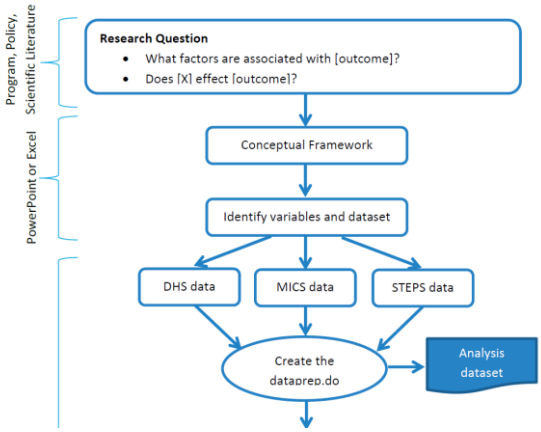
Data: Navigate the DHS

Topics: Navigate DHS documentation and identify DHS variables for analysis

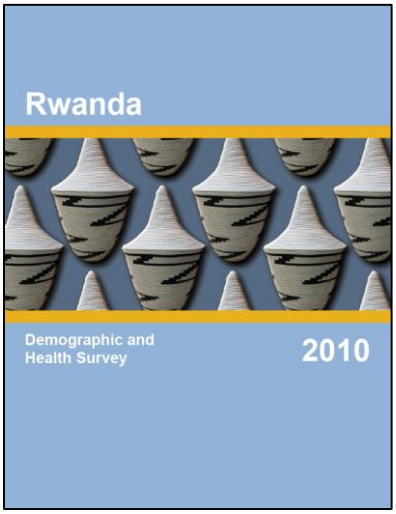


At this point in the analysis process, you should have a clearly-defined, answerable research question and a conceptual framework that helps you identify which variables you specifically need to answer your research question. You have looked at the available datasets via the national statistical agency website or another databank and determined that the Demographic and Health Surveys (DHSs) are appropriate for your research question. Now it is time to identify variables in the DHS datasets and generate variables for your analysis in a step that I call "data preparation."

If you have any questions about your analysis workflow, watch the Analysis Workflow lecture at www.populationsurveyanalysis.com.

This lecture on Navigating the DHS introduces four key documents to help you identify the right DHS variables for your analysis, and fully understand DHS variable definitions.

1. Questionnaire in the DHS Final Report



The first document is the DHS Final Report. In the development of your research question and conceptual framework, you probably read parts of a relevant DHS report. The last Appendix of the report contained the questionnaires that were administered in the survey, with the exact wording of each question, and the grouping of questions by module. A DHS dataset will contain multiple modules of data, and the variables will be organized in the roughly the same order as the questionnaires. Importantly, however, the variables in the Stata dataset will have different question numbers and labels. This is because after each DHS survey, responses are recoded so that variable IDs and labels are common across all DHS datasets, as described in the What Is The DHS? lecture. While I would not recommend starting with the DHS Final report to identify specific variables in Stata, this is a helpful document to refer back to, especially to identify keywords to aid your search for a variable in the Stata dataset.

2. lookfor variables in Stata

The easiest and recommended place to start your search for variables is in the DHS dataset itself using the `lookfor` command. For this approach, open the appropriate DHS dataset. I will open the KR, or kid's recode file, in this example. If you need a refresher on the different DHS datasets, refer back to the What Is The DHS? lecture.

With the dataset open in Stata, use the `lookfor` command with key words to identify variables. For example, if I am looking for mother's age, I can type `lookfor age` in the command window. A number of variables have "age" in the variable label or a category label; but this list is much shorter than if we had searched the entire dataset manually via the Variables window.

```
. lookfor age

variable name   variable label
------------------------------------------------------------------------------
v012            respondent's current age
v013            age in 5-year groups
v014            completeness of age information
v152            age of household head
v212            age of respondent at 1st birth
v221            marriage to first birth interval (months)
v320            age at sterilization
v372            shown pill package -na
v372a           shown condom package -na
v439            height/age percentile
v440            height/age standard deviation
v441            height/age percent ref. median
v447a           women's age in years (from household
                  questionnaire)
v452a           under age 18 (from household questionnaire)
v480            syringe and needle from new, unopened package
v511            age at first cohabitation
v525            age at first sex
v531            age at first sex (imputed)
v538            how previous marriage or un-na
v730            husband/partner's age
b6              age at death
b7              age at death (months, imputed)
b8              current age of child
b13             flag for age at death
h13b            given pre-packaged ors liqu-na
hw1             child's age in months
hw4             height/age percentile
hw5             height/age standard deviation
hw6             height/age percent of ref. median
hw7             weight/age percentile
hw8             weight/age standard deviation
hw9             weight/age percent of ref. median
hw70            height/age standard deviation (new who)
hw71            weight/age standard deviation (new who)
slang           language of the interview
```

