Article:
Anthony, D (2015) Issues in data analysis. Nurse researcher, 22 (5). 6 - 7. ISSN 1351-5578

https://doi.org/10.7748/nr.22.5.6.s2

Commentary

Nurse researcher is the journal which publishes a wide range of methodology papers but in my experience most papers aim to give guidance to early researchers. Here I comment on two papers. **Qualitative case study data analysis: an example from practice** is certainly one of these and is clearly comprehendible by novice researchers. The second **Multiple imputation method for handling missing data: A case study of a secondary data analysis study** is a rather different case. This is a useful paper but deals with a complex issue in quantitative data analysis. If you read this paper and found it hard going then please read the following few paragraphs where I try to put this paper into context and then re-read Walani and Cleland.

To help explain this paper I am going to talk about some recent work in which I have been involved though the lessons learned could easily be taken from many other studies. In this work we attempted to prevent non-communicable diseases such as heart disease, lung disease, diabetes and some cancers but reducing four risk factors – tobacco use, low physical activity and poor diet. We did this in the general community, schools, workplaces and clinics/hospitals. For children we gave out a self-completed questionnaire to establish their beliefs, knowledge, attitudes and behaviour with respect to these risk factors. One outcome we were interested in was whether the children smoke. Another question we asked was whether the child was a boy or girl. We were interested in gender as there is a difference between boys and girls – especially in certain ethnic groups.

Not all children in the pilot stated whether they smoked. Why was this? These children were all about twelve years of age. If the children missed the question by accident or did not understand the question (say) this could be said to be random. Data from the smoking variable is missing, and is missing completely at random, which is how we could describe it (MCAR). However this seems unlikely. While we made it clear that all data were confidential a later focus group identified that these assurances were not accepted by all. The children confirmed they understood the question but were not necessarily willing to answer it. Now we know that smoking is more or less acceptable in different groups. These children in Leicester were predominately from families originating in India or Pakistan. We know that smoking is much more common and more acceptable in males than females in the ethnic groups from India. These missing data were therefore probably not randomly missing. A schoolboy who might smoke would probably not want to be identified but a schoolgirl from a family in this part of the city might be even more inclined to keep her smoking behaviour to herself if she did not trust us to keep this information confidential. These missing data are not completely at random, they are probably related to gender and ethnic group. However for schoolgirls in a given ethnic group there may be no other known reason for data to be missing and therefore within such a group the missing data may be said to be randomly missing. We call this missing at random (MAR) but not completely at random. I.e. the missingness of the outcome is related to the predictor – in this case girls are less willing to answer the smoking question than boys.

Nearly there.

If they are less likely to answer the outcome variable (did you smoke) if they had smoked then the missing data are not completely random, or random at all. They are missing not at random (MNAR). This is also called non-ignorable.

So to recap if your outcome variable is missing data dependent on its own value, e.g. if smokers are less likely to report they smoke then you have a problem and it is called MNAR. If your outcome is missing data and this missingness is related to another variable whereby data is more likely to be missing dependent on that variable (say gender) but missing randomly within each subgroup (e.g. girls less likely to report smoking but within the girls whether they answer the smoking question is random) then the problem is amenable to imputation and is MAR. If some data just happen to be missing but this missingness is not related to any variable including the outcome itself (smokers equally happy to answer than non-smokers, boys equally as girls) then it is MCAR and is also amenable to imputation.

So what does all this mean? Why does it matter?

In some cases it does not matter. If a few people do not answer a question on a survey then you can simply remove them from the analysis. However if you need to use lots of variables in your analysis you may run into problems. Say your respondents in a long questionnaire all miss the occasional question. This does not sound much of a problem. However if 100 subjects each miss a different question in a 100 question survey you would have no totally complete data on any person despite the fact that 99% of data had been collected. Some method of imputing the missing data would allow analysis to take place where otherwise in this extreme case you would not have any data at all. But how do you impute (guess) the values that are missing? There are several methods but how would you know how accurate your imputed values were? One method is to stick in a bit of random variation and repeat the imputation a few times to end up with a few datasets. You can then do whatever analysis you like and with the results of all of them you can get a range of possible values which gives you therefore a plausible confidence interval for values – i.e. you have essentially guessed the missing values, used these imputed (guessed) values in your analysis but you know that from several such imputations that likely value is between this and that.

Having said all this if you have a lot of missing data but still have a reasonably large dataset subjects with incomplete data removed then I would suggest you just remove them and then do the analysis. Listwise deletion (i.e. an entire record is excluded from analysis if any single value is missing) which is available in statistics packages such as SPSS gives accurate results whether data are MCAR or MAR. There are packages for imputation (and statistics package, for example SPSS, have such a facility – sometimes as a additional component) but the results of imputation vary and while some packages give good results others give “terrible” ones (Allison, 2000). MNAR is a more problematic area and if your data are MNAR you probably are not going to be able to do a satisfactory analysis in any event.

Now to the qualitative paper.

There is a debate as to whether computer-assisted qualitative data analysis software (CAQDAS) such as NVivo is useful or not. Some qualitative researchers for whom I have considerable respect consider them redundant. While myself a predominantly quantitative researcher, I have completed some qualitative work as it is more appropriate to answer some research questions. I have employed NVivo in such a study and found it was not useful. It is necessary to code items of text for analysis when entering them into the package. I found it as easy to use pen an ink on the printed page to code items and then generate themes. However very respected colleagues of mine find use of these packages useful if not essential.

One would not even consider conducting quantitative work without a package such as SPSS (or SAS, PSSP etc.) unless it was very simple work involving (say) descriptive statistics or some very simple inferential test like chi square. It would be horrible to conduct a regression analysis and probably impossible to conduct in one's lifetime some iterative approaches such as artificial neural networks (don't panic, I'm not going to talk about these).

However it is less obvious that qualitative software is as necessary. Houghton et al. make a case for using a qualitative package and I have no particular objection to anything they say – and they report their experience very clearly.

I would simply give the same advice that I give to all my students who want to undertake qualitative work – try the CAQDAS. Whereas I would stipulate that quantitative work needs a software package (and I would add reference management needs a reference management product such as EndNote, Procite, Refman etc.) I am not at all convinced about the need for qualitative software. However these authors are, and I am not saying they are wrong. So try it. I did and I decided it was not useful. You may find it more useful than me.

Allison, P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research, 28*, 301-309.