MSE. It is often the lowest when sample size is small. Therefore, EI can be used with the mentioned distribution.

4. Side Imputation (SI) is an estimation method that can be used with uniform distribution (0,1), binomial distribution (50, 0.2), normal distribution (0,1) and real data situation. The mentioned distribution is not uniform distribution (0, 1), the lowest MSE will be gained in 50% of missing data when data size is greater than or equal to 30 (≥30). For binomial distribution b(50,0.5),b(50,0.8), it is found that high error is provided when *n* is small. This method is therefore not suitable. However, for every types of distribution in this research, SI will provide low MSE when there is 50% of missing data. If *n* is large, MSE will be low and vice versa.

5. Side Mean Imputation (SMI) is a suitable estimation method for normal distribution (0,1) more than other methods used in this research and provides low MSE when being used with data having uniform distribution (0, 1) In real data situation with 5%, 10%, and 20% of missing data, this method can be used as well. For binomial distribution b(50,0.2), b(50,0.5),b(50,0.8), it is found that SMI provides higher values than other methods (approximately similar to MI) and very high MSE when missing data is 50%.

6. New Multiple Imputation (Multi-Im) is the most remarkable estimation method among the others in this research as it provides low MSE in every case of tests especially when sample sizes are more than 30. It is more suitable than any other methods for binomial distribution b (50, 0.2), b (50, 0.5), b (50, 0.8) and real data situations as there are cases of minimum and maximum MSE compared to the others.

This is corresponding to the research of Sudsila P, Thongtheerapap A., and Chomtee B. (2017) who found that MI is the most efficient when sample sizes is 500 and the study of Schlomer, Bauman, and Card (2010) who found that MI provides good results in missing data estimation by giving allowable errors. However, it has conflicts against the study of Nilpattarachat P. (2016) who found that MI will be efficient when regression coefficient of missing independent variables is low and the sample size is small.

Conclusion

Development of missing data estimation techniques with statistics methods can be concluded according to research objectives as follows:

1. To study solution process for missing data

Missing data effect on researches, both in analysis and interpretive conclusion. Severity level of such effects depends on several components such as size and type of missing data, importance of variables that have missing data in the research, and missing data management techniques. Therefore, researchers try to find the most

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

efficient solutions. There are several of them these days such as further data collection, Listwise deletion, or substitution. Each method is differently suitable depending on research contexts.

Most researchers like to use statistics process for substitute calculation. Therefore, there are various estimation techniques statistically, such as Mean Imputation (MI), K-Nearest Neighbor Imputation (KNN), and relationships between variables etc.

2. To develop new data analysis techniques and statistics estimation methods by using data for missing data

Concept of mean application basis is developed in this study, for example, mean and median values are used for developing data analysis techniques and new statistics estimation methods such as Extreme Imputation (EI), Side Impultation (SI), Side Mean Imputation (SMI), and New Multiple Imputation (Multi-Im).

Research Suggestions

Suggestions for development of missing data estimation techniques with statistics methods are as follows:

1. Data with other types of distribution or further parameters should be studied.
2. Concepts of missing data estimation should be applied for writing estimation programs for greater convenience.
3. Concepts from other theories should be applied for further studies.

Reference

Kaewrattanapat N., Kasemtheekarun P., and Manochayakorn C. (2012). *Comparison of Data Mining Techniques Efficiency for Missing Data Substitution.* Meeting Documents for 8th National Academic Computer and Information Technology, Information Management, Suan Sunandha Rajabhat University, Bangkok.

Piyaporn and Sukon Prasitwattanaseri. (2006). Missing Data and Its Management. *Data Management & Biostatistics Journal, 4*(3), 52-61.

Sudsila P, Thongtheerapap A., and Chomtee B. (2017, June). *Comparison of Missing Data Estimation Methods of Dependent Variables in 2-Group Logistics Regression Analysis.* Presentation Documents for 2nd UTTC Academic Meeting, Academic Year 2017, Bangkok.

Nilpattarachat P. (2016). *Comparison of the Estimation Methods for Nonignorable Missing Data in Logistics Regression Analysis.* Master of Science Thesis, Chulalongkorn Universtiy.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/FC4JZ

Sukchareon W., (2015). Missing Data Operations. Rom Pruek Journal, *Krirk University, 33*(2), 12-32.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 2*(19), 153-176.

Jonsson, P., & Wohlin, C. (2004). *An evaluation of k-nearest neighbour imputation using likert data*. Paper presented at the 10th IEEE International Software Metrics Symposium, Los Alamitos, CA.

Raymond, M. R. (1986). Missing Data in Evaluation Research. *Evaluation & the Health Professions, 9*(4), 395-420.

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

Schmidt, C., Fizet, J., Properzi, F., Batchelor, M., Sandberg, M., Edgeworth, J. A. et al. (2015). A systematic investigation of production of synthetic prions from recombinant prion protein. *Open Biology, 5*(12), 1-7.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24*(1), 53-62.

Uchenna, P. O., & Nduka, E. C. (2012). Methods of analysing missing values in a regression model. *Indian Journal of Science and Technology, 5*(2), 2013-2016.