- Module Research & Biostatistics

Medical statistics I
Sampling & types of data & Descriptive statistics


Year
Second
year
Medical statistics I

CONTENT

Definition of medical statistics
Population & samples & sampling techniques
Diff between data & information
Types of variables
Measures of central tendency(mean –median-mode)
Measures of dispersion

Learning Outcomes

At the end of the lecture, the students should be able to:

A. Knowledge & understanding

A1 Differentiate between types of samples
A2. Differentiate between different types of variables
A3. Recognize measures of central tendency
A4. Recognize measures of dispersion

B. Intellectual:

B1. Select appropriate sampling technique
B2. Select summary measure in quantitative data
B3. Choose the best summary measure for qualitative data

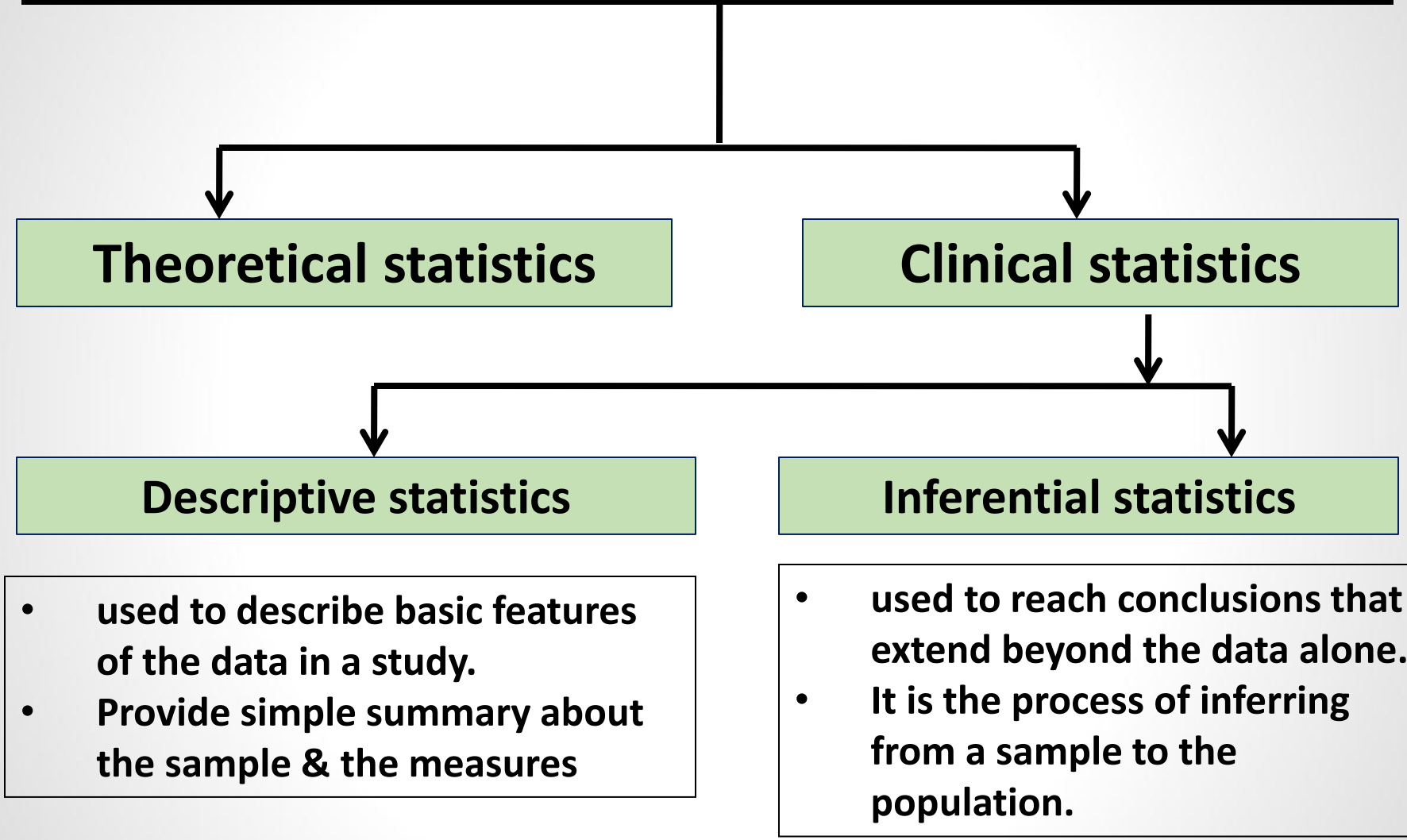
c. Practical

C1.Solve problems related to sampling techniques
C2. Calculate simple measures of central tendency
C3. Interpret summary measures

- **Def of medical statistics**

- **It is the study of methods of *collecting, presenting* (descriptive statistics), *analysing* and *evaluating* conclusions from data (inferential statistics).**

Science of Statistics



Importance

- It presents facts
- It simplifies mass of figures
- It reduces the volume of data
- It facilitates comparison
- **It helps in:**
- formulating and testing hypothesis
- formulation of suitable policies.
- measuring the standard of health.

Outcome1

Sampling

- **Sample is a subset of population that is used to gain information about the entire population.**

A good sample :representative-adequate-unbiased

Why Sampling?

