# Statistics Summative Assessment – Fejiro Gbagi

1. Variable types and unique examples of the corresponding variable in the dataset.

* Discrete Data – Gender
* Continuous Data – Age
* Qualitative – Marital Status (Nominal) and Type of Farm Entity (Ordinal)
* Quantitative – Household Size (Ratio)

1. The study design is a sample study with households chosen randomly within districts. The population was sampled by **multi-stage stratified random sampling** with 11 counties selected, then districts selected from countries, then selecting of villages within districts, and finally randomly picking farms with selected villages. The method used is appropriate for this application. A sources of bias is the selection of Countries from each sub-region based on “*formal expression of interest from respective institutions*” within countries concerned with managing climate change impacts - this is an example of **self-selection bias**. Other sources of bias include convenience bias (e.g. household head availability) and cost restriction affecting interviewers from conducting multiple survey visits.
2. 2 key research questions that can be studied based on the dataset provided.

* Is there a relationship between the number of separate land areas used as farmland and the total net household income from farm activities?
* Do more educated families earn higher total net household income from farming activities?

1. Statistical analysis, null hypotheses, alternative hypotheses, tests used, and significance.

* Is there a relationship between the number of separate land areas used as farmland and the total net household income from farm activities? – fplts vs. incfarm\_n
* Null Hypotheses – Any correlation between the number of separate land areas used as farmland and the total net household income from farm activities is purely due to chance. That is, the variables fplts and incfarm\_n are independent.
* Alternative Hypotheses – The correlation between the number of separate land areas used as farmland and total net household income from farm activities is significant and meaningful. That is, the variables fplts and incfarm\_n are dependent on each other.
* I would use the **simple linear regression** because my hypotheses is exploring the relationship between an independent (number of separate land areas used as farmland) and dependent variable (total net household income from farm activities).
* Significances will be determined based on the sample correlation coefficient and the sample size. In this example, the significance level chosen can be 0.05.

1. Result Interpretation

* Significant effect resulted – There is a correlation between the number of separate land areas used as farmland and the total net household income from farm activities. In this case, the null hypotheses is rejected.
* Results not significant – Any correlation between the number of separate land areas used as farmland and the total net household income from farm activities is purely due to chance. In this case, the null hypotheses is accepted.
* If a significant effect occurs we can consider the alternative hypotheses of there being a relationship between the variables fplts vs. incfarm\_n. However, I believe the result of this hypotheses will not be significant because the number of separate land areas used as farmland cannot be the only factor to determine the total net household income from farm activities. Other independent values (such as education level of farmers, cumulative area of farmland, etc.) will need to be added to the hypotheses to get a truly significant result.

1. Three types of visualizations.

* Histogram – Type of Farm Entity can be virtualized as a frequency histogram with 3 bars – Small Scale, Medium Scale, and Large Scale.
* Boxplots – Total net household income from farm activities can be virtualized as a boxplot to provide more information on the data’s range (25th percentile, median, 75th percentile) and the presence of individual outliers.
* Scatterplots – Can be used to show the relationship (such as positive and negative correlation) between number of separate land areas used as farmland and the total net household income from farm activities.

1. The comparison of data on fertilizer use, pesticide use and the irrigated area with the national average values from the FAO and the World Bank is important. This comparison can help the selection of a statistical significances (p-value) – i.e. probability you are willing to accept if wrong or the amount of risk you’re will to assume if wrong.