Basics of biostatistics

DR BHAVIN J PATEL

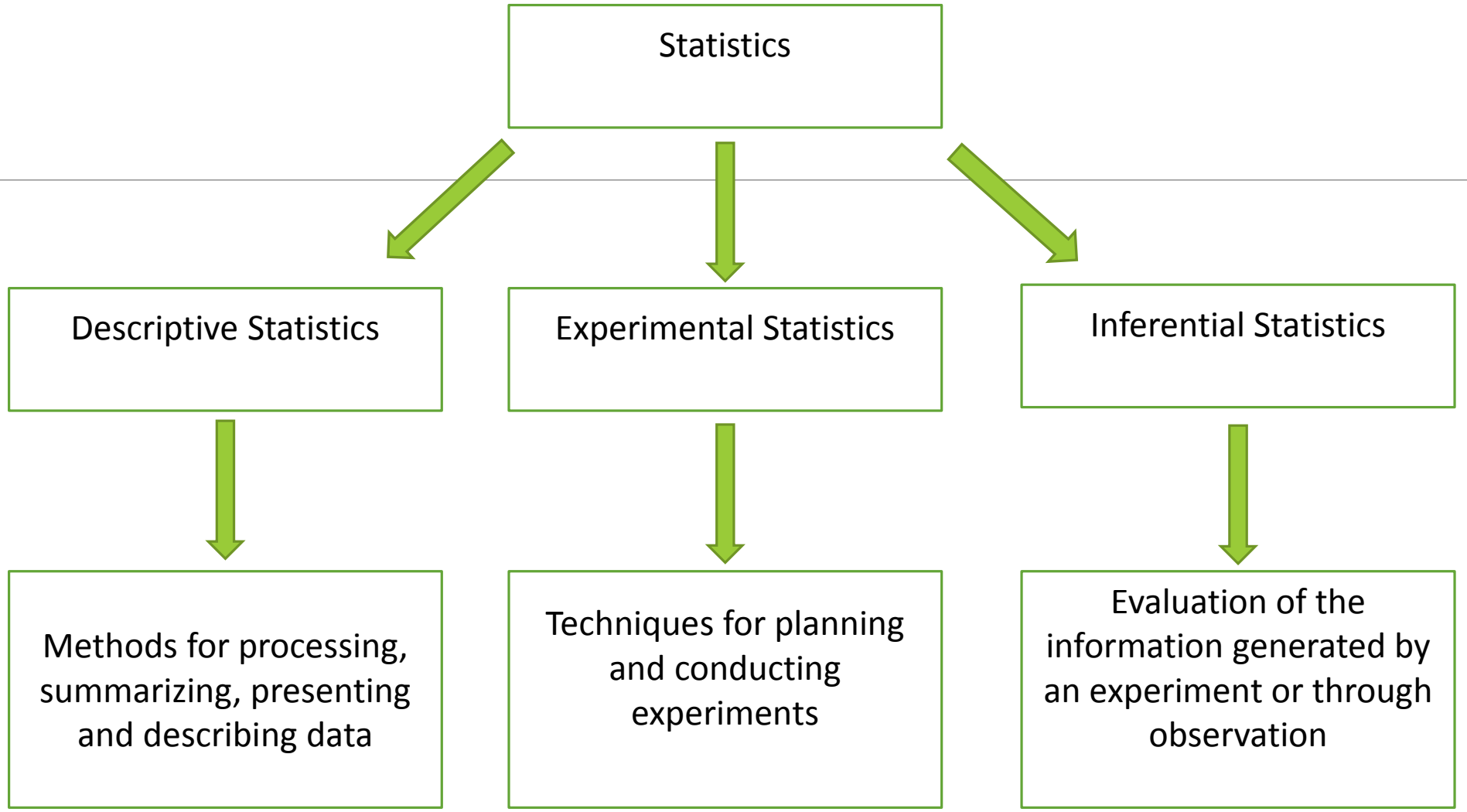
DM NEUROLOGY RESIDENT

MBS HOSPITAL, KOTA.

What is biostatistics?

- It is the science which deals with development and application of the most appropriate methods for the:
  - Collection of data.
  - Presentation of the collected data.
  - Analysis and interpretation of the results.
  - Making decisions on the basis of such analysis

- The methods used in dealing with statistics in the fields of medicine, biology and public health.


Statistics
Descriptive Statistics
Experimental Statistics
Inferential Statistics
Methods for processing, summarizing, presenting and describing data
Techniques for planning and conducting experiments
Evaluation of the information generated by an experiment or through observation

Descriptive statistics

- Summarizing and describing the data
- Uses numerical and graphical summaries to characterize sample data

DATA

- **Definition:-**

A set of values recorded on one or more observational units. Data are raw materials of statistics.

