The Prevalence of Missing Data in Survey Research

Hamzeh Mohd Dodeen
UAEU
United Arab Emirates

Abstract

The credibility of surveys relies significantly on the completeness of the data collected from representative samples. Missing data is a serious problem in survey research. The existence of variables with missing information negatively affects the research results and findings. This study examines the prevalence of missing data in surveys, and additionally compares its incidence between genders. A total of 119 relevant surveys from different countries represented the sample of this study. Results indicated that, on average, 38% of data was lost in the surveys analyzed. Males and females were very similar with respect to the extent of missing data, with an average of 37% and 38% respectively. Overall, results show that only 62% of the initial sample size was available at the end of the data collection stage.

***Keywords*:** missing data, survey, survey research, survey data, gender

Although missing data in research is a common problem, it is often ignored (McKnight, McKnight, Sidani, & Figueredo, 2007). Research results or findings are negatively affected by lost data in several ways (Acock, 2005; Bonder, 2006; Horton & Kleinman, 2007; Little & Rubin, 2002; McKnight, et. al, 2007; Wayman, 2003). First, missing data poses a threat to the validity of scientific inquiry. In fact, both the internal validity of causal inference, as well as, the external validity (generalizability) is negatively affected by missing data. Second, missing data decrease the statistical power of tests because of the reduction of the final sample size available for statistical analysis. Third, missing values make some data sets (or at least some variables) inappropriate for particular statistical analyses. For example, the minimum sample size required for conducting some statistical procedures or tests may be unavailable. Fourth, the final sample may no longer represent the actual population from which the sample was originally selected.

Generally, data can be missing in survey research for different reasons. Some reasons could be due to the design of the study (e.g., the study required too much of the participants' time), while others may be due to chance (Horton & Kleinman, 2007). For instance, a number of variables may be collected from only some subjects; while some items may be ambiguous or interpreted differentially by respondents; certain items may seem offending; some items may be inapplicable; and some subjects may not be willing to answer particular questions, or may quit before completing the survey; or errors may occur in any research step (Cool, 2000). Therefore, it is not uncommon to obtain either missing or unusable data in survey research (Cool, 2000; Raaijmakers, 1999; Raymond, 1987,Witta, 1994).

When conducting a survey research, it is usually recommended to increase the required sample size anticipating for possible missing information (Light, Singer, & Willett, 1990). However, the final sample

still ends up, in many cases, with a large proportion of lost data. The central goal of making valid inferences regarding the population of interest is threatened by missing data. This is more likely to occur when missing data make the sample different from the population that it was drawn (Wayman, 2003).

The main goals of this study were to estimate the incidence of missing data in surveys and the its prevalence with respect to the respondent's gender, and analyze how surveys' sample size is affected by this problem.

Literature Review

As missing data is a serious issue in surveys, it is useful to determine the prevalence of this problem in real surveys. It is also valuable to determine the proportion of missing data in surveys with respect to the gender of respondents for several reasons. First, it is rare to survey people using the responses of only males or only females. Having a reasonably equal ratio of males and females is one of the conditions that researchers usually like to have in order to improve the degree of the study representation and generalizability. Second, gender is the most common demographic variable that appears in almost all surveys. So it is necessary to estimate the prevalence of missing data with respect to this variable. This does not mean that the other variables have no effect; however, gender is evidently more common than others. Third, gender differences exist in most societies. Even in countries that promote equity between males and females in most life aspects, there are still some biological, psychological, and social factors that affect males and females differently. Consequently, these factors are likely to influence a person's response to a given variable or question.

In the literature of missing data a distinction is made between "unit missing data" which refers to data missing from a unit of analysis (e.g., a person or participant), and "missing values" which refer to scores or values on a particular variable (e.g., survey item or question) that are missing (McKnight, et. al, 2007). This study focuses on the missing values in survey research.

A typical survey consists of several questions or items about one (or more) topic of interest. Survey questions can be classified (based on the nature of the collected data) into two types: demographic questions that ask about personal facts and questions that ask about the survey topic. Questions on personal facts usually refer to physical and social details. For example, age, gender, level of education, race, income, occupation, marital status, number of children, place of birth, and number of working hours per day. In the second type, questions explore the participants' opinions, feelings, positions, or thoughts toward a specific subject of interest.