Lower cost

Saves time

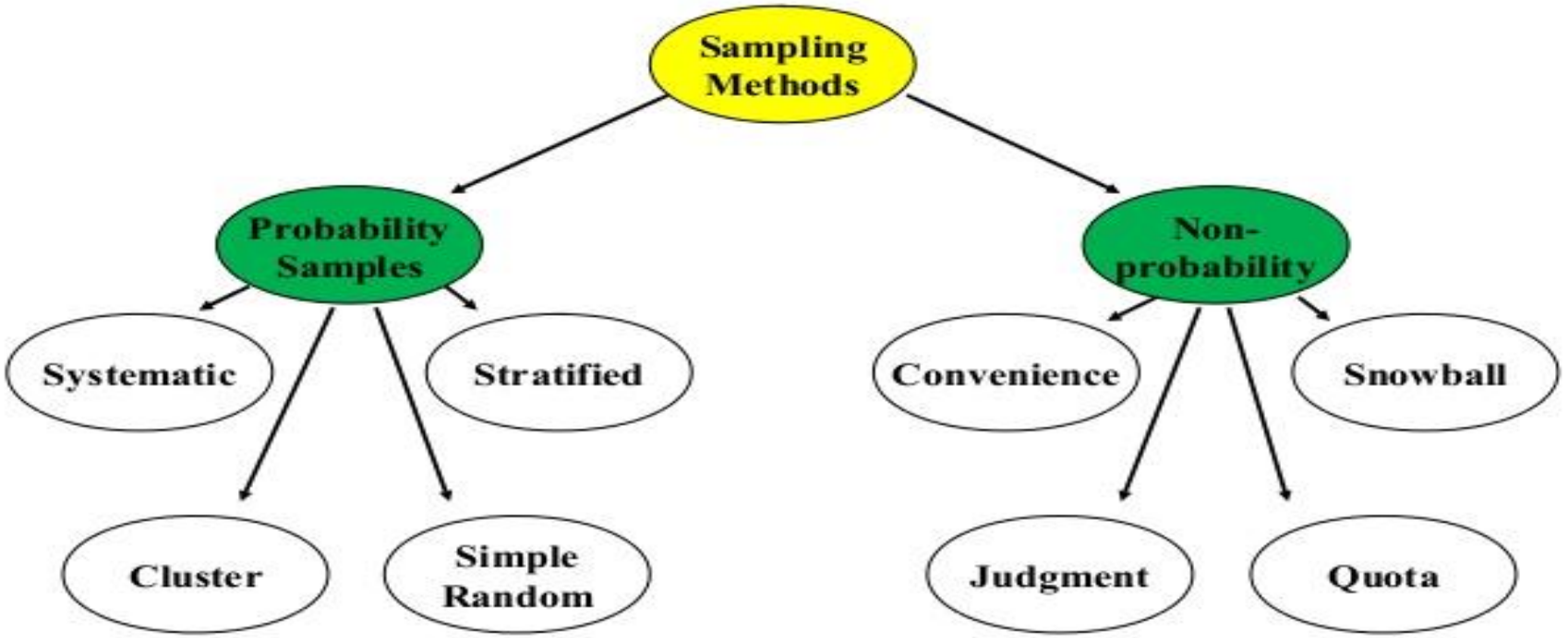
Provides more intensive and accurate investigations and information.

What happens when there is no sampling ?

Selection Bias (non representative sample):

systematic difference between the characteristics of the people selected for a study and those who are not.

Classification of Sampling Methods



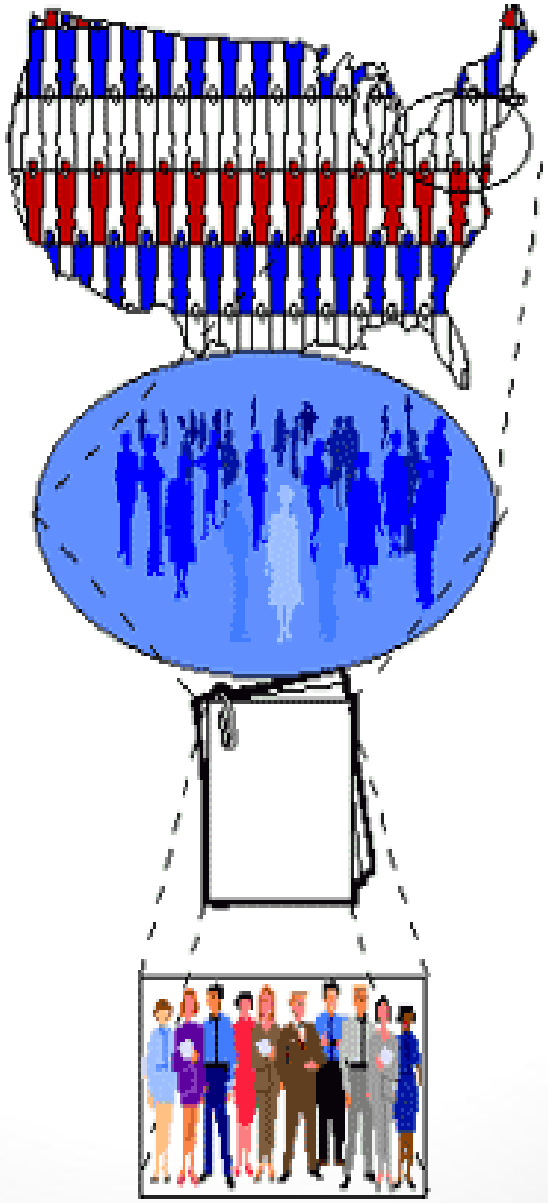
Every unit in the population has a chance (greater than zero) of being selected in the sample

Some units of the population have ***no*** chance of selection or where the probability of selection can't be accurately determined.

Population & sample

Who do you want to generalize to?

The Theoretical Population



What population can you get access to?

The Study Population

How can you get access to them?

The Sampling Frame

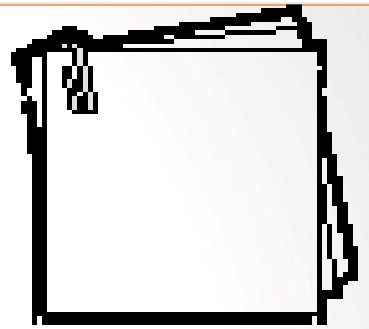
Who is in your study?

The Sample

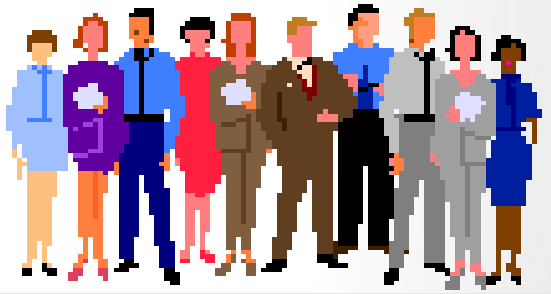
Population characteristics	Appropriate sampling technique
I. Population is a homogeneous mass of individuals	**Simple Random Sample**
II. Population is heterogeneous, consists of definite strata each of which is different, characteristics	**Stratified Random Sample**
III. Sample unit is a group not an individual •They are selected randomly from all groups of same type •All members of selected group will be included in the study	**Cluster Sample**
IV. Population is a confined community	**Systematic Random Sample**
V. Population is distributed over a large geographical area as in national surveys	**Multistage Random Sample**

