**ODK/SurveyCTO Lab**

**Data quality and management**

**Prepare Stata**

1. Open Stata. (If you do not have access to Stata, please email [researchsupport@poverty-action.org](mailto:researchsupport@poverty-action.org) so we can help you get a temporary license.)
2. Type the following command into Stata: **ssc install specialexp**. Wait for the download and installation to successfully complete.
3. Type the following command into Stata: **ssc install odkmeta**. Wait for the download and installation to successfully complete.
4. Open your USB, and unzip the “bcstats.zip” file and follow the installation instructions there. (Please note the bcstats package is also available on Box at <https://ipastorage.box.com/bcstats>; make sure to download the most current package from Box if you are completing this work after March 2015.)

**Familiarize yourself with the folder’s files**

1. Open the “\USB\SurveyCTO\exercises\4\_data\_quality\1\_lab\_work\files” folder from your USB. The files are numbered simply to help keep things linear – this is not a requirement for these do-files to work.
2. Open the file called “1.0 survey, paper version.html”. This is a print version of an xlsform.
3. Now, open the file called “1.1 survey, odk xlsform.xlsx”. This is the xlsform we are using in the scenario.
4. Next up, please see the files “1.2 survey, survey worksheet csv.csv” and “1.3 survey, choices worksheet csv.csv”. You should notice that these files are simply csv versions of the xlsform’s survey and choices worksheets. This will become relevant in the next section.
5. Next up, I just put a placeholder file there which is just a link to the SurveyCTO Server. In reality, at this point, you would be launching the SurveyCTO Client in order to download your data.
6. But since we are in a session and not the reality of your projects… “3.0 survey data.csv” actually contains dummy data from the aforementioned xlsform. You’re in luck! I’m not going to make you create your own data today. “3.0 survey data-consented-projects.csv” is simply the second raw data file, created because of a repeat group in the xlsform.

**Import your survey data using -odkmeta-**

1. -odkmeta- is a Stata program that imports your raw SurveyCTO csv data, and does all sorts of great things to it (e.g. labels, splits select\_multiple variables, merges repeat groups). We’re going to learn how to use odkmeta today. Please note you can also generate SurveyCTO Stata templates using the SurveyCTO Client to accomplish largely the same thing.
2. Type **help odkmeta** into Stata and briefly look over the helpfile. There is a large amount of information in the helpfile about -odkmeta-.
3. Open “4.0 create odkmeta do-file for survey data.do”. This will open the do-file in Stata’s do-file editor. Your task now is to replace the “???” text in the file with working code!
4. First up, set your current directory in the line beginning with cd (replace the “???” with your real current directory location). The current directory is the folder where you have saved all of the homework files beginning with “1.0 survey, paper version.html”…
5. Next, replace the three ??? portions of the odkmeta command with the appropriate filenames. In this case, you are looking for the names of the three files in your current directory that are your data file, your survey csv, and your choices csv.
6. Once you have done the above, run the do-file. This should create a totally new and separate do-file called "4.1 odkmeta do-file for survey data.do". I recommend you open this do-file to see how much cleaning/labeling work -odkmeta- has done for you automatically; an RA might spend months writing a similar do-file to prepare data for analysis.
7. Run "4.1 odkmeta do-file for survey data.do". This will create nice Stata .dta files for you! Check out “3.0 survey data.dta” especially, and see that -odkmeta-has created a wonderful, analysis-ready Stata data file for you and your team.