An important issue in studying missing data is to know the probabilistic process by which data become missing (missing data mechanism). In the literature, there are three types of missing data mechanisms (Acock, 2005; Little & Rubin, 2002). The first is when data are missing completely at random (MCAR). In this case, the probability that an observation is missing is unrelated to the value of any variable in the data set. For example, data on the variable "age" would not be considered MCAR if females were less likely to report their ages than males (prevalence of missing data is correlated with gender). The second type is when data are missing at random (MAR). Here, cases with missing data differ from cases with

complete data, or missingness on a variable depends on the values of other variables. For example, if a test is administrated before a survey administration session, then participants with lower scores on the test may be less likely to complete the survey. In this case, missing data are due to some external variable rather than the variable where data are missing. When data are MCAR or MAR the analysis is not biased, however, the problem of missing data still exists (Howell, 2007).The third type is when data are missing not at random (MNAR), also called non-ignorable. Dealing with missing data in this case is much more difficult than the other two types.

Missing data mechanisms are not the only factors to consider when we need to analyze or handle missing data problems. Other factors are whether missing values are with the dependent or independent variables, and whether missingness is due to the design of the study or due to chance. For example, for various reasons, some questions may not be asked to all the subjects of the entire sample. One more influential aspect to bear in mind is the size of missing data. Whether it is numerous or a few values and whether the information is missing from a few subjects or many of them, it is still a significant issue to consider and manage.

Methodology

Objectives

A valid and accurate assessment is the main goal in social research which heavily uses surveys for collecting data. Marie (1997) stated that improving the psychometric properties of the survey instrument is necessary since many survey instruments are poorly designed and interpreted. Often in data analysis of survey results, missing data cannot be excluded nor ignored. This study aimed at determining the prevalence of missing data in surveys in general and with respect to the gender of respondents in particular. Additionally, the study examined the effect of missing data on decreasing the final sample size and changing the ratio between the number of males and females with and without missing data.

The Survey Data Sets Used in the Study

Through an intensive process of search on the Internet, more than 250 authentic survey data sets were accessed and downloaded. These surveys were originally conducted in many countries to collect data information related to a wide range of issues including politics, family, environment, minorities, religion, health, job and customer satisfaction, youth, business, media, war, social security, and others. Data of these studies were raw and have not been treated by any method of missing values. Through initial analysis of each survey, many data sets were excluded in this stage for various reasons. For example, some data sets did not have variable labels and/ or variable values. Also, since the study compares missing values between male and female respondents, if the data file did not include the gender of the respondent it was not used. Some studies were excluded because they were designed for only one gender such as women and mothers. Additionally, some surveys collected only facts or personal information from respondents. At the end of this analysis, only 119 survey data sets were found appropriate for further analysis. These data sets which represented the study sample were originally from the following countries: USA (104), UK (8), Mexico

(5), Taiwan (1), and Japan (1). Examples of these data sets are: USA Social Capital Community Survey (2006), The U.S. Citizenship, Involvement, Democracy Survey (2005), National Survey of America's Families (2000), National Survey of Latinos Education (2004), British Election Study (2001), Japan National Survey on Family and Economic Conditions (2000), Political Tolerance in Taiwan (2004), and Pew Hispanic Center Survey of Mexicans Living (2006).

Results

To estimate the prevalence of missing data in surveys, the variable with the maximum number of missing data for each survey data set was identified. Then the percentage of missing data on this variable was calculated. Table 1 summarized the results of this analysis for five surveys selected as examples.

Table 1
Missing Data for Five Examples of Survey Data Sets

Survey	Sample size	Variable with max missing data	No. of missing data	Percentage of missing data
National Survey of Adolescents in the US-1995	4023	Do you think school programs are helpful?	1339	33%
Youth Vote Survey-USA-1999	807	Which political party is more conservative?	366	45%
British Crime Survey-2000	2561	Should Britain stay member of EU?	302	12%
Winthrop University Student 1996Religion Survey-	306	Human beings evolved from earlier species of animals?	214	70%
Smoking Survey University of Oregon-2002	776	Current thoughts about quitting smoking?	677	87%

The same analysis was conducted to the 119 surveys. The distribution of the percentage of missing data was represented in the following histogram (Graph 1). The average of the percentages of missing data over the 119 data sets was found to be 38%. This means that, on average, surveys have more than one third of their data missing.

Additionally, and based on the value of this percentage, survey data sets were categorized into three categories that represent less than 25%, 25-50% and more than 50% missing information. These three categories represent the states of having small, moderate, and large amount of missing data, respectively (Dodeen, 2003). Results are summarized in Table 2. A quick look at this table reveals that, in general, missing data is prevalent in surveys. Out of 119 survey data sets used in this study, 54 (45.4%) surveys have less than 25% missing data; while 49 surveys (41.2%) have more than half of their data missing.

