Subject: Correct weight for a sub-sample
Posted by katv on Thu, 09 May 2013 16:09:18 GMT
View Forum Message <> Reply to Message

Hi,

I am working with the Cambodia DH surveys. My sample consists of the children for whom there are both anthropometric (hw variables) as well as nutritional (v414 or v469, depending on the survey year) information. The sample boils down to under 24-month olds who are the youngest child in their household. What is the appropriate weight to use in this case? v005?

Thank you in advance!

Subject: Re: Correct weight for a sub-sample
Posted by Bridgette-DHS on Tue, 28 May 2013 14:40:26 GMT
View Forum Message <> Reply to Message

Using sample weights. DHS sample weights are used in almost every tabulation in DHS final reports. The few unweighted tables are clearly labeled. Sample weights are described fully in the Guide to DHS Statistics but briefly, weights are used in all analyses to make sample data representative of the entire population. There are different weights for different sample selections/units of analysis:
Sample weights in DHS datasets
Unit of analysis Variable
Households hv005
Household members hv005
Women or children v005
Men mv005
Domestic Violence d005
HIV test results hiv05

like other variables in DHS datasets, decimal points are not included in the weight variable. Analysts need to divide the sampling weight they are using by 1,000,000. Examples:

In Stata:
generate wgt = v005/1000000
tab var [iweight=wgt]
In SPSS:
COMPUTE WGT = V005/1000000.
WEIGHT by WGT.

In SPSS:
WTVAR=V005/1000000
WEIGHT BY WTVAR

Subject: Re: Correct weight for a sub-sample
Posted by adi.greif@yale.edu on Wed, 29 May 2013 19:13:47 GMT
View Forum Message <> Reply to Message

I would also like an answer to this question! I am interested in looking at rural women who rank poorest on the wealth index for various datasets.

I know that if you are using STATA, you can use svyset (and set the sample weight for the overall sample to v005/100000), and then the subpop option to specify the group you are interested in looking at. STATA will handle the weighing of the subpopulation. However, in your case, if you are looking at all non-missing answers to certain variables, I think you don't need to specify a subpopulation. I am basing this answer on the examples given here:
http://www.stata.com/features/survey/svy-survey.pdf

Does anyone know how to handle weighing sub-samples without using STATA's subpop command?

Subject: Re: Correct weight for a sub-sample
Posted by Reduced-For(u)m on Thu, 30 May 2013 01:42:22 GMT
View Forum Message <> Reply to Message

I think it totally depends on what sub-population you are looking at. If you are looking at the bottom quintile of the asset index, I'm not sure that that is the kind of sub-population you want to weight in terms of probability of sampling. They are by definition (I think) the lowest 20% of scores from a principal component analysis. I don't think those are weighted in any probability sense when computed (meaning, if you tab out the quintiles, I think you actually just see 20% in each bin, unweighted - though in some of the newer surveys I think they do this differently between rural and urban households, but that is adding on another layer of complexity).

So I'm just not sure that there are "Nationally Representative Weights for the Bottom Quntile of Household Asset Index". What would "nationally representative" mean in that context?

As for just the STATA question though, one thing you could do is something like this, which would preserve the relative probabilities implied in the DHS weights for your sample.

```
***begin
gen preweight = v005/100000
keep if assetquintile==5
egen weightsum = total(weight)
gen newweight = preweight/weightsum
```

*now you have weights that add up 1 for the group you wanted, proportional to their original weights. I'm not sure the interpretation is just what you wanted, but I'm not sure there is a perfect interpretation of what you want either.

*now, for your regressions, you can either just set this as the new weight in the same svyset manner
*Another option would be to directly specify the estimating procedure using [pweight=weight] and an appropriate clustering level.

***End

Interestingly, I think this answers your other question too, about cross-country stuff. Since the DHS weights sum to N (sample size) you need to re-normalize them so they all add up to 1 for each country...or, depending on what you are trying to do in that cross-country thing, maybe re-scale them again so that they sum up to Population. Depends on the parameter you are trying to estimate and the assumptions you're willing to make. We can pick this up in the other thread you commented on if you'd like.

