Introduction to Statistics

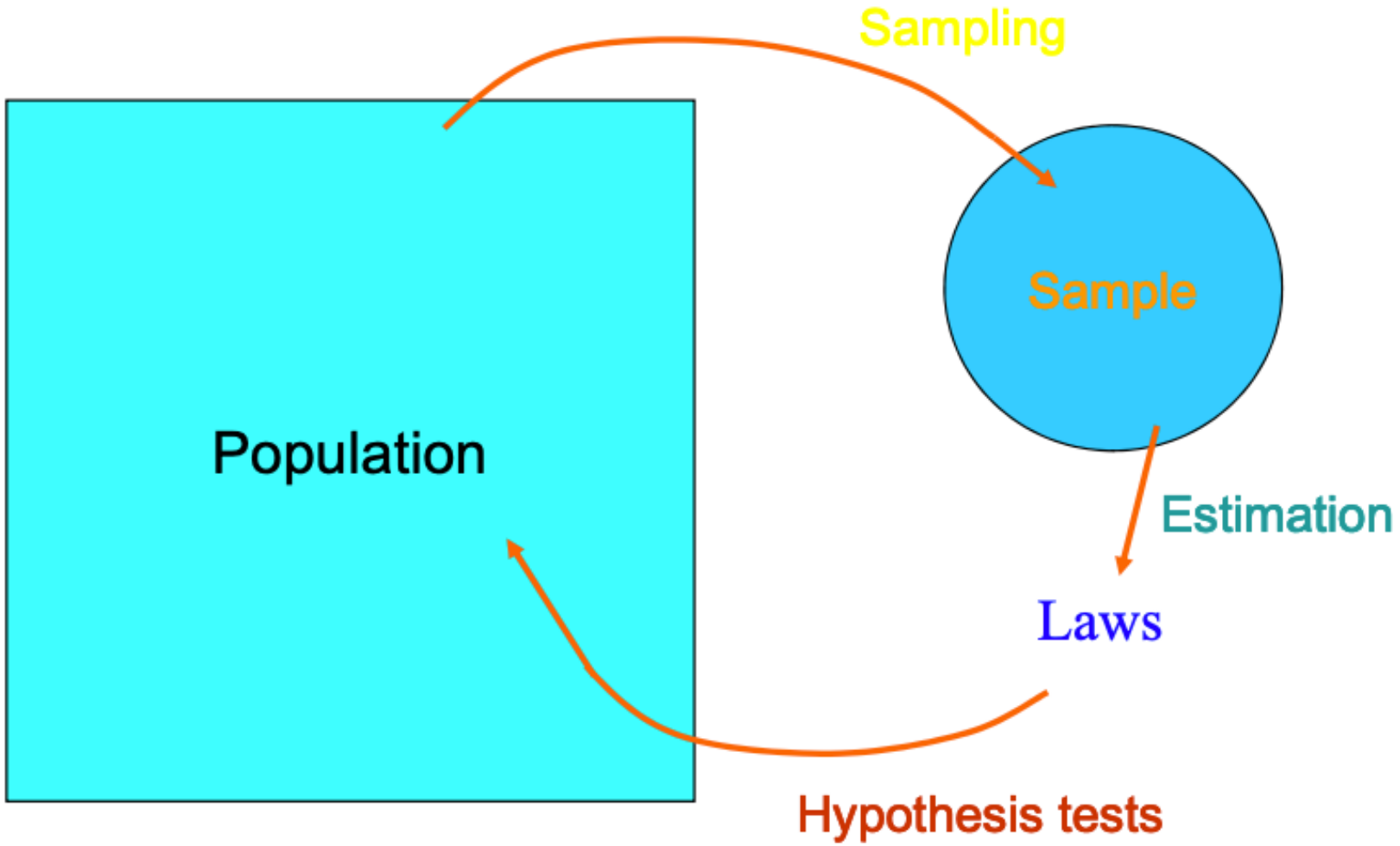
Road Map

- What is Statistics?
- Sampling models
- Statistical data
- Coding variable
- Organizing data