1. Simple random sample

List of Clients



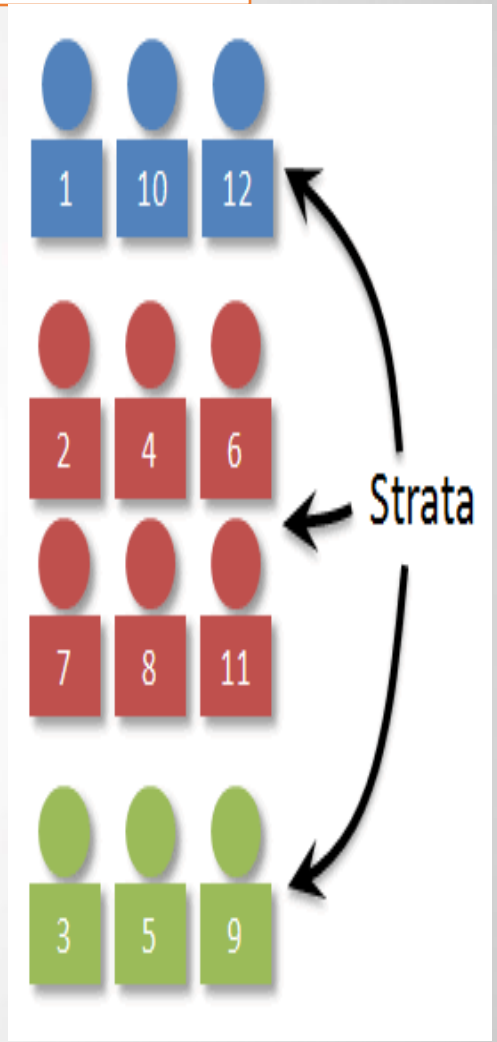
Random Subsample



Using random number table or computer programs

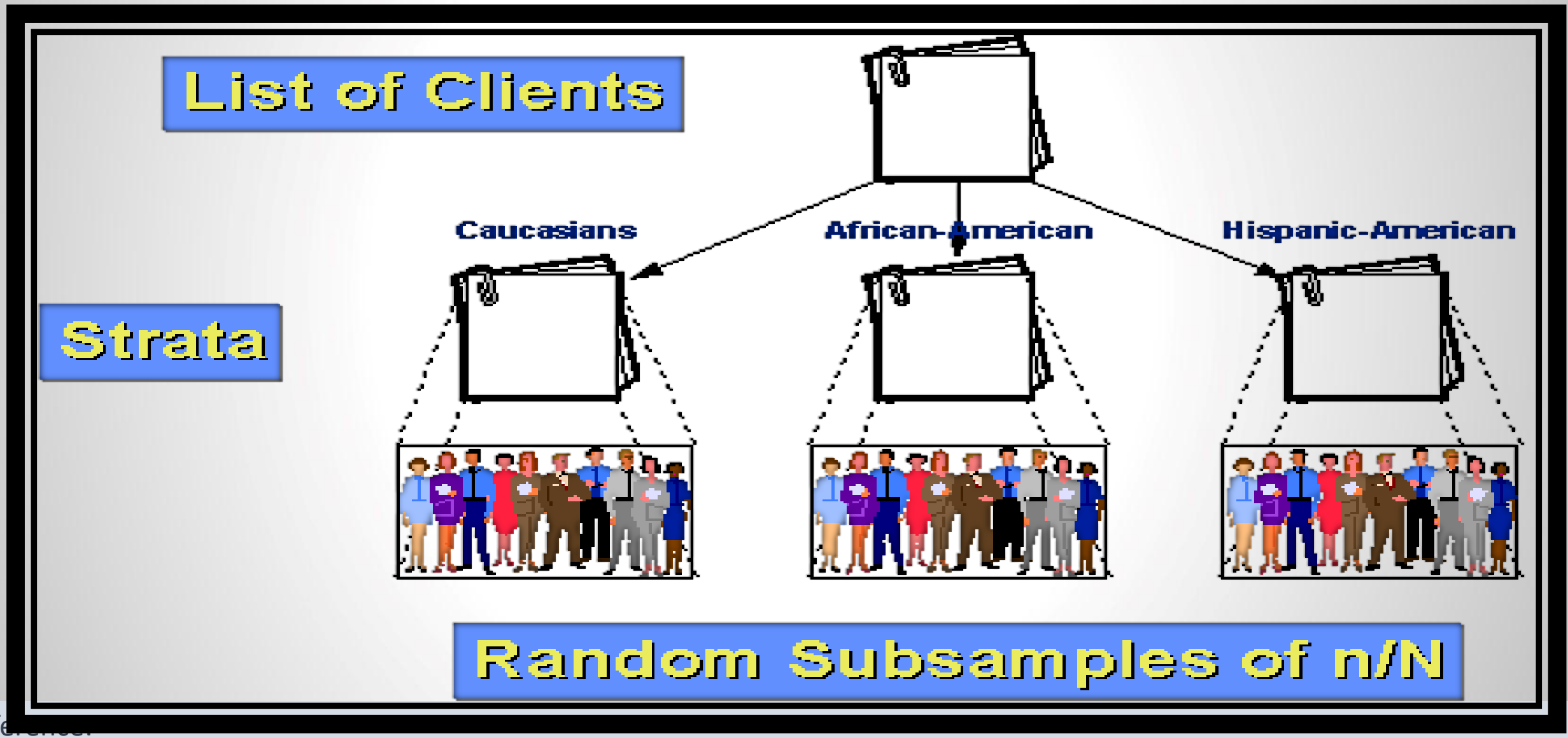
2.Stratified Random Sample

A stratified sample is obtained by separating the population into non-overlapping groups called *strata* and then obtaining a proportional simple random sample from each group.



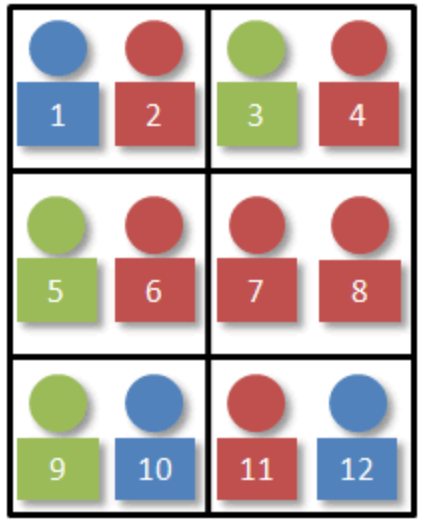
The individuals within each group should be similar in some way. Stratification according to sex, place of residence(urban-rural), regions of country; year of study

Stratified Random sampling
(HIV prevalence in USA)



3. Cluster sampling:

- Sample unit is a group not an individual (family, school class, department ….)
- They are selected randomly from all groups of the same type
- All members of the selected group will be included in the study.



Cluster Population

4. Systematic Random sampling

A list of sampling units (sampling frame)
Select sample units at regular intervals from this list every 3rd or 5th ot 10 th
<<<<The start is randomly >>>>