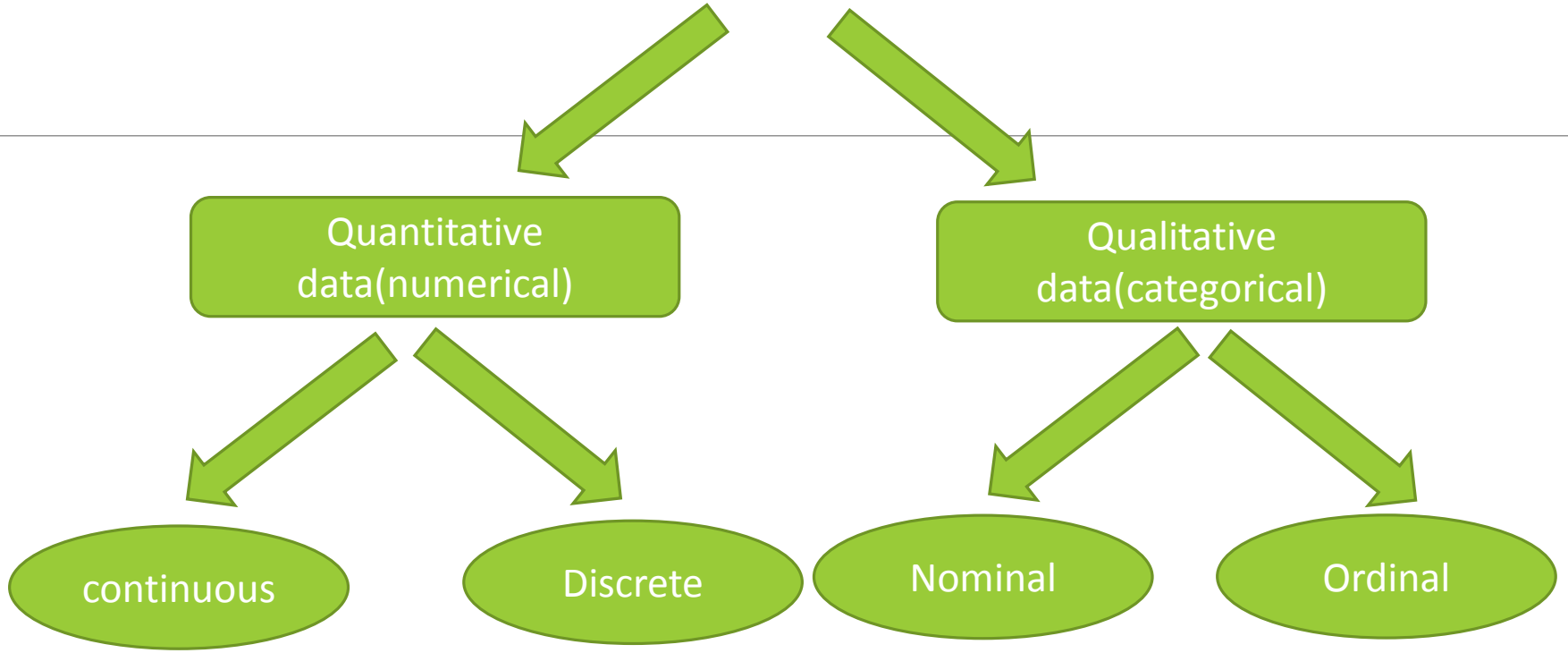
- **Sources of data:-**

Experiments

Surveys

records

Types of data



Quantitative data

- **Discrete**: Reflects a number obtained by counting—no decimal.
- **Continuous**: Reflects a measurement; the number of decimal places depends on the precision of the measuring device.
  - **Ratio scale**: Order and distance implied. Differences *can* be compared; has a true zero. Ratios *can* be compared.
    Examples: Height, weight, blood pressure
  - **Interval scale**: Order and distance implied. Differences *can* be compared; no true zero. Ratios *cannot* be compared.
    Example: Temperature in Celsius.

Qualitative data

- **Ordinal data**
  - **Difference** and **order** are implied BUT, **intervals are no longer equivalent**.
  - Example :- ranking syatem
- **Nominal data**
  - Only **difference** is implied.
  - Observations are classified into **mutually exclusive categories**.
  - Examples: Gender, ID numbers, pass/fail response

Methods of presentation of data

1 Tabular presentation

2 Graphical presentation

- **Purpose:** To display data so that they can be readily understood.

