Subject: Difference DTA and SAV
Posted by victor on Tue, 20 Dec 2022 18:52:23 GMT
View Forum Message <> Reply to Message

I have noticed that data is different between SAV and DTA datasets. Is this not an error?

Specifically looking at Gambia 2013, in household member recode. The DTA dataset has an additional value in the hv140 variable compared to the SAV dataset.

As a result, the mean calculation for registered children is different between these two datasets. With the DTA dataset it is possible to reproduce the figures in the official report. With the SAV dataset this value is a percentage point higher. It seems to me that however that SAV is a better approximation as with the DTA dataset the respondents with value 9 are used in the calculation.

For the DTA file

	hv140	n
	<dbl+lbl>	<int>
1	0 [Neither certificate or registered]	3294
2	1 [Has certificate]	8494
3	2 [Registered]	1961
4	8 [Don't know]	208
5	9	398
6	NA	38336

For the SAV file

	hv140	n
	<dbl+lbl>	<int>
1	0 [Neither certificate or registered]	3294
2	1 [Has certificate]	8494
3	2 [Registered]	1961
4	8 [Don't know]	208
5	NA	38734

Subject: Re: Difference DTA and SAV
Posted by Bridgette-DHS on Fri, 23 Dec 2022 12:50:51 GMT
View Forum Message <> Reply to Message

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

Before answering your question, we would like to ask whether you are reading the dta and sav files in R? Or are you reading them directly in Stata and SPSS?

Subject: Re: Difference DTA and SAV
Posted by victor on Sun, 01 Jan 2023 18:30:39 GMT
View Forum Message <> Reply to Message

Hi Tom, thank you for the reply.

I am reading both data files in R.

Subject: Re: Difference DTA and SAV
Posted by Bridgette-DHS on Tue, 03 Jan 2023 13:34:54 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff members, Tom Pullum and Trevor Croft:

In CSPro, the package in which the data files are originally constructed, we have two different codes:

1. Blank this is the not applicable code for use when a question is not applicable.

2. 9 (or 99, 999, etc.) this is the missing value code for use when a question is applicable, but a response was not given.

9 (etc.) is used very rarely nowadays with CAPI, but was needed for paper questionnaires, and was used by data entry staff when a particular question was not NA but the interviewer forgot to enter a response.

Usually in SPSS, blank is the system-missing value, and 9 is the user-missing value. In Stata, usually both cases are converted to missing.

We are surprised that the 9 shows up in the Stata file. Perhaps the DTA file for this survey was not created in our usual way, which would convert the CSPro 'missing' to a dot in Stata.

You are reading the Stata and SPSS files with R, and we believe the conversion to R is being handled in different ways. You will get agreement if, when you read the Stata file with R, you add a line to change the 9 to a dot.

Subject: Re: Difference DTA and SAV
Posted by victor on Tue, 03 Jan 2023 14:25:00 GMT
View Forum Message <> Reply to Message

Thank you for this explanation. Do I understand correctly that in this case the SAV file provides the data in the correct manner?

If so, that would mean that there is a mistake in the final report for Gambia 2013. On page 21, the total should be 58.1 not 57.1, 15.2 not 14.9 and 73.3 not 72.0.