Maxime Chabriel
ENS de Lyon/CERGIC
maxime.chabriel@proton.me

Introduction to Quantitative Methods
Directed Work
Univariate Analysis

Fall 2024
Master EEI

Objectives

The objectives of this session are:

- To learn the basics of Excel
- To review what we saw in Univariate analysis clas
- To learn how to implement the analysis we saw in class using Excel
- To develop good practices when it comes to handling datasets

Follow the instructions. I remain available if you have any question. But remember, computering is learned by searching your answers by yourself. Also, you live in the era of ChatGPT, this may the only class in all your studies where you will be incited to use it! You can work alone or in groups of 2. At the end of this session or before next class, send me your completed Excel file and your answers in a Word document in an email entitled "DW1 NAME1 NAME2".

Opening the dataset

We will be using a dataset containing country-level indicators of development.

- Download the country development database
- Open the data with Excel
  - **NB** a database is usually stored in the CSV (Comma Separated Values) format
  - If this is the case (and the data is not stored in a typical Excel .xls/.xlsx format), click right on the file → Open with → Excel
  - If the data appears mashed up in a single column, it means the separator of the CSV is not well defined
    - Select the column
    - Go to Data → Convert
    - Select delimiter → Next
    - Select either tabulation, comma or semi-column, depending on what works
    - Click finish
- Check the column names, the number of observations, are they coherent?

In our case, the database is pretty big, let's create a subset. A dataset is *too big* depending on the software you use (R is much more powerful than Excel for instance), the stats of your computer and how much you are getting confused by uncessary parasite data.

- Create a new Sheet. Rename both sheets (right click) *Data* and *Data subset*
- Go back to your original Sheet
- Create a filter
  - Select everything (Ctrl+A)
  - Clik Data → Filter
- We only want to keep the year 2015: Select the Year column → Unselect everything → Select 2015. Do not worry, we will be dealing with panel data later in this class

- Copy everything and paste it to the *Data subset* sheet. NB Copying your original data is a good practice as it prevents you from accidentally corrupting your original data. R has built-in saveguards, but in Excel Ctr+Z will not always save you...
- Save your Excel file: File → Save as. Create a new dedicated folder. Choose your folders wisely! A data folder can very quickly become messy if you do not pay attention when you save your files.

Now let's check if our dataset is clean.

- In the data subset we just built, what are the observations?
- Is there a specific column that uniquely identifies the observations?
- How many observations are there? Select the identifier column, on the bottom right appears the number of non-empty cells selected
- How many variables are there? Take a look, do you think they are relevant to study the development of countries? Do you wish other variables were available?
- Now let's check if there are missing values
  - Does the number of observations in the dataset checks out?
  - Without creating any new cells, count the number of missing values in the GDP variable. Is there a reason that may explain why the data is missing? Do you think the missing data comes from our end? From the collection procedure of the data? Or from the source of the data?
  - Just by looking at your data, you may notice a column is full of missing values. Go to your original dataset and filter for other years than 2015. Is the column still empty?
  - It seems like this column will not be any use for us today. Go back to your subset and delete the column (right click on its title). One down!
  - Select the column *Government expenditure on education (% of GDP)* (you can find elements in the Excel using Ctrl+F). How many missing values are there? Implement a filter on the table, just like we did on the original dataset, and filter out the non-missing values. Look at the observations, is there any logic to the missing values? Do you think the observations with non-missing values constitute a representative sample of your population? If not, delete it!
  - The year is not much use either, delete it
- Are the remaining variables all relevant to compare countries' development? Let's go other them one by one and discuss their relevance. Delete the unrelevant columns

Exploring the dataset

Unfilter your subset. We will now try to create a small overview of the dataset. There are still many variables, and it can be quite overwhelming when you look at them. Let's try using what we learned in class to quickly summarise them.