- **Principle:** Tables and graphs should contain enough information to be self-sufficient without reliance on material within the text of the document of which they are a part.

- Tables and graphs share some common features, but for any specific situation, one is likely to be more suitable than the other.

Tabular Presentation

- **Types of tables:-**

1. list table:- for qualitative data, **count the number of observations**

(frequencies) in each category.

A table consisting of **two columns**, the first giving an identification of the *observational unit* and the second giving the ***value of variable for that unit.***

Example : number of patients in each hospital department are

Department	Number of patients
Medicine	100
Surgery	88
ENT	54
Opthalmology	30

Tabular Presentation

2. Frequency distribution table:- for qualitative and quantitative data

- Simple frequency distribution table:-

Distribution of the studied individuals according to blood group

Blood group	Frequency	%
A	6	30
B	6	30
AB	5	25
O	3	15
total	20	100

Tabular Presentation

- **complex frequency distribution table**

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

Graphical presentation

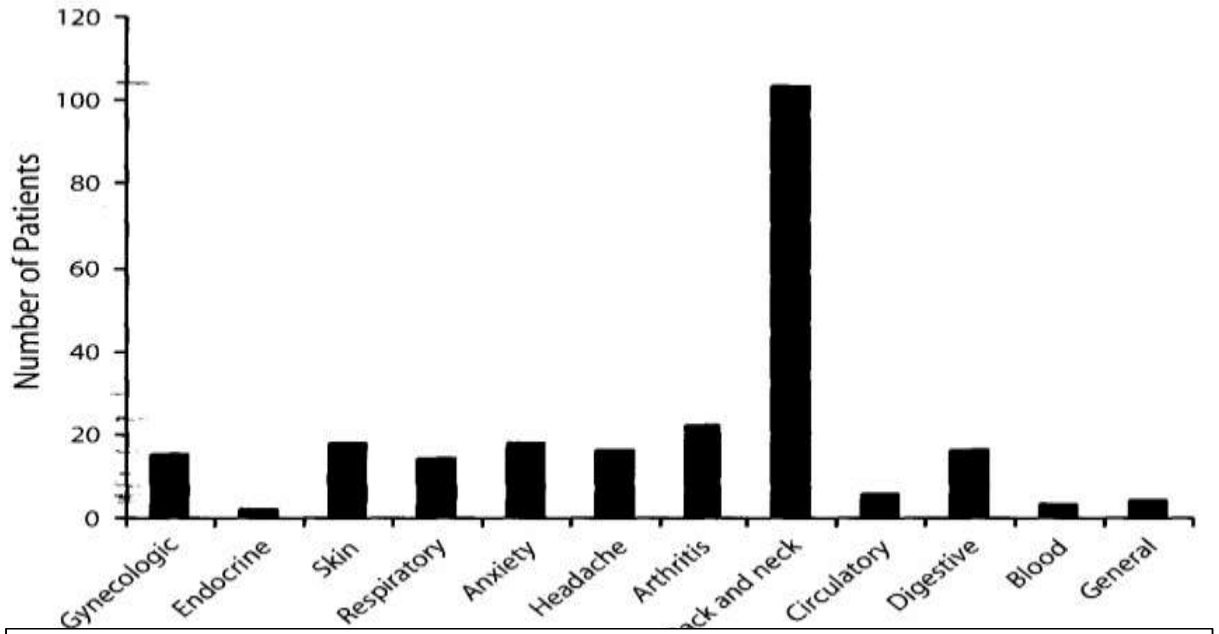
- For quantitative, continuous or measured data
  - Histogram
  - Frequency polygon
  - Frequency curve
  - Line chart
  - Scattered or dot diagram