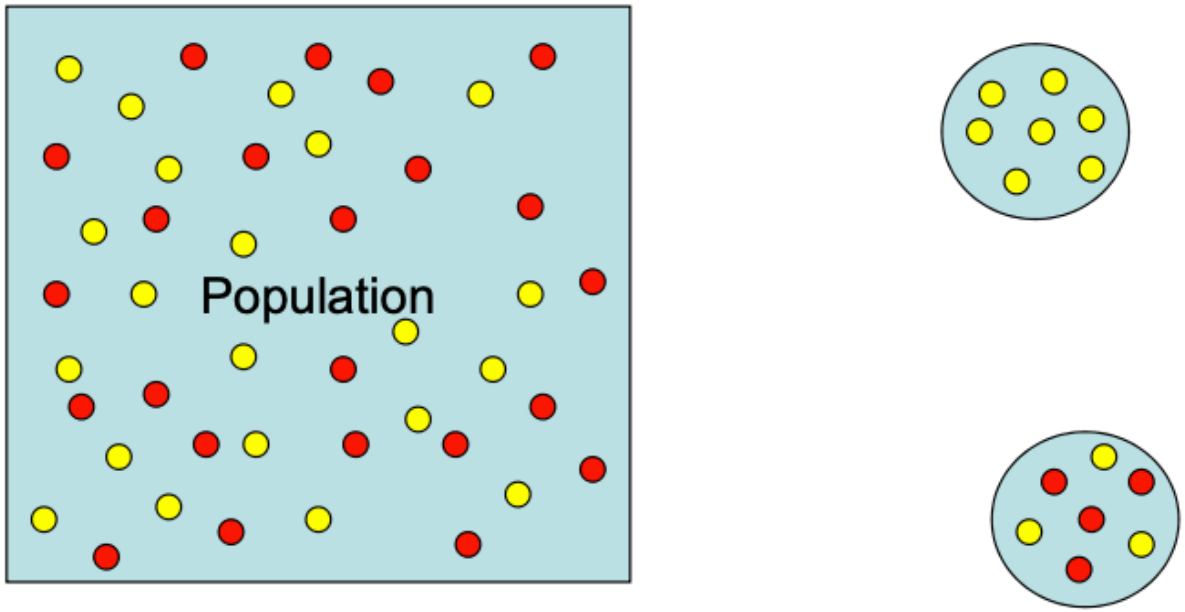
What is Statistics?

- Science of investigating population's laws
- (a) Population: The set of target objects of the study
  - Socio-demographic study: All citizens of 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

Population and Sample



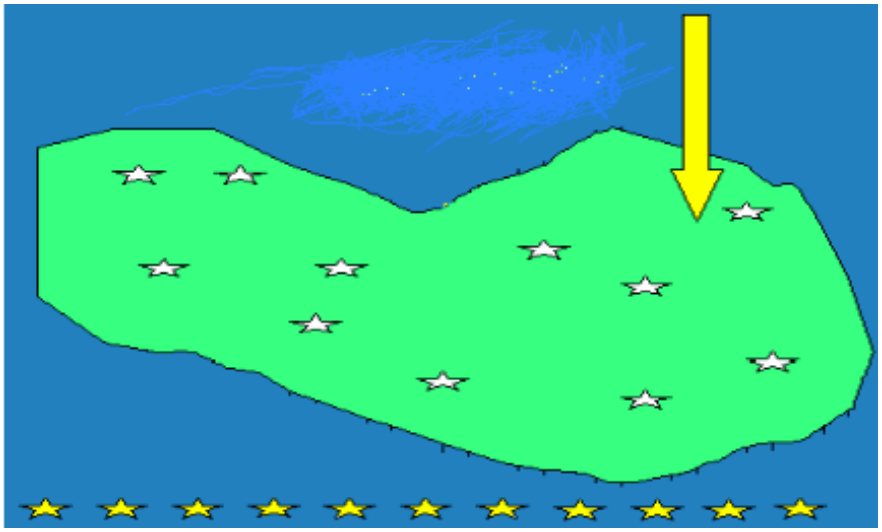
Sampling



- Sampling: process and method to select the sample
- How to do the sampling:
  - Representative for population of the study
  - Corresponding to the study target

Sampling models

- (a) One sample model:



