25 novembre 2024

Series 1 : Visual analysis of data

Exercise 1 : Data about scientific publications

Objective

The objective of this exercise is to train the different manners available for loading, manipulating and visualy analyze data.

A the level of Matlab the following specific functions are potentially of interest : livescript, readtable(), home/import data, table(), array2table, categorical(), varfun(), boxplot(), sort(), sortrow(), scater(), gscatter(), semilogy(), hold, grpstats(), publising parameters

Using data relative to the publication published by the world bank and available on Moodle, create a script to produce the following elements (on Matlab you can use the tool *Publish* to produce an output of your work)

1. Load data from Excel into a table ;
2. Add to the initial table the data relative to the population size
3. Look for aberration in the data using different types of graphics
4. Define the columns *code*, *level* as categories ;
5. Build a new table with the data aggregated by *region* ;
6. Create a bubble plot relative to the size of the population and the number of publications by regions
7. Create a scatter plot of the publication in regards to the size of the population, by country, by region
8. Create a box plot of the total publications per years ;
9. Order the data and produce a Pareto type bar plot for the first year and last year of the series ;
10. Analyze each graphic and make remarks that will appear when publishing your report
11. Change the parameter of the tool *Publish* to produce a pdf file without the lines of code.

During the exercise, take time to read the explanatory note (help) of the software to well understand and memorize the use of the different functions

Exercise 2 : Impact of a new drug

Objective

The objective of this exercise is to train the analysis of observational data so that to solve the Simpson paradox by a detailed cause-effect modeling.

The **ratio of healing** and **the post treatment blood pressure** of 700 patients who had access to a new drug are registered. The data is presented in the table below. Among the participants, 350 have taken the drug (with) and 350 did not take it (without). the row « low P » corresponds to the persons whose measured pressure is in the low category, since the row « high P » corresponds to the persons whose measured pressure is in the high category.

a) Analyze the data so that you can draw a causal diagram for the « treatment » and « self-healing » causes.

b) Write a comment based on the analysis of the data and specifying whether the drug is advisable or not.

	Without	With
Low P	81/87	234/270
High P	192/263	55/80
Together	273/350	289/350

	Without	With
Low P	93%	87%
High P	72%	69%
Together	78%	83%