**The DHS Program
Webinar on Analyzing DHS Data: Weights and Other Adjustments for the Survey Design**

**Wednesday, June 3, 2015, 10am EST**

**Script**

***Erica Nybro:***

Welcome, Everyone. I am Erica Nybro, the Virtual Communication Coordinator for The DHS Program. Thanks for joining us for DHS’s first-ever webinar on “Analyzing DHS data, Weights and other adjustments for the survey design”. We received dozens of questions on The DHS User Forum, and we are going to try to address many of them today. These questions have all informed the presentations that you’re about to hear from our 3 experts; other questions will be addressed directly after their presentations. In about 30 minutes we will also start taking follow-on and additional questions via the chat function in the Adobe Connect window. Please note that we will not be using the participant microphone function today, so please type any questions or comments into the chat window and we will do our best to respond.

If your question is not answered today, please rest assured that it will be answered on The DHS User Forum in the next week. A recording of today’s webinar, along with the PowerPoint presentation will All also be posted on the User Forum, the URL of which will be posted at the end of the presentation.

Before we get started, we would like to announce that we already have another webinar planned for July. The topic will be GIS data, spatial data. Please stay online with us until the end of the session when we will be opening up a survey window to ask for your feedback about today’s session and solicit suggestions for future webinars.

We’re going to start with a brief presentation from DHS samplers Drs. Ruilin Ren and Mahmoud Elkasabi on sampling and the DHS survey process.

***Mahmoud Elkasabi:***

Thanks Erica!!

Hello everyone. In the following minutes, before we answer your questions, I will share with you the main aspects of the sampling design in most of the DHS surveys.

In most of the DHS surveys, we use a Stratified Two-stage cluster sampling design.

In the first stage, we select a probability proportional to size sample of Enumeration Areas.

In the second stage, within the selected enumeration areas, a systematic sample of households is selected.

Now, you might ask, what is the enumeration area? The Enumeration Area is a geographical statistical unit created for a census and containing a certain number of households. A typical Enumeration area can be a city block or a village.

To select a sample of enumeration areas, we need a sampling frame. Usually we use a frame of Enumeration Areas prepared for the most recent population census. The selected EAs are called sample clusters.

After the selection of the sample clusters, we need to perform a Household listing process, in which all the residential households in the selected clusters are listed to create a frame of households for the second stage of selection.

Finally, after the households’ selection, we interview all the eligible members.

**Change slide**

Now moving to stratification aspect of the design; Stratification is the process by which the survey population is divided into subgroups or strata that are as homogeneous as possible.

In most of the DHS surveys, the cross-classifications between the country administrative regions or sub-divisions by urban and rural residence compose the sampling strata.

You always can check the sampling design in Appendix A in the DHS report to learn more about the sampling strata.

As you can see in the figure, we have four sampling strata; Within each stratum, the sample is designed and selected independently.

Now note that Strata should not be confused with survey domains. A survey domain is a population subgroup for which separate survey estimates are desired (for example, urban areas and rural areas).

Survey domains and strata may or may not be the same. For example, in the previous example, if the survey estimates are desired for each region, the country regions are considered as survey domains and enough sample size should be assigned for each region. However, if the survey estimates are desired for urban and rural areas separately within each region, the survey domains and the sampling strata will be the same.

**Change slide**

Introducing the survey domain concept naturally takes us to the sample size required to produce the survey estimates.

From our previous experience with the DHS surveys we know that to produce reliable Total Fertility Rate and Mortality Rates, we need about 800-1000 completed women per survey domain.

Other indicators in the DHS may require smaller sample size.

The total sample size should consider the number of domains.

Sample allocation has to take this minimum sample size into account. In other words enough sample size should be allocated for each design domain.

**Change slide**

In the following example, we will see how to allocate a sample of 4000 women over 4 sampling strata.

Assuming we have a country of 10 million households distributed as in the table.

For sake of simplicity, let’s assume that we have on average one woman per household and that we have a 100% household and individual response rate.