N = 100

want n = 20

N/n = 5

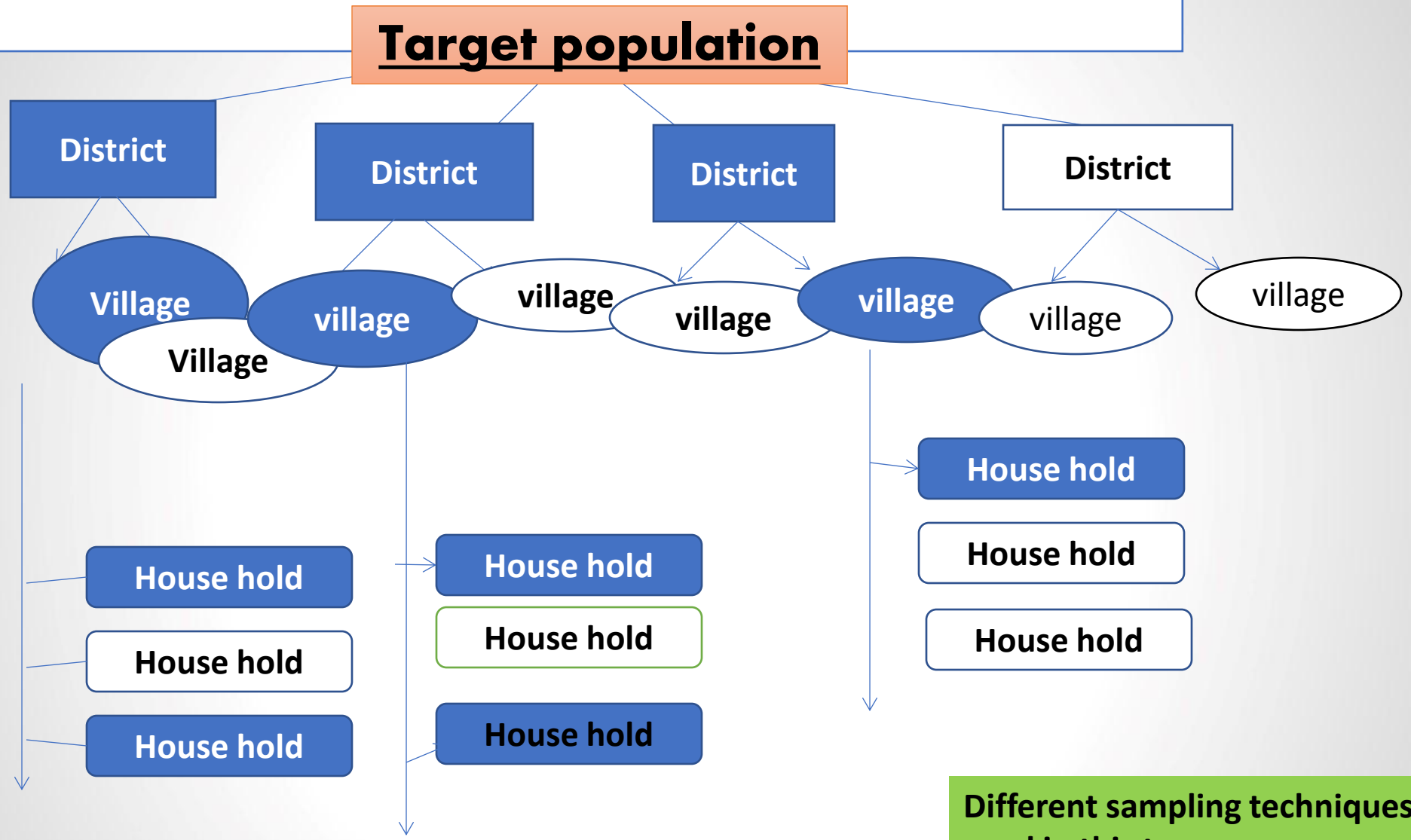
select a random number from 1-5: chose 4

start with #4 and take every 5th unit

http://www.socialresearchmethods.net/kb/sampprob.php

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

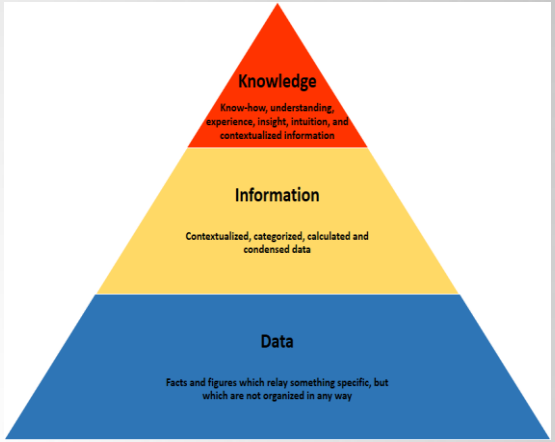
5. Multistage random sampling



Different sampling techniques can be used in this type

Reference:

Data and Information

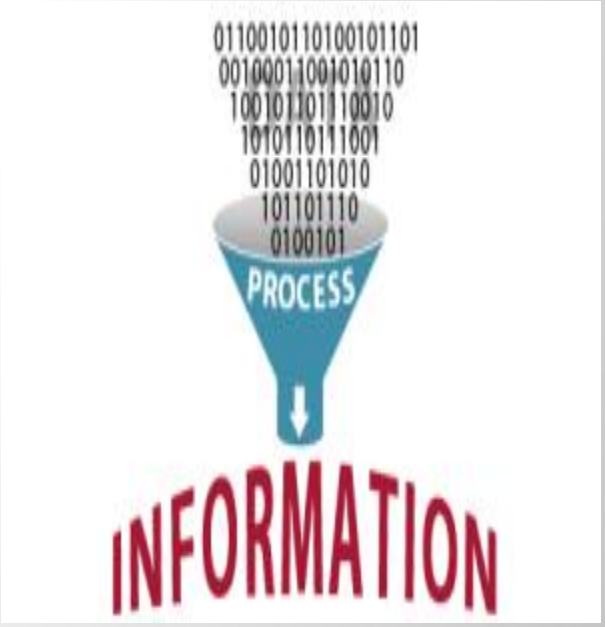


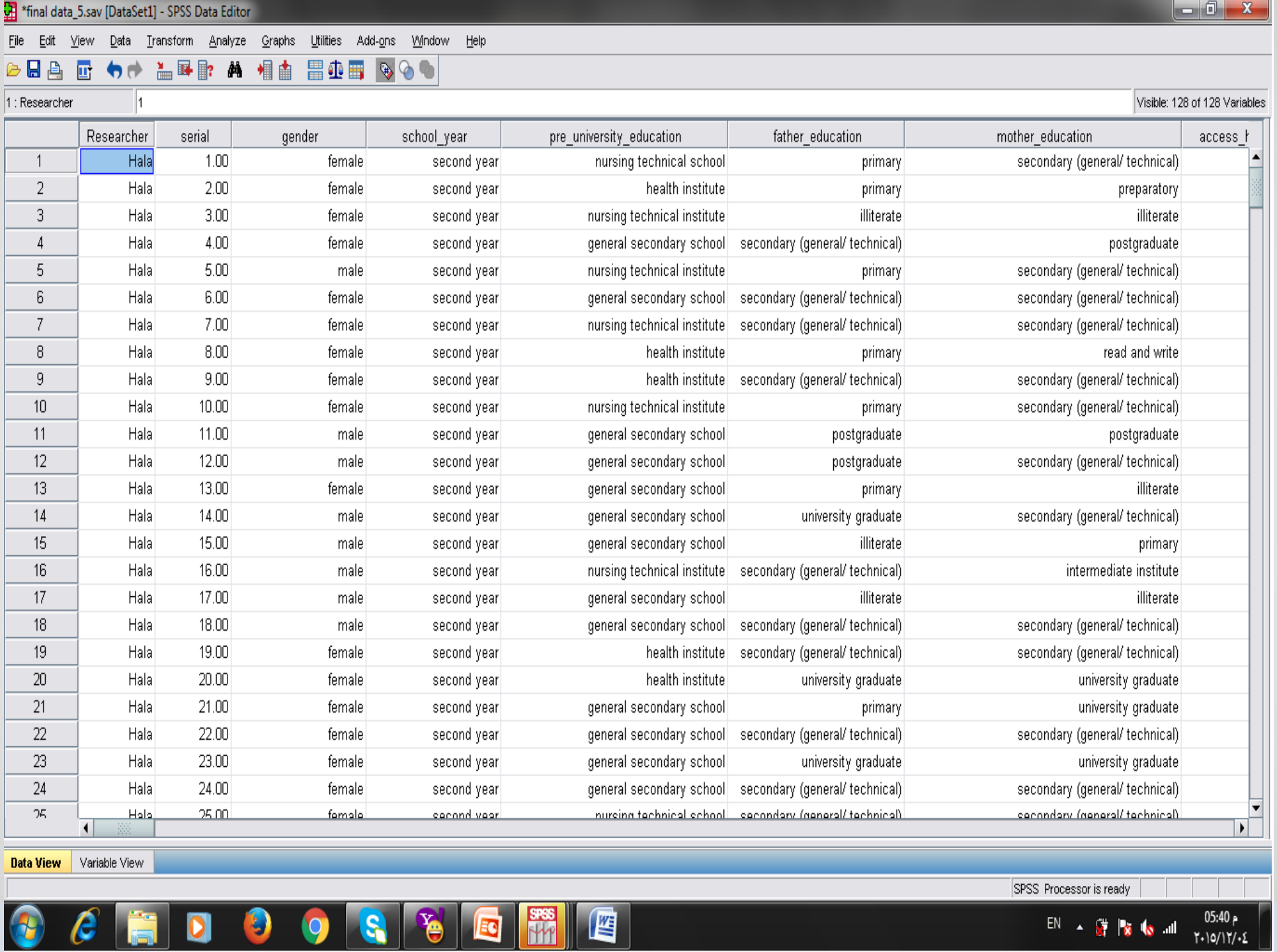
Information
=
Processed Data

Data consist of discrete observations of variables that carry no or little meaning when considered alone.

Data need to be transformed (manually or by computer programs) into information by reducing them and adjusting them for variations in age and sex and others.

Information support decision-makers, policy makers and planners to take proper action in their works.





*final data_5.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Researcher | 1 | Visible: 128 of 128 Variables

	Researcher	serial	gender	school_year	pre_university_education	father_education	mother_education	access_l
1	Hala	1.00	female	second year	nursing technical school	primary	secondary (general/ technical)	
2	Hala	2.00	female	second year	health institute	primary	preparatory	
3	Hala	3.00	female	second year	nursing technical institute	illiterate	illiterate	
4	Hala	4.00	female	second year	general secondary school	secondary (general/ technical)	postgraduate	
5	Hala	5.00	male	second year	nursing technical institute	primary	secondary (general/ technical)	
6	Hala	6.00	female	second year	general secondary school	secondary (general/ technical)	secondary (general/ technical)	
7	Hala	7.00	female	second year	nursing technical institute	secondary (general/ technical)	secondary (general/ technical)	
8	Hala	8.00	female	second year	health institute	primary	read and write	
9	Hala	9.00	female	second year	health institute	secondary (general/ technical)	secondary (general/ technical)	
10	Hala	10.00	female	second year	nursing technical institute	primary	secondary (general/ technical)	
11	Hala	11.00	male	second year	general secondary school	postgraduate	postgraduate	
12	Hala	12.00	male	second year	general secondary school	postgraduate	secondary (general/ technical)	
13	Hala	13.00	female	second year	general secondary school	primary	illiterate	
14	Hala	14.00	male	second year	general secondary school	university graduate	secondary (general/ technical)	
15	Hala	15.00	male	second year	general secondary school	illiterate	primary	
16	Hala	16.00	male	second year	nursing technical institute	secondary (general/ technical)	intermediate institute	
17	Hala	17.00	male	second year	general secondary school	illiterate	illiterate	
18	Hala	18.00	male	second year	general secondary school	secondary (general/ technical)	secondary (general/ technical)	
19	Hala	19.00	female	second year	health institute	secondary (general/ technical)	secondary (general/ technical)	
20	Hala	20.00	female	second year	health institute	university graduate	university graduate	
21	Hala	21.00	female	second year	general secondary school	primary	university graduate	
22	Hala	22.00	female	second year	general secondary school	secondary (general/ technical)	secondary (general/ technical)	
23	Hala	23.00	female	second year	general secondary school	university graduate	university graduate	
24	Hala	24.00	female	second year	general secondary school	secondary (general/ technical)	secondary (general/ technical)	
25	Hala	25.00	female	second year	nursing technical school	secondary (general/ technical)	secondary (general/ technical)	

Data View | Variable View

SPSS Processor is ready

EN 05:40 م ٢٠١٥/١٢/٠٤

Reference:

Sources of Data

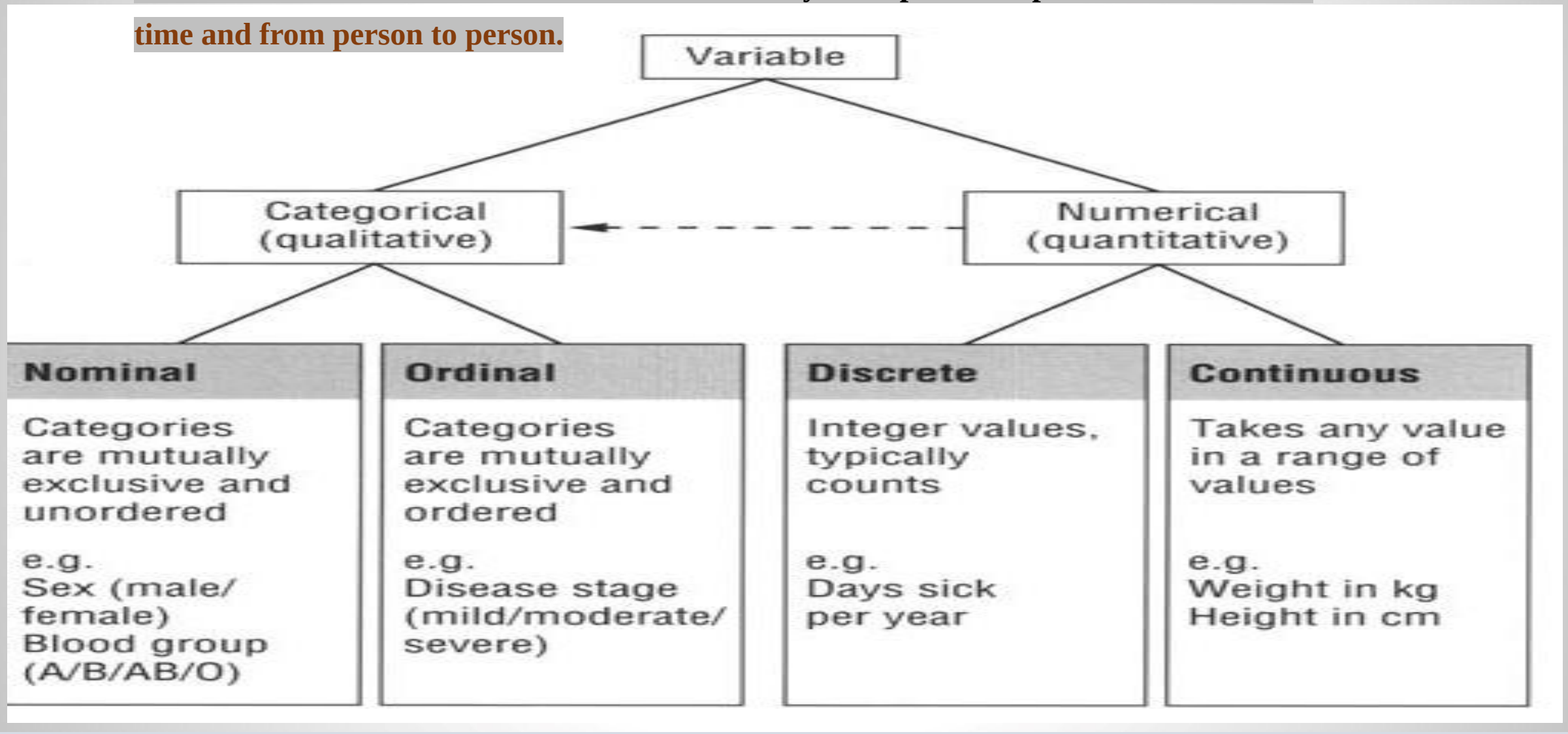
1. Population Census
2- Registration of vital events e.g. Births and deaths, marriage
3-Notification of diseases (Disease Registers)Communicable and non-communicable diseases.
4- Hospital Records
5- Epidemiological surveillance
6- Health Service records
7- Environmental Health data
8- Health Surveys (100 million health)
9- Published articles and reports

Egypt

Demographic Indicators	2011	1995	2005	2015	2025
Population					
Midyear population (in thousands)	82,080	58,945	72,544	88,487	103,742
Growth rate (percent)	2.0	(NA)	2.1	1.8	1.4
Fertility					
Total fertility rate (births per woman)	3.0	(NA)	3.2	2.8	2.5
Crude birth rate (per 1,000 population)	25	(NA)	27	23	19
Births (in thousands)	2,022	(NA)	1,927	2,026	2,006
Mortality					
Life expectancy at birth (years)	73	(NA)	71	74	76
Infant mortality rate (per 1,000 births)	25	(NA)	32	22	15
Under 5 mortality rate (per 1,000 births)	30	(NA)	39	26	18
Crude death rate (per 1,000 population)	5	(NA)	5	5	5
Deaths (in thousands)	396	(NA)	364	422	516
Migration					
Net migration rate (per 1,000 population)	-0	(NA)	-0	-0	-0
Net number of migrants (in thousands)	-17	(NA)	-17	-17	-17

Types of variables

Definition: Is a characteristic or attribute that vary from person to person, from time to time and from person to person.



Reference:

Analyzing &Summarizing qualitative data (by number or frequency and percent)

Table (2) : Barriers to hand hygiene among dentists (nominal variable)

Barriers	(no)Frequency	%
Lack of facilities	20	20
Priority to patients needs	30	30
Fear of dry hands	10	10
Forgetfulness	40	40
Total	100	100

Analysis of quantitative data

Measures of central tendency or averages.

- **Mean**
- **Median**
- **Mode**

II) Measures of dispersion (spread)

Range

Mean deviation

Variance **Standard deviation**

III) Measures of location **Percentile** **Quartile**

1.Mean

Mean (Average): Is obtained as sum of all values divided by the no. of values.

Mean = Σ x/n

The Arithmetic Mean

$$\bar{X} = \frac{\Sigma X}{n}$$

Example, the body weight for five boys may be 34, 41, 37, 32, 36 kg.
Arithmetic mean will be:

$$\bar{X} = \frac{\Sigma X}{n} = \frac{34+41+37+32+36}{5} = \frac{180}{5} = 36 \text{ kg}$$

Mean

Advantages

- Used in quantitative continuous data

Disadvantages

- affected by extreme values
- & it should not be used for non parametric or

Importance

- Best summarizing value for normally distributed data

Reference:

2. The Median
(50th percentile)

The median is the value that lies in the middle of the ordered observations.

A) When sample size is odd number:

1- The observations are ordered according to an ascending or descending magnitude.

2- Determine the rank of the median given by

$$\frac{n+1}{2}$$

3- Using the obtained rank and referring back to the ordered or arranged observations and find the value of median.

b) When sample size is even:

- — In a distribution with even no. of total values: Such a distribution has 2 middlemost values; median is the average of two middlemost values when arranged in an ascending or descending order of values.
- **Median = Mean (average) of (n /2)th and (n/2 + 1)th value in ascending order**
- Practical application

Advantages for median :

1- It can be used with quantitative & qualitative ordinal variables (e.g. median number of patients in cancer stages).

2. It is useful for summarizing data with extreme values as it is not affected by extreme values,

Disadvantages for median :

It **cannot** be used with qualitative nominal variables.

2- It is not easy to be used in statistical analysis.

3.The Mode

The mode is the most frequent observation.

This is done by finding the observation which has the highest frequency. e.g. weight of five children as follows : 9, 8, 12, 7, 8 kg.

It is seen that eight is the observation of highest frequency.

The mode = 8 kg

A similar procedure can be used for finding the mode from qualitative data.