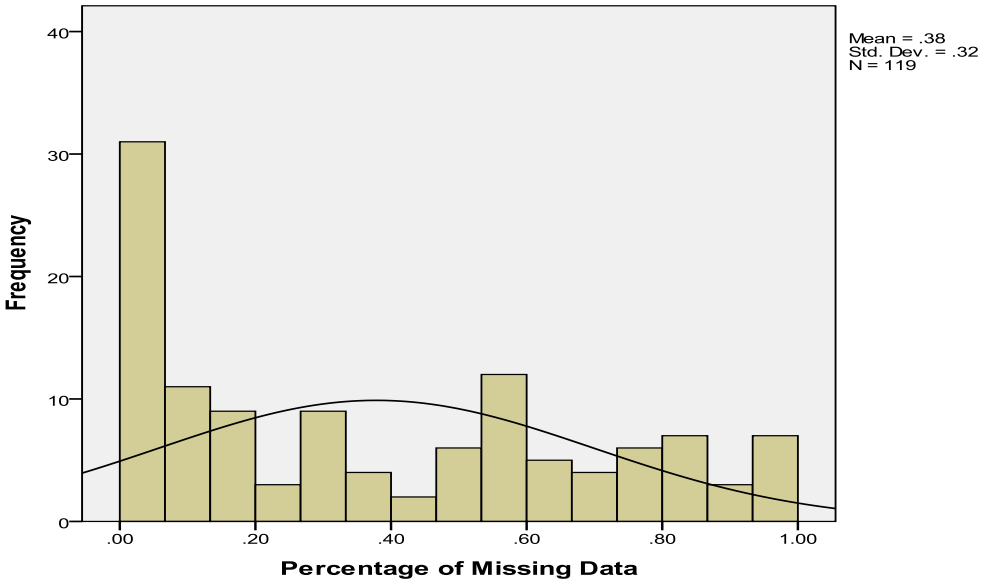
Table 2

Percentage of Missing Data in Surveys

Percentage of missing data	Number of surveys	Percentage of surveys
Less than 25%	54 (45.4%)	45.4%
25% - 50%	16 (13.4%)	13.4%
More than 50%	49 (41.2%)	41.2%
Total	119 (100%)	100%

Graph 1:

Distribution of the Percentages of Missing Data in Surveys



To evaluate the results between males and females, the percentage of missing data by males for each data set was calculated and compared with that of females. To illustrate this analysis, Table 3 summarized the results of the five selected survey examples.

Table 3

Comparing Missing Data between Males and Females for Five Surveys Examples

Survey	Sample size	Initial No. of Males	Initial No of Females	Percentage of Missing Males	Percentage of Missing Females	Ratio M to F Without	Ratio

						Missing	M to F with Missing
National Survey of Adolescents in the US-1995	4023	1142	2881	43%	30%	.40	.32
Youth Vote Survey-USA-1999	807	401	405	45%	46%	.99	1.01
British Crime Survey-2000	2561	1143	1418	5%	17%	.81	.92
Winthrop University Student Religion Survey-1996	306	84	221	42%	24%	.38	.29
Smoking Survey, University of Oregon-2002	776	345	431	83%	90%	.80	1.41

Although in Table 3 percentages of missing data between males and females are not very close to each other in some examples, results indicated that, overall, males and females are very similar with respect to size of missing data. The average percentage of missing data for males was 37%; while it was 38% for females (standard deviations were also very close with .32 for males and .33 for females). Also, the effect of missing data was analyzed by comparing males' to females' ratio with and without missing data. From all the surveys analyzed in this study, the average of the initial ratio between both genders (males to females) was 0.91. This means the number of male participants was less than that of females. Because of missing data, this average ratio was changed to 1.08. To further understand the change in genders' ratio, the number and percentage of data sets, in which the ratio has been reserved, was calculated. Results indicated that the ratio between males and females have been reversed in 23 (19.3%) data sets.

To analyze the effect of missing data in surveys, the loss in sample size for each data set was estimated by comparing the initial sample sizes (without missing data) with the final available sample size (with missing data). The ratio was 62% over the 119 surveys, which means that on average; only 62% of the initial sample size was available at the end of the data collection stage.

Discussion

Missing data is a common reality of research in general and survey research in particular. The problem still represents a significant challenge for social scientists due to the lack of understanding of its importance and prevalence (McKnight, McKnight, Sidani, & Figueredo, 2007). The purpose of this study

was to investigate the prevalence of missing data in surveys, and the magnitude of missing data for males and females separately, and the effect of missing data on sample size.

