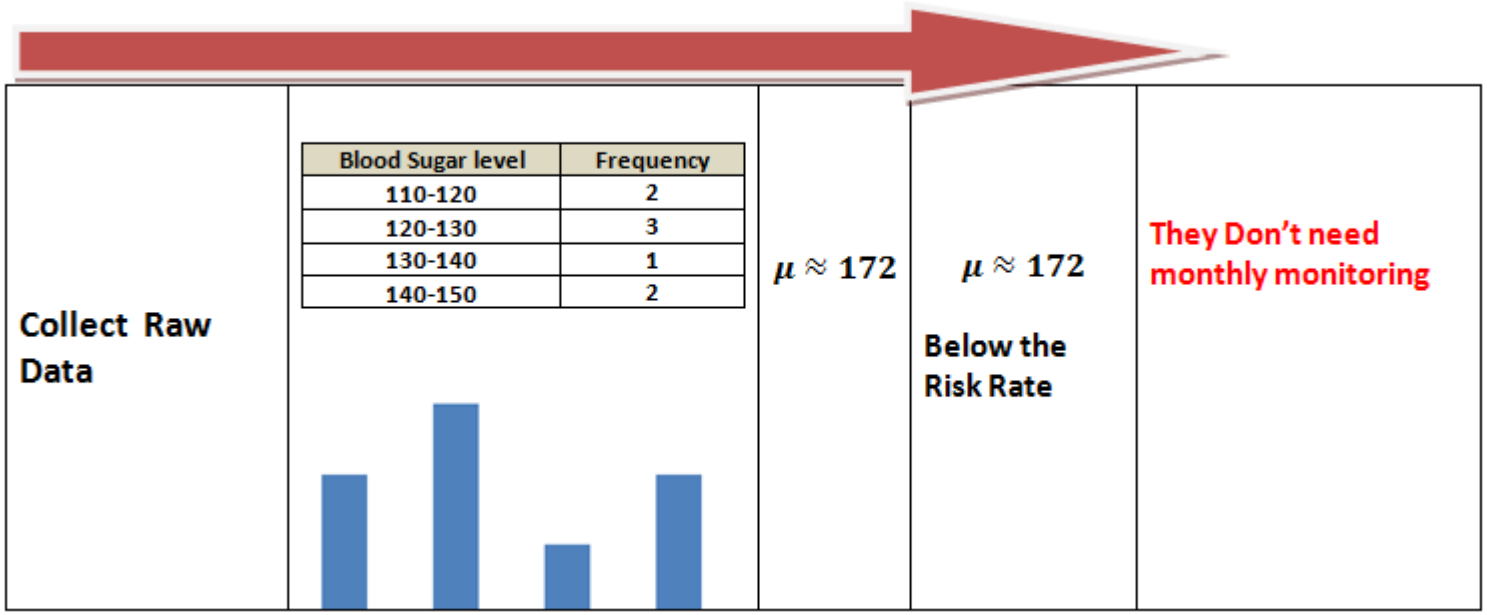
Statistics Models and Methods

By

Dr. Gullanar M Hadi

Statistics Models and Methods

Statistics: is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusion from data.

