Subject: Probable errors in Data and its rectification
Posted by Shekhar_GS on Mon, 21 Sep 2020 07:27:35 GMT
View Forum Message <> Reply to Message

Hi everyone,

I want to know what %age of data from NFHS4 survey could be erroneous? What is a standard way (if any) to deal with it?
Example: I found that a person aged 50years having 45 children, and his youngest child is aged 25, while his age when the first child was born was 24.
This data is from IAMR74FL - Mens recode.

Specific questions that I need answers to are as following:
1. Is it uncommon to find such errors?
2. Are they (errors) supposed to be treated as outliers for analysis or does DHS recommends any method to treat those errors?
3. Can the data sets be considered as cleaned-data or DHS expects the researchers to further clean them before any analysis?

Thanks in advance.

Subject: Re: Probable errors in Data and its rectification
Posted by Trevor-DHS on Wed, 14 Oct 2020 19:11:19 GMT
View Forum Message <> Reply to Message

In response to Shekhar_GS,
1) It is not unusual to find errors in survey data. While the DHS Program does a fair amount of data cleaning for each survey, we do not check or edit all possible issues in datasets, and in fact for many issues we do not make corrections to the data. However a lot of data have been checked for consistency, at the time of data collection, and during editing of the data, and many issues are resolved that those stages. The data are considered reasonably clean, but no survey data will ever be completely clean (if it is then you know it has been fabricated).
2) Yes, it is expected that analysts will make rational decisions about the treatment of outliers in their analysis.
3) The data is generally pretty clean, but we do expect researchers to review and decide for themselves whether they need to make decisions for their analysis which might include further cleaning or might include treatment of outliers.

The case you mention is indeed impossible, but it is 1 case in more than 112,000 and will have very little effect on your analysis. There are a few others that look questionable too for similar reasons. A simple solution would be to exclude these cases from your analysis. Alternatively, you can make decisions about how you want to edit the data. Perhaps you don't believe that the person had 45 children, but they also said they had 25 boys and 20 girls. They also said that they had their first child at age 25 and the youngest is 24, which is clearly impossible, but perhaps they misunderstood the question and gave the age of their oldest rather than their youngest. This is just to note that there are lots of things to consider when making decisions about further cleaning. The simplest and easiest to defend is simply to exclude extreme outliers from your analysis.