3.The Mode

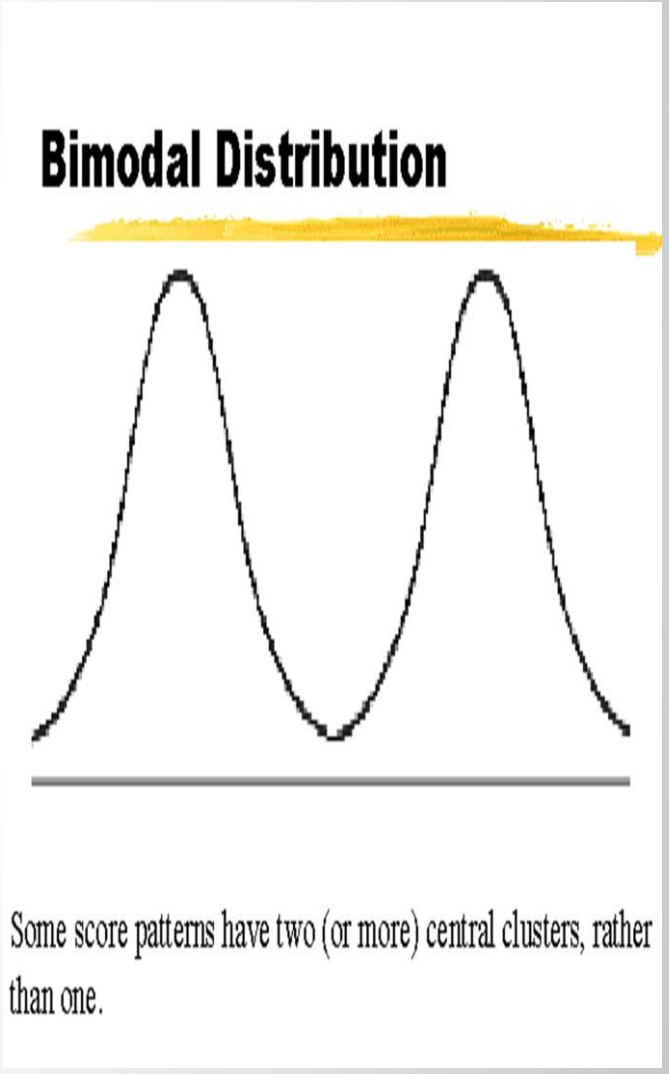
Advantages:

- 1. It can be used in all types of variables
- 2. It is not affected by extremes or out-lying observation

Disadvantages:

1. Sometimes the mode cannot be determined, this happens when all observation have the same frequency (i.e. **uniform distribution**).

2. Sometimes we may obtain two modes (bimodal) or more (multimodal) from the same group of data. e.g. 22, 24, 26, 28, 24, 26 Mode= 24 & 26



II. Measures of Dispersion

1.Range

2.Mean deviation

3. Variance

4.Standard deviation

Measures of Dispersion

Using measures of central tendency is not enough to describe completely a mass of data.

For example if we have five persons with age 30, 34, 32, 36 and 28 years, the mean age is 32 years.

We get the same mean age of 32 years for other five persons have their ages as 12, 30, 8, 62 and 48 years but **the two groups are totally different**.

1 . Range:

It is a simple measure of dispersion and by definition range is difference between the biggest and smallest observation.

From the above two examples range for first group = 36 - 28 = 8 years and for second group = 62-8 = 54 years.

2. Mean Deviation

$$\frac{\sum | x - \bar{x} |}{n}$$

3. Variance (S^2)

$$\frac{\sum (x - \bar{x})^2}{n - 1}$$

Calculations for understanding only

4. Standard Deviation (S, SD, σ):

It is **the commonly used measure of dispersion** and generally the best.

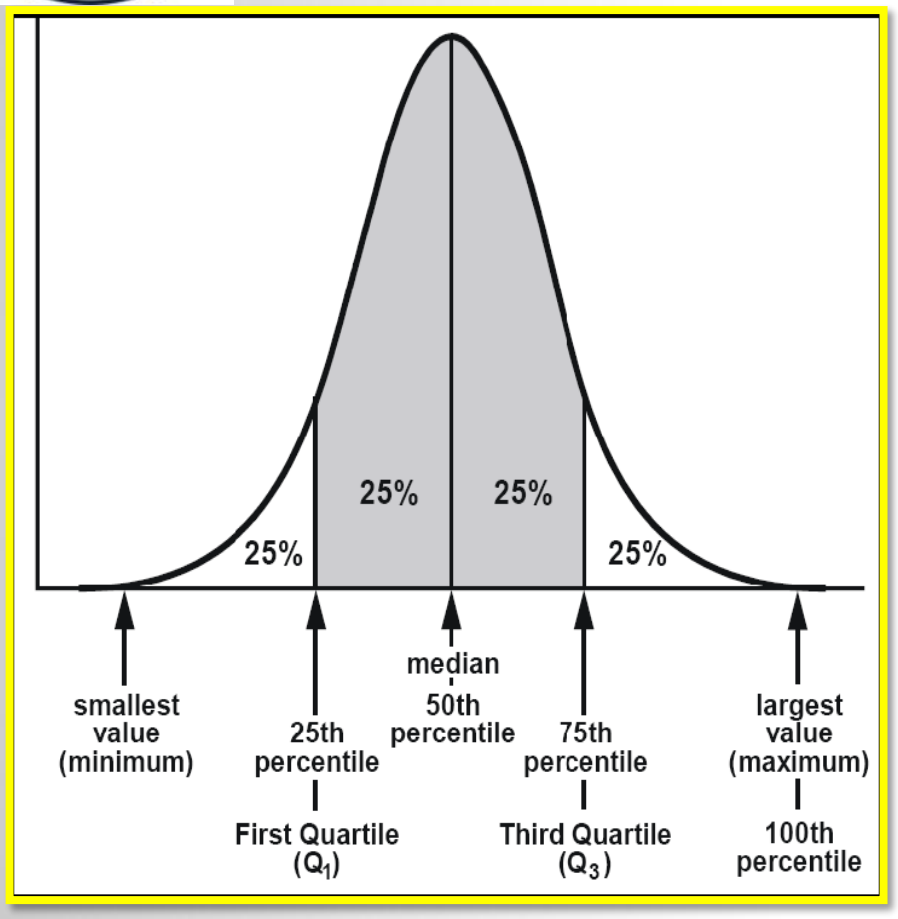
It measures the deviation of observations from the arithmetic me

1. obtaining the deviation of each value from the arithmetic me
2. square the deviation from the mean.
3. The squared deviations are summed and divided by the number of observations minus one (n-1) to get the variance (S^2)
4. The square root of variance (S^2) gives us the standard deviation(S).