**Change slide**

Using the proportional allocation, the sample will be allocated proportional to the size of each stratum. If the survey estimates are desired for the two regions only, this allocation will be satisfying.

However, if the survey estimates are desired for urban and rural areas within each region, the survey domains and the design strata are completely coincide. And so the proportional allocation will not be enough to achieve the survey objectives.

**Change slide**

To guarantee enough sample size for each survey domain, we should “under-sample” the large domains, the first two, and “over-sample” the small domains, the last two. This can be done by the Power allocation, where the sample allocation is proportional to a powered value of the household population, the power can be any value between 0 and 1.

As you can see in the table, now all the domains have enough sample size to produce the DHS key indicators.

**Change slide**

Assuming that we will select a fixed number of households, 20 households per cluster, the following will be the cluster allocation over the sample strata.

**Change slide**

So far, we have covered the main aspects of the sampling design. now we will move to the sampling weights, which are reflection of the sampling design.

To understand how the weights work, we should note that each selection stage is accompanied by selection probabilities; these probabilities should be considered in calculating the sampling weights. As we can see in the figure, we have a first stage selection probability P1 and a second stage selection probability P2; we also have a household response rate, which can be considered as a “response probability” and should be used in the weights calculations.

**Change slide**

Once we select the sample clusters, in the first stage, and select the sample households, after the household listing, in the second stage, we will have selection probabilities P1 and P2 for each sample cluster.

The multiplication of P1 and P2 is the overall selection probability P

**Change slide**

The inverse of the overall selection probability P is the design weight D. this weight can be calculated even before the field work.

During the field work, some of the selected households may refuse to participate in the survey or might not be at home during the survey time. To adjust for the impact of not interviewing these households, we divide the design weights by the stratum level response rates.

**Change slide**

Finally, as a DHS tradition, we normalize the weights. The main objective of the normalization is to make sure that the sum of the weighted cases equals the sum of the actual completed cases.

Note that the normalized weight is a relative weight, not valid for estimation of totals, and not valid for pooled data.

Finally, note that the weights are Case-dependent not variable-dependent, which means that the survey weights can be used to produce any kind of indicators using any study variables. However, you need to use the proper weights, normalized or un-normalized.

***Tom Pullum:***

This part of the webinar will overlap a little with what you have just heard, but now we are mainly interested in how to adjust for the design effects. I am a Stata user, and will use some Stata terminology, but will try to be more general.

Question: In the “Guide to DHS Statistics” it is not recommended to do weighted analysis if one is attempting regression analysis, but I have come across another point of view….

DHS has a long history. When the “Guide to DHS Statistics” was prepared, the main statistical packages could not include weights in all types of statistical analysis, and could not adjust at all for clustering and stratification. Please disregard that comment in the “Guide to DHS Statistics”. The advice in the next version, which is being prepared now, will be much different.

Most users of DHS data, including analysts at DHS, always adjust for weights, clustering, and stratification. Some users, usually economists, prefer not to use weights. They have reasons for this, but we do not share their point of view.

There are exceptions. If your purpose is to examine the sample, in its own right, rather than to use the sample to make population estimates, then the design effects can be ignored. For example, data quality analyses will usually give equal weight to every observation. When checking recodes, looking for extreme values, verifying skip patterns, etc., these adjustments do not need to be made.

**How a DHS survey differs from a Simple Random Sample, and what the adjustments do**

Question: Several people have asked about how DHS data differ from simple random samples (SRS) and how the adjustments correct for that.

This is really the fundamental issue for the “analysis” part of this webinar.

All standard statistical estimation and testing procedures are based on the SRS assumption. A simple random sample (SRS) has two defining characteristics. First, all cases in the population have the same probability of appearing in the sample. Second, each case is sampled independently of all other cases. A multi-stage cluster sample such as a DHS sample deviates from these assumption.

My answer will be in three parts, on the weight adjustment, the cluster adjustment, and the stratification adjustment. I will try to include some user questions within this overview, but then will return to some additional questions.

