Ethical Concerns in Handling Missing Data

Angela Gunn

Introduction

There is a surge of interest in Big Data technologies and using the wealth of data available in various company, government and research databases. What is not discussed often is what happens when data sets are incomplete and data is missing. Missing data can come in the form of a few pieces of information left out of a record or entire records missing from the complete dataset. How missing data is handled is concerning on an ethical platform in how conclusions from incomplete data are used to drive decisions.

In 2004, the Centers for Disease Control and Prevention (CDC) issued a report stating the concerning levels of lead in Washington DC's residential water systems did not pose any harm to the public (Renner, 2009). At that time, Washington DC was seeing spikes in the lead content in the water system caused by the use of chloramine for treating the water and an abundance of lead pipes. While the CDC stated explicitly that all data was used in its report, it was later discovered thousands of records of children's blood tests were lost and so were not included in the analysis. It was determined three times as many children had elevated lead levels than what was reported. According to the Mayo Foundation for Medical Education and Research, lead poisoning in children can

contribute to developmental delays, learning disabilities and other severe mental and physical ailments (Mayo Clinic, 2014).

The missing data was found when Congressional staffers investigated the CDC's explanation of the missing data and contacted the labs for the data. Over 4000 lab results were recovered. When included in the analysis, the lead levels in children were double what the CDC originally reported. It was also found that many of the children were not in housing areas with lead pipes as previously proposed, concluding that the source of the lead was Washington DC's drinking water (Renner, 2009).

This report explores the ethical implications of missing data. Both missing data within record attributes and missing outcome values as a result of attrition is examined. A look at the mechanisms of missing data and the methods used to handle missing data in data sets is explored. Next, ethical implications of missing data and how it should be reported is discussed. The report concludes with general recommendations for working with missing data and a final look at the CDC's conduct in the Washington DC lead water crisis.

Why is Data Missing?

The most common type of missing data most analysts will come across is data that is missing within a record in a data set. Attribute data can be missing from a record for many reasons, including the failure to input the data during data entry, the field for data

collection may not be available at the time the data was collected, or the data was simply not known at the time (NHTSA, 2015).

Attrition occurs when the outcome variable is not observed. Attrition is most often associated with people leaving studies before final outcomes can be recorded. Twisk and de Vente found attrition in longitudinal research studies is particularly common. While usually observed in the final measurements, missing data is also seen in intermediate measurements when subjects miss a measurement but return for follow-up measurements (Twisk & de Vente, 2002).

It is important to understand why the attrition in the data set exists because it will impact how analysis ofthe results are completed. Gerber and Green identified reasons why the final outcome measurement may be missing. This includes subjects dropping out of the experiment, an unwillingness to complete questionnaires, an inability to measure the outcome due to circumstances preventing researchers from accessing the data, or the data is intrinsically unavailable. Removing all subjects with missing data will bias the results; it is possible people who leave the study have shared characteristics different from those who stayed in the experiment. This suggests the unobserved results may not be the same as those that were recorded (Gerber & Green, 2012).

Jones et al. identify that an additional reason which deserves its own mention is attrition due to subject mortality. It is impossible to know if the subject would have completed the experiment or would have voluntarily dropped out. When a participant drops out of a study, the result is missing data, but when the participant dies the result is truncation of

follow-up. Consider a study on quality of life: would a subject who dies have a low quality of life? It is possible if the subject was ill they would have a lower quality of life, but what if the death was accidental? This example highlights the importance of understanding why participants drop out of studies (Jones, Mark, Mishra, & Annette, 2015).

Classifying Missing Data Mechanisms

Donald B. Ruben is credited with classifying missing data into three categories: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) . Classifying missing data helps to understand the effectiveness of different missing data techniques and resulting bias (Rubin, 1976).

Missing data is classified as MCAR if the underlying reasons for the data missing is independent of other attributes in the record. This means there is no correlation between whether a value is missing and any other attribute in the record(de Goeij et al., 2013). MCAR is the only classification which does not have a bias if missing values are ignored.