Remember, the mother is the respondent to the questionnaire, and we see two potential variables for her age: v012, respondent's current age, and v013, her age in 5-year groups. To explore the difference between these two variables we can look at the raw data values with a `list` statement, or a summary of the data values with `tabulate` statements. Refer to the Use and Explore lecture for a review of these commands.

```
. list v012 v013 in 1/15

     +-------------+
     | v012    v013 |
     |-------------|
  1. |   35   35-39 |
  2. |   35   35-39 |
  3. |   32   30-34 |
  4. |   32   30-34 |
  5. |   30   30-34 |
     |-------------|
  6. |   22   20-24 |
  7. |   22   20-24 |
  8. |   33   30-34 |
  9. |   26   25-29 |
 10. |   26   25-29 |
     |-------------|
 11. |   21   20-24 |
 12. |   36   35-39 |
 13. |   36   35-39 |
 14. |   26   25-29 |
 15. |   27   25-29 |
     +-------------+
```

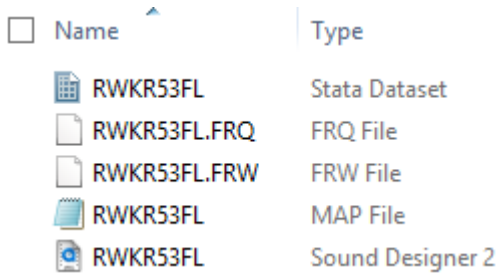
```
. tab1 v012 v013

-> tabulation of v012
```

respondent's current age	Freq.	Percent	Cum.
16	4	0.04	0.04
17	12	0.13	0.18
18	52	0.58	0.76
19	77	0.86	1.61
20	135	1.50	3.11
21	177	1.97	5.08
22	303	3.37	8.44
23	419	4.65	13.10
24	431	4.79	17.88
25	547	6.08	23.96
26	611	6.79	30.75
27	576	6.40	37.15
28	574	6.38	43.52
29	521	5.79	49.31
30	560	6.22	55.53
31	397	4.41	59.94
32	447	4.97	64.91

Cont…

```
-> tabulation of v013
```

age in 5-year groups	Freq.	Percent	Cum.
15-19	145	1.61	1.61
20-24	1,465	16.27	17.88
25-29	2,829	31.43	49.31
30-34	2,118	23.53	72.84
35-39	1,383	15.36	88.20
40-44	755	8.39	96.59
45-49	307	3.41	100.00
Total	9,002	100.00	



We see that v012 is a continuous variable of the mother's age in years, and v013 is her age in a 5-year age groups. If you have already decided during the conceptual framework stage how you will define mother's age in your analysis, then you can note the one variable whose values are best suited for your analysis. If you have not yet decided on a variable definition, then note both variables.

2a. Variable IDs

Here is a quick refresher about recoded variable IDs. The first letters of the variable [displayed on previous page] will tell you which module in the questionnaire the variable is from. All variables related to the woman respondent start with the letter "v". Notice that there are over a dozen variables in this dataset related to the woman's age – her age at 1st birth, her age at sterilization, her height-for-age percentile, and so on. Then there are variables starting with "b" from the "birth history" including the child's age at death or current age. And there is a section starting with "hw" from the height/weight anthropometry module for children under age 5. It will be important to recognize the difference between women variables and child variables later in this lecture.