- Go to the last line of your dataset. You can quickly do so by clicking any cell of the identifier column, then Ctrl + bottom arrow
- The column titles are out of your view, this is cumbersome. Go back to the top (Ctrl + top arrow) and click on the cell B2. Click on Display → Freeze the panels. Wherever you go in your sheet, you will always see your id and variable names!
- Skip a line under the last country and write Mean
- Now in the same line and in the column GDP, you will write your first Excel formula
  - A formula always begins with the "=" sign
  - You then type a function. Here it is the function MEAN(.). If your function is in French, it will be MOYENNE(.). Then type the range of the data whose mean you want to compute. The easy way is to drag your mouse on the GDP values

  - Excel is smart, you do not need to type the formula for each column. Once you have typed the formula, drag the bottom right corner along the line. It will carry forward the function to the folowing columns

- What is the average country population? The average GPD PPP per capita?
- How do you interpret the mean of the last column?
- Now below the Mean and for each column, compute:
  - The median
  - The variance
  - The standard deviation
  - The first quartile
  - The third quartile
  - The 10th centile
  - The 90th centile
  - The smallest value
  - The biggest value
  - The interdecile ratio
  - The number of missing values
  - The share of countries with a value above the mean
- Before we continue, cleanup time! Put your titles in bold. Put boxes around the cells.
- What is the variable the least equitably shared amongst countries?
  - No need to squint your eyes over your numbers, use formulas again!
  - In the cell S197 type "Extreme value"
  - Then under it type a formula that detects for each *relevant* statistic the value indicating the distribution with the highest spread
    * NB You can only compare statistics that are not dependent of distribution scale. If you divide/multiply all your variables by 10, does it change your statistic? If yes, then this means it is scale dependent
    * For some you might want to find the highest values. For others, the lowest
  - For instance, for Variance, type "= MAXIMUM(the line of your variance results)"
  - You can get the name of the column using the following formula (I did not invent it, ChatGPT found it for me): "=INDEX($E$1:$Q$1, EQUIV(T199, E199:Q199, 0))"

Plotting data

Now let's explore one of the variables in details. We will first look at the GPD PPP per capita. Create a new sheet and copy the identifier column along with the variable of interest. Rename the sheet. Let's begin with a categorical variable.

- Create a new column entitled "Income level"
- For each observation, compute the income level. Find the formula using ChatGPT. In 2015 we had:
  - 0-1045$ is lower income
  - 1046-4125$ is lower middle income
  - 4126-12745 is upper middle income
  - 12746$-above is upper income

- Build a pie chart representing the share of each income share in the total number of countries. To do so, select the data column, click on Insert → Pie plot
  - Add a title
  - Show the labels
  - Add notes on the graph showing the precise number of each share
- Do a bar plot (do not confuse with a histogram!). Make it clean
- Precisely, what type of variable is the income level? Check that the bars are in the right order

Now let's study a quantitative variable:

- Create the histogram of the GPP PPP per capita
- Change the bin size to 2000$. To do so, double click on the x axis. On the appearing panel on the right, select Axis option → Interval size
- Are you satisfied by the readability of your graph? What seems to be the issue?
- Let's delete some outliers. Copy the GDP PPP per capita to a new column, sort the data from big to small and delete the 10 biggest observations. Replot the histogram using this new column
- Clean the histogram. Precise in the title that you deleted the 10 biggest outliers
- Assume that the histogram represents a density. Describe it
- Now let's plot a graph to summarise the distribution of your data (spoiler: it's a boxplot)

Now let's compare the distribution of some variables.

- Select the GDP PPP per capita and Access to electricity variables. Create a boxplot. Are you satisfied by this graph? What seems to be the issue?
- Let's rescale and center the variable. Scaling is when you change the unit of a variable. Centering is when you change the mean of the variable to 0.
  - To do so, let's create new variables. Skip a column and copy paste the variable names.
  - In the first cell (let's call it Y) of your first variable, create a formula so that the matching cell (let's call it X) in the original variable is substracted its mean, and is divided by its standard deviation. You want: $Y = (X - E(X))/\sigma(X)$
  - Spread your formula to all the other cells
- Now replot the box plots. Clean the plots
  - Which distribution has the most outliers (as detected by Excel)?
  - Which distribution is the most unevenly distributed? the most evenly distributed?