Subject: Two or Three Sampling Stages
Posted by lukassg on Tue, 01 Apr 2014 18:13:47 GMT
View Forum Message <> Reply to Message

Hey!

Could someone explain me exactly what is the difference between three-stage and two-stage sampling?

I very well understand the two-stage sampling:

1. I select a certain number of enumeration areas based on a previous census or national survey.
2. I sample a certain number of households from these EAs/clusters.

For the three-stage sampling:

As I understand it, this stage is before the EAs/clusters are selected. So we have even larger areas in the first stage, from which we then select the EAs/clusters in the second stage and the households in the third stage? Why do we have this additional stage in the beginning?

Finally, I am pooling always two different DHS survey for one country to get an over-time estimator of certain effects. I know that I have to adapt the stratification and PSU specifications. Is that enough in the case of a country, e.g. Tanzania, where I pool the years 1999 and 2010, where one year has 3 stages and the other 2 stages? Or does that make it impossible to analyse?

In this context, I also stumble upon probability proportional to size (pps). Can someone explain me this concept as well? :)

Thanks for your help!

Lukas

Subject: Re: Two or Three Sampling Stages
Posted by lukassg on Wed, 02 Apr 2014 21:35:35 GMT
View Forum Message <> Reply to Message

I have a quick on-top question on my previous post:

Strata are mostly multi-level with region at the first-level and urban-rural at the second-level. What exactly is the difference then between a survey domain (population subgroup) and a strata (subgroup of homogeneous units)?

Thanks

Subject: Re: Two or Three Sampling Stages

Posted by Bridgette-DHS on Fri, 04 Apr 2014 12:32:49 GMT
View Forum Message <> Reply to Message

Following is a response to msg_1710, from DHS Senior Sampling Specialist, Ruilin Ren:

1. About the sampling procedure
Most of the DHS surveys use a two-stage cluster sampling procedure and use the latest population census frame. The primary sampling unit (PSU) is the census "Enumeration Area" (EA). After the selection of the EAs, there is a household listing operation. The household listing operation aims to visit all the EAs selected in the first stage and construct a complete list of all residential households in the EA. In the second stage, a certain number of households are selected from the updated list of households. All the selections are random selections.

In the case where a census frame is lacking, like in the Democratic Republic of Congo, there is often a need to select the sample in three stages. In the first stage, a certain number of PSUs (communes) will be selected. After the selection of the PSU, a list of villages (SSU, second sampling unit) in the selected communes will be established. At the second stage, one village will be randomly selected from the list. Then proceeds the household listing in the selected village. In the third stage, the sample households are selected. In summary, when there is no census sampling frame available, the sampling procedure might be multistage, that is, more than two stages.

2. About pooling two surveys
I do not think the sampling procedure matters if pooling different surveys together. But attention should be paid to the sampling weight if the surveys use normalized weights, such as for all the DHS surveys. Since the normalized weight has no unit, it is survey specific. You need to de-normalize the sampling weight first before pooling. De-normalizing means to divide the sampling weight by the overall sampling fraction (number of units selected over the total number of units in the target population), for both household weight and the individual weight, respectively. For example, divide the weight by the ratio of households in the sample for the survey to households in the population, and similarly for women.

3. Probability proportional to size (pps)
This is a sampling procedure in which the chance of a unit being selected in the sample is proportional to the measure of size of the unit. For example, the DHS sampling procedure selects the EAs in the first stage often using PPS sampling, and the measure of size of the EA is usually the number of households in the EA. This means that EAs with a large number of households will have a larger chance to be selected in the sample compared to EAs having a small number of households. This sampling strategy aims to increase the efficiency of the sampling and reduce the sampling errors.

Subject: Re: Two or Three Sampling Stages
Posted by Bridgette-DHS on Fri, 04 Apr 2014 12:33:51 GMT
View Forum Message <> Reply to Message

Following is a response to msg_1787, from DHS Senior Sampling Specialist, Ruilin Ren:

Stratification means to group similar units into a group (sampling stratum) in the sampling stage, and design and select an independent sample for each sampling stratum. This is aimed at strengthening the representativity of the sample with a given total sample size. Region crossed by urban and rural are frequently used as stratification for DHS surveys. Stratification is for sample selection purposes. There is no restriction on the stratum sample size once it is bigger than 2.
A survey domain is a subpopulation (geographically defined or social-economically, or demographically defined) where most of the survey indicators will be reported for the domain. Regions are usually survey domains in the DHS survey. Survey domain is for reporting purposes. A survey domain can cover many sampling strata, like the regions covering two strata (urban and rural). Since we pay attention to the precision of the indicators calculated/reported, usually there is a minimum sample size requirement for a survey domain. For example, for the DHS surveys, in high fertility countries, DHS requests at least 800 women interviews per domain in order to produce reliable estimates of fertility and childhood mortality at domain level.

