Subject: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Rean on Mon, 06 Jan 2025 06:10:28 GMT
View Forum Message <> Reply to Message

Dear DHS team and community members,

I have a question regarding applying sample weights (hv005) in the DHS datasets, particularly when calculating a representative Wealth Index (WI) for each cluster.

As I understand, the sample weights are provided at the cluster level, meaning all households within the same cluster share the same weight. However, this presents a challenge when trying to calculate a single WI value to represent the cluster. Without specific information about the sampling strategy or the relative importance of each household within the cluster, it seems that a simple arithmetic mean of the household WI values is the most straightforward approach.

However, this method assumes equal importance for all households within the cluster and does not account for potential variance within the cluster itself. As a result, it might overlook important intra-cluster disparities and may not fully represent the overall socioeconomic context of the cluster.

Given these limitations, I would like to ask:

1. Is there a recommended approach to aggregate household WI values into a single, representative cluster value while respecting the sampling design?
2. Are there additional resources or considerations regarding how intra-cluster variance can be incorporated into such calculations?

I would greatly appreciate any guidance or advice on this matter. Thank you for your time and for providing such valuable data for research.

Subject: Re: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Janet-DHS on Wed, 08 Jan 2025 19:11:22 GMT
View Forum Message <> Reply to Message

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

Yes, the sample weight is the same for all households in the same cluster. If you want a single number to describe the cluster, it would be the unweighted mean. The best number to describe dispersion is probably just the standard deviation. The following analytical study could be helpful: https://www.dhsprogram.com/pubs/pdf/AS76/AS76.pdf.

Note that the wealth index is calculated with households as units. I would recommend calculating the cluster-level means and standard deviations within the HR file, and then copying those numbers into the other files you are using.

Subject: Re: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Rean on Mon, 13 Jan 2025 05:14:59 GMT
View Forum Message <> Reply to Message

Thank you so much for your response and the helpful guidance regarding calculating the Wealth Index (WI) at the cluster level. Based on some challenges I encountered during my analysis, I have a couple of follow-up questions.

1. Overlapping Cluster Areas:
Given that GPS coordinates provided for clusters can have up to a 10km positional error, we often use a 10km × 10km square area centered around the given coordinates to ensure that the actual sampling points fall within this boundary. However, after defining these 10km squares, I observed that many clusters have spatially very close coordinates, resulting in significant overlap--sometimes as high as 80-90% between their corresponding areas. Despite this overlap, these clusters' mean WI values often differ significantly. This raises concerns about the representativeness of using the mean WI as the cluster-level indicator. I want to ask if DHS considered such overlapping cluster areas during the survey design. If so, how such scenarios are typically handled to ensure the validity of cluster-level WI values?

2. Weighted vs. Unweighted Mean:
In my work on predicting cluster-level WI using remote sensing data, I noticed that using the weighted mean of WI values often leads to better results than the unweighted mean. However, you mentioned previously that the unweighted mean is the recommended approach for cluster-level WI calculations. Could this observation be a coincidence, or does it suggest that the weighted WI might still have some relevance or utility at the cluster level, despite being theoretically less representative in this context?

I greatly appreciate any insights or advice you could provide on these issues. Thank you for your time and support in helping researchers like me better understand and utilize DHS data.

File Attachments

1) overlap_summary.txt, downloaded 8 times
2) cluster_means_corrected.csv, downloaded 8 times

Subject: Re: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Janet-DHS on Wed, 15 Jan 2025 13:51:43 GMT
View Forum Message <> Reply to Message

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

These are good questions. The sample design does not take into account the locations of the clusters within the strata. They are not required to be spaced apart, for example, with a minimum distance between them. As a result, sometimes they are closely spaced. This would usually reflect variations in population density within the stratum. When the clusters are displaced, they are kept within the same level 2 areas. What you are finding is quite possible and "legal" within the

sampling design. Not sure how you can handle it....

I don't think weights should play into the construction of the cluster-level WI. Say that you calculate the mean value of hv270 for the households in a cluster. Since all households in the cluster have the same weight, that mean will be the same whether or not you use weights. But then when you use the cluster-level mean in your analysis, you need to weight it by hv005 for the cluster.

I can imagine analyses in which the number of households in the cluster could be important. Could that account for the difference in fit that you are getting?

Subject: Re: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Rean on Fri, 17 Jan 2025 10:30:08 GMT
View Forum Message <> Reply to Message

Thank you so much for your insightful response!

Regarding the use of weighted and unweighted means for representing the cluster-level Wealth Index (WI), I would like to confirm my understanding. Since all households within a cluster share the same weight, I am using two approaches to derive a representative value for the cluster's wealth level for machine learning training labels. The two methods are:

1. Unweighted mean WI: The average WI for all households within the cluster.
2. Weighted mean WI: The average WI of all households within the cluster, weighted by the cluster's associated weight (hv005).

In my machine learning experiments, I observed that the model's performance was slightly better when using the weighted WI as training labels rather than the unweighted mean WI. Could the analysis you mentioned, which involves weighting by the cluster's weight, explain this result? In other words, using the weighted mean WI as the label for the clusters' overall wealth status might provide a more representative and useful feature for training the model.

I greatly appreciate your time and valuable insights, which have truly helped me a lot regarding this issue!

Subject: Re: Calculating a Representative Wealth Index for Clusters Using DHS Sample Weights
Posted by Janet-DHS on Tue, 21 Jan 2025 21:06:32 GMT
View Forum Message <> Reply to Message

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

I'm not sure that I understand the difference between your two calculations, but I will suggest a requirement or criterion: the weighted mean of the RWI should be the same as the weighted mean

of hv270, in each cluster and in the sample of households as a whole. Here are the Stata commands I would use, illustrated for the Kenya 2022 survey, and within that, for cluster #1.

```
use "...KEHR8CFL.DTA" , clear
keep hv001 hv002 hv005 hv270

* Construct RWI
egen RWI=mean(hv270), by(hv001)
list if hv001==1, table clean nolabel

* compare the weighted means of hv270 and RWI within a specific cluster
summarize hv270 RWI [iweight=hv005/1000000] if hv001==1

* compare the weighted means of hv270 and RWI in the entire sample
summarize hv270 RWI [iweight=hv005/1000000]
```

If you are doing something other than this, I wonder what it is. If you are just comparing weighted and unweighted estimates, then it is not surprising that there will be some differences in any analysis. My preference would be for weighted estimates. Here I do not use weights to calculate RWI but DO use weights in any statistical analysis, exactly as I would do with hv270.
