Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 14(1) 2021: 68-78
http://ejournal.undip.ac.id/index.php/media_statistika

MULTIPLE IMPUTATION FOR ORDINARY COUNT DATA BY NORMAL DISTRIBUTION APPROXIMATION

Titin Siswantining, Muhammad Ihsan, Saskya Mary Soemartojo, Devvi Sarwinda, Herley Shaori Al-Ash, Ika Marta Sari
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia

e-mail: titin@sci.ui.ac.id

DOI: 10.14710/medstat.14.1.68-78

Article Info:
Received: 17 April 2021
Accepted: 24 June 2021
Available Online: 30 June 2021

Keywords:
count data; generelized linear model; missing value; multiple imputation; poisson regression; rubin's rule.

Abstract: Missing values are a problem that is often encountered in various fields and must be addressed to obtain good statistical inference such as parameter estimation. Missing values can be found in any type of data, included count data that has Poisson distributed. One solution to overcome that problem is applying multiple imputation techniques. The multiple imputation technique for the case of count data consists of three main stages, namely the imputation, the analysis, and pooling parameter. The use of the normal distribution refers to the sampling distribution using the central limit theorem for discrete distributions. This study is also equipped with numerical simulations which aim to compare accuracy based on the resulting bias value. Based on the study, the solutions proposed to overcome the missing values in the count data yield satisfactory results. This is indicated by the size of the bias parameter estimate is small. But the bias value tends to increase with increasing percentage of observation of missing values and when the parameter values are small.

1. INTRODUCTION

Data can be carried out in various ways such as field surveys, interviews, experimental activities in the laboratory and so on. But in its implementation, especially in survey activities, it is often to be found the non-response items or items that are not answered by respondents and cause missing values. For example, a respondent refused when asked about age, number of children, marital status, etc. because they were considered private so that the required data was not obtained from the respondent. Missing values can be define as a failure to obtain observational values from several sample units which then cause problems in data analysis (Wiegand, 1968). Missing values or non-response items in data collection activities will then form a missing data.

In general, missing data or data containing missing values can cause two problems, namely loss of efficiency or information and cause bias (O'Kelly, 2014). This problem arises because the sample size used is reduced than it should. This action in overcoming missing values is the initial stage in the process of data analysis or commonly known as pre-processing data. At this stage, data must be prepared so that it is feasible to enter the analysis phase.

One easy way to handle this case of missing data is to delete observations related to the missing value. This certainly has an impact on the sample size that will shrink. The deletion of cases is the worst method and is not recommended in practical applications (Enders, 2017; Wilkinson, 1999). Another way to overcome the missing data is to fill the missing value with the estimated replacement values based on the values that were successfully observed. The method is then known as imputation. Imputation is a work process used to determine and assign substitute values for missing values (Rubin, 1987). The imputation method is important if the percentage of observations with missing values is relatively large.

The Imputation method has also developed a lot, for example (Akmam et al., 2019) discusses about multiple imputation with predictive mean matching method for numerical missing data. (Christopher et al., 2019) also explain about imputation for numerical data using Fractional Hot Deck. (Aristiawati et al., 2019) introduced the missing values imputation based on fuzzy CMeans algorithm. In addition, there is imputation for multivariate missing data using Sequential Regression Multivariate. For the mixed datasets, (Anwar et al., 2019) introduced the method of imputation with K-Harmonics mean algorithm.

In this study, we will focus on multiple imputation as our technique to overcome missing value. Because multiple imputation is a technique for overcoming missing values by replacing each missing value with more than one substitute value, called m values ($m > 1$) that resulting m complete data. Each of the complete data will be analyzed using the appropriate analysis method (based on cases) to obtain parameter estimations and the results will be combined based on Rubin's rules (Rubin, 1987). Multiple imputation has advantages over single imputation, which are; increases estimation efficiency, better accuracy of parameter estimation and provides more valid statistical inference (van Buuren, 2018).

