Stata Exam[[1]](#footnote-1) (v3)

# General Instructions

The exam is a hands-on Stata exercise. Based on the provided dataset, we ask you several questions ranging from easy to hard. Note that each task has questions of varying difficulty, thus, there is nothing to worry if you are not able to solve all questions of task 1, for example. We expect to receive the answers in the provided Excel template and one (!) clearly structured and commented Stata do-file. The do-file should only reference to your local folder in one line at the beginning of the file so that we can easily change and run your code on our machine. All your output files should be prefixed with “FirstnameLastname\_”. You do not need to submit them because we will re-run your code.

While the task is not timed, we expect you to complete it within 2 – 3 hours. So please do not spend 12 hours or more to get a perfect solution but rather submit what you have in reasonable time.

We suggest to verify whether your response email was sent by checking your Sent folder and also ensuring that all attachments (do-file and Excel results file) are valid. Any issues with the submission of the results will reduce the grade.

When finished, please submit summary the following files:

* Your do file named FirstnameLastname.do
* Your Excel results file based on the template named FirstnameLastname.xlsx

We hope you enjoy the exam and it will also help you to assess how proficient you are with Stata.

Finally, we ask you to solve the exam alone (of course, you can use the Stata help files and online resources as long as they are not interactive). While cheating on the exam might help you to get an interview, we are confident based on past experience that this will be quickly identified during the interview leading to direct and indisputable rejection from this and future positions.

# Stata Instructions

Please do not use “//” or “/\*” comments but only “\*” comments. Comments must start at the beginning of a line. Make sure you clearly identify which task you are solving in the code, e.g. “\* Task 1”. Any output entered in the Excel sheet must be an output from a Stata command. Thus, we will check whether we can see this output in the corresponding block of code for the respective task. For each such output, comment in the preceding line the question number like “\*Output 1.1”. Any manipulation of the data must be done with Stata commands. Thus, you cannot use the graphical interface to solve a task unless you record the Stata command into the do file.

Note that the do-file must be generic. If a task asks you to identify the maximum value and use this maximum value in follow-up questions, you should not hard code the maximum value, e.g. ’34.234’, but use macros to ensure the code still runs if we apply it to a slightly different dataset (with, for example, a different maximum value).

# Tasks

You have two datafiles called hh.dta (household dataset) and hhitem.dta (consumption dataset). The first file contains one household per row. The second file contains the consumption of households for different items. Thus, each row includes the consumption for one household and one item. One household can have multiple rows if the household consumed multiple items. Note that items which were not consumed by the household are not listed in the file.

Tasks with a “\*” require a response in the Excel sheet. Remember: those responses must be calculated by the do-file.

Each task is independent of the previous task. Thus, it is recommended to reload the original dataset at the beginning of a task.

## Task 1: Estimation

Open the household dataset.

* Rename hsize to hhsize
* Label the variable cluster as “Cluster”
* Arrange the cluster variable between the variables strata and hhsize
* \*1.1 How many households are there?
* Check whether the household weight is constant in each cluster (only use the integer portion of the weight).
* \*1.2 What is the average household size (without using sampling weights)?
* \*1.3 What is the average household size when using sampling weights?
* \*1.4 What is the standard error of this estimate?
* Now, let us consider that the sample is not a pure random sample but utilizes a two-level cluster approach. The primary sampling units are clusters stratified by the strata variable.
* \*1.5 What is the average household size when considering the sampling weights and the stratification of the sampling methodology?
* \*1.6 What is the standard error of this estimate?
* [Nothing you need to do for this exam but it is worth thinking about these observations. This could be a great question for an interview.]

Well done, that was the first task. As mentioned before, some questions in this task are not easy. If you did not manage to solve all of the questions, don’t worry but proceed with the next task.

## Task 2: Outlier Correction

Open the consumption dataset. Note that we do not use sampling weights in this task.

