Statistics Without Borders: Tanzania LFS

John Campbell

September 2021

1 Materials and Methods

1.1 Data Collection

The dataset is from the Survery of Household Welfare and Labor in Tanzania (SHWALITA). This survey is stratified across rural and urban zones across multiple regions in Tanzania. Households were selected via simple random sampling, with a small number of households being replacements upon refusal to participate in the survey. 4032 households were surveyed in total across three experiments.

The data was downloaded from the World Bank Microdata Library at
https://microdata.worldbank.org/index.php/catalog/3549/study-description

These files were individually downloaded as csv files and then processed in a Jupyter notebook in Python primarily using the pandas package. The data files used were hh_roster, hh_nonfood, hh_frqnonfood, and hh_info.

1.2 Data Processing

Every household and member has a unique id in the dataset given by the combination of a Regional, District, Cluster, and Household ID which is common across all datasets. Members of a household also are given an id within a given Regional, District, Cluster, and Household ID. The primary level of disaggregation was derived from the hh_roster dataset, which gives information about each member of a household. The NONE classifier was given to households with no members age 60+. The SNGL, DBLE, and OTHR classifiers were given to households with 1, 2, or 3+ members respectively with at least one being age 60+.

The metric to assign every group is the ratio of households with large household expenditure on health. This is defined by household expenditure as a ratio of 10% or 25% of either total household expenditure or income. Both income and total expenditures are able to be derived from this dataset, but expenditures is likely the less variable feature for a household. The total household expenditures is given from the hh_nonfood and hh_frqnonfood datasets. Because the recall period for hh_frqnonfood is two weeks while the recall period for hh_nonfood is either one month or a year, each expenditure was scaled to a

two-week period before totals were calculated. The survey codes in hh_nonfood used for health expenditures were 2022, 2023, and 2024.

The ratio of health expenditures for each household is weighted by number of household members when turned into averages. The equation for the averages is:

In this equation, i refers to an individual household, m is the number of members in the household, and w is the weight assigned to the household. Below is a summary of the counts of the data at the simple household level of disaggregation.

Simple Household Level	Status	Counts 10%	Counts 25%
TOTAL	0	2629	3622
	1	1403	410
NONE	0	1979	2689
	1	959	249
SINGL	0	71	83
	1	32	20
DBLE	0	74	109
	1	74	39
OTHR	0	505	741
	1	338	102

1.3 Secondary Level of Disaggregation

There are two potential secondary levels of disaggregation that were given in the Statement of Work. The first is whether the household is in a rural or urban. The areas that were chosen for the surveys were purposely stratified to show urban and rural trends, but those variables are not explicitly given in the dataset and thus would need to be derived in order to implement in the data, which is against the Statement of Work directions. The second possible metric is the ethnicity of the household. This is also not given, with the closest possible metric being the tribe of members within a community. Therefore, the secondary level of disaggregation was not applied to this dataset.

1.4 Statistical Analysis

Given the relatively low sample sizes for some of the household levels, bootstrapping was chosen to provide for more confident margins of error. 1000 independent bootstrapped sets for each group were created with replacement according to the disaggregated group size. For example, there are 148 DBLE samples in the dataset, so every set will have 148 samples chosen from the DBLE samples with the possibility for the same sample to be taken multiple times. Statistical summaries are given for both overall and disaggregated sections. 95% margins of errors are given for the proportion estimate given by 1.95 times the standard error.

The following table gives the final statistical figures for the data.

	level	samples	mean 10%	SE 10%	MOE 10%	mean 25%	SE 25%	MOE 25%
0	TOTAL	4032.0	0.347932	0.007410	0.014524	0.101608	0.004750	0.009311
1	NONE	2938.0	0.326551	0.008345	0.016357	0.085036	0.005013	0.009826
2	SNGL	103.0	0.309272	0.044196	0.086623	0.193893	0.039265	0.076959
3	DBLE	148.0	0.501838	0.041504	0.081348	0.265392	0.036438	0.071419
4	OTHR	843.0	0.400730	0.016559	0.032455	0.120346	0.011242	0.022035