Acta Sci. Pol.
Oeconomia 20 (2) 2021, 95–103
ISSN 1644-0757 eISSN 2450-047X DOI: 10.22630/ASPE.2021.20.2.19
acta_oeconomia.sggw.pl

ORIGINAL PAPER

Received: 10.07.2021
Accepted: 20.08.2021

THE OCCURRENCE OF MISSING DATA IN SURVEYS

Wioleta Zdebska

ABSTRACT

The purpose of this article is to discuss issues related to the analysis of missing data. Why do missing data occur in a data set? What percentage of the collected data constitutes missing data? What is the nature of missing data that emerges during data collection? The above questions are extremely important in assessing conducted surveys or in evaluating the quality of the collected data. A lack of reflection on the aspects mentioned above may lead to false conclusions and recommendations. This article presents not only an overview of the literature regarding missing data, but also shows how in a practical way an analysis of the randomness of missing data can be performed. The analysis presented in the article is based on data collected as part of the Polish General Social Survey carried out in 2008. The main recommendation of the author is to conduct an analysis of the randomness of missing data before analyzing the collected data.

Key words: survey, analysis of missing data, MCAR, MAR, NMAR

INTRODUCTION

It has been extremely popular for many years to conduct marketing research and to give recommendations based on drawn conclusions in order to better manage organisations or to reach potential customers. However, in order to be sure that the conducted marketing research is reliable and accurate, it is also necessary to look at the methodology of the research. Issues that the researcher should consider as a part of the research process include ways of selecting the research sample, data collection techniques, and correct translation of business questions into research questions and hypotheses.

However, in addition to the aspects mentioned above, an extremely important yet often overlooked issue, is analysis of missing data. It is important to do so because the existence of missing data entails potential risks. Due to the occurrence of missing data, reduced sample sizes may lead to biased estimators, lower statistical power of the analyses, problems in estimating confidence intervals, or type I or II errors in statistical inference [Collins et al. 2001, p. 330).

Recognising the gap in the area of missing data analysis in surveys, the author of this article would like to present an analysis of data deficiencies based on data collected as part of the Polish General Social Survey in 2008.

TYPES OF MISSING DATA

Missing data that occur during the collection of data for analysis can be divided into a few categories. The first category is data deficiencies, which in the literature is called unit non-response. Their occurrence results from the inability to conduct a questionnaire interview with a person selected to participate in the study. One reason for this may be that the potential respondent refuses to participate in the survey. The way to deal with this type of situation is the weighting procedure [Durrant 2005, p. 3].

✉wioleta.zdebska@gmail.com

The second reason for the occurrence of missing data may be the omission of potential respondents in the sampling frame. In this case the weighing procedure can be also used to deal with the existing missing data [Brick and Kalton 1996, p. 216].

The third type of data deficiency is referred to as item non-response. In this case, the person drawn to participate in the study agreed to participate but for various reasons decided not to answer individual questions in the survey questionnaire. Ultimately, the analysed data set contains various variables with missing data for individual observations. In this case the answer to this situation can be an analysis of complete observations or the application of different methods of imputing the missing data [Durrant 2005, p. 3, Shaharudin et al. 2020, p. 646).This means using such statistical methods that allow to replace the missing data by some value calculated on the basis of the data set held, or by the actual value of an observation that occurs in the data set [Dunn 1997, p. 161]. Multiple imputation of missing data is a method of imputation commonly used in research [Siddique et al. 2012].

When analysing data deficiencies, two aspects can be distinguished: patterns of data deficiencies and mechanisms for generating data deficiencies. The first of these refers to the configuration of values and missing data observable in the data set. Mechanisms, on the other hand, describe the relationship between variables and the probability of missing data [Enders 2010, p. 3–4].

In Figure 1 shows the types of missing data occurrence. The rectangles denote variables in the data set. Black colour indicates missing data occurrences.

