What is Mathematical Statistics ?

Science of investigating population's laws

a) *Population*: The set of target objects of study

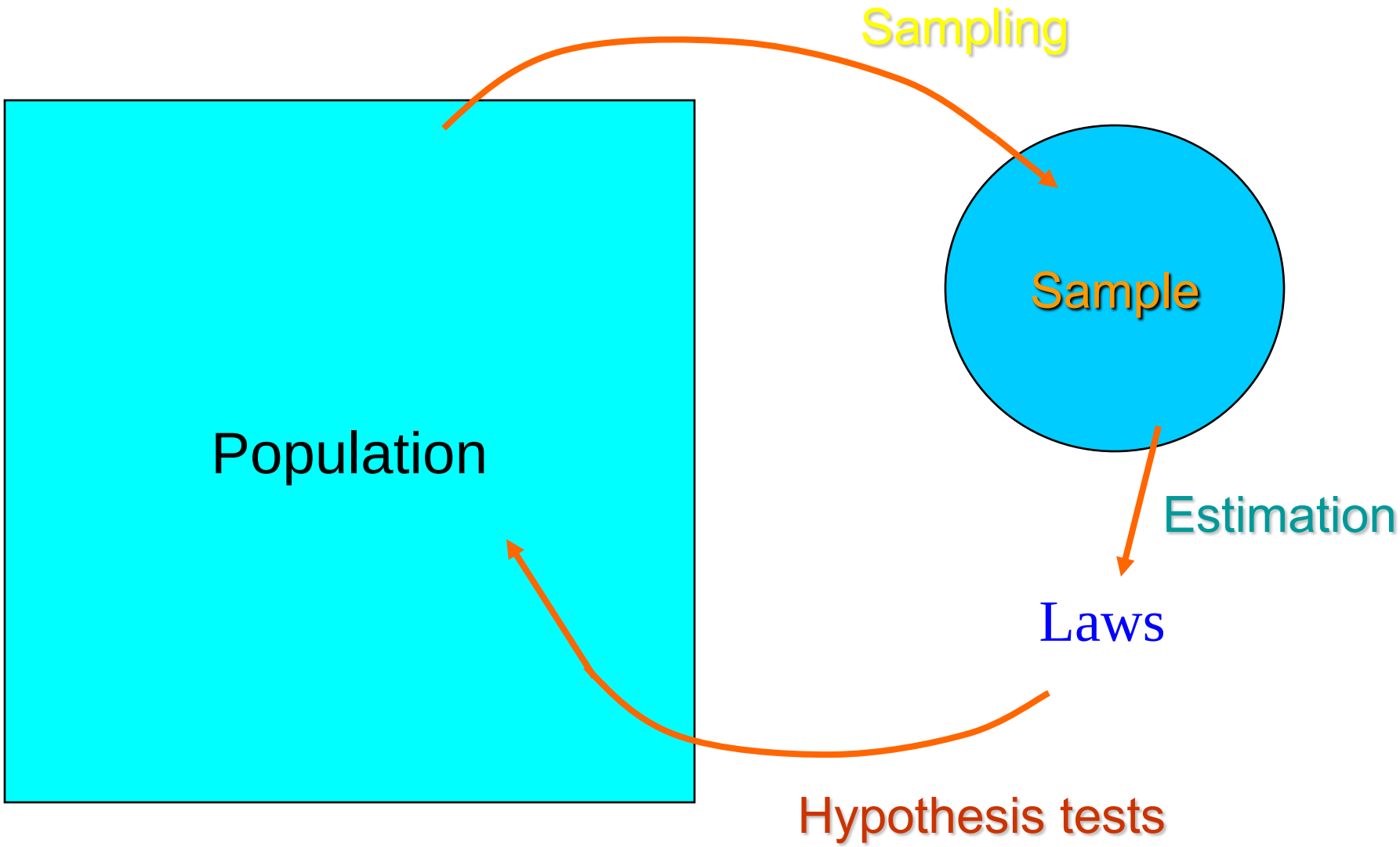
- ***Socio-demographic study:*** ***all citizentsof a given country***

- ***Forestry survey:*** ***All trees in a study region***

- ***Quality control:*** ***All product issues of a factory***

b) Sample

A reasonable small amount of individuals picked out from a given population for a specific study



Data - Coding

DATA: Information, usually numerical or categorical

a) ***Variable*:** (*quantity, characteristic, etc*.)