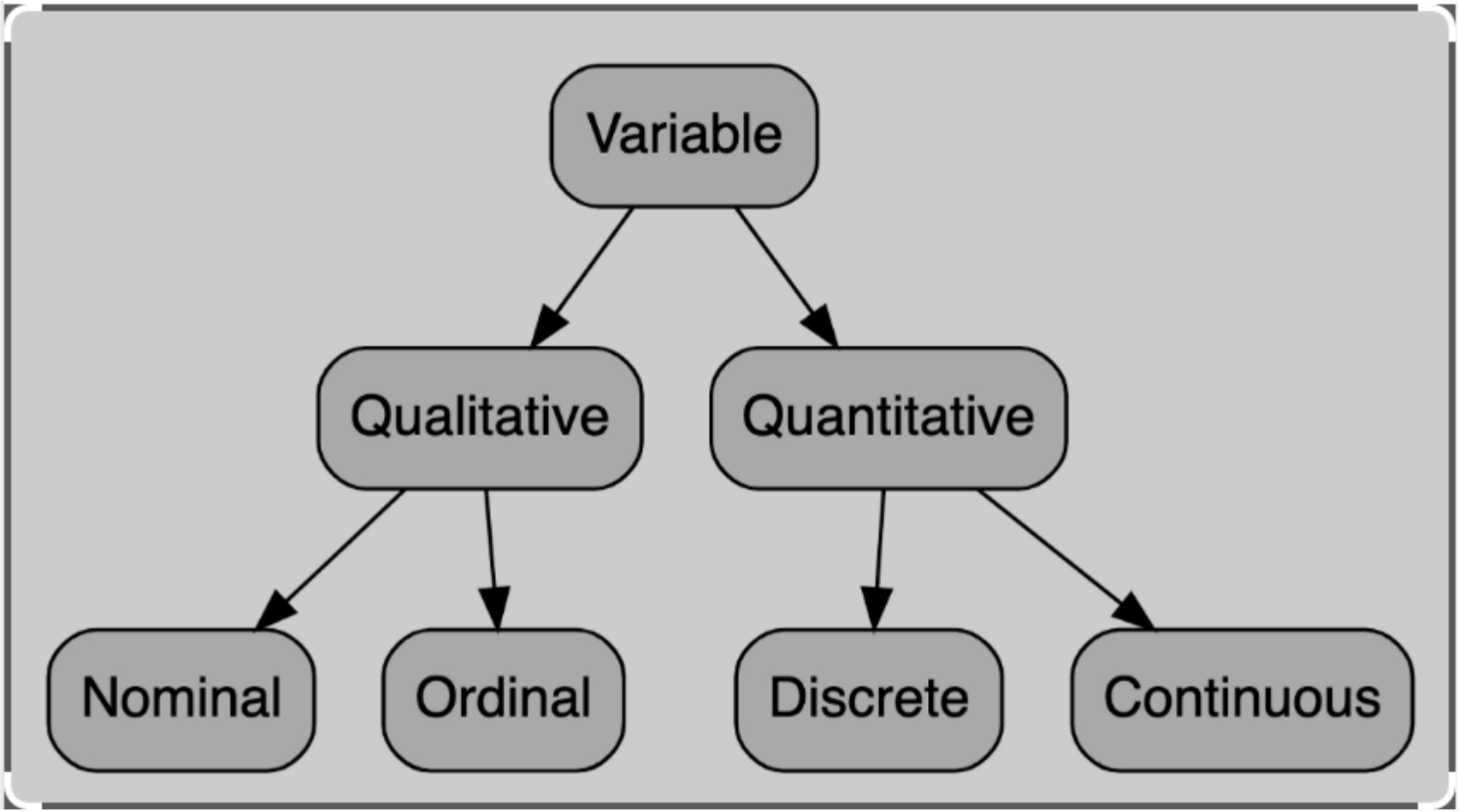


Some Definitions

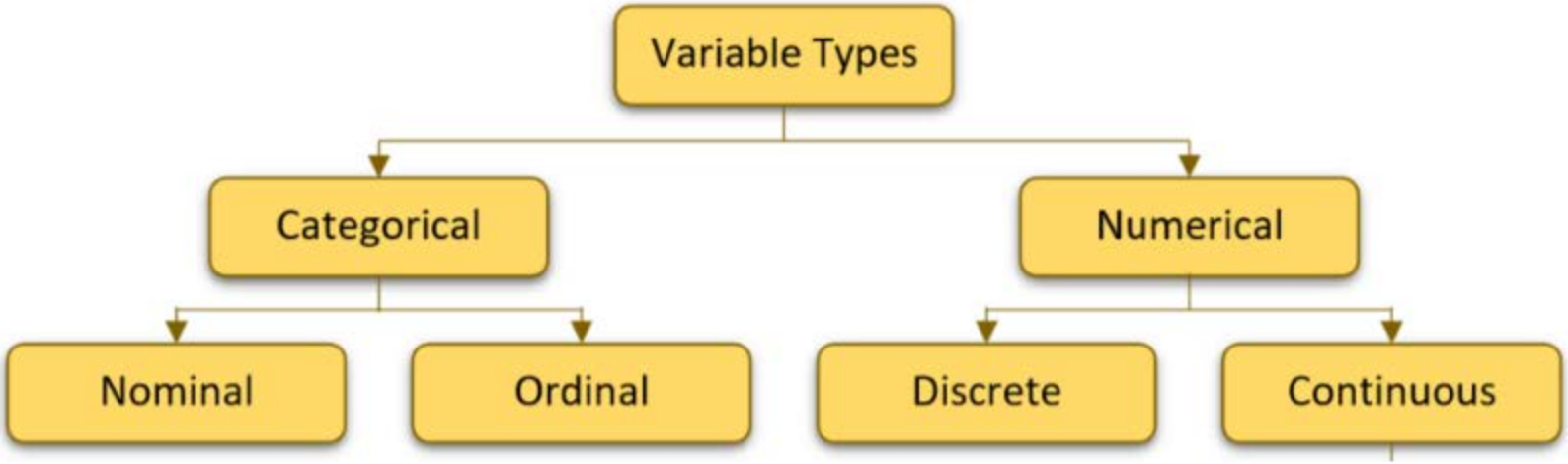
- **Variable: is a characteristic or attributes that can assume different values Age, Name, Gender, Weight, Number of Students, ID**
- **Gender can take different values within the population or within the sample M or F**
- **Data: are the values that the variables can assume**

ID	Name	Gender	Age	⇦ Variable
192384	Ali	M	18	
193870	Ahmed	M	19	⇦ Data
189009	Aya	F	20	

Types of Variables



Types of Variables



We must know the variable if it is qualitative or quantitative because the algorithm which we worked on in machine learning don't work if the target is qualitative it must be quantitative and vice versa to make classification or regression

Qualitative variables are variables that can be placed into distinct categories according to some characteristics or attribute

Variable	Data
Grade	A, B, C, D, F
Gender	M, F
Eye color	Blue, brown, green, Hazel
Nationality	Iraqis, Chinese
Rating scale	Poor, good, excellent
Major field	Mathematics, Computers
Ranking of tennis players	First, second, third

Quantitative are numerical and can be ordered or ranked

they can be arranged sequentially, but not in order of smallest and largest, The reason is that these values do not have a unit of measure. Any number that does not have a unit of measure is treated as a name or categorical qualitative

SID	Name	Age	Gender	Level	Courses	Grade
S01	Ahmed	22	M	1	1	23.5
S02	Ali	19	M	3	3	50
S03	Ashti	24	F	2	4	34
S04	Amna	20	F	4	5	55
S05	Mustfa	21	M	3	6	70
S06	Fatima	23	F	2	2	52.5

