IE266 Engineering Statistics

Case Study I

Due date: 21 April 2024, 23:59

In answering all questions, please state your assumptions clearly.

The impact of human activities on environmental degradation is still an important issue to consider. One of the sectors that causes environmental degradation and climate change is agriculture, the science or occupation of cultivating the soil, producing crops, and raising livestock. According to Climate Change 2022: Mitigation of Climate Change report of The Intergovernmental Panel on Climate Change (IPCC), GHG emissions from the agricultural sector are primarily due to land use change, enteric fermentation, and rice cultivation. The other issue about agriculture is water withdrawal during land use. According to The Food and Agriculture Organization (FAO), agriculture accounts for about 70% of global water withdrawals. Especially, producing meats and cereals requires a significant amount of water.

Retailers, which are the connection between farmers, manufacturers, and consumers, start to aim to reduce the environmental impact of agriculture. They try to widen their product offerings with products from sustainable operations such as organic, agroecological, or regenerative farming. These operations reduce the need for synthetic fertilizers and pesticides, which can further contribute to lowering GHG emissions associated with agricultural inputs.

Part I – Descriptive Statistics

FAO collects and stores historical data on global agricultural activities and countries' environmental impacts. This dataset encompasses rice, cereal, and meat production data from 1995 to 2020 for 158 countries. It includes information on production, water withdrawal, and GHG emissions associated with each agricultural product (rice, cereal, or meat) for each country in a given year. You can find the dataset in "Agricultural_Impact_Data.xlsx".

Here is a detailed description of the features below:

- **Country**
- **Region** – Region of the country
- **Subregion** – Subregion of the country
- **Development** – Development Level of the country (Underdeveloped, Developing-and Developed)

- **Year -** Year the data was recorded
- **Population -** Population of the country (Million people) in the related year
- **Agricultural land** – Total agricultural land of the country (hectare) in the related year
- **Item** – rice, cereal, or meat
- **Emissions** – Total amount of GHG emissions due to related production of the item (converted to their CO2 equivalent, measured in kilotons (CO2eq kt))
- **Production** – The production amount in a specific year (tons)
- **Freshwater withdrawals -** The water amount used to produce the item (kiloliters)

Note that, to summarize data, you are expected to use appropriate descriptive measures (measures of location, measures of dispersion, measures of position, etc.) and descriptive graphs (box plot, histogram, Pareto chart, bar chart, cross-tabulation, scatter diagram, etc.) while answering questions.

a) For each item (meat, rice, cereals), use a barplot to show the relative frequency distribution of the total production amount of the item between 1995 and 2020 over the subregions. What types of production are prominent in the subregions?

b) Determine the top countries whose GHG emissions due to rice in 2020 account for 80% of total emissions in the world due to rice, in 2020. Use the appropriate chart.

c) For the years 1999, 2009, and 2019, give the boxplot of the cereal production amount per capita of the developed countries. Are there any outliers? Interpret the changes over the years.

d) List the top 12 countries in the world for total rice production between 2011 and 2020. Show the total rice production amount and GHG emission amount (in terms of tons) from rice of these countries.

Calculate each country's average GHG emission amount from rice between 1995 and 2020. Then, plot a histogram after removing countries that have the 20 lowest GHG emissions and the 20 highest GHG emissions. Comment on the shape of the histogram.

e) Let define emission efficiency as the production amount per GHG emission amount. For each item (rice, cereals, and meat), compute the emission efficiency around the world for each year separately. Which item seems the most efficient? Is there any change over the years in efficiency values?

Consider the top 100 countries for total rice production in 2020. Find their emission efficiencies. Is there any relation between the development status of the country and emission efficiency? Discuss using a contingency table.

f) For rice production in 2020 all around the world, examine the relationship between production amounts, GHG emissions, and freshwater withdrawals. Is it possible to say that production amounts and GHG emissions are correlated? Is it possible to say that freshwater withdrawals and GHG emissions are correlated? Use appropriate descriptive statistics (graphical and numerical tools) to support your answer.