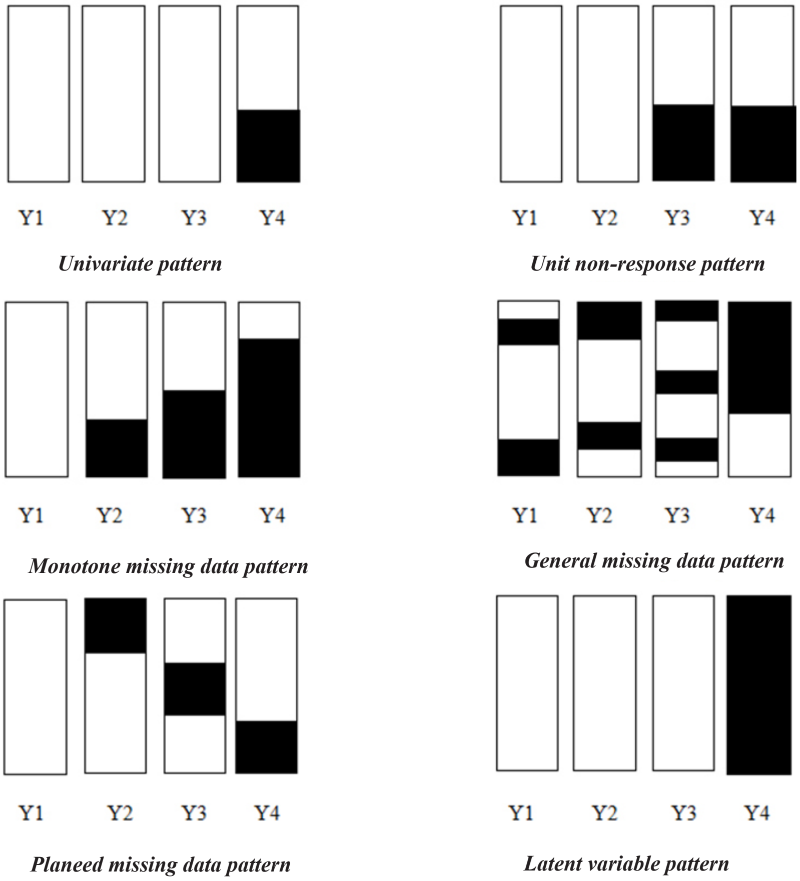

Y1
Y2
Y3
Y4
Univariate pattern
Y1
Y2
Y3
Y4
Unit non-response pattern
Y1
Y2
Y3
Y4
Monotone missing data pattern
Y1
Y2
Y3
Y4
General missing data pattern
Y1
Y2
Y3
Y4
Planeed missing data pattern
Y1
Y2
Y3
Y4
Latent variable pattern

Fig. 1. Patterns of occurrence of missing data

Source: Prepared by [Enders 2010, p. 4].

The first pattern is the univariate pattern, which illustrates a situation where missing data occur in a data set in a single variable. It may appear in the case of experimental studies [Enders 2010, p. 4–5].

In the second pattern, missing data do not appear in variables Y_1 and Y_2, as all respondents answered both questions. However, missing data appear in Y_3 and Y_4. This pattern is called the unit non-response pattern and often occurs in survey research [Enders 2010, p. 4–5].

Another pattern, monotonous missing data, represents a situation where, for some reason, participants stop participating in the study over time. It often occurs in the case of longitudinal studies [Enders 2010, p. 4–5]. In the literature this problem is referred to as panel depletion [Babbie 2004, p. 124].

General missing data is a pattern in which missing data occur in different variables of a data set. At first glance, the missing data generation mechanism appears to be random, although in reality it is not necessarily so [Enders 2010, p. 405].

Another pattern, called by Enders planned missing data pattern, illustrates the case where missing data are planned and controlled by the researcher. The purpose of planned missing data is to reduce the burden on the respondents to participate in the survey, as well as to reduce the cost of conducting the survey [Enders 2010, p. 4–5]. One way to create planned missing data is to use the three-from design method proposed by Graham et al. The study asks four sets of questions: *X*, *A*, *C* and *D*. The three versions of the survey questionnaire are the result of combinations of these groups of questions. However, each version of the questionnaire includes the *X* question set, which contains the most important questions of the research [Graham 2012, p. 282].