MAR implies missing data is related to a known data attribute. For example, if there are more missing values in an attribute for one gender, but within that gender it is random as to whether the values are missing or not (de Goeij et al., 2013).

MNAR describes missing data in which the underlying reason for the missing data depends on another unobserved variable. If a survey is on the quality of life for hospital patients, and the patients become so ill they cannot complete the survey, then the missing data is a result of the low quality of life (the outcome variable). It is also possible an attribute that could explain a missing value was not collected (not observed). It is often difficult to establish the difference between MAR and MNAR - to establish MNAR there must be knowledge of the missing observation (de Goeij et al., 2013).

Imputation

Imputation methods allow us to logically substitute missing data with possible values. Some researchers “equate imputation as an unethical practice of fabricating data” (Enders & Gottschall, 2011). Imputation is “making up” the data by estimating what the value could be (de Goeij et al., 2013; Graham, 2009). Some imputation methods, particularly single-imputation methods based on means or last observation carried forward, create biased parameter estimates and small standard errors. However, multiple-imputation methods have been shown to produce accurate estimates under MCAR and MAR datasets and correct for adjustments to standard errors (Enders & Gottschall, 2011).

If not addressed, missing data within data records can lead to statistical bias. *Listwise deletion*, and *Pairwise deletion* are two methods that ignore missing data when computing attribute statistics. Listwise deletion, in which the only records considered in

analysis are complete-case with no missing values, introduces bias estimates. In the case of a survey, the responses from people who do not answer all questions are likely different than the responses provided by people who answer all questions, perhaps due to demographics or socioeconomic profiles.

Pairwise deletion is best explained by visualizing a data set that has student marks for two different courses. Not all students have taken both courses. In averaging the course marks, the resulting means would be based on different subsets of students. Correlations cannot be performed in this situation (Graham, 2009). Variances on data with missing values that are not treated will be understated, causing hypothesis tests to have narrow confidence levels that suggest the data is more reliable than it actually is; the data will have less variance and smaller standard errors (Census, 2004).

Missing data values can be addressed with imputed data, eliminating the bias of excluding records due to the missing values. Adding the imputed data allows analysts to use more records in their analysis, increasing the statistical power of their calculations. While filling in the missing values with imputed data sounds like the solution to missing values, caution must still be used. It is important to remember that imputation does not increase the amount of information available. To say it adds information is to suggest copying rows of data increases your sample size. Attributes with imputed values, such as one for sex, are useful for generalized inquiries like “how many men versus women are in front-end accidents,” but not for more specific inquiries like “how many men were driving a convertible” (NHTSA, 1993).

Finally, a paper on missing values would not be complete without suggesting that a missing value, particularly on an attribute, may not be an actual error in the data. For example, in a survey that asks for a person's hair color, a respondent who is bald may skip the question. Survey questions should be constructed such that an option like "none" or "don't know" can be selected (Han, Kamber, & Pei, 2011).

Brief Description of Imputation Techniques

Methods for handling missing data within records range from ad hoc conventional solutions to methods that are more complex but have theoretical support as appropriate for MAR data variables. Of course the most accurate method is to track down the actual missing values, but this is often impractical from a logistical and cost basis.

Conventional Methods

Conventional methods include *mean imputation*, and *univariate imputation*. Mean imputation involves replacing all missing values with the average value of values that are known for that attribute. Univariate imputation takes all possible values and randomly assigns them to the missing values (NHTSA, 1993). Both methods distort estimates of variation and result in loss of statistical power (de Goeij et al., 2013).

Single Imputation Methods

Single imputation methods used in longitudinal data include *last observation carried forward (LOCF)*, *averaging available items* and *maximum likelihood*. LOCF substitutes the last observed value for the missing value; for example, carrying forward the last blood pressure value for a patient who misses a scheduled reading. LOCF assumes the last reading will be constant for the remainder of the study rather than fluctuating over time. It also assumes the probability of dropping out of the study is completely random. It is illogical to think either of these assumptions can always hold true if the subject drops out of the study and has no further follow-up measurements (Molnar, Hutton, & Fergusson, 2008).