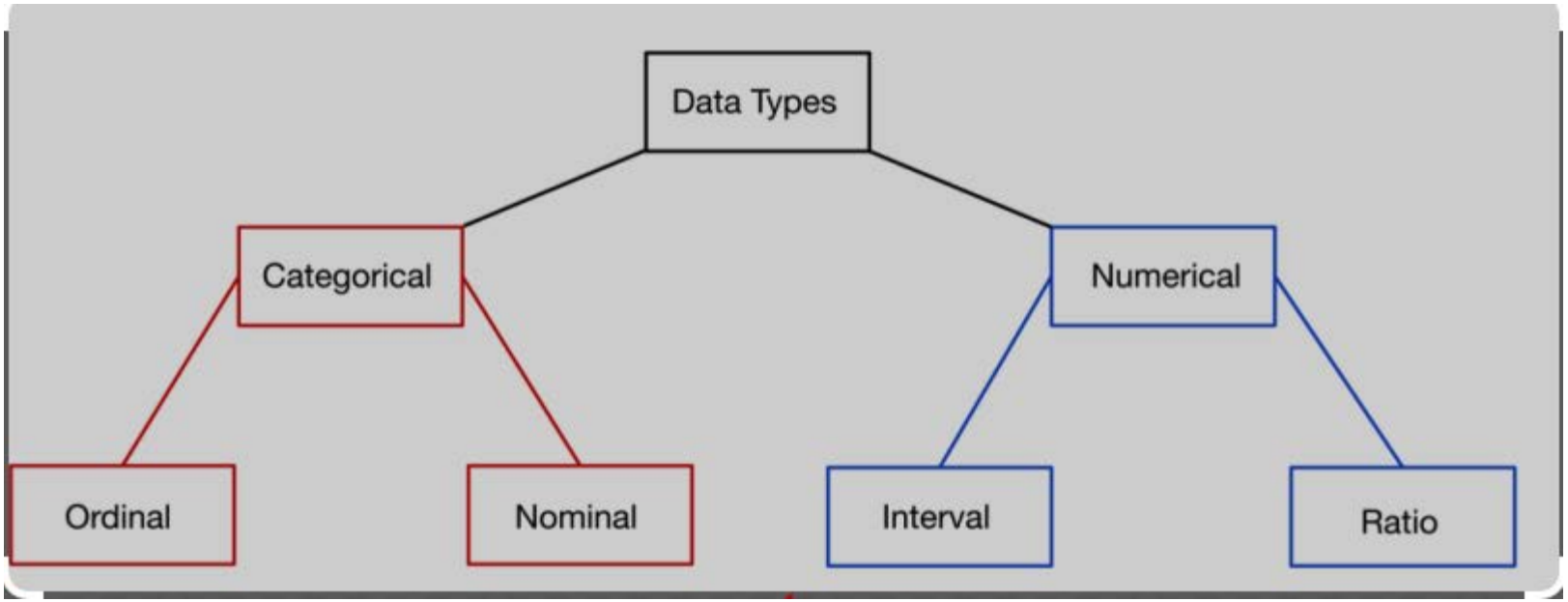
Variable	Data
Length	15 centimeters
Temperature	32 degrees
Time	0.43 seconds
Income	200 ID

Quantitative Discrete Variable assume value that can be counted

Variable	Data
Number of cars	20 cars
Number of Students	40 Students

Quantitative Continuous Variable can assume an infinite number of values between any specific value obtained by measuring. they often include the factions and decimals

Level of Measurements



Qualitative variables

(Descriptions will be through if it can OR cannot be arranged in order of preference (الافضلية)

Qualitative variables

(Descriptions will be through if it can OR cannot be arranged in order of preference (الافضلية))

Nominal level of measurement classifies data into mutually exclusive categories in which no order or ranking can be imposed on the data like Gender, Eye color, Nationality, Major field

Ordinal level of measurement classified data into categories that can be ranked in Grade (A,B,C,D,F),Rating scale (poor, good, excellent)