The last pattern is the latent variable pattern, in which latent variables are treated as those with missing data. An exemplification of such a variable can be a latent construct in factor analysis, which explains the relationships between observable variables. However, the factor itself has missing data for all observations because it cannot be measured [Enders 2010, p. 4–5].

MECHANISMS FOR GENERATING MISSING DATA

Missing Completely at Random (MCAR)

Rubin distinguishes three types of missing data. The first of these is called Missing Completely at Random. Missing data in this case are generated by a completely random mechanism. MCAR-type missing data occurs when the existence of missing data in an observation for a particular variable is not linked to the value of that observation, nor to the value of any other variable [Allison 2002, p. 3, Padgett et al. 2014, p. 2]. Hence, it can be said that the probability of missing data is completely independent of the data [Newman 2014, p. 376].

If the above condition is met for all variables in the data set under consideration, the data set is a simple random sample of such a data set in which all observations that are missing would have specific observable values [Baraldi and Enders 2010, p. 7].

Thinking about the MCAR missing data generation mechanism, one can imagine generating missing data in a way similar to tossing a coin, where heads would indicate the occurrence of a certain value in a variable and tails would indicate the occurrence of missing data [Graham 2012, p. 13].

Table 1 contains three variables: respondent's ID, job experience and income. The original data were supplemented in such a way as to present the mechanisms of generating missing data: MCAR, MARand NMAR.

If missing data are generated by the MCAR mechanism then the probability of missing data does not depend on the respondent's income.

$P\ (X|D = 2500) = 1/4$

$P\ (X|D = 3000) = 2/8 = 1/4$

$P\ (X|D = 5000) = 1/4$

In this situation, the probability of missing data occurrence does not depend on the length of job experience either.

$P\ (X|S = 0) = 2/8 = 1/4$

$P\ (X|S = 5) = 2/8 = 1/4$

In order to rule out the hypothesis that missing data are generated completely at random, specific

Table 1. Data set with information on job experience and income of the respondent, presenting different mechanisms of generating missing data

Respondent's No.	Job experience (S)	Income (D)	MCAR	MAR	NMAR
1	0	2500	2500	2500	2500
2	0	2500	2500	2500	2500
3	0	2500	×	2500	2500
4	0	2500	2500	2500	2500
5	0	3000	3000	3000	×
6	0	3000	×	3000	3000
7	0	3000	3000	3000	×
8	0	3000	3000	3000	3000
9	5	3000	×	3000	3000
10	5	3000	3000	×	3000
11	5	3000	3000	×	3000
12	5	3000	3000	3000	3000
13	5	5000	5000	×	5000
14	5	5000	5000	5000	×
15	5	5000	×	×	×
16	5	5000	5000	5000	5000

Source: According to [Allison 2002]. Own prepared in cooperation with prof. Szymon Czarnik.

statistical tests can be used. The first of these, relating to individual variables, is the Student's *t*-test for independent samples. The test is performed by dividing the observations of a variable into observed values and missing data, and then checking the difference in means between groups within another variable. A non-significant Student's *t*-test means that there are no statistically significant differences in means between groups, which is consistent with the assumption that missing data are generated by a completely random mechanism [Enders 2010, p. 18–19].

Another way to check whether the missing data generation mechanism is of the MCAR type is to perform the Little's Test. This is a general test and its calculation takes into consideration differences in the mean for all the variables in the analysis [Enders 2010, p. 19–21].

The next way to assess whether missing data are randomly generated is to check the correlation of missing data for selected pairs of variables. If the correlation between the variables is low, then it can be assumed that missing data are random [Tsikriktsis 2005, p. 56].

The last way to assess the randomness of missing data is to use a regression model [Piggot 2001, p. 360]. Some authors claims, however, that there is no possibility to assess the randomness of missing data at all [Lang and Little 2014, p. 6–7].

Missing at Random

The second mechanism for generating missing data mentioned in the literature is Missing at Random (MAR). Missing data are considered to be of MAR type if the probability of missing data on variable *Y* is uncorrelated with the unknown value of observations on variable *Y*, while controlling for other variables in the analysis [Allison 2002, p. 4]. The randomness of missing data occurs in association with categories of other variables in the data set. This means that there is a systematic relationship between the variable in which the missing data occur and the probability of missing data [Padgett et al. 2014, p. 2].