The characteristic measured or observed when an experiment is carried out or and observation is made, including

- **Characteristics:** Nationality, sex, occupation, etc

- **Measures** Weight, height, age, monthly income, …

- **Answers** to interview questions

- **States, forms** of companies, of study objects, etc.

b) <u>Observation</u>: *(individual, sample unit)*

The set of values of all variables denoted at a given observation, an object, a person or a sample , etc.

c) <u>Value set of variable:</u>

The set of all available values of a given variable

Example: variables Name, Age, Sex, Height, Weight, Housing

VSET(Name) = {A ,. . ., Ba , . . ., Tien , . . ., Yen , . . . , Xuan , . . .}

VSET(Age) = { 1 , 2 , . . . , 100 , . . . },

VSET(***Sex***) = {Male, Female},

VSET(Height) = [0.6 m , 2.30 m],

VSET(***Weight***) = [2 Kg , 150 Kg] ,

VSET(Housing) = {thatched house, brick house, appartment, villa}

2. Variable types

a) Quantitative variables: (measures)

- Continuous variables

Example: Weight, Temperature, Density of a chemical substance in water

- Discrete variables

Example: Income, Salary, Price,

- Integer Variable

Age, Amount of children in household

b) Qualitative variables (norminal or categorical variables)

Charateristics of study object, usually with non-number values

Example: Sex (male-female), Residence place
Reason of borrow (for Health care, for Education, etc.
Occupation (Farmer, Worker, Vender
Transport (by foot, by boat, bicycle, motorbike, car, etc.)

- *Ordered qualitative variables*:

Values of variable can be ordered in certain way, presenting their importance levels.

Example: Housing, Water source, Transport mean, etc.

- *Unordered qualitative variables*: (nominal variables)

Values of variable can not be ranged in order

Example: Ethnic, Occupation, Reason of migration, etc.

c) *Independent variables*

Reasons or factors impacting on studied process

d) ***Dependent variable***:

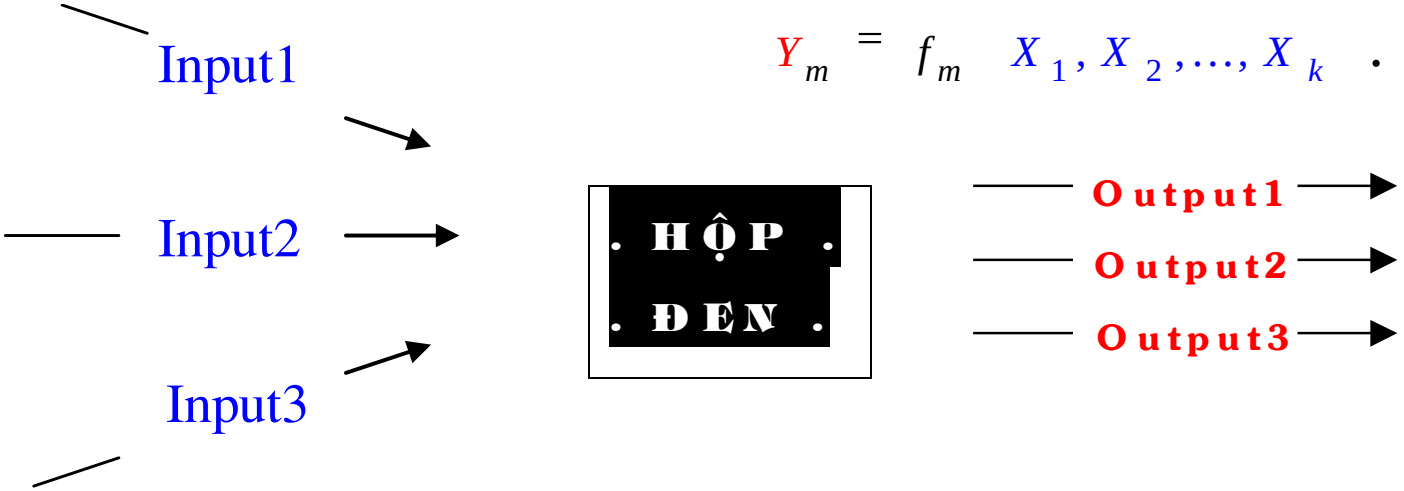
(response variables) – Results, consequences