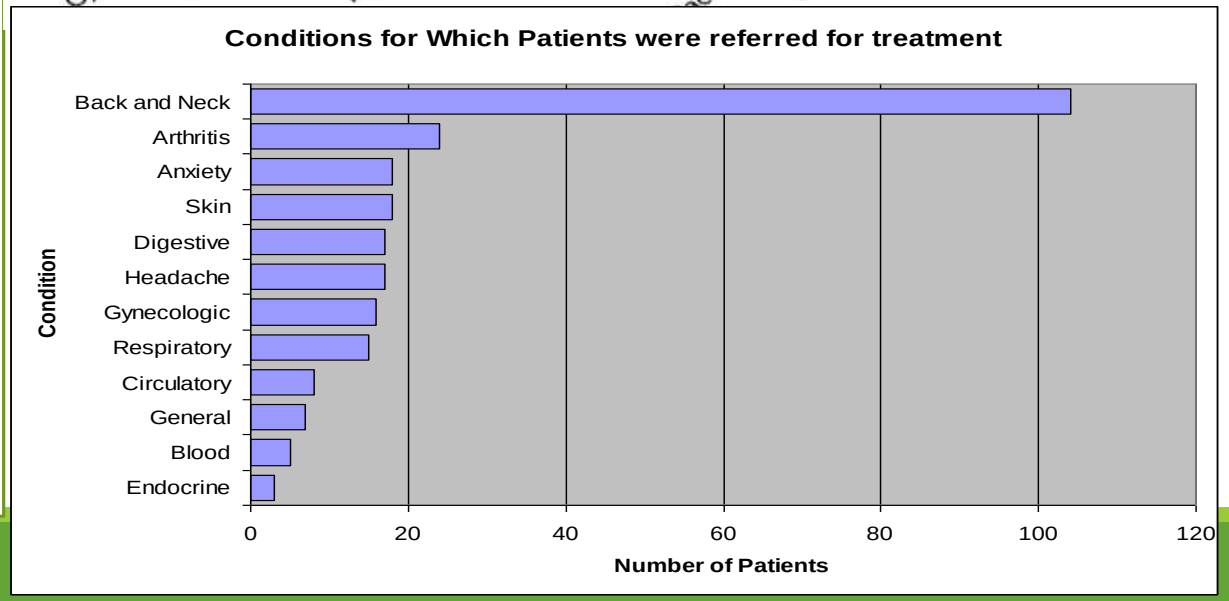
- For qualitative, discrete or counted data
  - Bar diagram
  - Pie or sector diagram
  - Spot map

Bar diagram

- It represent the measured value (or %) by separated **rectangles** of constant width and its lengths proportional to the frequency