3. MAP file

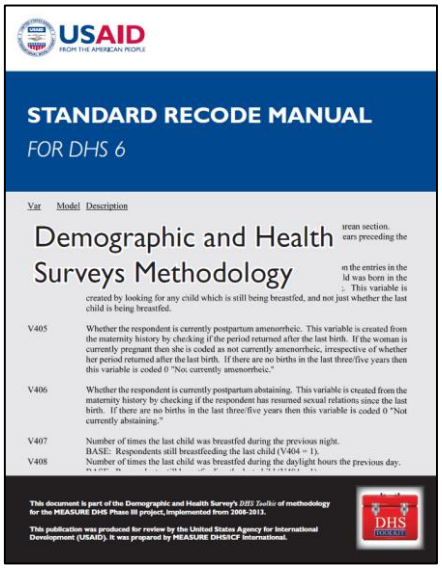
The second document, the MAP file, is a summary of all variable IDs, variable labels, and variable response categories or values. Stata's `lookfor` command is very good for searching key words, so you might never need to refer to the MAP file. However, for complex variables, especially variables that include a mix of continuous and categorical responses, the MAP file is a useful reference.

The MAP file is included in the zip file with the dataset that you downloaded from the DHS website. To view its contents, right click and open it with notepad. I recommend using your ctrl+F key to find key words. For example, I will search for the word "time" to show you an example of a variable that

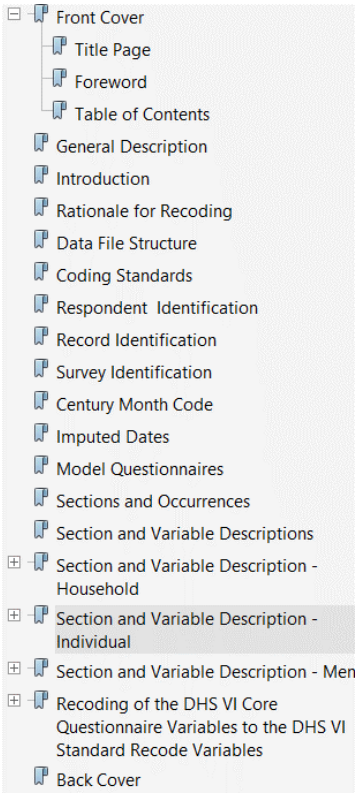
```
-------------------------------------------------

Item Name              Item Label
-------------------------------------------------
V115                   Time to get to water source
                              0:900
                              996    On premises
                              997    Not dejure resident
                         (m) 999    Missing
                        (na)        NotAppl
```

mixes continuous and categorical values. We see that V115, time to get to water source in minutes, has values 0 to 900 minutes (continuous values), as well as additional categories for respondents who have water “on premises”, “not a dejure [regular] resident”, “don’t know”, or “missing”. The MAP file has additional information about the formatting of the data file which you can ignore for now.

4. Recode manual



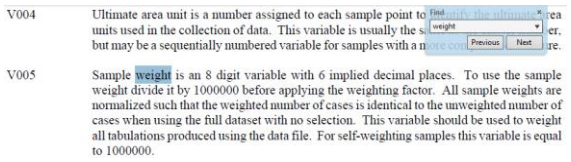
The third document, the Recode Manual, lists every variable ID in every recoded DHS dataset, with a detailed definition of that variable. To download the Recode Manual go to www.measuredhs.com, the publication search, and search for the keyword “recode”. Since the MeasureDHS Project is funded through the US Government, it is revised every couple of years under a new contract, and hence you will find several versions of the Recode Manual. Just use the most recent version – it includes historical notes.



We see that the Recode Manual is an enormous reference document. Usually you will have opened the Recode Manual with a certain variable in mind that you have a question about. Use the ctrl+F keys to navigate to the variable. To further simplify your search, navigate to the relevant questionnaire; for example, I will navigate to the women’s questionnaire to learn about variables in the kid’s recode file. Here are a few examples.

4a. Example: sampling weight

Let us say that you opened the recode manual to learn about the sampling weight variable. You use the ctrl+F keys to find the word "weight", and advance to the section the manual that provides descriptions of the variables. We find V005 which is the sampling weight in the woman's questionnaire and is the sampling weight in the kid's recode file.



V004 Ultimate area unit is a number assigned to each sample point to identify the ultimate area units used in the collection of data. This variable is usually the same as the cluster number, but may be a sequentially numbered variable for samples with a more complex structure.