Referring to the example presented in Table 1, it can be said that if the missing data are generated by the MAR mechanism, the probability of missing data can (and in this case does) depend on the amount of income.

$P\ (X|D = 2500) = 0$

$P\ (X|D = 3000) = 2/8 = 1/4$

$P\ (X|D = 5000) = 2/4 = 1/2$

The probability of missing data may (and in this case does) depend on the job experience.

$P\ (X|S = 0) = 0$

$P\ (X|S = 5) = 4/8 = 1/2$

However, the probability of missing data cannot both depend and not depend on the amount of income while controlling for the job experience variable.

For job experience of 0 years, the missing data are equally likely to occur for respondents earning PLN 2500 and PLN 3000; and for job experience of five years, the missing data are equally likely to occur for respondents earning PLN 3000 and PLN 5000.

For job experience of zero years:

$P\ (X|D = 2500) = 0$

$P\ (X|D = 3000) = 0$

For job experience of 5 years:

$P\ (X|D = 3000) = 2/4 = 1/2$

$P\ (X|D = 5000) = 2/4 = 1/2$

Comparing the mechanisms for generating missing data discussed so far, it is worth mentioning that MCAR-type missing data meet the assumptions for MAR missing data [Allison 2003, p. 545].

Not Missing at Random

The last mechanism for generating missing data described by Rubin is the Not Missing at Random mechanism. Missing data will be identified as generated by a completely non-random mechanism when the occurrence of missing data for a specific variable is related to the value of observations on this variable, even when controlling for other variables [Enders 2010, p. 8, Padgett et al. 2014, p. 2].

In the example presented in Table 1, the missing data mechanism is NMAR when the probability of missing data depends on income.

$P\ (X|D = 2500) = 0$

$P\ (X|D = 3000) = 2/8 = 1/4$

$P\ (X|D = 5000) = 2/4 = 1/2$

The probability of missing data does not depend on the job experience of the respondent.

$P\ (X|D = 3000) = 2/8 = 1/4$

$P\ (X|D = 5000) = 2/8 = 1/4$

In this case, even when controlling for job experience, the probability of missing data depends on the amount of income.

For a job experience of zero years, the probability of missing data is different for incomes of PLN 2500 and PLN 3000.

$P\ (X|D = 2500) = 0$

$P\ (X|D = 3000) = 2/4 = 1/2$

For a job experience of five years, the probability of missing data is different for incomes of PLN 3000 and PLN 5000.

$P\ (X|D = 3000) = 0$

$P\ (X|D = 5000) = 2/4 = 1/2$

The problem here, which is related to the existence of non-random missing data in the analyses, is the generation of loaded estimators which differ from the parameters in the population by a systematic error [Graham 2009, p. 553].

METHODOLOGY AND RESULTS

When reviewing the available data sets, it can be observed that some variables generate more missing data than others. The question generating potential problems related to the existence of missing data in the PGSS concerns the respondent's income. In 1995, the question about average monthly income was not answered by 1.9% of the respondents. In the last years of the survey, the number of missing data increased to 5.6% for the 2008 survey and 6.2% for the 2010 survey, respectively[1].

An analysis of missing data was performed on the example of the variable income from work of the respondent in the PGSS conducted in 2008. In order to check whether the missing data in the income variable are completely random, the Student's t-test for independent samples was performed. The income variable was divided into two sets of data – those respondents who answered the income question and

[1] Respondents who were not working during the survey were excluded from the missing data analysis.