**The weight adjustment**

The population is stratified into geographically defined areas for which separate estimates of the outcomes are desired. Usually the strata are all urban and rural combinations of the “region” variable—in other words, all combinations of hv024 and hv025 (or v024 and v025). Then, relatively speaking, the larger strata (in terms of population size) will be under-sampled and the smaller strata will be over-sampled. That is, the distribution of the sample, across the clusters, will be more equal, or uniform, than the distribution of the population.

To compensate for these unequal probabilities of selection, the over-sampled areas should be “weighted down” and the under-sampled areas should be “weighted up”. If you do not weight, then all the means, proportions, rates, regression coefficients, etc., will be biased toward the over-sampled areas. But if weights are used, then means, etc., will be unbiased estimates of the population values.

As described earlier, by Ruilin and Mahmoud, the weights depend on more than just the strata, but all components of the calculation are intended to eliminate bias from the weighted estimates.

Question: Could you please talk about weighting for different types of outcomes? For example, for diarrhea or for malnutrition….

Weights are specific to the cases. They have nothing to do with the outcomes or predictors. This is a common misunderstanding of the weights. To repeat: the weight is a characteristic of a case, and is not linked to the variables, except for the way in which certain variables may only apply to certain units (household, woman, man, child, couple, etc.).

Question: How to use the weights when sub-sampling for analysis—for example, only using women of a specific age range?

The same weights would be used for a subsample that are used for the full sample. All of the design effects apply in the same way. In Stata, the weights are always re-normalized so that the sum of the weighted cases is the same as the sum of the observed cases. You do not have to do that step.

**The cluster adjustment**

Clusters are enumeration areas (EAs), generally equivalent to villages in rural areas or neighborhoods in urban areas. The reason for using clusters as the primary sampling units (PSUs) is to reduce the cost of data collection. However, when clusters are used, the SRS assumption of independent observations is violated. The clusters are independent but the households within the same cluster are NOT independent. They tend to be similar. There tends to be less variability in a cluster sample than in a simple random sample. For example, because the clusters are sampled from within the strata, and the strata are all urban or all rural, everyone in the cluster will be urban or everyone will be rural—there is NO within-cluster variation in this characteristic. There may be relatively little variation in type of source of drinking water, for example. For other variables there may be a lot of variation within the cluster, but there are almost no variables for which the variability is unaffected.

The adjustment for clustering is equivalent to scaling the non-SRS sample down to a comparable, but smaller, SRS sample. I think of this adjustment as reducing to the “effective” sample size. In the “smaller” or “effective” sample, the values of the means, proportions, rates, coefficients, etc. will be unchanged. The only effect is on the standard errors and confidence intervals. What you USUALLY will see, when you add the cluster correction, is a tendency for the following changes in the results:

--standard errors will increase;

--test statistics will be closer to zero;

--p values will increase (become less significant);

--confidence intervals will get wider.

The cluster id code is v001 or v021. (These are exactly the same.) In the various recode files there will be a prefix h or m. The couples file includes v001 and mv001 but they are exactly the same and you can use either one. The code is embedded within hhid and caseid. In the AR file the name is hivclust.

**The stratification adjustment**

A stratified sample is actually better than a Simple Random Sample, because you force the distribution of the sample across strata to match the distribution of the population across the strata (strictly speaking, to match the distribution of the sampling frame). This is not affected by the under-sampling or over-sampling. You are making the sample more efficient, effectively increasing the sample size. The adjustment will not alter the estimates of means, coefficients, etc. As with the cluster adjustment, the only effect will be on the standard errors of estimates. However, the effect of the stratification adjustment is THE OPPOSITE of what was listed above for the clustering adjustment. The stratification will USUALLY tend to REDUCE the standard errors.

