EXERCISE 1. USING THE IPUMS-DHS WEBSITE

IPUMS-DHS Training

These exercises are designed to illustrate the many tools available in IPUMS-DHS that facilitate research with DHS data. We'll start by getting some practice using data discovery and documentation on the IPUMS-DHS public website.

1. Getting Started

The Final Reports are an important source of information and provide a way for researchers to check that they are handling the data appropriately. The following is a section of Table 3.1 from page 30 of the 2008-09 Kenya Final Report:

Table 3.1 Background characteristics of respondents

Percent distribution of women and men age 15-49 by selected background characteristics, Kenya 2008-09

	Women			Men		
Background characteristic	Weighted percent	Weighted	Unweighted	Weighted percent	Weighted	Unweighted
Ethnicity						
Embu	1.4	120	145	2.1	70	80
Kalenjin	13.2	1,115	750	13.3	432	297
Kamba	10.9	923	666	11.6	378	274
Kikuyu	19.4	1,642	1,504	17.5	569	545
Kisii	6.9	579	447	7.0	228	179
Luhya	16.3	1,373	1,266	17.7	578	539
Luo	13.0	1,098	1,113	13.0	425	458
Maasai	1.3	113	124	1.2	39	42
Meru	4.9	415	367	5.1	168	155
Mijikenda/Swahili	5.1	430	717	4.0	131	240
Somali	2.8	240	679	2.1	69	202
Taita/Taveta	0.9	79	124	1.1	37	48
Other	3.7	317	542	4.2	136	197

On the IPUMS-DHS website, at www.idhsdata.org, the complete report can be found by clicking on "Source Documents" in the left-hand column.

Do not try to download the full report now, as that will put too much stress on the Internet connection, but in the future, this is the place to find the reports.

2. Data discovery, variable names, frequencies, weights

At **www.idhsdata.org**, click on "Select Data" at the top of the page.

Choose "Women" as the unit of analysis. Click on "Select Samples." Check the boxes in front of Kenya and Uganda and then click on "Submit Sample Selections." Now the documentation will show only information related to these two countries.

Find the ETHNICITYKE variable, using either the "Search" function or the drop-down Topics menu. (Hint: ETHNICITYKE is treated as a "Demographic" variable.)

a. Click on "Original DHS variable names" at the top of the screen. What is the original DHS name for ETHNICITYKE? _____ Knowing the original name can be useful if you ever need to combine DHS and IPUMS-DHS data files. Note that the codes may be different across the two datasets even when the variable name is the same.

Now click back to IPUMS-DHS variable names. Click on the name of the variable, ETHNICITYKE, to pull up its documentation. There are a series of tabs across the top of the screen.

b. On the "Codes" tab, you will see the Codes and Frequencies for the ETHNICITYKE variable. An X means that that a given value appears in a sample; a dot means that value is not found in a country-year sample.
    1) Choose "Case Count view." What, in your opinion, is the first year in which you could conduct an analysis of Somalis? _____
    _____
    _____
    2) Now compare the unweighted frequencies for this variable on the IPUMS-DHS website to the unweighted frequencies (in the fourth column, headed Women unweighted frequencies) shown on Table 3.1 of the 2008 Final Report above. Do they match? _____ Compare the unweighted numbers in column 4 to the weighted numbers in column 3. Does weighting affect results? Explain briefly. _____
    _____
    _____
    _____
    3) Now click on "Select Data" at the top of the page to return to the drop-down list of topics. Select the "Technical" variables. What is the name of the variable you would use to weight the ETHNICITYKE variable appropriately, to produce the weighted frequencies shown in Table 3.1? _____

3. Comparability tab

Next, we will explore the AIDKNOWONE variable. Click on the "Change Samples" box and limit the samples to those from Kenya and Uganda. Now find the AIDKNOWONE variable. (Click on Select Data at the top of the page, then use either the drop-down menu of "HIV/AIDS -> AIDS Knowledge" variables or the Search tool).

a. Click on the Comparability tab for AIDKNOWONE and read how the question wording differed across surveys. Which surveys are likely to generate more "Yes" responses based on question wording, in your opinion? Explain. _____
_____
_____

b. Now click on the Codes tab and look at the ratio of yes to no responses in the (unweighted) frequencies. Which samples have the highest ratio? Based on your answer to 3a, do you think these are "real" difference or differences caused by measurement? Are you aware of any contextual factors, such as AIDS information campaigns, that could affect the ratios?