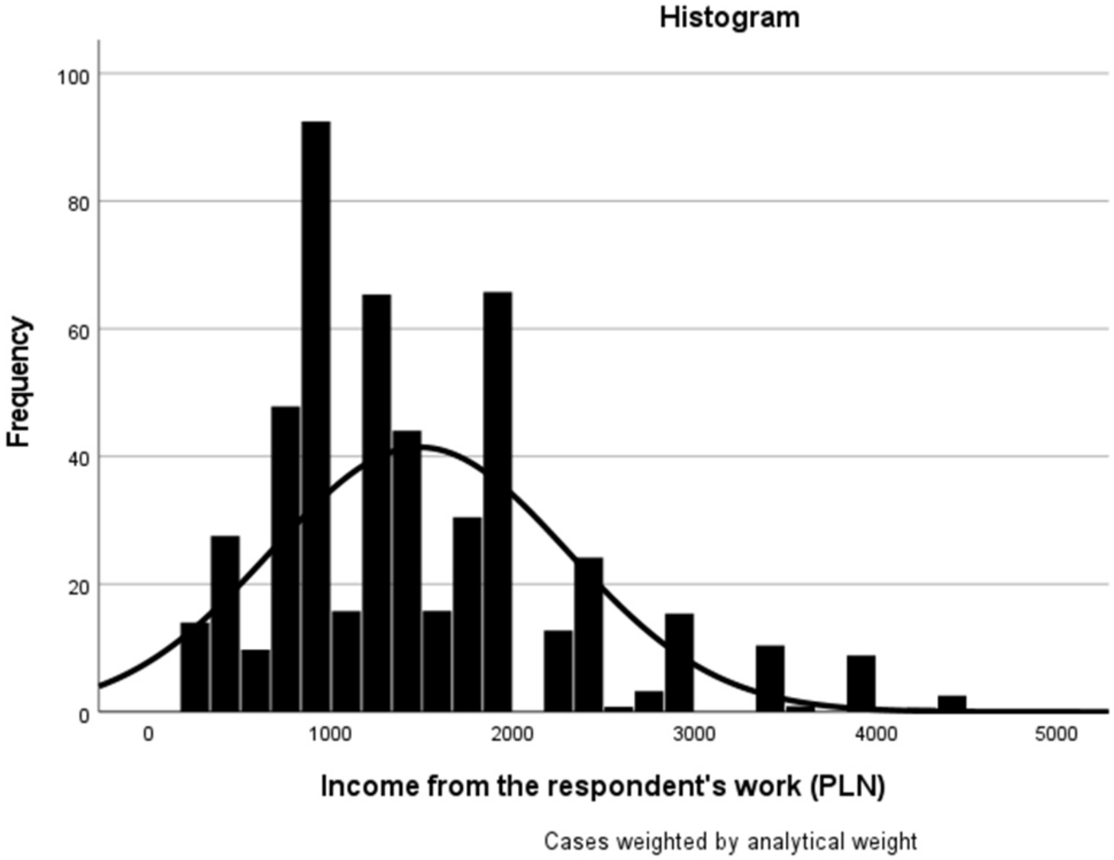


Fig. 2. A histogram for the variable income from the respondent's work

Source: Author's own compilation.

those who did not answer the question. The responses of the respondents who did not work during the survey, and thus had no income, were excluded from the analysis. Nine respondents whose income differed from the income of other respondents were also excluded from the analysis.

In order to check the average value for both groups on another variable that is correlated with the income variable, correlations were checked between the income variable and continuous variables that could be correlated with the income variable, i.e. the number of years of schooling, the number of years of income-generating work since the age of 14 and the size of the respondent's dwelling.

The correlation turned out to be statistically significant for only one variable. This variable was the number of years of schooling of the respondent.

The group of respondents who answered the question about income from work was significantly larger than the group of respondents who did not answer this question. Taking into account this fact and the assumption of the Student's *t*-test of the equality of the analysed groups [Bedyńska and Cypryańska 2013, p. 163], in further analyses observations with missing data were used and a sample was drawn from among the respondents who answered the question about income.

The differences in the mean for the two groups analysed were found not to be large, nor was the difference between the standard deviations (See Tab. 2).

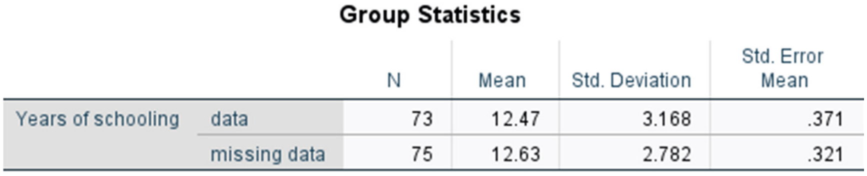
A non-significant Student's t-test for independent groups means that it is impossible to reject the null hypothesis that missing data are generated by a completely random mechanism (See Tab. 3). In this situation, missing data can be treated as random.