Missing value on ordinary count data is considered to be the main focus in this study. By Central Limit Theorem, the problem of missing value on ordinary count data that assume Poisson distributed will be solved with approximation of normal distribution as the assumption of large sample size was fulfilled. Generalized Linear Model Poisson Regression is selected to be analytical method because we will assume that ordinary count data as response variable has relationship with a predictor variable. In the end of this study, we will show a result of numerical simulation with few parameters (λ) and large sample size (n).

2. MATERIAL AND METHODS

2.1. Count Data

Before we go further about the imputation task, here will be described about count data because this study is focused on missing value for count data. In statistics, count data is a statistical data type, a type of data in which the observation can take only the non-negative integer values {0, 1, 2, 3, ...}, and where these integers arise from counting rather than ranking (Barbur et al., 1994). Examples of count data include the number of children in a family, the number of crimes committed or the number of someone recovering from the illness (van Buuren, 2018).

The simplest probabilistic model for count data is the Poisson distribution that introduced by Siemon D. Poisson in 1837. A random variable is Poisson distributed with parameter $\lambda > 0$ is notated by $Y \sim \text{Poisson}(\lambda)$ and this random variable has pmf (probability mass function) as

$$f(y) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!}; y = 0,1,2,\ldots \\ 0 \quad ; y \text{ elsewhere} \end{cases} \tag{1}$$

Poisson distribution has equidispersion property about its parameter. This property can be proved by moment generating function of Poisson distribution using Equation 1 and this property can be witten as

$$E(Y) = Var(Y) = \lambda \tag{2}$$

Equation 2 shows that the expectation and the variance value of Poisson Distribution is equal (Hogg Allen T Craig, 1978). Next will be shown the pmf (probability mass function) plot of Poisson distribution for few parameters.

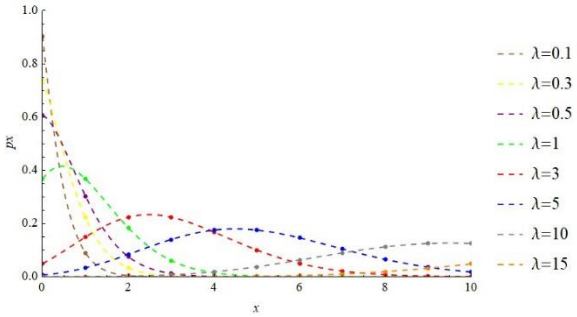

Figure 1. Pmf Curve for Poisson Distribution in Few Parameters

Based on Figure 1 we can see that the greater value of the Poisson distribution parameter, the pmf curve of the Poisson distribution is increasingly symmetrical which indicates the fulfillment of normal properties.

Sometimes if we have one response variable that has Poisson distribution and the other predictor variable that assumed fixed, and we wish to know the relation about the variable we can construct the Poisson regression that modeled as follows

$$y_i = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) + \varepsilon_i \tag{3}$$

The Poisson regression model in Equation 3 can be used to know the relationship about predictor variable/s to a response variable that Poisson ditributed. The method to get the parameter estimation on Poisson regression model can be found by maximum likelihood method and iteratively by Newton-Raphson method (Barbur et al., 1994).

2.2. Multiple Imputation

Classically, the technique to overcome missing value is delete the observation of the missing value itself or namely case deletion. It's very easy to do but it will decrese the size of sample and larger bias of parameter estimation will be produced. The another way to dealing missing value is imputation. Single imputation is filling in each missing value by one imputation value only.

Another technique as a solution to overcome missing values is multiple imputation. Multiple Imputation was first introduced by Rubin (1987). By definition, multiple imputation is similar to single imputation, but substitute values that produce more than one value. So that in multiple imputations some complete data will be formed from the estimation results through this technique. Then each complete data is analyzed using standard analysis which will then produce a combined parameter for the final estimate. Multiple Imputation fills in missing values by assuming that the missing value also comes from the Posterior

distribution which is the same as the observed value so that the obtained results are valid (Rubin, 1987). The Figure 2 ilustrates the workflow of multiple imputation. There are 3 stage multiple imputation, namely imputation stage, analysis stage, and pooling stage.