Averaging available items can be used on data records where all attributes are a value on a one-dimensional scale. It takes the average of responses in that record and uses that average for missing values in that record (Enders & Gottschall, 2011). For example, if a person answered four out of five questions in a survey, where all questions are answered on a scale from one to five, the missing value would be substituted with the person's average response in the other four questions.

MCAR-based data may use listwise or pairwise deletion methods. Listwise methods remove all records that are not complete, regardless if the attribute of interest has its value missing or not, while pairwise deletion removes only records where the attribute of interest have values missing. Both methods still introduce some bias in the results, and

should only be used if it is absolutely certain the data falls in the MCAR classification (Graham, 2009).

Maximum likelihood, also known as full information maximum likelihood or hot-decking, uses different combinations of attributes in multiple iterations until all unknown values are populated. For example, to determine age in the National Automotive Sampling System (NASS) General Estimates System (GES), the missing value for “sex” of a person involved is determined by correlating the data elements age, hour, day of week, violations charged, person type, seating position, drug & alcohol involvement and number of occupants and vehicles involved (NHTSA, 2015). Maximum likelihood, when used in large sample sizes, can produce approximately unbiased estimates that are close to the actual values. However, it is difficult to determine which and how many attributes should be used for the process, and typically specialized software is required to do the analysis (Allison, 2002). If used in only a single iteration, it performs minimally better than means substitution with the same problem of underestimating the standard error (de Goeij et al., 2013).

Multiple Imputation

Multiple imputation uses algorithms to audition different combinations of attribute values until a model is determined that best fits the data (highest log likelihood value). It results in unbased estimates, maximized statistical power and standard errors that are not underestimated. The process typically involves three steps - imputation, analysis and pooling. The imputation step replaces each missing value with m>1 simulated values,

based on a distribution provided by the analyst. In the analysis, *m* copies of the data set are made for each of the simulated values and analyzed for parameter estimates and standard errors. In the last step, pooling, the analyses are integrated to yield a single set of results based on interval estimations, and likelihood-ratio test statistics (Enders & Gottschall, 2011; Peng, Harwell, Liou, & Ehman, 2006).

Imputation for MNAR

MNAR based data can be analyzed with a *selection model* and a *pattern mixture model*. The selection model combines two models of regression equations to predict the response probabilities for each missing outcome variable. The pattern mixture model identifies cases with the same missing data pattern. Pattern-specific estimates can then be computed on the model parameters, and then the weighted average of the estimates to yield the results. These MNAR models require assumptions to be made in their implementation and relies heavily on multivariate normality to ensure the estimates are not distorted (Enders & Gottschall, 2011).

Discussions on Missing Data and Ethics

One of the best ways of handling missing data is to prevent it from being missing at all (Enders & Gottschall, 2011). However, particularly in a human-subject research study, it is unethical to force participants to stay in any study. This would go against the applications of the Belmont Report's principle of informed consent to research, which

states that participants have the right to withdraw from the research at any time (Belmont, 1979). One way to at least prepare for the potential of participant withdrawal is to measure the intent to drop out. While some people will indicate they will drop out and do, others who indicate they will drop out will stay for the duration of the study. This latter group may give clues to imputing the output variables of the people who do actually drop out (Graham, 2009).

Following-up on data collection is also an option to filling in missing data. This is often impossible or impractical depending on the nature of the data. Where the limitation is the practicality of following up on every missing record, it may be possible to instead follow-up on a random subset of the missing records. Inferences on the data collected in this subset will also give clues to imputing the values of other missing data, particularly if the missing data mechanism at play is MAR (Graham, 2009).