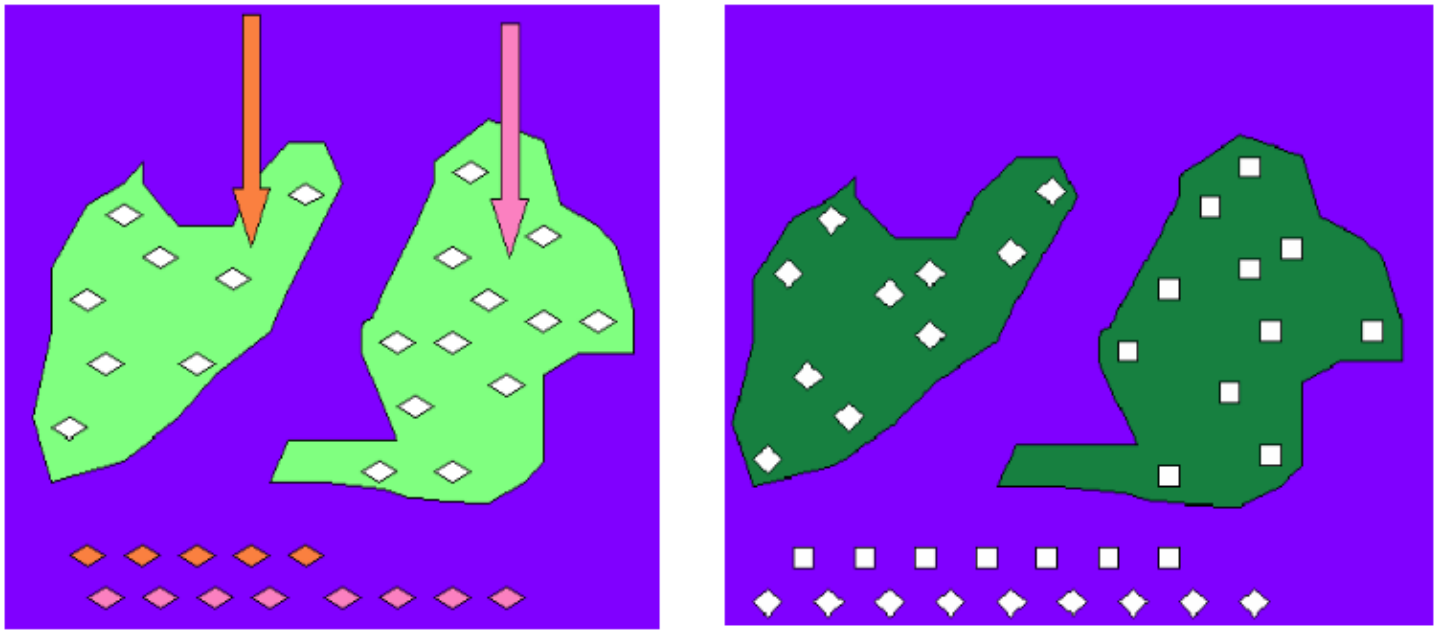
  - usually concerns with an intervention on population: If the intervention should make some change in the population?
  - Choose individuals from the population randomly to perform a sample

Sampling models

- Perform one sample model to investigate:
  - Example 1: If in Ha Dong 90% motorcyclists use helmets
  - Example 2: If in Ha Noi proportion of girl students less than 50%
  - Example 3: If in Viet Nam bred breeding is popular among more than 70% women

Sampling models

- (b) Two independent sample model:



  - Model of two groups of objects with different:
    - Intervention levels
    - Individual proper

Sampling models

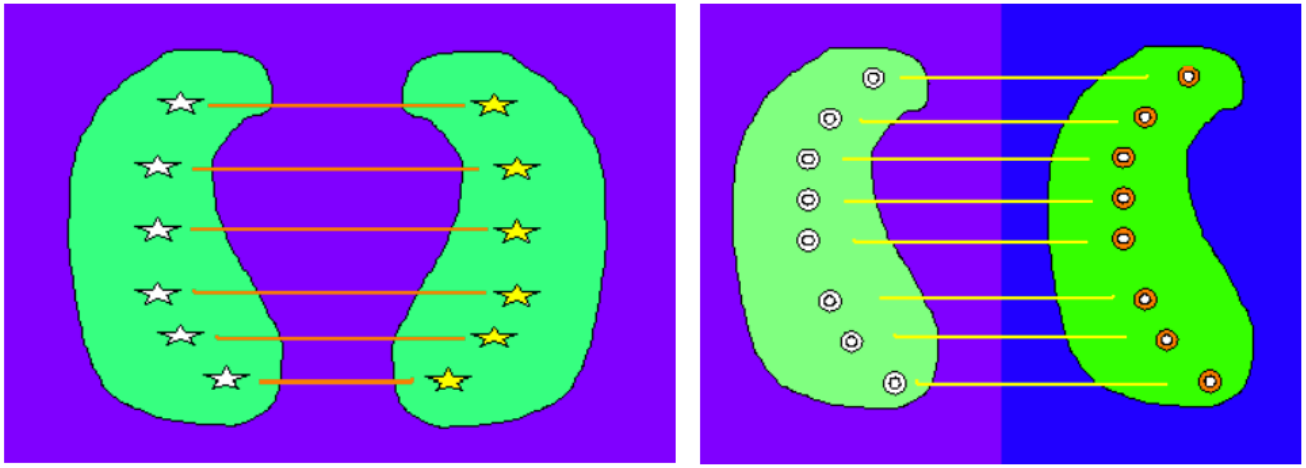
- In the model of two independent samples:
  - Numbers of observations in two groups (sample sizes) may be different
  - Observations of each group are independent from those of the second group
  - Sampling: Observations must be randomly selected from each of two groups

