(base) C:\Users\Bea\Documents\Python Scripts\pre-work>python pre-work-math-n-stats.py

==Experimental Design==

i - Statistical units are the surveyed adults both in 1971 and in 2006.

Sampling method unknown, hopefully random, but answering a health survey

usually implies being concerned about your health, so some bias is to be

expected. Also, the non-response rate regarding phisical activity is

noticeable, making it unclear whether the results are representative.

ii - Response variables: BMI

Explanatory variables: intake level, activity level, survey year

iii - It is observational, since no intervention by the scientists was made.

iv - No.

v - I guess there are more explanatory variables than those taken into account.

Or maybe the definition of the same variables has changed.

==Single Variable==

i - Mean: 12.22

Median: 11.50

Mode: 11 (4 times)

ii - Standard deviation: 5.96

IQR: 9.00

iii - Skewness: 0.21

Kurtosis: -0.35

iv - Distribution of frequencies using 9 bins:

[4 4 5 9 4 6 7 0 1]

Cummulative frequencies using 9 bins:

[ 1. 8. 13. 22. 29. 35. 39. 39. 40.]

v - I could have let the software decide, but since we had all integers

and the max value was 28, I decided to use 27 bins of size 1 for a first

run, but the results were too noisy. Then, I decided to divide the number

of bins by three (i.e. bins=9), in the aim of obtaining bins stretching

from integer to integer, so that the obtained classes were simpler to read

and also the result had a much better look, so I kept that value at 9.

vi - Aproximated Mean: 12.21

Aproximated Median: 10.89

Aproximated Mode: 10.89 (9 times)

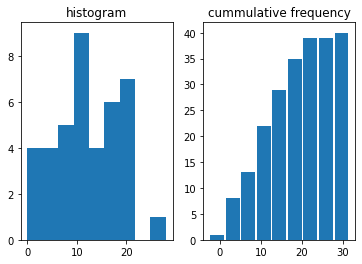
Vii - Aproximated Standard Deviation: 6.33

Aproximated IQR: 9.33

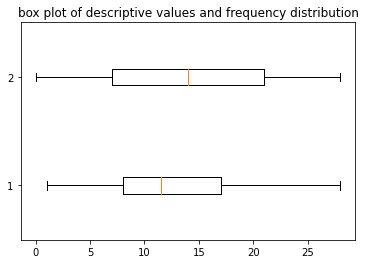
viii - Aproximated Skewness: 0.02

Aproximated Kurtosis: -0.83

ix - See plot



x - See plot



xi - No, they depend on the bin size.

xii - They are pretty similar, but of course not the same since we

rely on an approximation.

I would use the aproximated values since they give me a little bit of

independence from the observations and because they use less computing

power (in the end, they are less numbers).

xiii - The distribution is bimodal, right-skewed and platykurtic.

xiv - As corresponds to a right-skewed distribution, the mean (12.21)

is higher than the median (10.89). Most observations most fall within

IQR (9.33), with some outliers falling in the higher bin

Using median and std deviation allows for more robust results than

mean and std deviation.

==Bidimensional Distribution==

i - The variables have a relationship, their association is linear, strong

and positive.

As per correlation, apparently there is, but we are lacking

information to affirm anything beyond a spurious one.

ii - The linear correlation coefficient is positive.

iii - Given that we have identified a relationship between the variables

with a positive correlation, we could calculate the probability of the

dependent variable falling within a certain range given a certain value

of the independen variable.

I have no idea how to do that so far, but I hope to be able by the end

of the Bootcamp :-\)