Table 2. Pearson correlation coefficients between respondent's income from work and other variables included in the Polish General Social Survey 2008

Specification	Size of the respondent's dwelling	Years of schooling	Years of work experience since the age of 14
Respondent's income PLN correlation	0.69	**0.329**	0.083
Statistical significance	0.129	**0.001**	0.063
N	478[2]	509	507

Source: Author's own compilation based on [ADS].

Table 2. Presentation of statistics – mean, standard deviation for groups: missing data and observable values

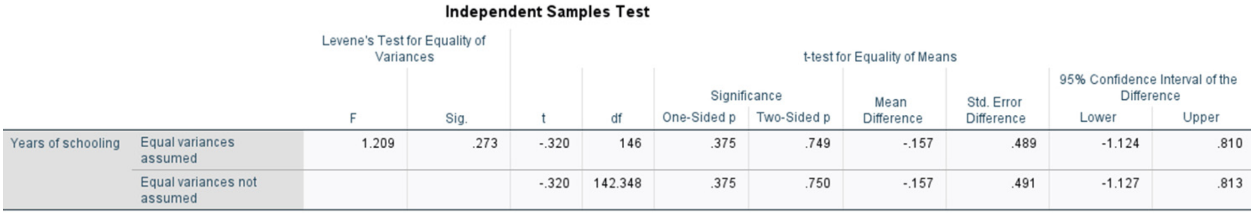


Group Statistics

		N	Mean	Std. Deviation	Std. Error Mean
Years of schooling	data	73	12.47	3.168	.371
	missing data	75	12.63	2.782	.321

Source: Author's own compilation.

Table 3. Student's *t*-test for two independent groups (missing data and observable values on the respondent's work income variable) for the respondent's length of job experience



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance		Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p			Lower	Upper
Years of schooling	Equal variances assumed	1.209	.273	-.320	146	.375	.749	-.157	.489	-1.124	.810
	Equal variances not assumed			-.320	142.348	.375	.750	-.157	.491	-1.127	.813

Source: Author's own compilation.

CONCLUSIONS

The existence of missing data is a challenge for many analysts and researchers conducting surveys. The best way to avoid large missing data is to design your survey carefully [Allison 2002, p. 2–3].

If, however, a significant number of missing data is present in a data set, the next step should be to verify the randomness of the missing data. For this purpose, it is worth using statistical tests, i.e.: Student's *t*-test for independent samples and the Little's Test. If the mechanism for generating missing data is not completely random, it is worth considering potential variables that are correlated with the variable having missing data. The theory or results from the existing studies may help to identify these variables [Enders 2010, p. 17].

The omission of data gap analysis in research and methodological reports may lead to inadequate research conclusions and recommendations. Even if

[2] Any outliers were excluded from the analysis.

there appear to be few missing data in the data set, i.e. less than 3%, very often the purpose of the analyses performed is not to calculate statistics for a single variable but to perform multivariate analyses. In this case, it may turn out that, out of the whole sample, the analysis will be carried out on a much smaller number of observations that are in the data set [Allison 2002, p. 2].

ACKNOWLEDGMENTS

I would like to thank Professor Szymon Czarnik for his cooperation while preparing a simulation presenting the differences between the three mechanisms of generating missing data: MCAR, MAR and NMAR, as well as Professor Adam Sagan for valuable comments on earlier drafts.

REFERENCES

ADS. Lista dostępnych zbiorów danych. Instytut Studiów Społecznych UW, Instytut Filozofii i Socjologi PAN. Retrieved from http://ads.org.pl/lista-zbiorow-danych.php?co=Polskie%20Generalne%20Sondaże%20Społeczne [accessed: 15.06.2021].