I encourage you to test the effect of these adjustments for specific models. First, for example, do a regression with no adjustments for the design. Repeat with just the weight adjustment. Repeat with just the cluster adjustment. Repeat with just the stratification adjustment. Repeat with combinations. Compare the results in terms of coefficients, standard errors, the significance of covariates, and the overall measures of fit. This exercise will give you an idea of the impact of the adjustments.

Some more specialized questions:

Question: How to make these adjustments in Stata?

The weight adjustment can be made by itself, preferably with “pweight”. For example, “regress y x [pweight=v005]”. With pweight you do not need to divide by 1000000. It will automatically normalize the weights to have a mean of exactly 1.

Some Stata commands, such as “summarize”, do not allow pweights. For them you can use iweight, which must be divided by 1000000, but you can do that within the command, as in “summarize y [iweight=v005/1000000]”

You can also get weighted means with fweight, as in “summarize y [fweight=v005]”. If you use fweight=v005, then the means, proportions, rates, and coefficients will be correct. Even standard deviations will be correct. However, n’s must be divided by 1000000. Standard errors must be multiplied by the square root of 1000000, i.e. by 1000. (Remember that the standard error of a statistic is inversely proportional to the square root of the sample size.) Confidence intervals, etc., must be re-calculated.

The clustering adjustment can be made by itself, as an option (after the comma). For example, “regress y x, cluster(v001)”. Weighting and clustering can both be done, with “regress y x [pweight=v005], cluster(v001)”

Ignoring the strata, you will get exactly the same results as above if you use the svyset command. For example, “svyset [pweight=v005]”, followed by “svy: regress y x” will give exactly the same results as “regress y x [pweight=v005]”.

If you want to adjust for the strata, you MUST use svyset. The syntax could be, for example, “svyset v001 [pweight=v005], strata(v023)”

Question: Several users report that svyset will not work properly; the run terminates with an error message.

Answer: Try one of the following “singleunit” options:

svyset v001 [pweight=v005], strata(strata) singleunit(centered)

svyset v001 [pweight=v005], strata(strata) singleunit(scaled)

svyset v001 [pweight=v005], strata(strata) singleunit(certainty)

I generally use the “centered” option but am sure that arguments could be made for any of the options.

Question: When we consider PSU clustering, can we also consider clustering at lower levels (i.e. households and women)?

Answer: There is no question that women in the same household tend to be similar, children in the same household tend to be similar, children with the same mother tend to be similar, etc. You can include a hierarchy of levels in the same model. For example, you could be looking at height and weight measurements of children under five in the PR file. The children would be level 1, households would be level 2, and clusters would be level 3. In Stata, enter “help me” (here “me” refers to “mixed effects”!) for the syntax.

Question: Several users have asked about which weight to use—for example, the household weight, woman’s weight, etc., when there is more than one version of the weight variable.

Answer: The weight variable is always v005, sometimes with an additional letter at the beginning, such as h. Usually there is no ambiguity about which weight to use, but when you merge files you may get more than one weight. The general rule, in such cases, is to use the weight from the units that tend to have higher nonresponse rates. Specifically:

For households (HR file) and persons in households (PR file) use hv005.

For women (IR) and children of women (BR) and (KR) use v005.

For men (MR) and couples (CR) use mv005. Men tend to have more non-response than women, so you need to use the correction for men’s non-response when studying couples.

When using any variables from the domestic violence module, use dv005.

When using any variables for individual from the HIV testing (in the AR file), use hiv05. You need to use the correction for higher rates of non-response for testing.

When using any variables for couples from the HIV testing (using a merge of the CR and AR file, for example to study HIV discordance), use hiv05 from the man’s record.

Question: Several users have asked whether estimates at the district level, etc., are acceptable.

Answer: Estimates at any geographic level--or for any subpopulation--will be unbiased. The reason why we discourage estimates for small geographic areas is sample size, especially in terms of a small number of clusters. At the district level, it can easily happen that there are NO clusters at all. If there are clusters, there may only be one or two. The clusters may not be very representative, even though there is no reason to think they are systematically too high or too low on any of the indicators. Some districts may have many clusters, enough to produce reliable estimates, but we strongly advise against trying to produce estimates for all districts.