$$Y_1 = f_1(X_1, X_2, \ldots, X_k),$$

$$Y_2 = f_2(X_1, X_2, \ldots, X_k),$$

$$\ldots\ldots$$

$$Y_m = f_m(X_1, X_2, \ldots, X_k).$$



Example: 1. Education study

- **Dependent variable:** Examination scores

- **Independent variables:** Age, Sex of students, Age, Sex, Teaching methods, seniority of teachers

2. Rice production study

- **Dependent variable:** Rice yield

- **Independent variables:** Land area, Amount of fertilizer used, Water quantity, Air temperature, Season, Region

CODING

Turning collected information into numerical form suitable for computing process

i) *Coding quantitative variables*

Values of quantitative variables are measures

→ The measures are taken directly as codes of variables

ii) Coding qualitative variables

- *For ordered qualitative variables*:

Take integer numbers as codes for ordered levels of a given variable

- *For unordered qualitative variables:*

+ 1-st way : *Coding in the same way as for ordered variables*, Each value of variable → one integer number

+ 2-nd way: From a given variable perform new auxiliary binary variables (impuls variables), each of those takes only two values 0 -1

Example:

a) Coding ordered qualitative variables

"Transport means"

~ By foot → 0
~ By bicycle → 1
~ By motorbike → 2

"Housing"

~ Homeless → 0
~ Thatched house → 1
~ Wooden house → 3
~ Appartment → 5
~ Villa → 6

b) Coding unordered qualitative variables

"Borrow reason": Production, Shoping, Health care, Education, Wedding

1-st way:

~ Production → 1
~ Shoping → 2
~ Health care → 3
~ Education → 4
~ Wedding → 5

2-nd way : Perform 5 new auxiliary binary variables

Main variable	Variable 1 **Production**	Variable 2 **Shoping**	Variable 3 **Health care**	Variable 4 **Education**	Variable 5 **Wedding**
Production	**1**	0	0	0	0
Shoping	0	**1**	0	0	0
Health care	0	0	**1**	0	0
Education	0	0	0	**1**	0
Wedding	0	0	0	0	**1**

4. Organizing data

Data matrix:

- Columns → **variables**,
- Rows → ***Observations***

Example: Demographic survey

	Name	Age	Sex	Income	Height	Weight	Whatching TV	Housing
Person1	*Vân*	*27*	*Female*	*650000*	*1m55*	*55Kg*	*Every day*	*Hired*
Person 2	*B- ờng*	*46*	*Male*	*980000*	*1m68*	*67Kg*	*Rarely*	*Brick H.*
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Person 40	*Việt*	*31*	*Male*	*775000*	*1m73*	*58Kg*	*Every day*	*Wooden*
Person 41	*Canh*	*77*	*Female*	*325000*	*1m49*	*46Kg*	*Never*	*Thatched*

↓

1	VAN	27	2	650	1.55	55	2	0
2	BUONG	46	1	980	1.68	67	1	5
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
40	VIET	31	1	775	1.73	58	2	3
41	CANH	77	2	325	1.49	46	0	1

Exercise

- Determine the list of variables (quantitative, qualitative – ordered – unordered) present in the survey questionaire
- Determine the set of possible values of each variable in the above list
- Make the coding for the mentioned variables