Introduction to epidemiology

Using Python

Iván Andrés Trujilllo

Facultad de Minas
Universidad Nacional de Colombia

Introductory Epidemiology in Python

Python is *programming language* created by Guido Van Rossum. there are two version of python 2.x and 3.x, now the major references are avaliable for python3 , then we wille use it.

Python considerations

python is more flexible and therefore there are diferentes ways of doing the same thing, and this could be considered by someone as difficult and others can be consder it powerful, in this course will try uses the most popular libraries and syntax to standaridized the management of data.

it is easy to know! python is a powerful programming language, not only is a statistics program it is all you can code! you can programming you own medical assitance!

installing Python

Linux

Python comes integrated with some distros of linux as debian, xubuntu and others, if you want installing a ”Enviroment of desktop” of a some linux distribution to do data analysis the xfce version is advisable due is ligth and low in system resource comsuption.

usually In Mac Python 2.x ,come integrated with Mac.

installing python

Windows

in windows you can download of

www.pyhton.org

Python is a black screen!

yes, python is a black screen! dont have buttoms, only code! however there are Integrated Development Enviorement that is designed to improve your work with python, you can downlad Spyder that is free or work with IDLE, that come by default with python.

IDIE

in linux you can install IDLE with $ sudo apt-get instlal idle

Python is only community support!

if you want analyze data with only download python, maybe you cant get much, because of python by default dont support data by analasys, conversely pepople have been created the code to do it.

[language=Python] print("Hello World") we must in agreement with the syntax available to understand better the code. we will write code as is customary written in the Web!

Most popular libraries by statistical analysis

- pandas
- matplotlib
- seaborn
- os
- statsmodels
- scipy
- numpy

Installing libraries

linux

you only need the following code sudo apt-get install python3-library-name

Installing libraries

windows

with terminal pip install package-name

Installing libraries

Mac

open your terminal and type: pip3 install library-name

Installing libraries

Linux

you only need the following code sudo apt-get install python3-library-name

Installing libraries

windows

with terminal pip install package-name

to open a script, when you enter a IDLE you can press **ctrl** + **n** and appear a white block, here que can save all our code, and repeat the tasks, for all databases or other research.

let's star to programming

you must type all that appear in the following theme, is very useful uses the script. [language=Python] import stats print(str(python for all))

STATISTICS library

statistics library is integrated by default with python installation it is useful to calculate, mean, median , mode and other statistics. import statistics data=[1,2,3,4,5] statistics.mean(data) arithmetic mean

statistics functions

1. mean()
2. median()
3. mode()
4. stdev()
5. variance()

OS library

integrated by deafult

OS is a library used to uses acces to the Operative System: Linux, Mac, Windows, etc... [language=Python] import os

Working directory

the working directory is a path that indicates a python where input and otput files for instance datasets, images or documents. to know by default what is the current workign directory: [language=Python] import os print(os.getcwd()) dir=os.getcwd() print(dir)

Change Current Working Directory

[language=Python] os.chdir(”Path”) Note that in python it is a commentary My work space is in the path ”/home/ces/pygy/” os.chdir(”/home/ces/pygy/”) for instance with windows and linux you can go to the folder and press ctrl + l and this get the path to the folder [language=Python] path=os.getcwd() print(path) there must appear your path!

Pandas

pandas is a library that allow us work with a spreadsheet of have some function that are useful to data analisys and management. the class of objects that generate pandas is also known as dataframe.

pandas

[language=Python] import pandas as pd a=[1,2,3,5] type(a) a=pd.DataFrame(a) type(a) **as** put a short name to pandas in this case pd

Pandas

[language=Python]
df=pd.read$_e xcel$(”$file_e xcel_n ame.xls$”)
take in mind that you need to load pandas library, and take care about the name of your excel file, due python is case senstive not is equal ”FILE.xls” that ”file.xls”.

pandas

Describe your data

[language=Python] df.head() is any integer number df.describe()
Describe your numerical data

Simple Merge

[language=Python] joint=pd.merge(base1,base2, on="identifier") if we want concatenate two datasets it is neccesary that in both bases appear a indentification variable, could be a number, or a string.

Simple append

[language=Python] joint=masterbase.append(usingbase) masterbase it is the base we want stack using base.

Epidemiology Research

Hierarchy of causality

- Cases
- case - series
- cohort sectional
- case-control
- experimental
- Meta-Analys

The scope of this material

We introduce to the introdcutory statistics used in the basic epidemiology research, we cover basic topics used in case series, cohort sectional and case-conntrol studies.

Normality

The normality is a concept derived of nature, that was mathematically encripted in some symbols:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Normality usually it is a necessary condition

to statistical inference the normality it is a condition for the assert of some hypothesis, that are very common in clinical research, and other fields of sciences as biology, economics, and physisc.

the values of variables

there's a concept that deriver from the regular law of some variables called as normality, what means that a variable be normal?

Histogram

what is the distribution of the variable regarding it self? plt.hist()

Visual relation between two variables

scatter

```
plt.scatter(x,y,data=”dataname”)
```

Numerical measure of assocation

there are two important distiction so far, we above talk about of linearity, now we can introduce the concept of monotonic function.

Linear function

$$y = \beta + \beta_1 x \tag{2}$$

thus means that changes in x always have the same effect on y. therefore this effect is β_1

$$\frac{dy}{dx} = \beta_1 \tag{3}$$

that in this way we can read as for a unity of change in x, then y growth in β_1 unities.

Monotonic function

there are two reasons to establsih the matter of this.

$$y = f(t) \tag{4}$$

in this case for instance y is the number of bacteria and t is the time transcurred in observation, dont matter how they chage (observed that the baceteria growth in a exponential way) always thath the time trasncurred the number of bacterias grown, in other words a monotic function imply

$$\frac{dy}{dt} > 0 \tag{5}$$

$$x_1 > x_2$$
$$f(x_1 > f(x_2)$$

the following is a linear model with two variables;

$$y_i = \beta_0 + \beta x_i + u \tag{6}$$

y_i is knowed as dependent variable while x_i as independent variable, u is denomined as error term. This way implied something, the variable implied the other.

in this moment β_1 is known as the slope, and means the effect that have the incresea of one unity of the variable x_i on y_i.

naive definition of probability

the probability of a succes or event not is more that the number of favorables events divided the number of possibles outcomes in a experiment.

Odds ratio

usualy used as a mesaure of risk, the odds ratio is only the ratio between two probabilities,