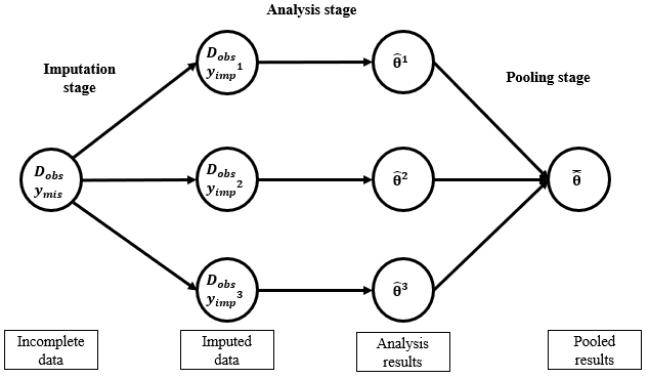

Figure 2. Scheme of main steps in multiple imputation with m=3 (van Buuren, 2018).

2.2.1 Imputation Stage

Normal linear model can be used to approximation model in imputation stage. It means for every modeled data the properties of normality should be fulfilled (we use Kolmogorov-Smirnov test to check the normality of ordinary count data for large sample). Consider the following model.

$$\boldsymbol{y} \sim \boldsymbol{N}(\boldsymbol{X\beta}, \boldsymbol{\sigma^2 I}) \tag{4}$$

Equation (4) is a basic model that considered in this study. Based on Central Limit Theorem, data that has Poisson ditributed can be approximates with normal distribution as the sample size is large. The final goal at this stage is to estimate the missing value based on bayesian inference which is assumed that missing value y_{mis} comes from the posterior distribution which is equal to the value that was successfully observed y_{obs}. It will be denoted as

$$\boldsymbol{y_{mis}} \sim \boldsymbol{N}(\boldsymbol{X_{mis}\beta_*}, \boldsymbol{\sigma_*^2 I}) \tag{5}$$

Based on Equation (5), y_{mis} is vector of estimated missing values and X_{mis} denotes the covariate matrix that pairs with the response y that considered as missing value. By Bayesian method, $\boldsymbol{\beta_*\sigma_*^2 I}$ can be estimated so the missing value of response variable can be calculated.

Rubin (1987) and Buuren (2018) formulates the imputation task for n sample size and q parameters as follows.

a. Draw a random variable that has $\chi^2_{(n-q)}$ denote by g then Then calculate the variance for estimating missing values using the following equation

$$\sigma_*^2 = \hat{\sigma}^2(n-q)/g \tag{6}$$

b. Then calculate $\boldsymbol{\beta_*}$ by adding up the estimation value of $\boldsymbol{\beta_*}$ and an uncertainty value vector $q \times 1$ which is drawn from standar normal distribution denotes **z**. Mathematically formulated as follow

$$\beta_* = E(\beta_*) + uncertainty\ value \tag{7}$$

$$= \widehat{\boldsymbol{\beta}} + \boldsymbol{\sigma}_*[\boldsymbol{V}]^{\mathbf{1/2}}\boldsymbol{z}$$

c. Finally the missing value can be predicted by

$$\boldsymbol{y}_{\boldsymbol{miss}*} = \boldsymbol{X}_{\boldsymbol{miss},}\boldsymbol{\beta}_* + \boldsymbol{z}\boldsymbol{\sigma}_* \tag{8}$$

Where z denotes a random vector that drawn from standar normal distribution by simulation. If we want to do as many as m imputation simulations, then the steps above are repeated m times.