Many of the imputation methods look at auxiliary variables to cluster like records to determine the values of missing data. It is tempting to collect as many auxiliary variables as possible to ensure the most accurate imputation results, however doing so may not be ethical. Collecting more data than necessary from participants invokes privacy concerns. Even if the information collected is not deemed personally identifying, such as gender, city of residence, or hair color, it may be possible to identify the participant from the combination of all auxiliary variables. Additional variables included in a survey also adds to the burden on participants are given as part of the study and may lead to unintended negative consequences. For example, particularly long surveys may frustrate a respondent and lead to more missing values, not less. Participant

fatigue or boredom may also decrease the integrity of the data. It is important to balance the number of variables collected for comprehensive evaluation with the ethical concerns of how the additional variables impact participants (Enders & Gottschall, 2011).

One solution for collecting additional auxiliary variables is to use a planned missing data design. The idea is that researchers purposely vary the questions on the surveys such that each participant does not answer some of the questions. In a 3-form design, all participants would get the same set of base questions. The remaining questions are divided into three groups; each of the three surveys drops one of the question groups. This results in missing data for one third of these additional variables. Assuming the three different surveys are distributed randomly to participants, the missing data can be classified as MCAR, which has a reduced bias risk. Though this method will reduce power for some analysis, it allows researchers to collect more information through auxiliary variables without overburdening the participants with extra long surveys (Graham, 2009).

Extreme Attrition and Missing Data

How much missing data is too much? The answer depends on the data and the purpose of the study. If the variable with missing values is determined to be MCAR, missing 67% of the data is not really an issue since MCAR does not introduce bias if only available values are used. A 67% drop-out rate on the other hand for MAR or NMAR is

concerning. If the missing data is unrelated to the outcome variable (MAR), the resulting parameter estimates will be noisy, but may still be useful. If the missing data is related to the outcome variable (MNAR), the parameter estimates may be distorted. As stated previously, it is often impossible to determine the exact mechanism for the missing data, so it is hard to determine the effect the missing data will have on analysis (Enders & Gottschall, 2011).

If there is an extreme volume of missing data, particularly with attrition, is the research still valid? There are a number of ways of looking at this. First, all data collection takes time and has a cost. Consider a longitudinal study following participants over many years. To start the research over again is impractical, both for recruiting and for the duration of the study. Depending on the study, there may be reimbursements to participants or the costs of medication or supplies that add to the overall costs of the study (Enders & Gottschall, 2011).

To completely discard a study with concerning missing data rates must also be evaluated on ethical grounds. To file a study in a drawer and not share it with the academic community is unethical in itself. By sharing the results, knowledge can be gained on what auxiliary variables should be collected in future experiments to better handle missing data. Lessons learned from analyzing the data and the study's processes can be used to prevent issues observed during the course of the study and provide new ideas or areas of interest for future studies. Abandoning a study not only has a financial impact of sunk costs, but also has a cost to participants for the burdens they bared during the study (Enders & Gottschall, 2011). Consider a study on a new

medication with some negative side effects. Participants entering the study did so with the purpose of advancing medical knowledge and potentially receiving the benefits of the new drug. Those who suffer from the negative side effects but continue treatment to the study's conclusion deserve to be rewarded with the knowledge that they contributed to the current state of medicine. To deny them this reward due to other participants dropping out is unjustifiable.

A "cost-utility model" is used to determine the ethical ramifications of abandoning a study with high attrition. Before the study even begins, a researcher should consider the costs of doing the research against the utility (benefit) of conducting it. Research with high costs but low utility should not ethically be considered. The cost of *not* doing the research against the potential benefits that may be observed should also be considered. Enders & Gottschall quote Rosenthal and Rosnow in their argument that "the failure to conduct a study that could be conducted is as much an act to be evaluated on ethical grounds as is the conduction of a study" (p. 369).