* The consumption is in local currency, which we call X. Assume that 6X is equal to 1 USD. Convert consumption to USD (do not create a new variable).
* \*2.1: What is the single largest value for consumption of any one item by any household?
* You will observe that this value is very high. This indicates that there are still outliers in the dataset. Let’s correct them. [Note that we simplify the procedure here for the purpose of the exam.]
* \*2.2: Calculate the standard deviation across all consumption values and report how many consumption values are beyond 2 standard deviations of the mean.
* [The problem with this approach is that we use the standard deviation across products, which inflates the standard deviation. Imagine that one item like salt has an average consumption of 1 USD while meat might have 10 USD. In this case, the calculated standard deviation across all items will be too large for salt. Therefore, we will use an item-specific standard deviation approach.]
* Calculate the item-specific standard deviation of consumption in a new variable called xsd.
* Calculate the item-specific mean of consumption in a new variable called xmean
* \*2.3: How many records have consumption values beyond 2 times the item-specific standard deviations of the item-specific mean.
* Replace the outliers with missing values
* Each item has a 3-digit code. The first digit signifies the food category. Aggregate the consumption per household at the food category level.
* \*2.4: Identify the largest consumption across households and across food categories. Report the household id of the household with the largest consumption value for any single food category.

## Task 3: Democratic Food Shares Wrong

Open the consumption dataset. We’d like to know the share of the food items in the consumption bundle per household. There are different ways to calculate them. Using the democratic approach, we show how a small conceptual error can lead to wrong results.

* Create a new variable xtotal that contains total consumption per household.
* \*3.1: Report the household id of the household with largest consumption.
* For each household, calculate the share of consumption for each food item relative to total household consumption. Call the new variable xshare.
* Now we want to get the average consumption share per food item. Before aggregating, please incorporate the sampling weights from the household dataset to our current dataset so that we can weight the households correctly. Please make sure (in the code) that all records are linked to exactly one household record.
* Then calculate the mean of the consumption shares per item across households while considering the sampling weights per household. Now you should have one (average) consumption share for each item.
* \*3.2: Normalize the share per item so that the sum over all items becomes 1. Report the normalization factor (the sum over all food shares before normalizing).
* Save the table in Stata format (you will need it in the next task) with food items as rows and only one column for the share of the food item.

## Task 4: Democratic Food Shares Correct

Unfortunately, there is a conceptual mistake in the last aggregation step of the previous task: Some households might not have consumed a product. In this case, the correct food share at the household level for this product is 0. However, we are missing this information. Thus, the average over the food shares across households is biased upwards (as you basically miss many zeros in the sum for calculating the average). Let’s correct this.

* \*4.1: Reload the consumption data and report the number of rows in the dataset.
* Reorganize the data so that you have the item-specific consumption in columns (called xfoodFOODID) with one row per household.
* \*4.2: Report the mean consumption of item 101 across households (don’t consider sampling weights)
* Iterate over all the xfood variables and replace missing values with 0s.
* \*4.3: Again, report the mean consumption of item 101 across households (don’t consider sampling weights).
* Reorganize the dataset so that we have the same format as the original consumption data (with the same variables)
* \*4.4: Report the number of rows in the dataset.
* Now, follow the previous step from task 3 to create the shares, add the sampling weights to the dataset and aggregate across households.
* \*4.5: Report the sum over all food shares.
* Add the table you have saved in the last task (3) so that we can compare the difference between the ‘wrong’ and the ‘correct’ method of calculating food shares.
* \*4.6: Report the mean absolute difference between both methods. Thus, you need to calculate the absolute difference between the two shares for each item. Then take the average of the absolute differences over all items.
* Visualize the difference with items on the x-axis and the difference between the methods on the y-axis. The first item should be at x=1, the second item at x=2, and so on. Give the graph the title “Differences”.
* Save the graph as png file.

1. Prepared by Utz Pape (World Bank). [↑](#footnote-ref-1)