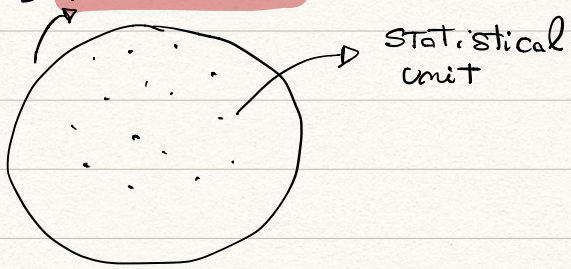
STATISTIS

IS A MATHEMATICAL SCIENCE PERTAINING TO THE COLLECTION, ANALYSIS, INTERPRETATION OR EXPLANATION, AND PRESENTATION OF **DATA**

DATASET is a sort of a collection of data and also the object of statistical analysis, representation, etc...

We fix a set of object that we want to describe

- **POPULATION**



We can arrange the population in a list

U_1
U_2
$\vdots$
U_i
$\vdots$
U_N

we can observe different attribut of each seubject

Weight	Height	Age	...	Sex
x_1	x_2	x_3	...	x_K

x_{i1} rapresent the attribute x_1 of the unit i

N = number of units in the population

Set of unit + Set of attribute = DATASET

DATASET X

STATISTICAL UNITS	ATTRIBUTES X_1	...	X_J	...	X_K
U_1	x_{11}	...	x_{1J}	...	x_{1K}
⋮	⋮		⋮		⋮
U_i	x_{i1}	...	x_{iJ}	...	x_{iK}
⋮	⋮		⋮		⋮
U_n	x_{n1}	...	x_{nJ}	...	x_{nK}

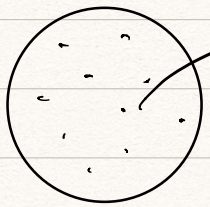
This can be can considered as a matrix in a matematical point of view

Can be seen as a table in an informatic point of view

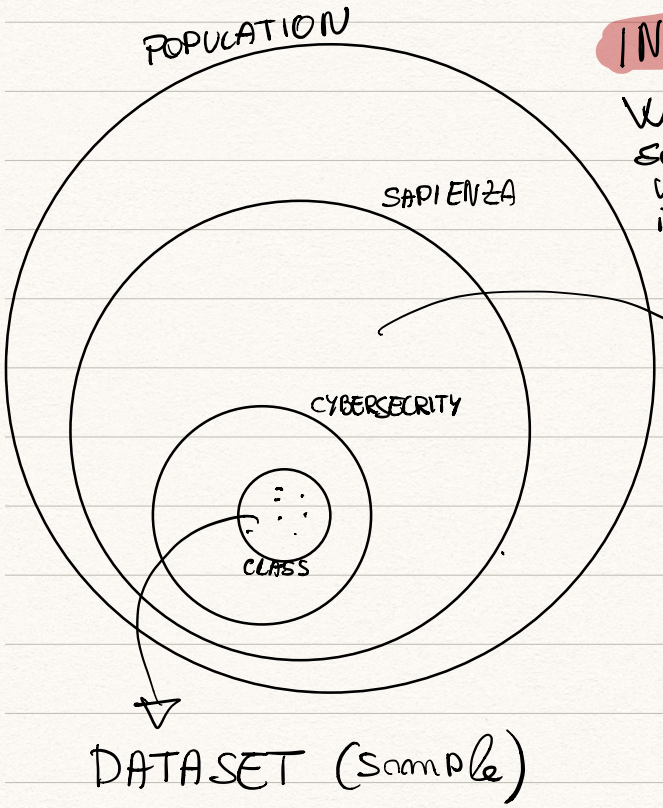
Population with some attribute so we have a dataset this is the easyest kind of setup and is called

DESCRIPTIVE STATISTICS

Because it is just a description of my population.