Investigating the prevalence of missing data in surveys is an important issue because surveys are a common and practical tool to collect information about or from people. Therefore, surveys should be constructed and conducted appropriately to give clear, accurate, and valid results. Having missing data threatens the validity of the results. Additionally, as the size of missing data increases, the validity of the results decreases. Hence, the credibility of the organization producing the report or study may be jeopardized (Witt, & Kaiser, 1991).

The results of this study showed that surveys, in general, have more than one third of lost data. Moreover, in more than 40% of surveys analyzed in this study, 50% of data or more was missing. This clearly indicates that missing data is a prevalent problem in surveys. Bodner (2006) attained similar results on analyzing a random sample of empirical research journal articles from the Psych INFO database. McKnight and his colleagues (McKnight, et.al. 2007) analyzed missing data in over 300 articles in a prominent psychological journal across a 3-year period, and also found that missing data is prevalent with an average amount that exceeds 30%. The frequency of missing data should raise the awareness of social scientists on the significance of this problem.

Overall, surveys analyzed in this study showed that males and females were similar with respect to percentages of missing data. This, of course, does not mean that in each survey, the percentage of missing data by males is equal to that of females, but it is the average over the analyzed surveys for each gender. An important point that could be concluded from this result is that there is no general tendency on one gender to miss data more than the other.

The effect of missing data on the survey sample size was clear and substantial. First, missing data causes a significant drop on the sample size. The final sample size was only 62% of the initial sample size for all the surveys. This big loss in sample size could be converted to loss in effort, time, and money that have been invested in collecting data. Furthermore, decreasing sample size has its strong negative effects on sample representation of the study population and its statistical power. Second, the missing data changed the actual ratio between males and females. Overall, females were more than males in the initial samples, but this ratio was finally reversed because of missing data.

The magnitude of missing data in surveys is serious and should be considered in all survey research and applications. In this regard, the American Psychological Association Task Force on Statistical Inferences strongly emphasized that researchers report complications like missing data and non-responses in their studies. Also, researchers should document how the actual analysis differs from the planned analysis (Wilkinson and APA Task Force on Statistical Inference, 1999).

Finally, as it is often true, prevention is better than cure, and treating missing data should start at the prevention level before remediation. All procedures to reduce the likelihood of obtaining missing data in each study or project should be evaluated, and the appropriate ones applied. This strategy should be part of every phase of research, starting from the development of the research question and design of the study, and continue throughout planning, piloting, implementation, monitoring, and data management and analysis (Hardy, Allore, & Studenski, 2009). The statistical procedures used to treat missing data in surveys

are not supposed to compensate for or replace the essential efforts to reduce missing data. Strategies range from preparing clear and credible instruments to the use of effective data collection methods.

References

Acock, A. C. (2005). Working with missing data. *Journal of Marriage and Family, 67*,1012-1028.

Bodner, T. E. (2006). Missing data: Prevalence and reporting practices. *Psychological Reports, 99*(3), 675-680.

Cool, A. (2000). *A review of methods for dealing with missing data*. Paper presented at the Annual Meeting of Southwest Educational Research Association, Dallas, TX.

Hardy, S. E., Allore, H., & Studenski, S. (2009). Missing data: A special challenge in aging research. *Journal of the American Geriatrics Society, 57*(4), 722-729.

Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician, 61* (1), 79-90.

Howell, D. C. (2007). *Treatment of missing data*. Retrieved from http://www.uvm.edu/~dhowell/StatPages/StatHomePage.html

Light, R., Singer, J., & Willett, B. (1990). *By design: Planning research on higher education.* London, England: Harvard University Press.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd Edition, New York: Wiley & Sons.

Marie, L. (1997). *The application of item response theory to employee attitude survey data Using Samejima's graded response model.* (Unpublished Doctoral Thesis).The University of Connecticut, Connecticut, USA.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction.* The Guilford Press: New York.

Raaijmakers, Q. (1999). Effectiveness of different missing data treatments in survey with Likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59,* 725-748.

Raymond, M. R. (1987). Missing data in evaluation research. *Evaluation & The Health Professions, 9,* 395-420.

Wayman, J. (2003). *Multiple imputation for missing data: What is it and how can I use it?* Paper presented on the Annual Conference of the American Educational Research Association (AERA), Chicago, IL.

Wilkinson, L., & The APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.

Witta, E. L. (1994*). Are values missing randomly in survey research*? Paper presented at the Annual Conference of Mid-South Educational Research Association, Nashville, TN.