Quantitative variables

Interval level of measurements

- There is no meaningful zero 0 ≠nothing

  temperature

Ratio level of measurement

- There is meaningful zero 0 = nothing

number of cars

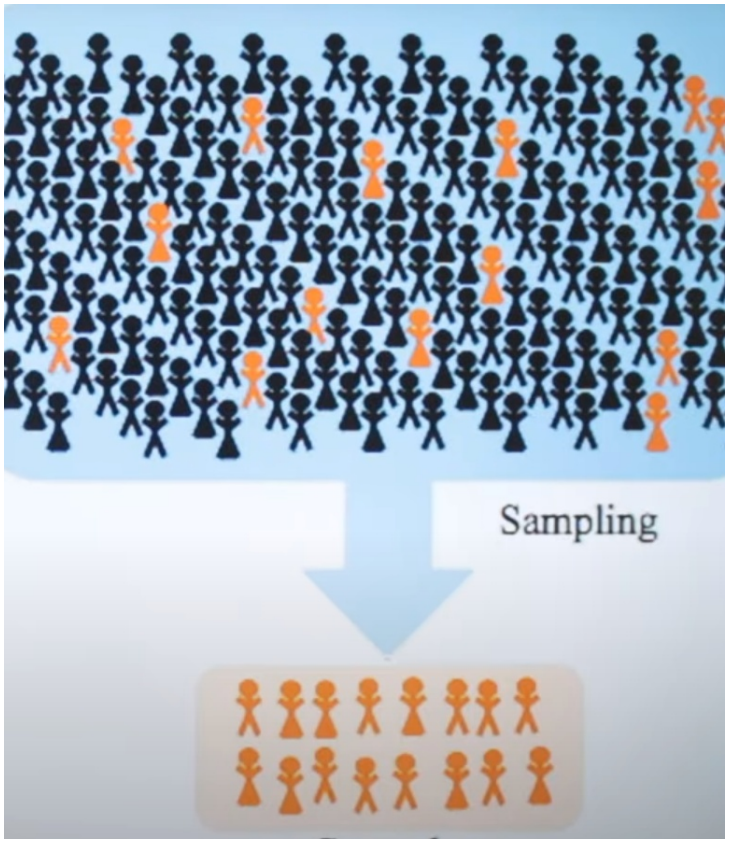
Independent (input) and dependent (target) variables

now we need to deal with the variable in another way, in Machine Learning or in Statistical Model , if we want to build a model for classification or prediction, we must deal with the variable in another way, which is the independent and the dependent

Independent (Input)	Dependent (Target)
Income	Expenditure
Age	Experience
Smoking	Cancer