Sampling models

- Perform two independent sample model to investigate:
  - Example 1: If women are better in foreign languages than men
  - Example 2: If there is any difference between Ha Noi and Ho Chi Minh City in immigration from rural areas
  - Example 3: If quality of coffee produced in Lam Dong is different than that in Dak Lak
  - Example 4: If number of traffic accidents per month in Ba Dinh district decreased after 31/12/2023

Sampling models

- (c) Model of two dependent (paired) samples:



  - Two dependent sample model is used in a study when:
    - Each object in the first sample is chosen together with a similar (paired) object in the second sample, or
    - Any object in the second sample is the same one in the first sample, but the measures in the two samples are taken under different conditions

Sampling models

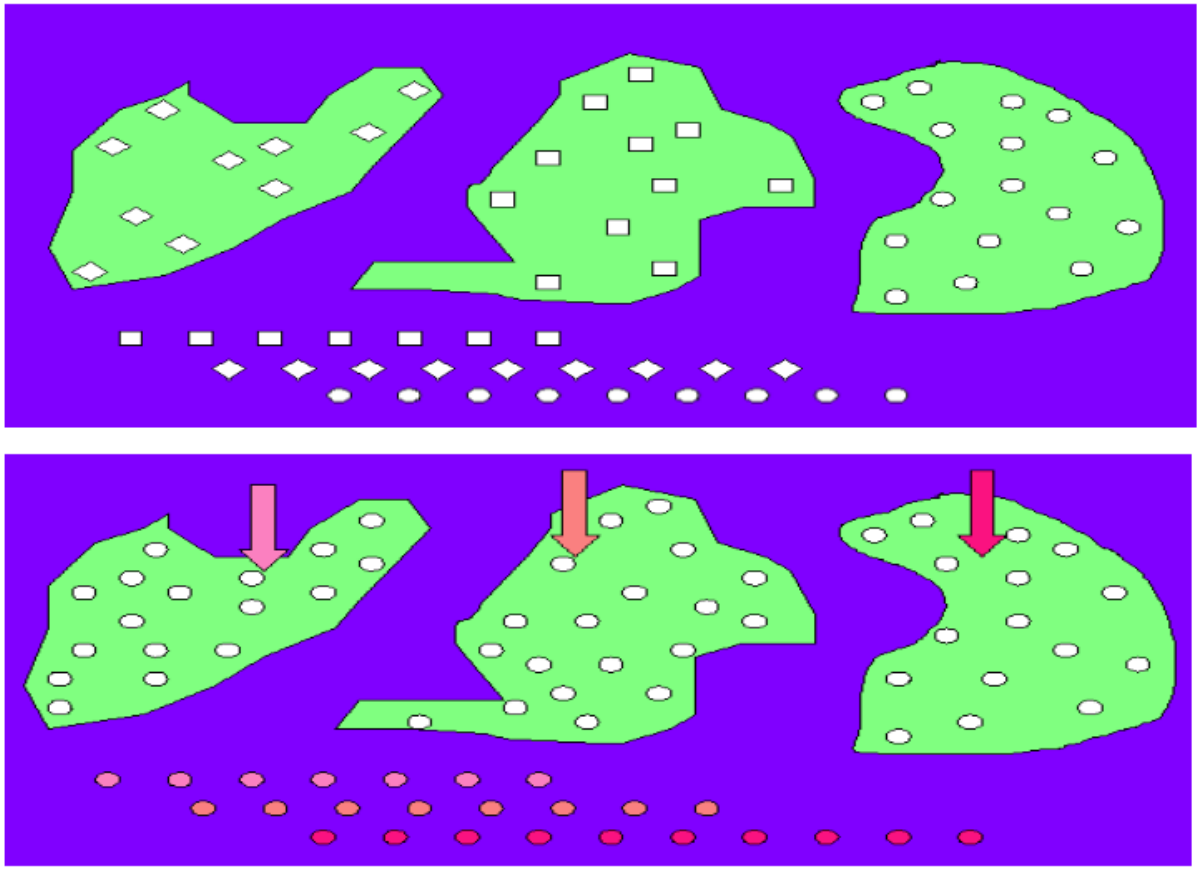
- In the model of two paired samples:
  - Observation amounts (sample sizes) of two samples are equal
  - Information taken from one observation is related with that of correspondingly paired observation
  - In pairing to perform the samples, all factors which may influence on study issues must be taken into account