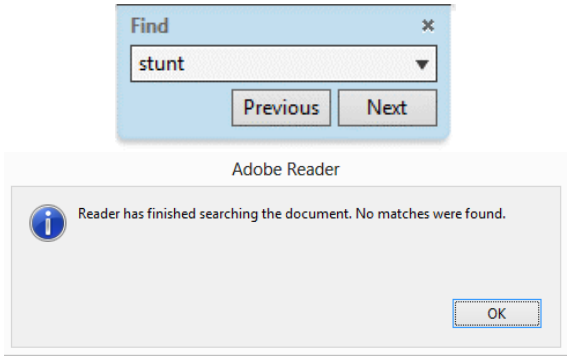
V005 Sample weight is an 8 digit variable with 6 implied decimal places. To use the sample weight divide it by 1000000 before applying the weighting factor. All sample weights are normalized such that the weighted number of cases is identical to the unweighted number of cases when using the full dataset with no selection. This variable should be used to weight all tabulations produced using the data file. For self-weighting samples this variable is equal to 1000000.

Here we learn that "Sample weight is an 8 digit variable with 6 implied decimal places. To use the sample weight, divide it by 1,000,000 before applying the weighting factor. All sample weights are normalized such that the weighted number of cases is identical to the unweighted number of cases when using the full dataset with no selection." And so on.

This is informative, right? The Recode Manual is telling us to divide V005 by 1,000,000 before we can use it in the analysis, and it is helpful to know that the weighted counts are normalized around 1 so that we can interpret them as weighted frequencies and weighted sample sizes. If the weights are not normalized around 1, then we cannot interpret weighted totals.

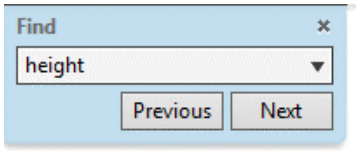
Sampling weight = v005 / 1000000

4b. Example: stunting (height-for-age)



Now let us use the Recode Manual to learn about stunting, a measure of child malnutrition derived from a child's height and age. If the child's height is <-2 standard deviations below the average height for his or her age group, then the child is considered stunted.

Using ctrl+F, let us look for "stunt". We do not find anything. From reading the DHS report, we know that stunting was measured and reported, so it *must* be in the recode manual – why do you think that we cannot find it?



V439 Height for Age percentile.
V440 Height for Age standard deviations from the reference median.
V441 Height for Age percent of reference median.
V442 Weight for Height percent of reference median based on DHS reference standard.
V443 Weight for Height percent of reference median based on Metropolitan Life or Foggarty reference standard.
V444 Weight for Height percent of reference median based on WHO reference standard.

The NCHS/FELS/CDC reference standard only contains data for children up to the age of 18 years. For all women aged 18 and over, the value of 215 months (17 years, 11 months) is used for their age, on the assumption that women are fully grown by the age of 18. Weight-for-age indices are not included as the weight of an adult woman is very dependent on her height. For the weight-for-height indicators, the CDC standard only applies up to a height of 137 centimeters, and almost all adult women are taller than this height. For this reason the weight-for-height Z-scores and percentiles are not available. However, three measures of percent of reference median are included, one based on the Metropolitan Life or Foggarty standard, the second based on the WHO standard and a third based on a DHS standard. These indices have been adjusted for pregnant women according to duration of pregnancy.

The anthropometric indices above are based on the CDC Standard Deviation-derived Growth Reference Curves derived from the NCHS/FELS/CDC Reference Population. The measures are presented with two implied decimal places (no decimal points are included in the data file). To produce the actual measure, divide the variable by 100. If either the weight or the height of the respondent is missing, then the corresponding measures above are set to the missing code 9999 or 99999. If either the height or the weight is outside of the acceptable range for the calculation of these measures, then the corresponding measures are set to code 9998 or 99998.

Let us try to find a different keyword. Since stunting is a measure of height-for-age, let us look up the word “height”. We find a number of variables related to height. V439 through V441 are specifically related to Height for Age, but these are maternal measurements and the variable starts with a “V” indicating that the data are about the woman, not her children.

HW4 Height for Age percentile
HW5 Height for Age standard deviations from the reference median
HW6 Height for Age percent of reference median
HW7 Weight for Age percentile
HW8 Weight for Age standard deviations from the reference median
HW9 Weight for Age percent of reference median
HW10 Weight for Height percentile
HW11 Weight for Height standard deviations from the reference median
HW12 Weight for Height percent of reference median

The measures above were calculated using the CDC Standard Deviation-derived Growth Reference Curves derived from the NCHS/FELS/CDC Reference Population. The measures are presented with two implied decimal places (no decimal points are included in the data file). To produce the actual measure, divide the variable by 100. If either the weight or the height of the child is missing then all of the above measures are set to the missing code 9999 or 99999. If either the height or the weight is outside of the acceptable range for the calculation of these measures then all of the above measures is set to code 9998 or 99998.