Reporting Missing Data

Since missing data is something that researchers cannot get away from, it is necessary to discuss how to report missing data in research reports. Burton & Altman performed a review of 100 cancer prognostic research articles and found 81% of them reported missing data in the covariates collected (Burton & Altman, 2004). This means 19 articles did not even address the fact they had any missing data; it is unlikely all of these studies

had no missing data at all. 32 articles discussed the methods used to handle the missing data, but some just stated imputation was used without giving the details of the actual method employed. Without this information, replicating the statistical results is difficult.

Burton & Altman also found only 21 of 100 articles discussed the mechanisms for the missing data. Considering the previous discussion on the effect of missing data on analysis and the bias it introduces, how can we trust the analysis and recommendations made by these articles? It is possible the researchers eliminated all incomplete cases (complete-case analysis) and hence the analysis did not include any missing data; this information should be included in the article.

Only 87 of 100 articles stated the total number of cases used in the analysis. Many also did not discuss the choice of covariates or even how many were collected and used in the imputation process. This makes it difficult to replicate research or to improve on it in future research endeavors.

When imputation methods are used to handle missing data, it is not enough to just list them. Spineli *et al.* looked at 190 systematic reviews published from 2009 to 2012. They found a majority of analysts describe the strategies used for handling missing data, but not the justification as to why each strategy was employed or the specifics of their implementation. Burton & Altman found 32/100 articles they examined discussed the methods used to handle the missing data, but some just stated imputation was used without giving the details of the actual method employed (Burton & Altman, 2004). As

previously discussed, MNAR imputation methods require some assumptions to be made in its implementation. The choices made here will influence the outcomes of the analysis. While there were improvements in the rates of reporting missing data in their review, there is still room for improvement (Spineli, Pandis, & Salanti, 2015).

The American Psychological Association (APA) Working Group on Journal Article Reporting Standards (the JARS Group) published a reporting standard (the JARS standard) that includes recommendations on how to report missing data. The standard recommends reporting the percentages of missing data, empirical evidence supporting the causes of the data that is missing (MCAR, MAR or MNAR), and methods of addressing the missing data. It also recommends summarizing any records that are deleted from analysis. Having a published reporting standard may encourage researchers to consciously consider missing data's impact on their analysis, how it should be handled and how it should be communicated. Published reporting standards also encourage consistency in how data is reported, which would make the data more useful in meta-analysis and make it easier for other researchers to duplicate or build on published research (Appelbaum, Cooper, Maxwell, Stone, & Sher, 2008; Enders & Gottschall, 2011).

The DC Drinking Water Crisis; Concluding Remarks

The conclusions released in the CDC's original report on Washington DC's residential water supply were celebrated not only by Washington DC council members and water

authorities but by officials in and outside the US. The report is cited by many different governing bodies to downplay the severity or serve as justification for their lack of action in not responding to high levels of lead in their water. The implications of the missing data on society is alarming (Renner, 2009). The mechanism for the missing data is MAR, which supports the bias results of the report's findings.

Why was the report released if data was missing? One reason is growing pressure on the CDC to release the report. Another reason is that the person who ultimately released the report was under the false impression that only results below CDC's level of concern were missing, which is an ethical concern in itself (Leonnig, 2010). Whatever the reason, the report authors did not put the effort in to understand the missing data or try to track it down. Even though they knew the data was missing, they did not mention it anywhere in the report, and infact made a point of explicitly stating all data available was used. Technically this is not a lie - they used all the data in their current possession - but it is certainly misleading. This misrepresentation of data could be avoided if the JARS standard was upheld, in particular the recommendation of including causes of missing data, whether theoretical or evidence based(Enders & Gottschall, 2011). The CDC did neither.

The missing data in the DC water crisis is different from the missing attribute values or missing outcome variables due to attrition since it is clearly attributed to a lack of standards by the authors of the report. It does, however, highlight the severity of consequences when data is missing. A phone call to the labs from which the missing data would have come from was all it would have taken to resolve the problem.