Subject: Re: Two or Three Sampling Stages
Posted by user-rhs on Wed, 16 Apr 2014 01:10:57 GMT
View Forum Message <> Reply to Message

Bridgette-DHS wrote on Fri, 04 April 2014 08:32Following is a response to msg_1710, from DHS Senior Sampling Specialist, Ruilin Ren:
[...]
2. About pooling two surveys
I do not think the sampling procedure matters if pooling different surveys together. But attention should be paid to the sampling weight if the surveys use normalized weights, such as for all the DHS surveys. Since the normalized weight has no unit, it is survey specific. You need to de-normalize the sampling weight first before pooling. De-normalizing means to divide the sampling weight by the overall sampling fraction (number of units selected over the total number of units in the target population), for both household weight and the individual weight, respectively. For example, divide the weight by the ratio of households in the sample for the survey to households in the population, and similarly for women.
[...]

Could someone at DHS reconcile the advice given above with the advice from Tom Pullum from an earlier posting (link: http://userforum.dhsprogram.com/index.php?t=msg&th=136&goto=260&S=608aa561007d1aea931e056a867781f1#msg_ 260)?

Bridgette-DHS wrote on Thu, 04 April 2013 14:28Here is a response from one of our DHS Stata experts Tom Pullum, that should answer your question.
[...]

The weights should be ok. Sometimes surveys from several countries are pooled, and then the weights may need to be changed by a different multiplier for each survey.
[...]

In both cases, the OP asked about pooling different survey waves from the same country (here: Tanzania 1999 and 2010, and in the linked posting Philippines 1998, 2003, and 2008). My understanding is that the weights add up to the sample size for that wave. I'm trying to think of the implications for using the weights as-is. In addition, what is a reliable source of age-sex-specific population size/number of households for the particular survey year needed to de-normalize the weights? Would the PRB have it? World Bank? DHS Program website?

Thanks,
RHS

Subject: Re: Two or Three Sampling Stages
Posted by lukassg on Wed, 16 Apr 2014 13:45:34 GMT
View Forum Message <> Reply to Message

Hey,

cool that you are onto a similar issue, I just posted a related question in the weighting forum, check here

http://userforum.dhsprogram.com/index.php?t=tree&goto=2028&S=3112776f03aedf45db3e342717b963e8#msg_2028

would be a lot easier if one can just use the weights as if. And as for a reliable source, I looked into it and think the respective censuses are the best approach since the DHS survey are mostly based on them

---

Subject: Re: Two or Three Sampling Stages
Posted by Trevor-DHS on Wed, 16 Apr 2014 15:07:25 GMT
View Forum Message <> Reply to Message

Both of the messages from Tom and Ruilin are saying essentially the same thing, just in different ways. When combining surveys, you need to adjust the weights so that they are comparable. The easiest way to do that is something like:

new Weight = survey Weight * (Population N / Survey N),

or as Ruilin wrote it "divide the weight by the ratio of households in the sample for the survey to households in the population, and similarly for women":

new Weight = survey Weight / (Survey N / Population N)

The two are equivalent.

The survey Weight and the survey N you have from the dataset. For the population N, you can choose your source of preference. For myself, I usually go to the UN's World Population Prospects: http://esa.un.org/wpp/unpp/panel_indicators.htm . If I am analyzing women, I would pick "Women aged 15-49", Medium variant, and pick the year most appropriate to my analysis for the country.

I hope this helps.

Subject: Re: Two or Three Sampling Stages
Posted by user-rhs on Wed, 16 Apr 2014 15:14:01 GMT
View Forum Message <> Reply to Message

Quote:And as for a reliable source, I looked into it and think the respective censuses are the best approach since the DHS survey are mostly based on them

Ideally, we would use the census figures, but in my experience, the amount and quality of information on the respective Bureau of Statistics/Ministry of Health websites varies by country. Sometimes the age range breakdown is 15-64 instead of 15-49. Also, the census won't be available for all the years that you need. Even in the U.S., the census is done every 10 years with intercensal estimations done annually.