2.2.2 Analysis Stage

After the complete m-dataset of the results of the multiple imputation stages in the previous section is obtained, the next step is to analyze each complete data from the results of the imputation with an appropriate analysis method. In this study, the analytical method that to be chosen is the Generalized Linear Model Poisson Regression analysis. Then the steps in this analysis phase are as follows;

a. Check that Y ~ Poisson (λ) for each complete dataset as we require for the analysis using Generalized Linear Model Poisson Regression by Kolmogorov-Smirnov test
b. Bulid Generalized Linear Model Poisson Regression for each complete dataset
c. Estimating parameters on Generalized Linear Model Poisson Regression for each complete dataset

All these stages above are flexible or can be adjusted to the analytical method chosen by the researcher (depending on the case that is the focus of the study) (Ibrahim et al., 2005). The end of this stage will produce parameter estimates from the analysis method selected for each complete data estimate. Estimates of the parameters in this simulation will be combined based on Rubin's rules which will be discussed later.

2.2.3 Pooling Stage

After parameter estimation is obtained for each complete data from imputation and analysis stages, the next step is pooling or combining parameters based on Rubin's rules (Rubin, 1987). This step aims to obtain a final parameter from the m parameters from m complete datasets. Following are the equations in Rubin's rule.

For example if there are M complete dataset of imputation results or for $m = 1, 2, \dots, M$ and the estimated k-regression coefficient ($k = 0, 1, 2, \dots, p$) on m-imputed data is denoted by $\hat{\beta}_k^m$. he pooling parameter for k-regression coefficient is obtained by calculating the average of $\hat{\beta}_k$ values from m-imputed analysis results. Mathematically, the pooling parameter can be written as

$$\begin{aligned}\bar{\hat{\beta}}_k &= \frac{1}{M}\left(\hat{\beta}_k^1 + \hat{\beta}_k^2 + \cdots + \hat{\beta}_k^M\right) \\ &= \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_k^m\end{aligned} \tag{9}$$

If the analysis just limited by one predictor variable, so we can obtained the $\bar{\hat{\beta}}_0$ and $\bar{\hat{\beta}}_1$. While the total variance associated with $\bar{\hat{\beta}}_k$ is formulated by

$$V_{\bar{\hat{\beta}}_k} = \bar{W}_{\hat{\beta}_k} + \left(1 + \frac{1}{M}\right)\bar{B}_{\hat{\beta}_k} \tag{10}$$

$\bar{W}_{\hat{\beta}_k}$ denoted within-imputation variance can be written as $\bar{W}_{\hat{\beta}_k} = \frac{1}{M}\sum_{m=1}^{M} Var(\hat{\beta}_k^m)$ and $B_{\hat{\beta}_k}$ denoted between-imputation variance can be written as $\bar{B}_{\hat{\beta}_k} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_k^m - \bar{\hat{\beta}}_k\right)^2$. In this study, we only focus on pooling parameter for regression coefficient on Equation (9).

2.3 Evaluation Measurement of Multiple Imputation

Evaluation measurement that can be used to measure the efficiency of multiple imputation technique is "bias" (Falcaro & Carpenter, 2017). Bias value can be defined as the difference between the average parameter estimate and the true value which is assumed to be represented by the estimated parameters obtained when the model is full. In the simulation that will be carried out in the next chapter, the bias value is defined as follows

$$\begin{aligned} bias &= \bar{\hat{\beta}}_k - \beta_k \\ &= \bar{\hat{\beta}}_k - \hat{\beta}_k \ (k = 0,1,2\ldots) \end{aligned} \tag{11}$$

$\bar{\hat{\beta}}_k$ denoted the average parameter for k-regression coefficient that already defined in Equation (9) and $\bar{\hat{\beta}}_k$denoted the parameter estimation of k-regression coeffiicient while data is fully observed (0% missing value) (Gupta & Grover, 2017). This will be clarified in numerical simulations which will be discussed next.

2.4 Missing at Random

For each method of analysis for missing data, statistical assumptions about how the missing value could occur must be made. A multiple imputation technique is suitable to be applied in the case of missing values that occur with the Missing at Random mechanism. The mechanism of Missing at Random or which can be abbreviated as MAR assumes that the probability or likelihood of a value in the observation data is missing depends on one or more other variables that have been successfully observed (Rubin, 1987). This assumption will be used on numerical simulation later.

