Statistics Summative Assessment

MATLOTLO MAGASA

Two of the variable types that I have identified in the survey are categorical variables and ordinal variables. For example, ages and genders are categorical data, while the farm sizes and income of the farmer households are ordinal data. Ordinal data is also a form of categorical data. Numeric variables are the quantities of the harvests in kilograms. These variables are continuous, while the number of members in a household is discrete.

The agricultural survey has been an experimental one as the conclusions drawn from the study has been determined empirically.

The sampling method employed for the survey was both a multistage stratified sample. The sample for the study first began by the researchers dividing Africa into groups of the sub-regions. The countries in each region have a shared climatic characteristic, hence stratified. Selecting districts were selected by ensuring each district represents agro-ecological zones and farming systems was stage 2. Stage three, the villages and farms within the districts were then randomly selected for performing the survey.

Random sampling is known as "the purest mode of sampling" as there is an equal chance of being picked in the population. However, in this study it's not a practical way to approach the research due to the very large population of Africa, the vast differences in the ecological, social and political conditions of the population. The chances of error will increase.

The method used (multistage stratification) is appropriate as it creates a more controlled form of studying a sample and it seeks to reduce sampling error.

There are a few possible sources of bias that can be identified within the survey:
Different countries using different definitions or policies in classifying farm size introduces a bias because a 'small' plot in South Arica may well just be a 'large' plot in Egypt. There is therefore less control on the data/ variables. Household self-assessing results in self-selection bias.
"Countries from each sub-region were selected based on formal expression of interest from respective institutions…"- self-selection bias.
Inconsistent ways of interviewing e.g. some countries' interviewers' names were anonymous to the interviewee others were not. Other households knew the date of the interview, others not. The blocking method could have been used for such issues, however was not implemented. Cost limitations can also introduce bias of convenience.

Hypothesis based on dataset:

1) Household/farms with an income of more than $600 per month yield more crops per km^2 than household with income less than $600.

2) Commercial farmers spend more money ($) per km^2 of cultivated land.

Statistical analysis plan: Identify hypotheses>> Select and access a dataset>> List inclusion/exclusion criteria >> Review the data to determine the variables to be used in the main analysis>> Select the appropriate statistical methods and software. (Centers for Disease Control and Prevention, 2013)

For hypothesis number 1), the null statement is: Income per household has no effect on crop yield per km^2. The alternative hypothesis is: Households with an income of more than $600 will produce more crops per squared kilometer.

To test the hypothesis I would use the significance test and p-value test.

If there is a significant effect in the results, that would imply that the p-value is significantly less than the alpha value or p-value < 5%. In this case, the null hypothesis can be rejected, meaning that households or farms with an income of more than $600 per month **will** produce more crops per km^2 (square kilometer).

If the results were not significant, the p value would be more than alpha and therefore the null hypothesis cannot be rejected.

Visualizations that can be used are the boxplots to evaluate the use of fertilizer and consumptions. These are better suited for showing more on variation.
Bar plots can be used to compare datasets in terms of highs and lows within the set, frequency or changes over time and can be used to compare, for example, the most used irrigation method across countries.
The pie chart can be used to represent what percentage of crops is planted on a given plot of land.

The comparison between survey and FAO results is important because the spread in result will help reveal where inferential statistics have gone wrong compared to empirical statistics and help create better methods to improve data.

References

Centers for Disease Control and Prevention, 2013. *Creating an Analysis Plan.* Atlanta: CDC.