Statistics Summative Assessment

A statistical survey was carried out for more than 9500 households in growing seasons in eleven countries. The purpose of the survey was to obtain data to undertake the Ricardian analysis of climate change impacts on agriculture in Africa.

Based on the data set provided, different types of variables can be identified, which in turn can be used to calculate different measures of central tendency and variability. An example is binary variables as can be seen under household characteristics, which is used to describe the genders (gender2, gender2 etc.). Another example is the nominal variable which is used to identify crop types, (c1 to c6).

The study design for this survey was based on random sampling of households within districts that represent key agro-climatic zones and farming systems. The method of sampling used was the multi-stage sampling. Based on the application of the data, the sampling method was the best choice to reduce the level of bias. Focusing on key agro-climatic zones, narrowing into districts and random household sampling within districts reduces the level of bias. There is some level of bias as data was collected based on predominant farm types.

From the data set provided, 2 additional key research questions can develop, and they are:
1. The level of education/literacy among farmers in different farm types (small scale and large scale) and how it affects their income.
2. The percentage of children of school age within farming communities who are out school and engaging in full time labor.

To analyze the level of education among farmers with respective to their scale of farming, we will look at the distribution of education in the varying farm types. For this research, we will start with a null hypothesis that literacy has no effect on net incomes, while the alternative hypothesis will be that literacy among heads of households is proportional to net incomes. The test to be used will be simple linear regression. This is the appropriate test because we aim to establish a statistically significant relationship between two variables. A probability of 5% will be used to determine significance.

A p-value lower than 5% will indicate that the null hypothesis can be rejected, and that literacy has a relationship with income, while a high p-value will be shown to say there is no sufficient result to nullify the null hypothesis.

Visualization techniques that can be used for this data set include:
1. Histograms.
   At a glance, the histograms will show patterns and distributions of the different data sets with respect to frequency, which will enhance interpretation. It will also show whether the data is skewed to the left or right

2. Scatter plots
   This will show relationships between two variable and will directly show trends in patterns in data.

3. Bar plots
   This will be ideal to show comparison between different farm types and the variables.

Data on fertilizer use, pesticide use, and irrigated area is not particularly necessary for the research that was carried out. This data can be useful for alternative research.