3. EXPERIMENTS AND RESULTS

In this simulation, 200 pairs of complete observational data are generated consisting of one response variable Y and one predictor variable X. The response variable generated is a type of discrete or count variable that follows the Poisson distribution. While the predictor variables raised are continuous variable types. The parameter values of the response variables used are 5,10 and 15 to approximate the normality properties.

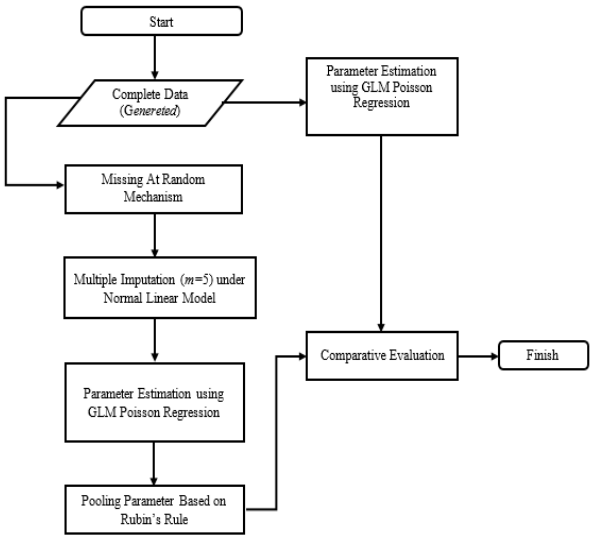

Figure 3. Flow Chart of Numerical Simulation

In this simulation, the Missing At Random (MAR) mechanism is carried out in three different scenarios based on the percentage of the number of missing values that is 10%, 20% and 30% for the purpose of comparison and evaluation of efficiency (Gupta & Grover, 2017).

First of all, we would check whether the dataset which we're generated were also Normal distributed. For it purpose, we construct Kolmogorov-Smirnov test (Conover, 2008).

H_0: Data come from the population which has distribution function $F(y)$ (Normal Distribution)

H_1 : Data did not come from the population which has distribution function $F(y)$ (Normal Distribution)

With significance level $\alpha = 0{,}05$ and test statistic

$$T = \sup_y |F_n(y) - F(y)|$$

Where $F(y)$is distribution function of normal distribution (theoretical distribution) and $F_n(y)$ is distribution function of data (empirical distribution). If p-value less than $\alpha = 0{,}05$ we reject the null hypothesis and conclude that data did not follow the theoretical distribution (normal distribution). This is quite important because our method is depend on normality properties, so the normality of data especially for response variable should be fulfilled.

Table 1 is the result of normality test of the data from few parameters λ.

Table 1. Normality Test of Data

Sample size (n)	Parameter value (λ)	P-value	Conclusion
200	5	0.587	H_0 is not rejected
200	10	0.665	H_0 is not rejected
200	15	0.704	H_0 is not rejected

Based on Table 1 above, H_0 is not rejected for each simulation parameter value in significance level $\alpha = 0{,}05$ and we conclude that data follows normal distribution which is

being our theoretical distribution for each simulation parameter value. Since normality properties were fulfilled for each simulation parameter value, we able to run the imputations under normal linier model.

Table 2. Comparison of Numerical Simulation

λ	MAR	Bias ($\hat{\beta}_0$)	Bias ($\hat{\beta}_1$)
	10%	0.007115	0.000413
5	20%	0.047002	0.00194
	30%	0.067196	0.004578
	10%	0.006396	0.000202
10	20%	0.024658	0.002506
	30%	0.05338	0.005668
	10%	0.000812	0.000228
15	20%	0.001866	0.000182
	30%	0.037408	0.002702

Table 2 shows bias value of $\hat{\beta}_0$ and $\hat{\beta}_1$ for each simulation MAR mechanism of each parameter λ. Based on that table we can conclude that the increasing value of the parameter λ, the parameter bias value for $\hat{\beta}_0$ and $\hat{\beta}_1$ tend to decrease for each percentage of MAR mechanism.