Subject: Re: Correct weight for a sub-sample
Posted by adi.greif@yale.edu on Thu, 30 May 2013 02:19:25 GMT
View Forum Message <> Reply to Message

Thank you so much for the fast response, it is tremendously helpful. I had considered the procedure that you wrote down but I wasn't sure it work--- I don't quite understand how STATA is estimating the standard errors using svyset, and was afraid this would introduce some sort of problem given all the warnings about using subpop rather than dropping observations.

I'll follow up on the cross-country question here -- again, thank you so much for offering to help. I am interested in estimating differences in variables (such as the average proportion of women working in agriculture) across former British and former French colonies. I'm taking one survey from each country to make weighing easier for me to understand. I want each survey to be considered as an observation (even though different countries have different population sizes). So I think in this case you are right that I need to renormalize the DHS weights so they all add up to 1, rather than using the total population size to renormalize. You wrote in another post that it depends on "What do you want the thing you are estimating to represent? An average of all the people that live in those five countries? An average that thinks of a country as an individual? An average that represents the people sampled in the DHS?"

I think I am going after the second estimation you mentioned "An average that thinks of a country as an individual" and that's why I should normalize to 1 rather than using the total population. Do you agree? I am not sure what you meant by it depending on the assumptions I am willing to make?

Many thanks,
Adi

Subject: Re: Correct weight for a sub-sample
Posted by Reduced-For(u)m on Thu, 30 May 2013 03:03:23 GMT

Hi Adi,

You are right that normalizing each survey so the weights add to 1 would end up treating each country as an observation weighting-wise (and would still preserve the within-country sampling probabilities, so each country would represent a 1 that is a weighted average of it's population - man, this stuff is always a mouthfull).

As for the assumptions thing - this whole weighting bit is really about two different things in a regression context (as opposed to a tabulate means context). First is population weighting - to make the survey nationally representative. The second is efficiency - if you have observations that are like means (say, a state-by-year panel where states have different populations) you might want to weight up the populous states not for representativeness but for smaller standard errors (efficiency). There is a good paper called "What are we weighting for" which is here if you have access: http://www.nber.org/papers/w18859

Basically, before I suggest any weighting scheme, I just want to know why people are weighting. In this case, I think if you are happy treating each country as an observation (in the weighting sense) then you are fine. It gets a bit metaphysical at times, and I don't have all the answers by any stretch, but I've been trying to figure out some of these issues in my own cross-country stuff, so I'm also trying to figure out what other people are thinking when they weight. Somehow to me the idea that one survey is weighted the same as another survey seems reasonable enough, but there would be lots of people who think that they should be population weighted (a weighted average of heterogeneous treatment effects), and others who would say to weight those surveys by Pop to get more efficient estimators (supposing homogenous treatment effects). I think as long as you are clear, they are all fine (with the "weighting for efficiency" being the most suspect).

Here's a little thought experiment: if every single person in every single survey had the exact same response to your X of interest - how would you want to weight? I think at that point, I'd just weight everyone with 1, because a person is a person. This is totally different than tabulating population means.

Subject: Re: Correct weight for a sub-sample
Posted by adi.greif@yale.edu on Thu, 30 May 2013 03:14:56 GMT

Thank you again for your very helpful and timely response!

Best,
Adi

Subject: Re: Correct weight for a sub-sample
Posted by bsayer on Mon, 08 Jul 2013 16:22:12 GMT

The use of subpop in Stata has nothing to do with weights and everything to do with intra-cluster correlation. You need one observation per stratum-PSU combination to accurately calculate this. The best thing to do is use all the observations and the subpop option.

There are some situations where there might be some alternatives. If for some reason you think you are in those situations, you should study the issue carefully. I doubt that you will get a completely correct answer in a forum.

For a variable that represents percentiles of an entire population, it should have been weighted when it was created. If you want to create a new percentile, then you will need to create a weighted version for the population that you are interested in. For example, if you want the percentile of women ages 20 to 25 that have never had a child, you would use that population and the corresponding weight for women. This is because different women have a different probability of being selected in the survey (typically urban women have a higher probability, for example). So if urban women have a higher probability of not having had a child, then we need to account for both the probability of selection and the probability of not having had a child.

I would suggest something like small area estimation for these types of problems.