→ DATASET = POPUCATION



INFERENTIAL STATISTICS

We observe a just a subset of something larger that is unknown but in which we are interested

POPULATION
unknown in is entirety

DATASET
just a subset of POPULATION

SAMPLE
usually is a random sample
(TAKEN WITHOUT BIASES)
hopefully

This distinction can be applied to both inferential and descriptive statistics:

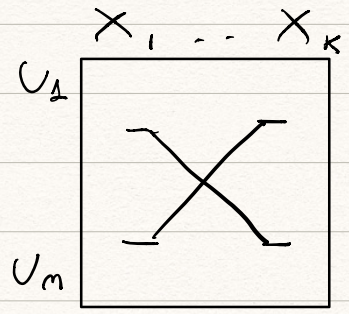
UNIVARIATE

MULTIVARIATE

The distinction is on the structure of the dataset

NORMALLY

x_{ij} con $j = 1 \dots n$ e $j = 1 \dots K$



When we consider only 1 attribute

$\Rightarrow K = 1$ we talk bout

UNIVARIATE STATISTICS

If we talk bout multiple attribute $K > 1$ we talk about

MULTIVARIATE STATISTICS

↳ we can have bivariate trivariate etc...

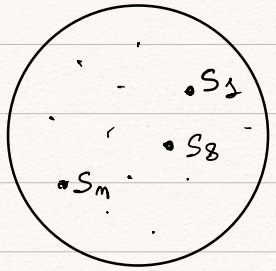
Usually Techniques developed for univariate case can be next extended to multivariate case using matrix mathematical Tools.

EXAMPLE

POPULATION = class

1 attribute

X_1 = Avg grade



50 students

We have a univariate DATASET

So we have just a list or a vector

DATASET

Stud 1	28.1
	27.9
	⋮
Stud m	29.5

⇒

X
x_2
⋮
x_m

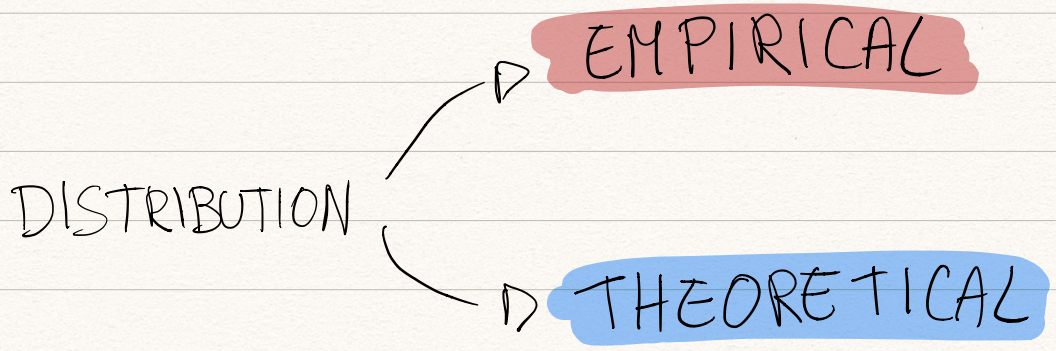
X	Y
x_2	y_1
⋮	⋮
x_m	y_m

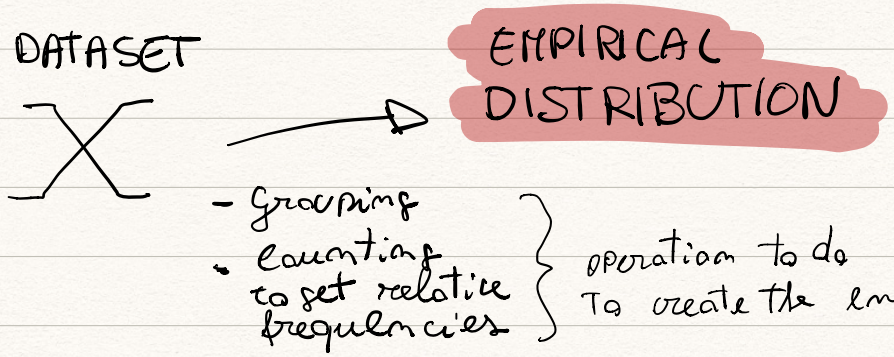
in case of MULTIVARIATE case

From a DATASET we can extract a DISTRIBUTION that is really crucial and central in statistics because from the dataset we have too much informations and we need to take conclusions.