Part II – Statistical Inference

Retailers are becoming more conscious of environmental issues and are carefully examining the operations of their farmers. They aim to collaborate with farms that have less negative impact on the environment. The top management team of a large retailer, which has many stores in different regions and countries, wants their sustainability department to prepare a report so that the top management team can introduce new targets on GHG emission levels from agriculture.

The sustainability department investigates the farmers in Region A and Region B. It is assumed that the climate and earth conditions of countries in the same Region are the same. In one region, more than 500 farms exist, varying in size. A sample of 20 farms in Region A and 25 farms in Region B is selected for inferences about the population. The data related to rice production of farms for a harvest season, 12 weeks. It is assumed that the farms' operations are independent of each other and each week. Every farm operates in one harvest season in a year, which means that after harvesting one area farmer can not crop the same area again. The month consists of 4 weeks.

A detailed description of the data is given below:

- **Farm**
- **Week**- Week the data was recorded
- **Harvest Amount**- Total amount of rice harvested from a specific farm in a specific week (t)

- **Agricultural land** – Total relevant agricultural land of the farm where harvested rice is planted (hectares)
- **Harvest-related Emissions –** The amount of greenhouse gas emissions due to rice harvesting in a specific week of the farm. For example, energy consumption by using Combine Harvester or other equipment, etc. (converted to their CO2 equivalent, measured in tons (CO2eq t))
- **Storage-related Emissions –** The amount of greenhouse gas emissions at the farm due to keeping/ storing the harvested rice in a specific week. For example, electricity consumption in the warehouse, etc. (converted to their CO2 equivalent, measured in kilotons (CO2eq kt))
- **On-Farm Loss** – The amount of production loss of rice while harvesting (t)

Yield is defined as "Harvest Amount - On Farm Loss". Yield is the total amount of rice that can be sold/used.

a) For Region A, comment on whether, in the first month, the total Harvest-related Emissions per farm (CO2eq kt/farm) is similar to that of the last month at a 0.05 significance level. (Before the analysis, statistically verify whether population variances are identical). Use confidence intervals for comparison. Do the same analysis for Storage-related Emissions per farm for the first and last month. Which assumptions are needed to make statistically sound comparisons? Support your assumptions with visual aids.

b) Compare the average Harvest Amount of Region A per farm for each week (t/farm) with that of Region B at a 0.05 significance level. Do the same analysis for each week's average On-Farm Loss/Harvest Amount ratio per farm. Finally, whether Harvest-related Emissions of Region A are greater than Storage-related Emissions in the same Region at a 0.05 significance level for each week. Use confidence intervals for comparison purposes. Which assumptions are needed to make statistically sound comparisons?

c) The retailer's sustainability department believes the farm's harvest and storage-related emissions/yield is a good indicator of the emission efficiency. A farm is indicated as efficient if its total emission (harvest and storage-related emissions)/total yield ratio in a harvest season is less than 1.25. The summary of the status of farms is provided below. They believe the proportion of efficient suppliers in Region A is greater than in Region B. Check whether there

is sufficient statistical evidence to falsify the sustainability department's claim. Use confidence intervals at a 0.05 significance level. Comment on your findings. State any necessary assumptions to make statistically decisive comparisons.

Format and Organization:

- We expect you to carry out your statistical analyses by using R for each question and to report your analyses and conclusions.
- Number and title report sections properly. You do not need to include an introduction, conclusion, and appendix section in the report.
- Please use 12 font size and 1.5 paragraph spacing with reasonable margins. The format and organization of the report will be considered in grading.
- Please use comments in your R scripts to make your codes readable.

You should upload a zip file including R script, and all the files you use to answer the questions (Excel files, etc.) to the **Case Study 1 Material Submission**. Your report should be uploaded to the **Case Study 1 Report Submission**. Please note that one group member will submit your work. **Report must be coherent within itself and must be comprehendible without the need to run the R code.**