Stat 220 Lab 3: Analyzing Global Data

Background

Education is a fundamental human right and a cornerstone for individual and societal development. Literacy rates and educational attainment levels are key indicators of a country's development status. Analyzing the distribution of these variables across different countries can provide insights into global inequalities and inform policy decisions aimed at improving education worldwide.

Scenario

You have been provided with a dataset from UNESCO containing information on literacy rates, years of schooling, GDP per capita, and other socio-economic variables for countries around the world. The dataset can be accessed at `https://richardson.byu.edu/220/global_data.csv`.

There are several variables in this data set. They are listed here:

Variable Name	Description
`country`	Full country name.
`iso3c`	ISO 3-letter country code.
`Access_Electricity`	Access to electricity (% of population). Measures the percentage of the population with access to electricity.
`CO2_Emissions`	CO2 emissions (metric tons per capita). Measures carbon dioxide emissions per person.
`Internet_Users`	Internet users (per 100 people). Measures the percentage of the population that uses the internet.
`GDP_Per_Capita`	GDP per capita (current US$). Gross Domestic Product divided by midyear population.
`Literacy_Rate`	Literacy rate, adult total (% of people ages 15+). Measures the percentage of people aged 15 and older who can read and write.
`Primary_School_Enrollment`	Primary school enrollment, net (% of primary school-age children). Measures the percentage of children of official primary school age who are enrolled in school.
`Access_Water`	Access to at least basic drinking water services (% of population). Measures the percentage of the population using at least basic drinking water services.
`Unemployment_Rate`	Unemployment rate, total (% of total labor force). Measures unemployment as a percentage of the total labor force.
`Life_Expectancy`	Life expectancy at birth (years). The average number of years a newborn is expected to live under current mortality levels.
`Fertility_Rate`	Fertility rate, total (births per woman). Measures the number of children a woman is expected to have over her lifetime.
`Population_Growth`	Population growth (annual %). Annual percentage growth rate of the population.

It is natural that there will be variations for these important values between countries. We may assume that natural variance between countries would result in a normal distribution of the quantity of interest. For example, if literacy rates are normally distributed, there would certainly be some countries with higher rates than others, but plotting them as a histogram would result in a bell curve.

There are a few reasons why data might not be normally distributed.

- Severe outliers
- Long tail in one direction or the other, causing a skewed distribution
- Data clustered in multiple spots, causing a bimodal distribution (i.e. two bell curves side by side)

If we determine that specific variables are not normally distributed, and we determine in what fashion the data is non-normal, we can perhaps recommend specific actions to determine why that data behaves this way.

As a team of statisticians working for an international education organization, your task is to analyze this data to:

A. Estimate the parameters (mean and standard deviation) of key variables using the concept of maximum likelihood estimation (MLE), assuming a normal distribution.

B. Assess the normality of these variables using both the empirical rule and graphical methods.

C. Interpret your findings in the context of global education, identifying patterns, anomalies, and potential areas for intervention.

Your Task

You will perform the following tasks:

1. **Data Acquisition and Preparation**
   - **Load the Dataset**: Access and load the dataset from the provided link.
   - **Understand the Variables**: Familiarize yourself with the variables included in the dataset (e.g., literacy_rate, years_of_schooling, gdp_per_capita).
   - **Data Cleaning**: Remove missing values to ensure we can complete the analysis
2. **Statistical Analysis**

   For each key variable:

   a. **Calculate MLE Estimates**:
      - Compute the maximum likelihood estimates for the mean and standard deviation under the assumption of a normal distribution.

   b. **Empirical Rule Verification**: The empirical rule of the normal distribution states that 68% of the data falls within 1 standard deviation of the mean and 95% of the data falls within 2 standard deviations of the mean. If something is normally distributed, we'd expect it to follow approximately that pattern.
      - Calculate the percentage of data falling within 1 standard deviation of the mean.

- Calculate the percentage of data falling within 2 standard deviations of the mean.
- Compare these percentages with the theoretical values of approximately 68% and 95%, respectively.

c. **Graphical Analysis**:

- Create histograms and overlay normal distribution curves.
- Generate Q-Q plots (quantile-quantile plots) to assess normality. A qq-plot is a common method used to determine if a theoretical and observed distribution match. It plots the theoretical quantiles of a fitted distribution with the observed quantiles. There is a function called qqplot in the statsmodels library. This reference can teach more: `https://www.geeksforgeeks.org/qqplot-quantile-quantile-plot-in-python/`

3. **Variable Normality**

- **Assessment**: Based on your calculations and graphical analyses, determine whether each variable can be considered normally distributed.
- **Interpretation**: For the variables that are not normal, determine if there is a reason for the non-normality, be it outliers, skewness, or multi-modality.
- **Justification**: Provide evidence and reasoning for your conclusions, citing specific results from your analysis.

4. **Contextual Discussion**

- **Implications for Education**:
  - Discuss what the distribution of literacy rates and educational attainment levels suggests about global education.
  - Identify any regions or countries that are significant outliers and hypothesize reasons for these deviations.
- **Policy Recommendations**:
  - Suggest ways in which international organizations can use this information to target educational interventions.
  - Consider socio-economic factors that may influence the distribution of educational outcomes.

Deliverables

Submit a report containing:

1. **Introduction**: Outline the purpose of your analysis.
2. **Methodology**: Detail your analytical approach, including any data cleaning steps.
3. **Results**: Present your findings with supporting tables, calculations, and graphs.
4. **Discussion**: Interpret your results in the context of global education, discussing patterns, anomalies, and implications.

Ensure your report is well-organized and clearly written, with all graphs and tables appropriately labeled.

Submission

- **Deadline**: October 1
- **Format**: Submit your report via GitHub. A python notebook would again be sufficient.
- **Collaboration**: Include the names of all group members and ensure that each member contributes to the analysis and report writing.