_____

_____

_____

c. Then return to sample selection, uncheck Uganda, and limit your samples to just those from Kenya.

4. Finding the most appropriate variable for your analyses

Table 3.1 Background characteristics of respondents

Percent distribution of women and men age 15-49 by selected background characteristics, Kenya 2008-09

	Women			Men		
Background characteristic	Weighted percent	Weighted	Unweighted	Weighted percent	Weighted	Unweighted
Highest level of schooling						
No education	8.9	752	1,242	3.4	112	171
Some primary	29.9	2,526	2,431	27.1	883	889
Completed primary	26.9	2,272	1,973	24.7	804	800
Some secondary	12.2	1,030	961	14.6	477	429
Completed secondary	14.7	1,243	1,123	20.5	666	612
More than secondary	7.3	620	714	9.7	316	355

The figures on Education in Table 3.1 of the DHS Final Report for Kenya 2008 are based on the variable EDACHIEVER in IPUMS-DHS. Bring up the full list of education variables available for the Kenya samples in IPUMS-DHS. (Hint: Education variables are considered indicators of Socioeconomic Status.)

a. Look at the Description tab for the EDUCLVL and EDACHIEVER variables. What is the difference between them? _____

_____

_____

_____

Which variable could you use to track change or disparities across all the Kenyan DHS samples, from 1989 forward? _____

b. Sometimes analysts want to conduct a multivariate analysis, to study how, for example, each additional year of schooling by the mother affects a child's chance of survival or height-for-age. Which IPUMS-DHS education variable could you use like this in a regression equation for Kenya?

_____

c. Literacy is another dimension of education covered by the IPUMS-DHS data. How did the DHS surveys change their method of collecting data on literacy over time? Click on LIT1 and LIT2 and explore the tabs to find the answer. _____
_____
_____
_____
_____

5. Learn how complex variables were constructed

Read the variable description for WEALTHQ (Hint, found in "Household characteristics"}. What is the wealth quintile variable based on: the household's take-home wage and salary income or on something else? Explain briefly in your own words. _____
_____
_____
_____

6. Fewer errors

As part of the data integration process, IPUMS-DHS staff link the household files to the files of individual women interviewed in those households. Researchers can therefore use information about the characteristics and possessions of the household where a woman was interviewed without having to link the household (HR) and woman's (IR) files themselves.

Let's explore the household variable TIMETOWTRHH (Time to reach water source and return, in minutes, from the household record) to see how IPUMS-DHS documentation can help researchers avoid inadvertent errors.

a. Look at the codes and frequencies tab for TIMETOWTRHH for the Kenyan samples, using the case count view. What evidence do you see of "heaping" or digit preference/rounding in the responses? What sorts of numeric responses are favored by respondents? _____
_____

b. Given your response to 6a, how would you recommend grouping the data into other smaller categories? Explain why you chose this approach to grouping.
_____
_____

c. Based on the variable description text, are there any respondents you might want to exclude when analyzing TIMETOWTRHH? Briefly explain. _____
_____

d. Examine the universe tab for TIMETOWTRHH. How does the sample vary across the years?
_____
_____
_____

e. If you wanted to examine time to fetch water for all women in all years, what value would you ascribe to women currently coded as "not in universe" for TIMETOWTRHH? How would you approach the universe difference in your research? _____
_____
_____
_____

f. Now click on the survey text tab for TIMETOWTRHH. The question wording looks similar across the Kenyan DHS samples. However, things appear more complicated if you click on the "text" link for each sample, to view the wording of the 2-3 preceding questions. Which samples asked about fetching drinking water only? _____ Which samples asked about non-potable water sources (e.g., for washing clothes)? _____ Which samples asked about the potable and non-potable water together? _____ Would these differences matter to you as a researcher? Explain. _____
_____
_____
_____

g. Why is information about the source of water and distance to that source important?
_____
_____
_____