Lab 1: Data visualization and basic Excel usage

These labs contain exercises to be completed using a spreadsheet software.

There are multiple providers of spreadsheet software, but the University of Glasgow provides students with a license to use Microsoft 365 which contains Excel. You can use Excel either on your computer or in the cloud. The GTA is using Excel in the cloud and it is **strongly recommended** that you also use Excel cloud during the Lab sessions, because it will look similar to what the GTA has. Note that if you use Excel locally on your computer, the interface is specific to your operating system and system language.

To go to Excel online, go to https://www.office.com/launch/excel?auth=2.

Exercise 1. Data on a single variable.

Open up the attached dataset, ”ages.xlsx”, on Excel. This is a data set about the (imaginary) ages of survey respondents, n=30.

Part 1. Summary statistics.

- Create a new tab on the spreadsheet.
- Compute the mean, median, min, max of ages using Excel functions. Make a table.
- Compute the mean age using only the following excel commands: SUM(), COUNT().
- Which one is higher, the mean or the median? What does this tell us about the shape of the distribution of the data?

Part 2. Plotting data.

- Create a new tab on the spreadsheet with the data from the original tab.
- Compute a frequency distribution table. Decide yourself the cutoff points.
- Compute a corresponding cumulative distribution table.
- Make a graph describing the frequency distribution. The title should be “Frequency distribution (histogram)”.
- Make a graph describing the cumulative frequency distribution. The title should be “Cumulative Frequency distribution (ogive)”.

Part 3 (**Optional**). Plotting data: Pie charts.

- Make a pie chart summarizing the age distribution. Which graph is more informative in this case, the pie chart or the histogram?

Exercise 2. Data on multiple variables.

Open up dataset incomes.xslx. This file contains (imaginary) data on two variables, income (inc) and years of schooling (educ).

Years of schooling is categorical and coded as follows:

9	Grade 9
10	Grade 10
11	Grade 11
12	Grade 12
13	1 year of college
…	
17	5 years of college
18	6+ years of college

Part 1. Correlation

- Open up the data and create a new tab with the original data.
- Compute the sample mean and the sample variance of both variables.
- Compute the coefficient of correlation between income and years of schooling. How do we interpret the correlation coefficient? Is it a large or a small coefficient?

Part 2. Scatterplots

- Create a new tab with the original data.
- Make a scatterplot of the data. This is a plot where each individual (each row in the data) is described as a dot, and the x-axis value shows the years of education, the y-axis value shows the income of the individual.
- How does the scatterplot reflect the correlation coefficient from Part 1?

Part 3. Binned plots

- The scatterplot can be hard to read when there are many observations. Different data visualization tools can make the data easier to interpret.
- Create a new tab with the original data.
- For each value of "educ", compute the **conditional** average of income. That means that for each level of education you calculate the average income (Hint: Use Excel formulas "UNIQUE()" and "AVERAGEIF()").
- Make a binned plot where x-axis is the years of education and y-axis is the average income for the given level of education. (Hint: There will be only one dot in the picture for each years of schooling).
- When can a binned plot be more informative than a scatterplot?

Part 4. (**Optional**). Percentiles and IF-statements on Excel

- Compute the 50th percentile of the data (the median) conditional on years of education. (Hint: you should calculate the 50th percentile for all values of years of education; use formula MEDIAN(IF()))

- Compute the 25th and the 75th percentiles of the data conditional on years of education. (Hint: use formula PERCENTILE(IF()))
- How do we interpret the table?