**Import your backcheck data using -odkmeta-**

1. This set of steps is essentially the same as above, except it’s so -odkmeta- imports your backcheck data instead of your survey data, so this will be brief.
2. Note that there is a paper version of the backcheck instrument at “5.0 backcheck, paper version.html”, a backcheck xlsform at “5.1 backcheck, odk xlsform.xlsx”, a csv version of the xlsform’s survey worksheet at “5.2 backcheck, survey worksheet csv.csv”, a csv version of the xlsform’s choices worksheet at “5.3 backcheck, choices worksheet csv.csv”. There is also backcheck data at “6.0 backcheck data.csv”.
3. Open “7.0 create odkmeta do-file for backcheck data.do”. Again, this will open the do-file in Stata’s do-file editor. Your task again is to replace the “???” text in the file with working code!
4. First up, set your current directory in the line beginning with cd (replace the “???” with your real current directory location). The current directory is the folder where you have saved all of the homework files beginning with “1.0 survey, paper version.html”…
5. Next, replace the three “???” portions of the odkmeta command with the appropriate filenames. In this case, you are looking for the names of the three files in your current directory that are your backcheck data file, your backcheck survey csv, and your backcheck choices csv.
6. Once you have done the above, run the do-file. This should create a totally new and separate do-file called "7.1 odkmeta do-file for backcheck data.do". I recommend you open this do-file to see how much cleaning/labeling work -odkmeta- has done for you automatically; an RA might spend months writing a similar do-file to prepare data for analysis.
7. Run "7.1 odkmeta do-file for backcheck data.do". This will create nice Stata .dta files for you! Check out “6.0 backcheck data.dta” especially, and see that -odkmeta-has created a wonderful, analysis-ready Stata data file for you and your team.

**Modify the high frequency check template**

1. Now, the fun begins. You are going to modify the high frequency check template to work with your data.
2. Please note again that while I have provided a copy of the high frequency checks package in the folder, you should always make sure to be using the most current version of the template by checking out [HFCs on Box](https://ipastorage.box.com/hfc-package).
3. Open “8.0 hfc original template.do” and poke around a bit. Read the comments especially to see what is available in the whole template, as today we are only going to implement a very tiny corner of the HFC universe.
4. Now open “8.1 hfc modified template.do”. Find the first “???” and replace it with your current directory.
5. Find your second “???” and replace it with your raw survey data, the Stata data file you created with odkmeta.
6. Find your third “???” and replace it with your unique ID variable.
7. Find your fourth “???” and replace it with your favorite fruit variable.
8. Run your do-file. Make sure to examine your “8.3 hfc log.log” file for the results of your check.
9. Now open “8.5 adv hfc modified template.do”. This is a more advanced (and realistic) high frequency check procedure.
10. In a separate window begin open “8.4 adv hfc data.dta” and get a feel for the mock survey data that you’ll be checking.
11. In the “8.5 adv hfc modified template.do” file follow the prompts and replace “???” in the code with working segments. To the best of your ability, answer the 9 comment questions sprinkled throughout the do-file. Read all the comments and really try to understand what quality checks each code block is attempting to perform on the data.

**Modify the bcstats template**

1. You are going to modify the bcstats template to work with your data.
2. Please note again that while I have provided a copy of the bcstats package in your folder, you should always make sure to be using the most current version of the program and the template by checking out [bcstats on Box](https://ipastorage.box.com/bcstats).
3. Open “9.0 bcstats original template.do” and poke around a bit. Read the comments especially to see what is available in the whole template, as today we are only going to implement a very tiny corner of the bcstats universe.
4. Now open “9.1 bcstats modified template.do”. Find the first “???” and replace it with your current directory.
5. Find the second “???” and replace it with the name of your CHECKED survey Stata data file.
6. Find the third “???” and replace it with the name of your backcheck Stata data file.

1. Find the fourth “???” and replace it with your unique ID’s variable name.
2. Make sure to read the explanation for T1 variables. And then find the fifth and sixth “???” and replace them with two variable names. For the sake of this exercise, use the age and title variables as T1 variables.
3. Make sure to read the explanation for T2 variables. And then find the seventh “???” and replace it with the favorite fruit variable, for the sake of this exercise.
4. Make sure to read the explanation for T3 variables. And then find the eighth and final “???” and replace it with the favorite fruit variable, for the sake of this exercise.
5. Run your template! Note that this creates two Stata data files, checked and deduped, for your survey and backcheck data. It also creates log files, and a csv file of differences between the two datasets; ignore this “bc diffs” file for today, because it’s been a long day!

\*\*\* Please note that the majority of this assignment has been “laid out” for you, with only minor understanding of the files required. You will, however, be expected to do much more complex work when working with your own data… Practice makes perfect. \*\*\*