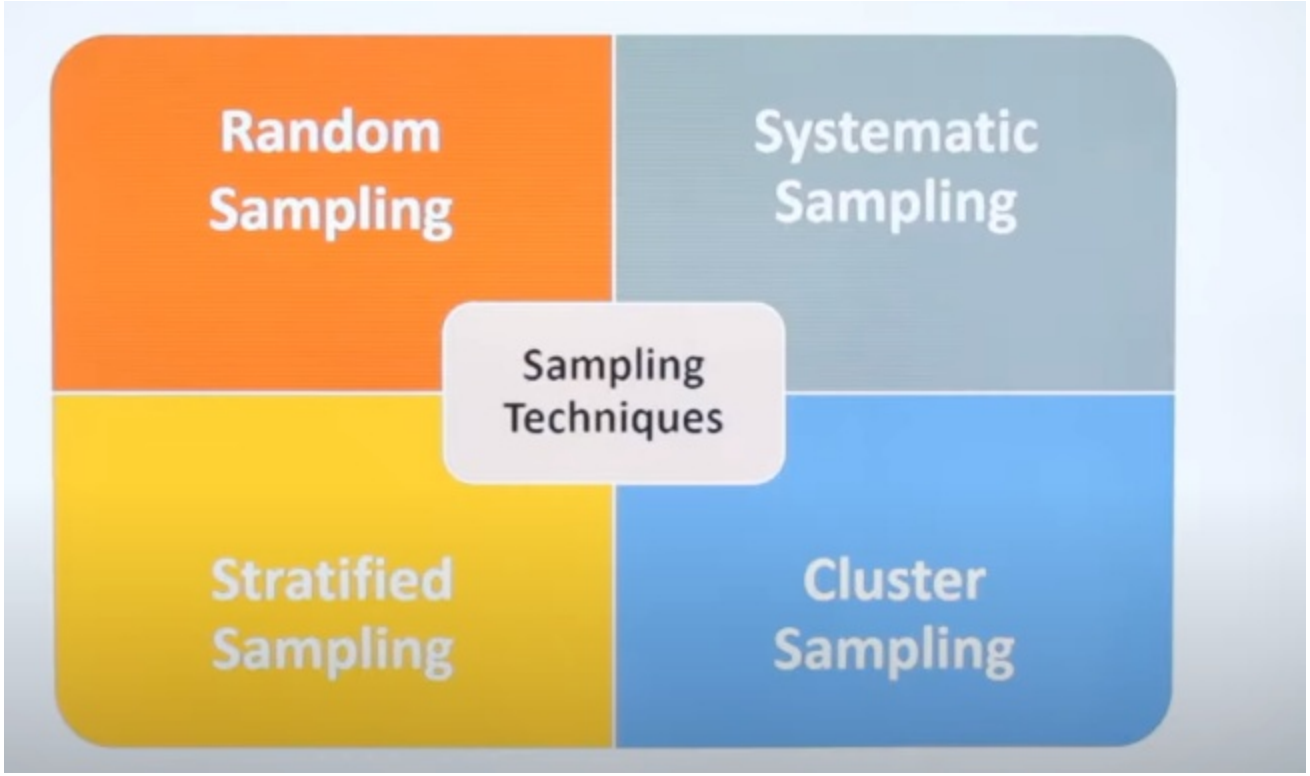
Population and Sample

- Population consists of all subjects (human or otherwise) that are being studied
- Sample is a group of subjects (subset) selected from that population



Sampling Techniques

- What are the best ways to choose the sample? To give it to the decision-makers before starting work

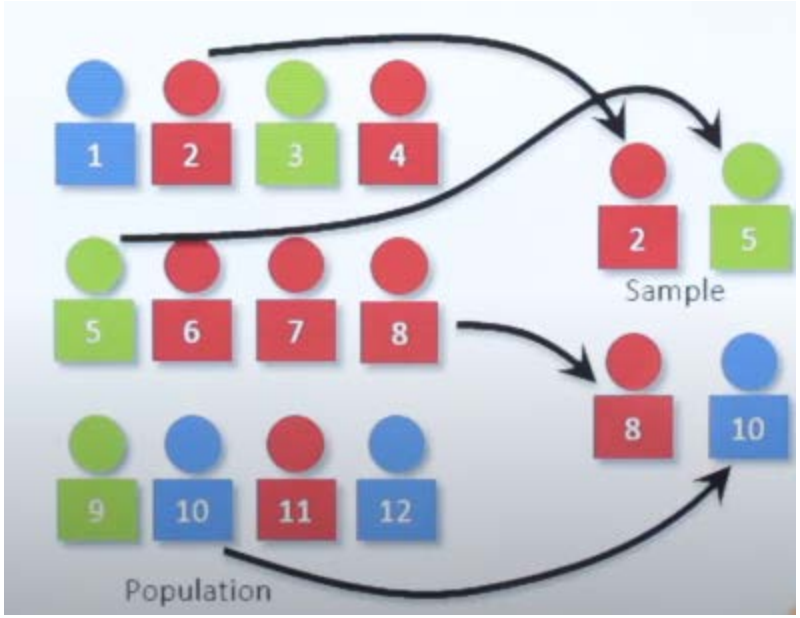


Random Sampling

is a procedure for sampling from a population in which

- the selection of a sample unit is based on chance
- every element of the population has a known, non-zero probability of being selected