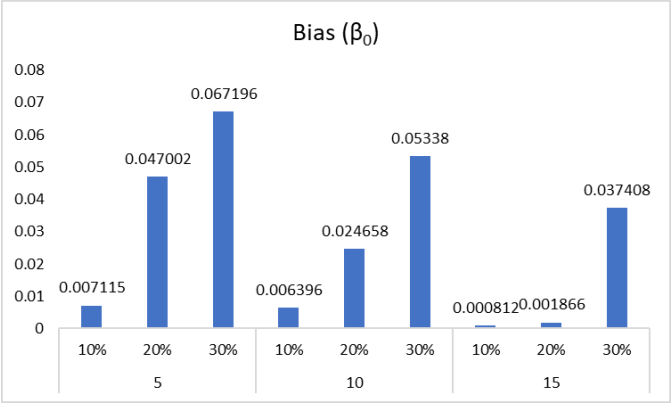

Figure 4. Graph for bias value of $\hat{\beta}_0$

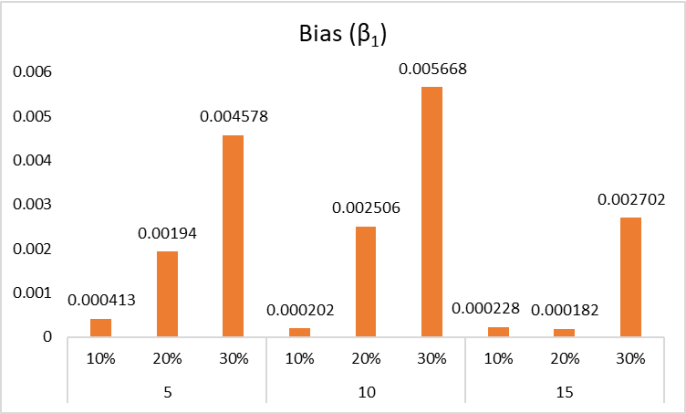

Figure 5. Graph for bias value of $\hat{\beta}_1$

Figure 4 and Figure 5 show the bias resulting from both estimated regression coefficients β_0 and β_1. It can be seen that for $\lambda = 5, 10, 15$ shows a larger bias in the random missing proportion of $10\% - 30\%$, so it can be shown that this method is suitable for small random missing proportions to obtain a minimum bias for both regression coefficients.

4. CONCLUSION

Missing value can be defined as a failure to get the information in the sample unit of research. Missing values must be overcome to get better statistical inference. One technique commonly used to overcome missing values is the multiple imputation. Missing values can occur in all types of data, including count data. One simplest probability distribution that includes count data is Poisson distribution with the assumption that the equidispersion is fulfilled. If there is a predictor variable that is thought to affect the response variable that has a missing value, then a regression analysis can be performed to estimate the missing values. Because the response variable discussed in this study is Poisson distribution, in overcoming this the Multiple Imputation technique with Generalized Linear Model Poisson Regression method is suitable to be applied.

Based on numerical simulations that have been done, the multiple imputation technique with the GLM Poisson regression analysis method as a solution to handle missing values in the count data gives satisfactory results especially if the percentage of missing values is relatively small and the value of parameters is large. This is indicated by the small value of bias of the estimated Poisson regression coefficient. However, for observations of relatively large missing values and small parameter values, the multiple imputation technique with the Poisson regression is not recommended to be implemented because it will produce larger parameter bias values of the estimated Poisson regression coefficient.

As this study also gives some suggestions for research or subsequent writing with topics that are relevant to what has been discussed in this study. These suggestions are; Using other methods such as Fully Bayesian to get the imputation equation in the general count data case, learn more about the imputation case for missing values in other types of count data such as the overdispersion model, the excess zero model and so on. Extending the research by examining cases of multivariate regression or multiple regression, combining sample sizes and the number of imputation simulations and implementing multiple imputation techniques on real data.

ACKNOWLEDGMENT

This research is supported by PITTA B 2019 Research Grant from Universitas Indonesia (ID number NKB-0676/UN2.R3.1/HKP.05.00/2019).