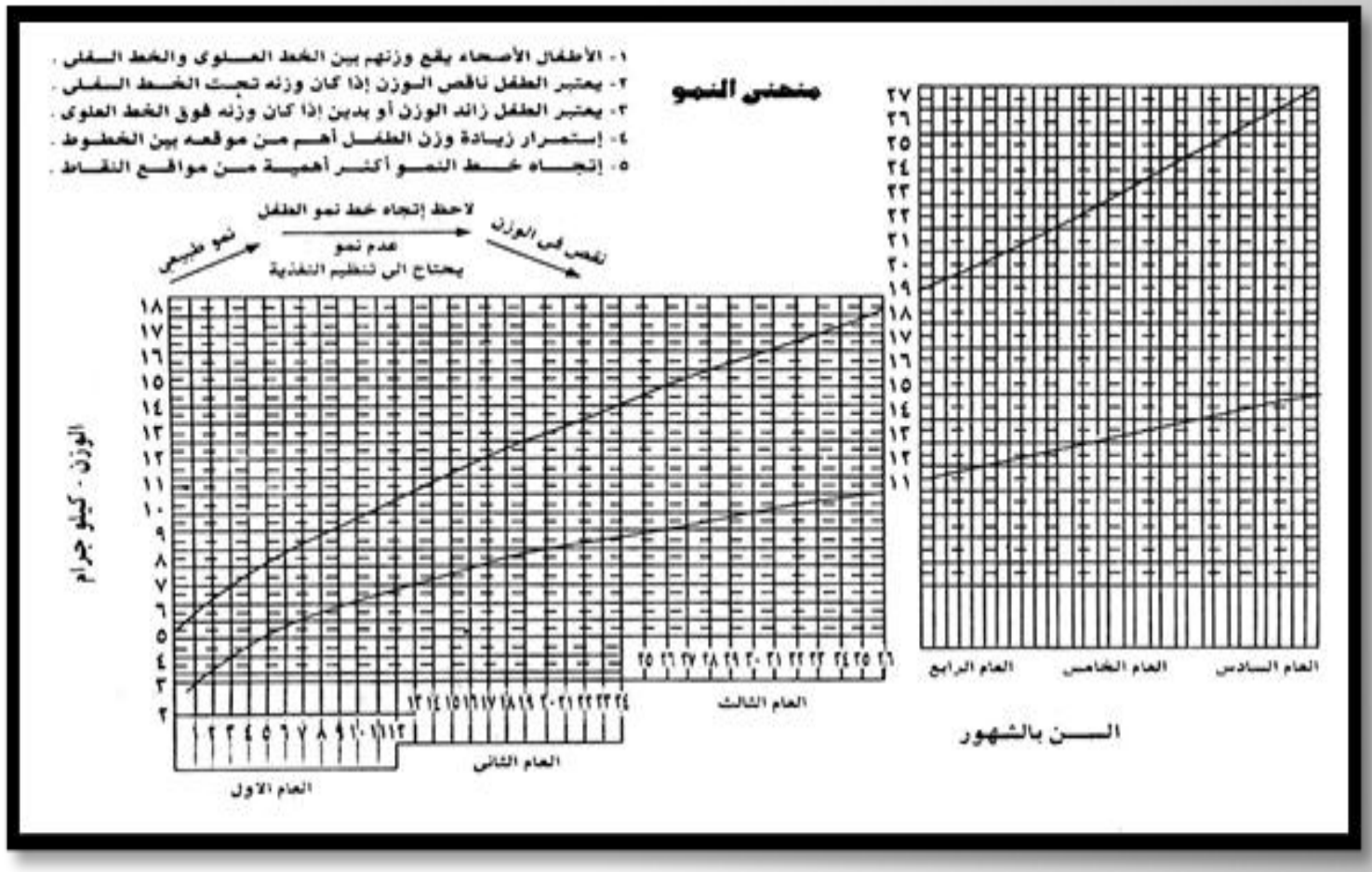
Calculation for understanding only

$$= \sqrt{\frac{\sum \left(x - \bar{x} \right)^2}{n-1}}$$

III) Measures of location



Quartile



Percentile

Questions
“sampling”

1.A computer generates 100 random numbers, and 100 people whose names correspond with the numbers on the list are chosen What is the type of sampling technique?

A.Simple random sample

B.Systematic random sample

C.Cluster sample

D.Multi-stage random sample

E.Stratified random sample

Questions

2. A researcher wishes to investigate covid 19 immunization coverage in Egypt. The best sampling technique is:

A.Simple random sample

B.Systematic random sample

C.Cluster sample

D.Multi-stage random sample

E.Stratified random sample

Answer is

Questions

3. Number of covid 19 vaccine doses for each health care workers in a set of data is considered as :

A.quantitative discrete

B.qualitative variable

C.quantitiative continous

D.nominal variable

E.ordinal variable

Questions

4.Heamoglobin variable in blood coded as (Aneamic & normal) was examined in two groups. This variable can be represented in tables by :

A. Mean and standard deviation

B. Median and range

C. Number and percent

D. Variance

E. Ratio

Questions

5. The best summary measure for body mass index variable in a set of cardiac patients when the variable is normally distributed is :

A. **Median**

B. **Mean**

C. **Mode**

D. **Frequency**

E. **Percentage**

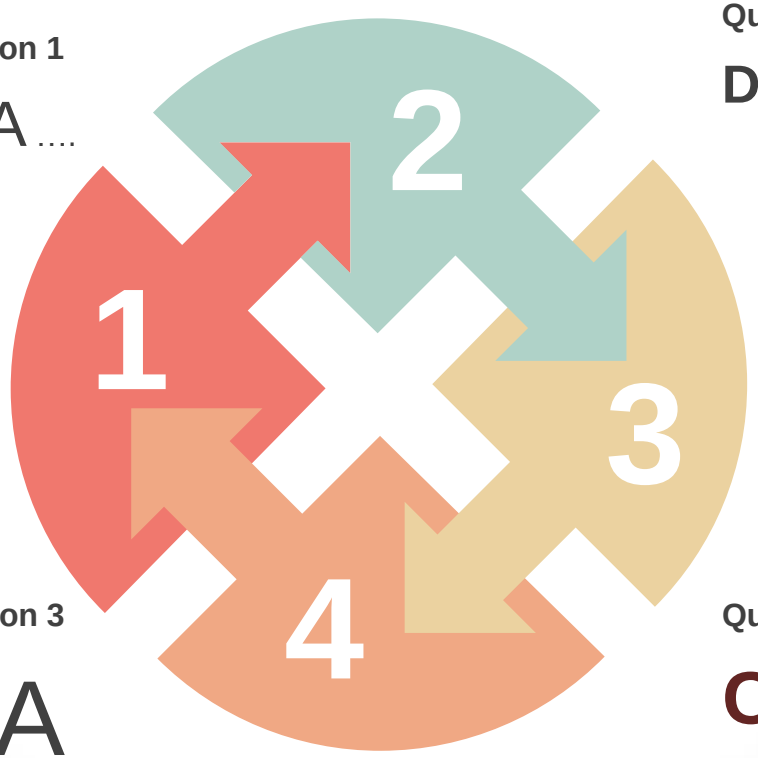
Questions

Question 1

A ....

Question 2

D.....



Question 3

A

Question 4

C

Question 5 ....B

References

1. Vivek Jain . Review of preventive and social medicine (including biostatistics) 7th ed., Facts and concepts from latest Edition of Park 2013.

2.Public Health & Community medicine department book (2018). Faculty of medicine .Mansoura University.

3 .Machin D,Campbell M,Walters S. Medical statistics A Textbook for the Health Sciences. Fourth Edition(2007) John Wiley& sons,Ltd.

ext