I wonder if DHS can make these figures available for all DHS surveys ever conducted, since the general advice is to de-normalize the sampling weight for pooled analysis. I haven't looked for census HH numbers (haven't had to), but my hunch is that it would be harder to come by than the # of individuals.

One limitation I can think of from relying on the census is that although the most recent census is used as the sampling frame, sometimes it is necessary to make adjustments in the field on the actual # of HH, especially if some time has passed since the census. In stable populations, this is probably OK, but where there is a lot of in and out-migration or population growth/loss, this might be an issue.

Subject: Re: Two or Three Sampling Stages
Posted by user-rhs on Wed, 16 Apr 2014 15:22:32 GMT
View Forum Message <> Reply to Message

Trevor, Tom said "The weights should be OK," which made it seem like no adjustments needed to be made, unless data were pooled from different countries.

Thanks for the UN link. What would you suggest as a source of household N?

Subject: Re: Two or Three Sampling Stages
Posted by lukassg on Wed, 16 Apr 2014 18:17:34 GMT
View Forum Message <> Reply to Message

two things:

1. Do you think it's better to use the UN data or the census data? This is for the case when I have a DHS survey from e.g. 2009, the census is from 2007 and the UN data from 2005 (or 2010).

2. Once I changed the weights with the given calculations and pooled the surveys, is there a follow up step with the pooled dataset? I.e. do I have to re-normalize the weights?

Thanks

Subject: Re: Two or Three Sampling Stages
Posted by Trevor-DHS on Wed, 16 Apr 2014 21:50:52 GMT
View Forum Message <> Reply to Message

To user-rhs:
I think the weights may be ok if you have two samples from the same country that have roughly the same sample size. If the sample size differed substantially between the two surveys or you are using data from two different countries then I would say that you need to adjust. Also note the caveats in Tom's note referenced below.

I rarely use a household N, but if you wish to you again have a number of possibilities. Some DHS samples are prepared with households as the measure of size and the total number of households may be reported in appendix A of the DHS report. If the measure of size was total population instead of households you could use the total population reported in the DHS report divided by the total de jure sample population from the dataset. You can also get an estimate of the total population from the UN web site that I referred to earlier. Counts of households aren't reported as widely as population counts, but there are some sources. For example, Wikipedia even has total households (http://en.wikipedia.org/wiki/List_of_countries_by_number_of_ households). The source for this appears to be the UN Demographic Yearbook (http://unstats.un.org/unsd/demographic/products/dyb/dybcensu sdata.htm (look for "households by type of household").

Also, I should mention an alternate source for population numbers - the US Census Bureau's International Data Base: http://www.census.gov/population/international/data/idb/info rmationGateway.php

To lukassg:
1) I don't think there is a right or wrong answer. I think you have to decide which is the most appropriate. You could decide to use the 2010 number as that is close to 2009, you could use the 2007 census number, or you could interpolate between the UN's 2005 and 2010 numbers. I don't think it will really matter in an meaningful way. Just document what you actually do in your analysis writeup.

2) You don't need to re-normalize weights - it will have no effect on the results of your analysis, other than producing much bigger Ns.

I hope this helps.

Subject: Re: Two or Three Sampling Stages
Posted by lukassg on Sun, 27 Apr 2014 10:51:30 GMT
View Forum Message <> Reply to Message

Hey, I have one follow up question:

1) When de-normalizing the weight, I need to use the number of women 15-49 interviewed in the survey. I am using the Birth Recode file.
Is it correct if I do 'codebook caseid' and then use the number of unique values as to total number of women interviewed? Or is it better to pick the number from the final reports? If so, is that from Appendix A, number of eligible women, correct?

Subject: Re: Two or Three Sampling Stages
Posted by Bridgette-DHS on Mon, 28 Apr 2014 14:37:32 GMT
View Forum Message <> Reply to Message

Following is a response from our Senior Sampling Specialist, Ruilin Ren:

The total number of women interviewed in a DHS is given in many places in the final report, for ex, in chapter 1, table: "Results of household and individual interviews", and in the table: "Background characteristics of respondents" in the individual chapter.

You can also get the total number of women interviewed by taking all cases where V015=1. You should use the number of women interviewed, not the number of eligible women.

Subject: Re: Two or Three Sampling Stages
Posted by lukassg on Mon, 28 Apr 2014 15:05:17 GMT
View Forum Message <> Reply to Message

thanks a lot for your reply, thats helpful :)