The final goal of statistics is describe the DISTRIBUTION of the population. This can be exact in the DESCRIPTIVE case and can be an assumption/deduction in the INFERENCIAL case.

Distribution can be done both in the UNIVARIATE case and in the MULTIVARIATE case



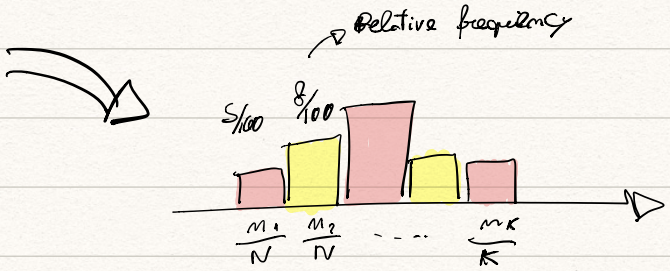


UNIVARIATE DATASET

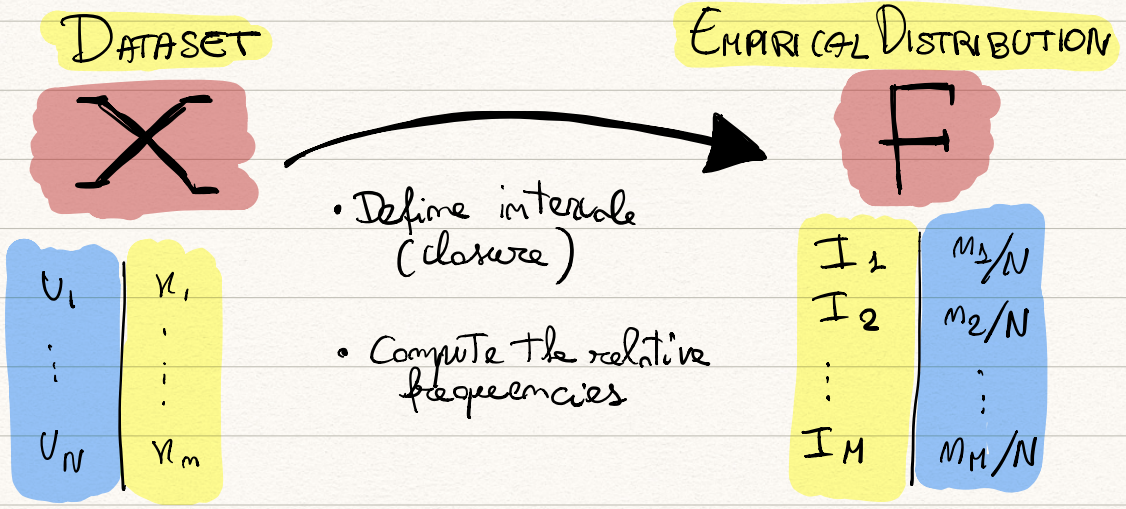
Set of statistics units : students in statistic course

X	Weight
x_1	75
x_2	81
⋮	⋮
x_i	65
⋮	⋮
x_N	92

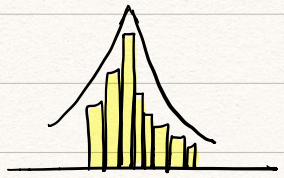
X : Weight



To have a better view of the dataset : create a distribution.
to create this there is an operation of grouping, group the stats units that have similar weight and counting how many there are of them.
I'm essentially determining the relative frequencies



Larger is the DATASET less important are the assumption to create the DISTRIBUTION



In statistics we are not interested in the association between units and the values. This association is discarded when we compute the distribution.

We lose information to gain more knowledge (we gain more Privacy)