all good sampling methods rely on random sampling

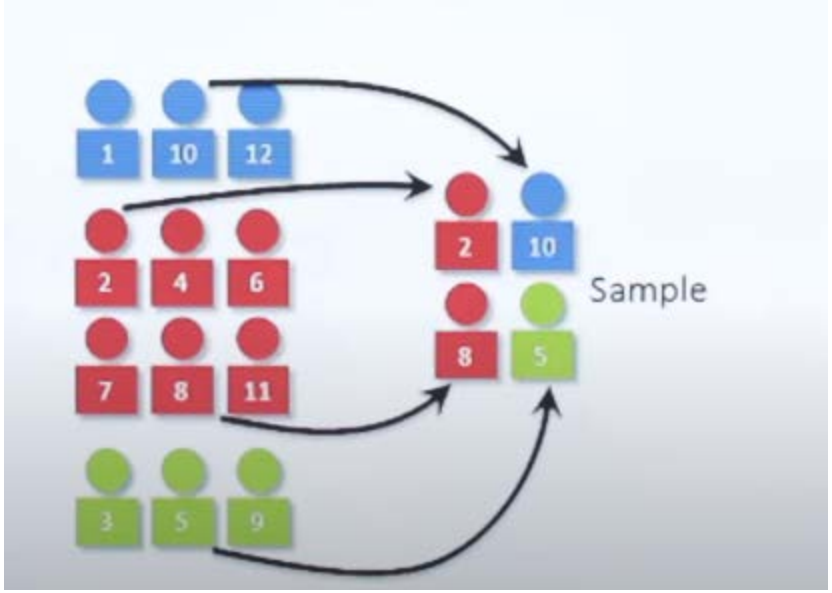


Systematic Sampling

- is a type of probability sample members from a larger population are selected according to random starting point but with a fixed, periodic interval. This interval, called the sampling interval is calculated by dividing the population size by the desired sample size

Stratified Sampling

- referred to a type of sampling method with Stratified sampling, the researcher divides the population into separate groups called strata. Then a probability sample (often a simple random sampl) is drawn from each group

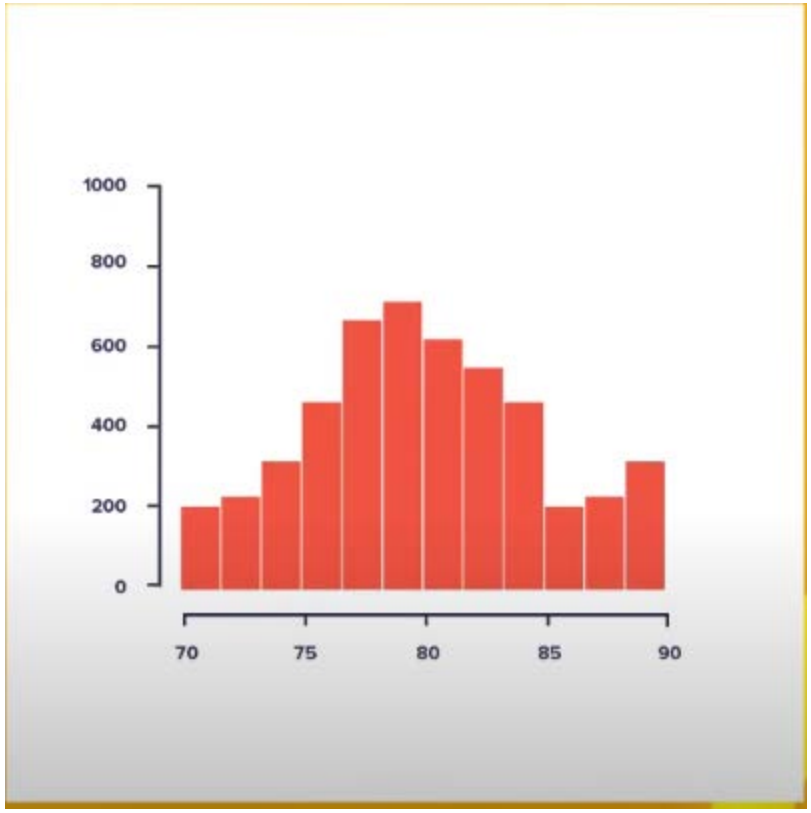


Cluster Sampling

- Cluster sampling is a sampling plan used when mutually homogeneous yet internally heterogeneous grouping are evident in a statistical population… in this sampling plan, the total population is divided into these groups (known as a clusters) and simple random of the groups is selected

Data Understanding

Data Visualization



Data Visualization