Sampling models

- Perform two paired sample model:
  - Example 1: To investigate the influence of cigarette on hypertension disease: perform two samples of smoking and non smoking people, each person from the non-smoking group is paired with one smoking person similar about age, sex, weight, height, occupation, etc.
  - Example 2: Comparing 2022 and 2023 to investigate if there is a changing in persons' opinion about Covid Vaccination

Sampling models

- (d) Model of multi-independent samples:



Sampling models

- In the model of multi-independent samples:
  - Numbers of observations in groups (sample sizes) may be different
  - Observations of each group are independent from those of the other groups
  - Sampling: Observations must be randomly selected from each of groups

Sampling models

- Perform multi-independent sample model to investigate:
  - Example 1: Compare examination results of several high schools in Ha Noi
  - Example 2: Compare salary in different economic sectors
  - Example 3: Compare water supplying of ethnic groups

Statistical data

- Data: Information, usually numerical or categorical
- Elements, Variables, and Observations in statistical data:
  - Elements (study units, objects): are the entities on which data are collected.
  - A variable is a characteristic of interest for the elements.
  - The set of variables' measurements obtained for a particular element is called an observation.
  - A data set with *n* elements contains *n* observations.

Variable types

- (a) Quantitative variables (measures)
  - Continuous variables (e.g. weight, temperature, density of a chemical substance in water)
  - Discrete variables (e.g. income, salary, price)
  - Integer variable (e.g. age, amount of children in household)

  → Quantitative variable data indicate how many or how much

  → Quantitative variable data are always numeric

Variable types

- (b) Qualitative variables (nominal or categorical variables)
  - Characteristics of the study object, usually with non-numeric values
  - Examples: Gender (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.)

Variable types

- Ordinal qualitative variables:
  - Values of variables can be ordered in certain way, presenting their importance levels
  - Example: Housing, Water source, Transport mean, etc.
- Unordered qualitative variables (nominal variables):
  - Values of variables can not be ranged in order
  - Example: Ethnic, Occupation, Reason of migration, etc.

Variable types

- Examples of variable types:
  - Given variables: Name, Age, Gender, Height, Weight, Housing

    VSET(Name) = {Ba, Hoa, Lan, …}

    VSET(Age) = {1, 2, 3, …}

    VSET(Gender) = {Male, Female}

    VSET(Height) = {0.6 m, 2.30 m, …}

    VSET(Weight) = {2 kg, 150 kg, …}

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

Coding variable

- 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

Coding variable

- (ii) Coding qualitative variables
  - For ordered qualitative variables
    - Take integer numbers as codes for ordered levels of a given variable
  - For unordered qualitative variables
    - 1st way: Coding in the same way as for ordered variables, each value of variable → one integer number
    - 2nd way: From a given variable perform new auxiliary binary variables, each of those takes only two values 0-1

Coding 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

Coding example:

- (b) Coding unordered qualitative variables
  - "Debt reason": Production, Shopping, Health care, Education, Wedding
  - 1st way:

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

  - 2nd way: Form up 5 auxiliary binary variables

Main variable values	Variable 1 Production	Variable 2 Shopping	Variable 3 Health care	Variable 4 Education	Variable 5 Wedding
Production	1	0	0	0	0
Shopping	0	1	0	0	0
Health care	0	0	1	0	0
Education	0	0	0	1	0
Wedding	0	0	0	0	1

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	Bường	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