Subject: Re: Sample Weight for Merged Dataset
Posted by Bridgette-DHS on Wed, 24 Nov 2021 18:11:43 GMT
View Forum Message <> Reply to Message

Following is another response from DHS Research & Data Analysis Director, Tom Pullum:

First, regarding the re-numbering of v021 and v023. Here's a clumsy way to do it. Say that your surveys are numbered 1 through 8, and that the codes for v021 and v023 are always less than 1,000 (you have to check that, and modify if there are exceptions). Then you can construct new variables, cluster_id and stratum_id, defined with cluster_id=survey*1000+v021 and stratum_id=survey*1000+v023. These new variables will have unique numbers for the clusters and strata in the pooled data file.

If you leave the weights as they are now, then the total weight for each survey will just be the total sample size for that survey. This is probably the least acceptable thing to do, because sample sizes are determined by many different arbitrary considerations, such as the budget for the survey. There are two alternatives. One is to multiply v005 by a scaling factor such that the total weight for a survey becomes proportional to the population of the country at the time of the survey. The UN Population Division website gives population estimates. But if you do this, you will find that the results are dominated by the largest country, sometimes overwhelmingly. The second alternative is to re-weight with a survey-specific factor that gives equal weight to each country. I personally prefer this but not everyone does. There is a larger problem with pooled estimates. The surveys were done at different times and we almost never cover all the countries in a region or even a sub-region. Within DHS, we only construct a pooled estimate with many countries when studying something that is relatively rare within the individual countries. (Pooling successive surveys from the same country is more defensible.)

A colleague, Mahmoud Elkasabi, points out that men are often subsampled, and they often have a different age range than women. You need to take that into account. If the men have been subsampled, you need to scale up mv005. For example, it's a 50% subsample, multiply mv005 by 2. If you explore the forum, you should be able to find old postings that discuss all these issues. Good luck.