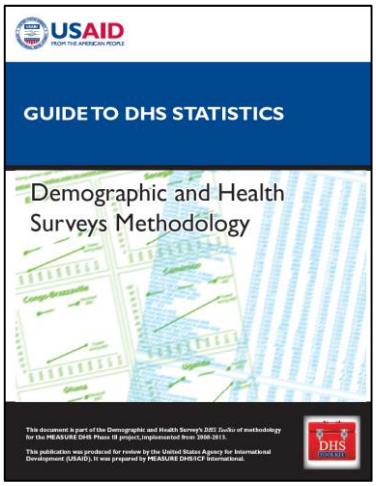
So we keep looking. Here we go – the Height and Weight module which contains information related to children. Variables HW4 through HW6 look like the variables we are seeking. However, from working with this dataset, I happen to know there are more relevant variables than these.

HW70 Height for age standard deviation (according to WHO)
HW71 Weight for age standard deviation (according to WHO)
HW72 Weight for height standard deviations (according to WHO)
HW73 BMI standard deviations (according to WHO)

The measures above were calculated using the new Child Growth Standards released by the World Health Organization on April 27 2006. The new Standards are the result of an intensive study initiated by WHO in 1997 to develop a new international standard for assessing the physical growth nutritional status and motor development in all children from birth to age five. WHO and its principal partner, the United Nations University, undertook a Multi-centre Growth Reference Study which is a community-based, multi-country projects involving more than eight thousand children from Brazil, Ghana, India, Norway, Oman, and the United States of America. The measures are presented with two implied decimal places (no decimal points are included in the data file). To produce the actual measure, divide the variable by 100. If either the weight or the height of the child is missing then all of the above measures are set to the missing code 9999 or 99999. If either the height or the weight is outside of the acceptable range for the calculation of these measures then all of the above measures is set to code 9998 or 99998.

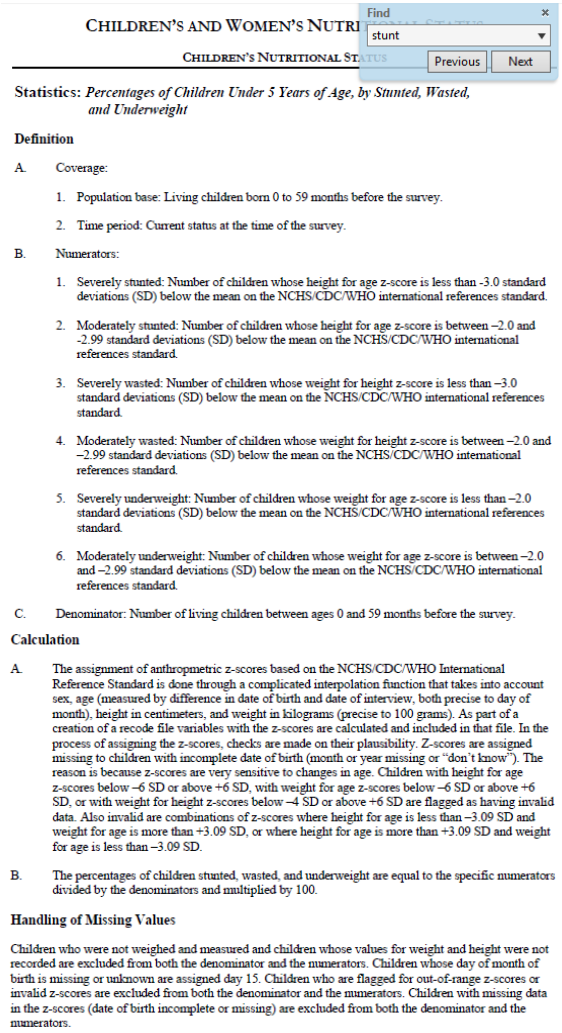
If you keep searching, you will find HW70, with a below explanation about updated Child Growth Standards released by the WHO in 2006. The Recode Manual briefly describes why it was important to update the international standards for median child height-for-age. Stunting should be calculated from this updated variable. We see that the measures are presented with two implied decimal places, so to produce the actual measure of height-for-age standard deviations, you must divide HW70 by 100.