Mentioning the missing data in the original report if timeliness was a concern then following up with complete information would have also been a better and more ethical approach. Just one sentence in a footnote would have been met a bare minimum requirement of disclosing the true nature of the data used. The recipients of such reports are not privy to the data used in the research, and nor should they be expected to question the results when the source is presumed to be reputable, as the CDC is assumed to be. It is the responsibility of the researchers to be forthcoming with all information on the quality of the data used, the responsibility of the authors of papers to ensure all facts are delivered and not misrepresented, and the responsibility of the academic, business and research communities to uphold and demand consistency and ethical practices are observed in all data analysis and resulting recommendations and conclusions.

Bibliography

Allison, P. D. (2002). *Missing Data*. SAGE Publications, Incorporated.

Appelbaum, M., Cooper, H., Maxwell, S., Stone, A., & Sher, K. J. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *The American Psychologist*, *63*(9), 839–851.

Belmont. (1979). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Retrieved from http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html

Burton, A., & Altman, D. G. (2004). Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer*, *91*(1), 4–8.

Census, 2000. (2004). *The 2000 Census: Counting Under Adversity*. (Panel to Review the 2000 Census, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, & National Research Council, Eds.). National Academies Press.

de Goeij, M. C. M., van Diepen, M., Jager, K. J., Tripepi, G., Zoccali, C., & Dekker, F. W. (2013). Multiple imputation: dealing with missing data. *Nephrology, Dialysis, Transplantation: Official Publication of the European Dialysis and Transplant Association - European Renal Association*, *28*(10), 2415–2420.

Enders, C. K., & Gottschall, A. C. (2011). The Impact of Missing Data on the Ethical Quality of a Research Study. In *Handbook of Ethics in Quantitative Methodology*.

Gerber, A. S., & Green, D. P. (2012). Attrition. In *Field Experiments: Design, Analysis, and Interpretation* (pp. 211–252). W W Norton & Company Incorporated.

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

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques: Concepts and Techniques*. Elsevier.

Jones, M., Mark, J., Mishra, G. D., & Annette, D. (2015). Analytical results in longitudinal studies depended on target of inference and assumed mechanism of attrition. *Journal of Clinical Epidemiology*, *68*(10), 1165–1175.

Leonnig, C. D. (2010, May 20). CDC misled District residents about lead levels in water, House probe finds. *The Washingto Post*, p. A01.

Mayo Clinic, S. (2014, June 10). Diseases and Conditions Lead poisoning. Retrieved December 12, 2015, from http://www.mayoclinic.org/diseases-conditions/lead-poisoning/basics/definition/con-20035487

Molnar, F. J., Hutton, B., & Fergusson, D. (2008). Does analysis using “last observation carried forward” introduce bias in dementia research? *CMAJ: Canadian Medical Association Journal = Journal de l’Association Medicale Canadienne*, *179*(8), 751–753.

NHTSA. (1993). *Imputation in the NASS General Estimates System* (Version DOT HS 807 985). U.S. Department of Transportation - National Highway Traffic Safety Administration. Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/807985.PDF

NHTSA. (2015). *National Automotive Sampling System (NASS) General Estimates System (GES) Analytical User’s Manual 1988-2014* (Version DOT HS 812 215). U.S. Department of Transportation - National Highway Traffic Safety Administration. Retrieved from http://www-nrd.nhtsa.dot.gov/Pubs/812215.pdf

Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. (2006). Advances in Missing Data Methods and Implications for Educational Research. In S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 31–78). Greenwich, CT: Information Age.

Renner, R. (2009). Troubled Waters, Part II: On the Trail of the Lost Data. *Professional Ethics Report (AAAS)*, pp. 1–3.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581.

Spineli, L. M., Pandis, N., & Salanti, G. (2015). Reporting and handling missing outcome data in mental health: a systematic review of Cochrane systematic reviews and meta-analyses. *Research Synthesis Methods*, *6*(2), 175–187.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies. How to deal with missing data. *Journal of Clinical Epidemiology*, *55*(4), 329–337.