Subject: Re: Appropriate handling of missing values in analysis
Posted by Bridgette-DHS on Thu, 08 Dec 2022 16:04:45 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

Glad the forum is helpful to you, and thanks for saying that!

The term "missing values" is ambiguous. In the data files, a blank (or a dot in Stata) means NA, or "Not Applicable". It is used, for example, if there is a skip or filter in the questionnaire such that the respondent was not asked a specific question. As used in the Guide to DHS statistics, "missing" means the variable is not NA, but there was not a valid response. Sometimes those responses are included in the denominator, but sometimes they are not.

Here is an example from the KR file for the 2016 DHS survey in Ethiopia. h11, diarrhea in the past two weeks, has a code 8 for DK. Say we want to calculate the proportion of children who had diarrhea in the past two weeks, using Ethiopia 2016. The unweighted distribution of h11 is as follows:

. tab h11

```
      had diarrhea |
         recently |      Freq.     Percent        Cum.
------------------+-----------------------------------
               no |      8,826       88.21       88.21
yes, last two weeks |      1,090       10.89       99.10
       don't know |         90        0.90      100.00
------------------+-----------------------------------
            Total |     10,006      100.00
```

"tab h11,m" will show that there were 635 NA cases. All of the NA cases are children born in the past 5 years who died before the survey. Clearly they should be omitted from both the numerator and the denominator. The denominator for the proportion would include all 10,006 cases, but the numerator would only include the "yes" responses. The "don't know" responses are in effect grouped with "no". The reason for grouping them with "no", as I think of it, is to avoid over-estimating the prevalence of the outcome. There could be other variables in which "don't know" would be grouped with "yes", but for the same reason, that we want to be conservative and avoid over-estimating the prevalence of an unfavorable outcome.

Actually, although this is what DHS usually does, someone else might want to drop the "don't know" responses from the denominator. I don't have strong feelings about that but I prefer to keep them in the denominator.

The distribution of h11 in the 2011 survey looks like this:

. tab h11

had diarrhea recently	Freq.	Percent	Cum.
no	9,068	83.90	83.90
yes, last two weeks	1,620	14.99	98.89
don't know	105	0.97	99.86
9	15	0.14	100.00
Total	10,808	100.00	

There are not supposed to be any 9's; the label for h11 does not include 9. However, for some reason, there were 15 children for whom the interviewer could not get a valid response. To get the proportion in this situation, I would prefer to drop the values with code 9. "Don't know" comes from the mother; perhaps the child is temporarily staying with the grandmother, for example, but "9" is completely meaningless. The remaining 10,793 cases would be in the denominator and (as with the 2016 data) the "yes" cases would be in the denominator. However, I'd have to go to the CSPro code to see what DHS actually did with those 15 cases. The Guide to DHS Statistics suggests that they were retained in the denominator, and I'm just saying that's not the only option.

For some variables, there will be a code 9 (or 99, etc.) that IS in the variable label but it means "refused", "inconsistent", etc. Usually I would omit them from the denominator, but this is a judgment call and I'm not 100% sure what is done during data processing.

My personal practice is usually to do a recode rather than dropping cases from the file or over-writing standard variables. For example, I would use these lines:

```
gen diarrhea=0 if h11<=8
replace diarrhea=1 if h11==1
```

This is a long answer but I hope it's clear. I make a distinction between what I would prefer to do personally and what DHS does (or may do) during data processing.