The two examples provided were of complicated variables, which is often the case when you need to refer to the Recode Manual, though you will see that most of the descriptions are very straight forward. For example V501 is the current marital status of the respondent and V502 is a simplified version of that variable where marriage categories are summarized as currently, formally, or never married.



5. Guide to DHS Statistics

The previous thee documents provided definitions of recode file variables. The fourth document, Guide to DHS Statistics, describes how to calculate common indicators for analysis from these recode variables. For example, the Guide to DHS Statistics has definitions of child stunting, acute respiratory infection, and exclusive breastfeeding, all of which follow an international standard definition, but not mother's education, age, or marital status which researchers often categorize differently in analysis. Find the Guide to DHS Statistics under the publications tab at www.measuredhs.com.


CHILDREN'S AND WOMEN'S NUTRI

CHILDREN'S NUTRITIONAL ST

Find: stunt — Previous — Next

Statistics: *Percentages of Children Under 5 Years of Age, by Stunted, Wasted, and Underweight*

Definition

A. Coverage:

1. Population base: Living children born 0 to 59 months before the survey.
2. Time period: Current status at the time of the survey.

B. Numerators:

1. Severely stunted: Number of children whose height for age z-score is less than -3.0 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.
2. Moderately stunted: Number of children whose height for age z-score is between –2.0 and -2.99 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.
3. Severely wasted: Number of children whose weight for height z-score is less than –3.0 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.
4. Moderately wasted: Number of children whose weight for height z-score is between –2.0 and –2.99 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.
5. Severely underweight: Number of children whose weight for age z-score is less than –2.0 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.
6. Moderately underweight: Number of children whose weight for age z-score is between –2.0 and –2.99 standard deviations (SD) below the mean on the NCHS/CDC/WHO international references standard.

C. Denominator: Number of living children between ages 0 and 59 months before the survey.

Calculation

A. The assignment of anthropmetric z-scores based on the NCHS/CDC/WHO International Reference Standard is done through a complicated interpolation function that takes into account sex, age (measured by difference in date of birth and date of interview, both precise to day of month), height in centimeters, and weight in kilograms (precise to 100 grams). As part of a creation of a recode file variables with the z-scores are calculated and included in that file. In the process of assigning the z-scores, checks are made on their plausibility. Z-scores are assigned missing to children with incomplete date of birth (month or year missing or "don't know"). The reason is because z-scores are very sensitive to changes in age. Children with height for age z-scores below –6 SD or above +6 SD, with weight for age z-scores below –6 SD or above +6 SD, or with weight for height z-scores below –4 SD or above +6 SD are flagged as having invalid data. Also invalid are combinations of z-scores where height for age is less than –3.09 SD and weight for age is more than +3.09 SD, or where height for age is more than +3.09 SD and weight for age is less than –3.09 SD.

B. The percentages of children stunted, wasted, and underweight are equal to the specific numerators divided by the denominators and multiplied by 100.

Handling of Missing Values

Children who were not weighed and measured and children whose values for weight and height were not recorded are excluded from both the denominator and the numerators. Children whose day of month of birth is missing or unknown are assigned day 15. Children who are flagged for out-of-range z-scores or invalid z-scores are excluded from both the denominator and the numerators. Children with missing data in the z-scores (date of birth incomplete or missing) are excluded from both the denominator and the numerators.

5a. Stunting (height-for-age)

Let us stay with this example of stunting to see how to define the variable for our analysis Using ctrl+F and the word "stunt", we find a page that describes how to calculate all three of the standard child nutrition indicators: stunting, wasting, and underweight. The coverage section tells us that these indicators are for all children age 0 to 59 months. The numerator section tells us the technical definition of who is considered stunted which is all children whose height-for-age z-score is between -2.0 and -2.99 standard deviations below the mean of the international reference population, and severely stunted children have a z-score below that. The denominator is all living children age 0 to 59 months. The next section describes the calculation that produced these z-scores. The section on Handling of Missing Values is handy – here we learn that children who were not measured, whose values were missing, or whose values were out of range or invalid should be excluded from the calculation. At the stage of deciding which variables to keep for the analysis, these details might not be important to you; however, as you start generating variables for your analysis, these documents will be essential.

6. DHS User Forum

Despite extensive documentation, you might still have questions that are not answered in the Final Report, MAP file, Recode Manual, or Guide to DHS Statistics. For these questions, check out the DHS User Forum at this web address: http://userforum.measuredhs.com/