Allison, P.D. (2002). Missing data. Quantitative Applications in the Social Sciences. Sage Publications, Hawthorne Series.

Allison, P.D. (2003). Missing data techniques for structural equation modeling. Journal of Abnormal Psychology 112(4), 545–557.

Babbie, E. (2004). Badania społeczne w praktyce. Wydawnictwo Naukowe PWN, Warszawa.

Ballard, J., Richmond, A., van den Hoogenhof, S., Borden, L., Perkins, D.F. (2021). Missing Data in Research on Youth and Family Programs. Psychological Reports 00332941211026851.

Baraldi, A.N., Enders, C.K. (2010). An introduction to modern missing data analyses. Journal of School Psychology 48, 5–37.

Bedyńska, S., Cypryańska, M. (2013). Statystyczny Drogowskaz 1. Praktyczne wprowadzenie do wnioskowania statystycznego. Wydawnictwo Akademickie Sedno, Warszawa.

Brick, J.M., Kalton, G. (1996). Handling missing data in survey research. Statistical methods in medical research 5(3), 215–238.

Collins, L.M., Schafer, J.L., Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 6(4), 330–351.

Dunn, G. (1997). Compensating for missing data in psychiatric surveys. Epidemiologia e Psichiatria Sociale 6, 3, 159–162.

Durrant, G.B. (2005). Imputation methods for handling item – nonresponse in the social sciences: a methodological review. NCRM Methods Review Papers. NCRM/002.

Enders, C.K. (2010). Applied missing data analysis. The Guilford Press, New York.

Graham, J.W. (2009). Missing data analysis: making it work in the real world. The Annual Review of Psychology 60, 549–576.

Graham, J.W. (2012). Missing data analysis and design. Springer, New York.

Lang, K.M., Little, T.D. (2018). Principled missing data treatments. Prevention Science 19(3), 284–294.

Newman, D.A. (2014). Missing Data: Five practical guidelines. Organizational Research Methods 17(4), 372–411.

Padgett, C.R., Skilbeck, C.E., Summers, M.J. (2014). Missing data: the importance and impact of missing data from clinical research. Brain Impairment 15(1), 1–9.

Pigott, T. D. (2001). A review of methods for missing data. Educational research and evaluation 7(4), 353–383.

Shaharudin, S.M., Andayani, S., Kismiantini, N., Binatari, N. (2020). Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia. International Journal of Advanced Trends in Computer Science and Engineering 9(1.4), 646–651.

Siddique, J., Harel., O., Crespi, C.M. (2012). Addressing missing data mechanism uncertainty using multiple-model multiple imputation: application to a longitudinal trial. Annals of Applied Statistics 6(4), 1814–1837.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. Journal of operations management 24(1), 53–62.

Zdebska, W. (2021). The occurrence of missing data in surveys. *Acta Sci. Pol. Oeconomia 20* (2), 95–103, doi: 10.22630/ASPE.2021.20.2.19

WYSTĘPOWANIE BRAKÓW DANYCH W BADANIACH ANKIETOWYCH

STRESZCZENIE

Celem niniejszego artykułu jest omówienie zagadnień związanych z analizą braków danych. Dlaczego braki danych występują w zbiorze danych? Ile procent zebranych danych stanowią braki danych? Jaka jest natura braków danych, które pojawiły się w trakcie zbierania danych? Jakie czynniki mogą wpłynąć na potencjalne pojawienie się braków danych? Powyższe pytania, na które autorka artykułu pragnie odpowiedzieć w jego ramach są niezwykle istotne w ocenie prowadzonych badań ankietowych lub ocenie jakości danych zastanych. Brak refleksji nad wspomnianymi powyżej aspektami może prowadzić natomiast do wyciągania fałszywych wniosków oraz rekomendacji. Stąd też analiza braków danych jest niezwykle istotnym, a często nadal pomijanym etapem analizy danych ankietowych. W ramach artykułu przestawiona została analiza losowości braków danych na podstawie danych zebranych w ramach Polskiego Generalnego Sondażu Społecznego w 2008 roku.

Słowa kluczowe: badania ankietowe, analiza braków danych, MCAR, MAR, NMAR