REFERENCES

Akmam, E. F., Siswantining, T., Soemartojo, S. M., & Sarwinda, D. (2019). Multiple Imputation with Predictive Mean Matching Method for Numerical Missing Data. *ICICOS 2019 - 3rd International Conference on Informatics and Computational Sciences: Accelerating Informatics and Computational Research for Smarter Society in The Era of Industry 4.0, Proceedings, February 2020*. https://doi.org/10.1109/ICICoS48119.2019.8982510

Anwar, T., Siswantining, T., Sarwinda, D., Soemartojo, S. M., & Bustamam, A. (2019). A

Study on Missing Values Imputation using K-Harmonic Means Algorithm: Mixed datasets. *AIP Conference Proceedings*, *2202*(December). https://doi.org/10.1063/1.5141651

Aristiawati, K., Siswantining, T., Sarwinda, D., & Soemartojo, S. M. (2019). Missing values Imputation Based on Fuzzy C-Means Algorithm for Classification of Chronic Obstructive Pulmonary Disease (COPD). *AIP Conference Proceedings*, *2192*(December). https://doi.org/10.1063/1.5139149

Barbur, V. A., Montgomery, D. C., & Peck, E. A. (1994). Introduction to Linear Regression Analysis. *The Statistician*, *43*(2), 339. https://doi.org/10.2307/2348362

Christopher, S. Z., Siswantining, T., Sarwinda, D., & Bustaman, A. (2019). Missing Value Analysis of Numerical Data using Fractional Hot Deck Imputation. *ICICOS 2019 - 3rd International Conference on Informatics and Computational Sciences: Accelerating Informatics and Computational Research for Smarter Society in The Era of Industry 4.0, Proceedings*, *February 2020*. https://doi.org/10.1109/ICICoS48119.2019.8982412

Conover, W. J. (2008). *857_1734* (Issue 1999).

Enders, C. K. (2017). Multiple Imputation as a Flexible Tool for Missing Data Handling in Clinical Research. *Behaviour Research and Therapy*, *98*, 4–18. https://doi.org/10.1016/j.brat.2016.11.008

Falcaro, M., & Carpenter, J. R. (2017). Correcting Bias Due to Missing Stage Data in the Non-Parametric Estimation of Stage-Specific Net Survival for Colorectal Cancer Using Multiple Imputation. *Cancer Epidemiology*, *48*, 16–21. https://doi.org/10.1016/j.canep.2017.02.005

Gupta, V. K., & Grover, G. (2017). Multiple Imputation for Gamma Outcome Variable Using Generalized Linear Model. *Journal of Statistical Computation and Simulation*, *87*(10), 1980–1988. https://doi.org/10.1080/00949655.2017.1300904

Hogg Allen T Craig, R. V. (1978). *Introduction to Mathematical Statistics Fourth Edition*.

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. In *Journal of the American Statistical Association* (Vol. 100, Issue 469, pp. 332–346). Taylor & Francis. https://doi.org/10.1198/016214504000001844

O'Kelly, M. (2014). Multiple Imputation and Its Application. James Carpenter and Michael Kenward (2013). Chichester: John Wiley & Sons. 345 pages, ISBN: 9780470740521. *Biometrical Journal*, *56*(2), 352–353. https://doi.org/10.1002/bimj.201300188

Rubin, D. B. (Ed.). (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470316696

van Buuren, S. (2018). Flexible Imputation of Missing Data, Second Edition. In *Flexible Imputation of Missing Data, Second Edition*. https://doi.org/10.1201/9780429492259

Wiegand, H. (1968). Kish, L.: Survey Sampling. John Wiley & Sons, Inc., New York, London 1965, IX + 643 S., 31 Abb., 56 Tab., Preis 83 s. *Biometrische Zeitschrift*, *10*(1), 88–89. https://doi.org/10.1002/bimj.19680100122

Wilkinson, L. (1999). Statistical Methods in Psychology Journals: Guidelines and Explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594