Question: What to do if the weights must be integers, as required in geoRglm, for example?

Answer: I said a little about this earlier. The DHS weights have been normalized to have a mean of 1. Then they are multiplied by 1000000 and then they are rounded to the nearest integer. This gives many significant figures, as well as an integer.

Never divide these weights by 1000000 and then round to an integer. Many weights would then round to 0, and those cases would be dropped from any calculations that used weights.

You could divide by 1000, say, and then round to an integer, but this will have a detectable influence on the results.

Many calculations with weights will work fine if you leave the weights alone—that is, use v005 as it is. However, calculations may produce overflows (too many significant digits). You could use double precision (or the equivalent) so that more significant digits are retained during the calculations. Contact us through the forum if you need more suggestions.

Question: I am doing a study of low birth weight using a 2011 DHS survey. More than 60 percent of birth weight is missing, so I want to impute missing values. DHS has suggested to use sample weight, so when do I do this? Before imputation or after?

Answer: I would be very reluctant to impute such a high percentage of missing birthweights. You could only do this by using covariates, and imputation would amplify or increase the association of birthweight with those covariates. However, if you insisted on doing this, I would recommend that you use weights in the imputation model.

Question: When data for multiple time points in the same country are combined, should I adjust the weights?

Answer: We analyze change in a country by appending the files from successive surveys (note: this is not the same as a merge). For example, to examine change in Rwanda between the 2000 and 2010 surveys, we can append the KR file from the 2000 survey to the KR file from the 2010 survey (or the other way around; you can start with either one). We would not use the pooled file to estimate pooled means, etc. The pooled file could be used for data processing convenience, to do separate calculations for the 2000 and 2010 surveys, but the main reason is to look at change. For simplicity, define a variable “survey” that is coded 0 for all cases in the 2000 survey and coded 1 for all cases in the 2010 survey.

How would you adjust the weight cluster, and strata variables?

Adjusting the weights. In Stata, pweights are automatically normalized, so that the total weight equals the number of cases and the mean weight is 1. That is, Stata will calculate the total weight (the sum of v005 across all n cases), call this v005\_total, and then replace v005 by v005\*(n/v005\_total). The normalized weights are proportional to the original v005 and they add to n.

When the weights in the pooled file are normalized, following this procedure, the weighted and unweighted totals will be the same for each survey separately and for the two surveys combined. In other words, you do not need to do anything to v005. The normalization that Stata will do automatically is sufficient. This is what I would recommend for studying the change between the 2000 and 2010 surveys.

Adjusting the cluster codes and strata codes. The 2000 and 2010 surveys tended to use consecutive number of clusters, starting with 1. Unless you re-number, Stata will think that cluster 1 in the 2000 survey and cluster 1 in the 2010 survey are the same. You must redefine with a command such as this: “egen cluster = group(v001 survey)”. It is remotely possible that some actual enumeration areas appeared as clusters in both surveys, but this is unlikely, and the sampling information required to check this is not accessible. Similarly, if, say, v023 is the stratum variable in each survey, you need to construct a new stratum variable for the two surveys combined with a command such as this: “egen newstrata = group(v023 survey)”.

You may want to pool surveys to get more power for a test of some relatively rare relationship. In this situation I would give equal weight to each of the surveys. Say, for example, that there are ten surveys, the unweighted number of cases in survey k is nk and the total number of unweighted cases in all ten surveys is N. In order to have the same weighted total in each survey, i.e. a total of N/10, you could multiply v005 in survey k by (N/10)/nk. You would in effect weight up the smaller surveys and weight down the larger surveys. This step would keep the factor of one million but you would then want to round the adjusted v005 to the nearest integer. I would be very cautious about pooling in this way, because the surveys have different dates of data collection and the combined reference population may not be meaningful. You would almost certainly want to include fixed or random effects in the analysis.