- Use:- discrete qualitative data
- Types:- simple
  multiple
  component

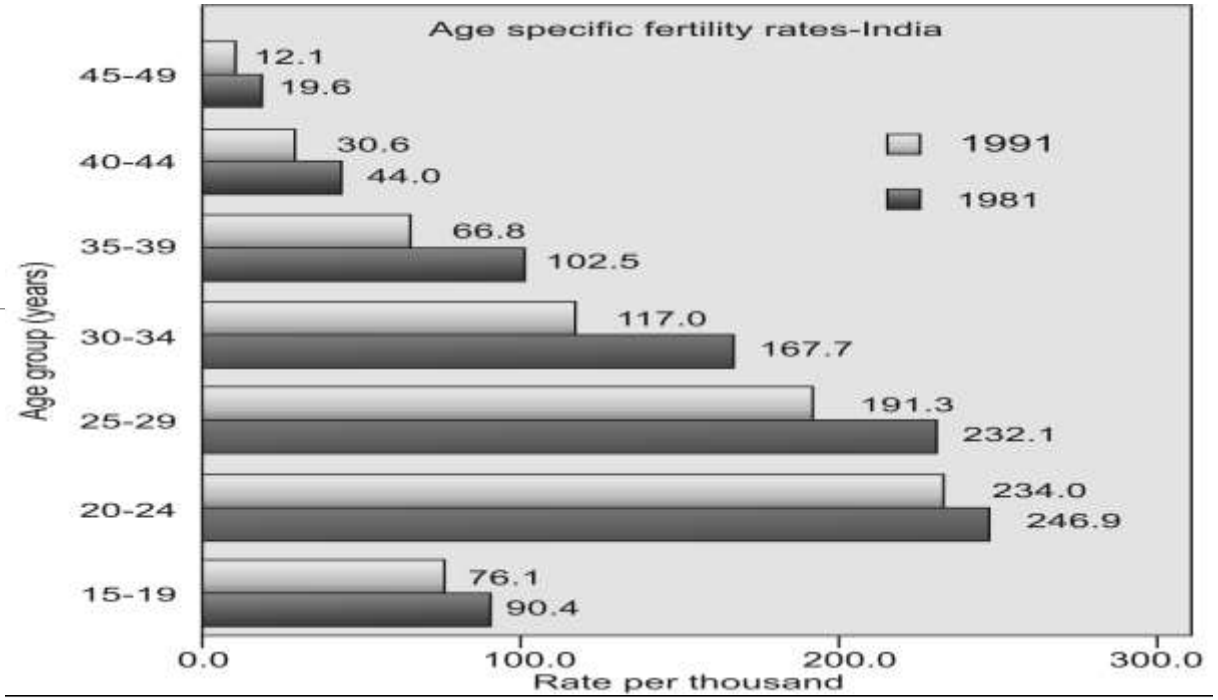




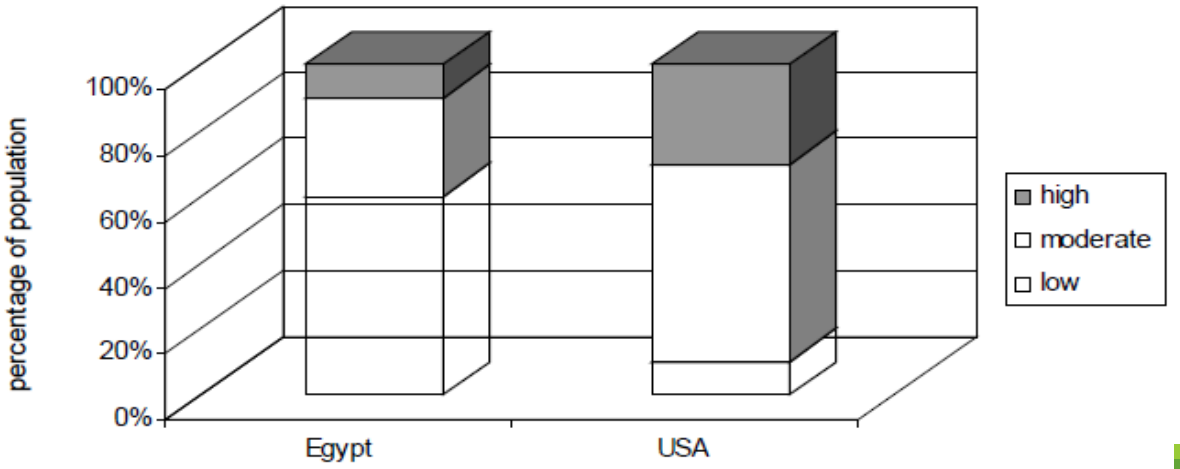
Bar diagram

- **Multiple bar chart:-** Each observation has more than one value represented, by a group of bars.

Component bar chart:-subdivision of a single bar to indicate the composition of the total divided into sections according to their relative proportion.

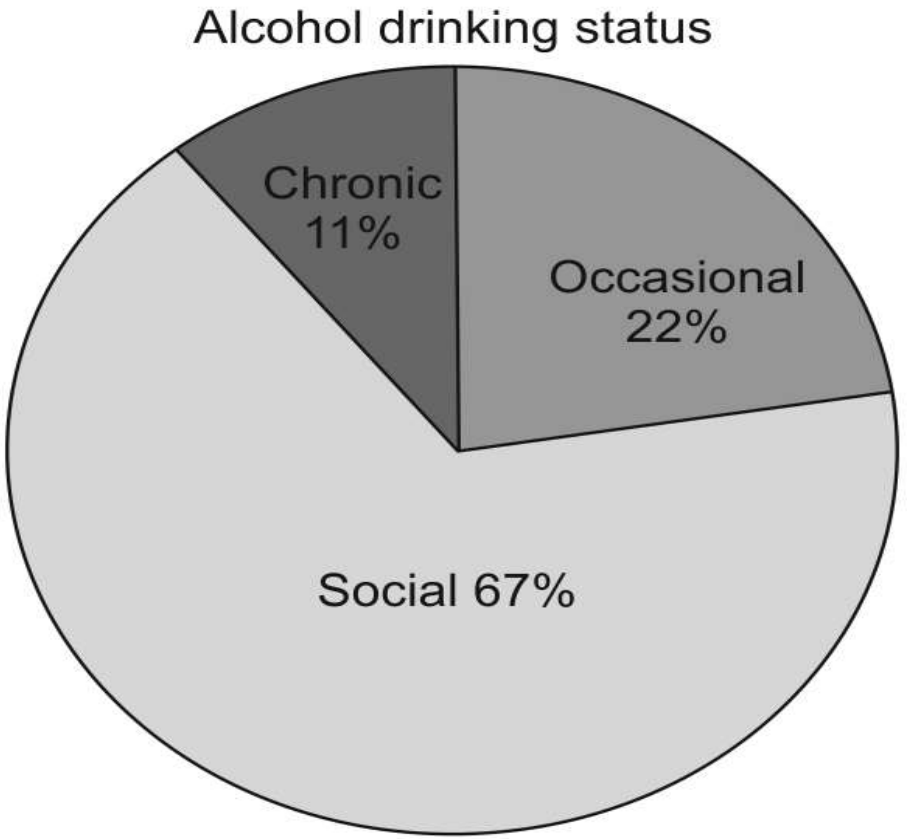


Comparison between Egypt and USA in socio-economic standard of living



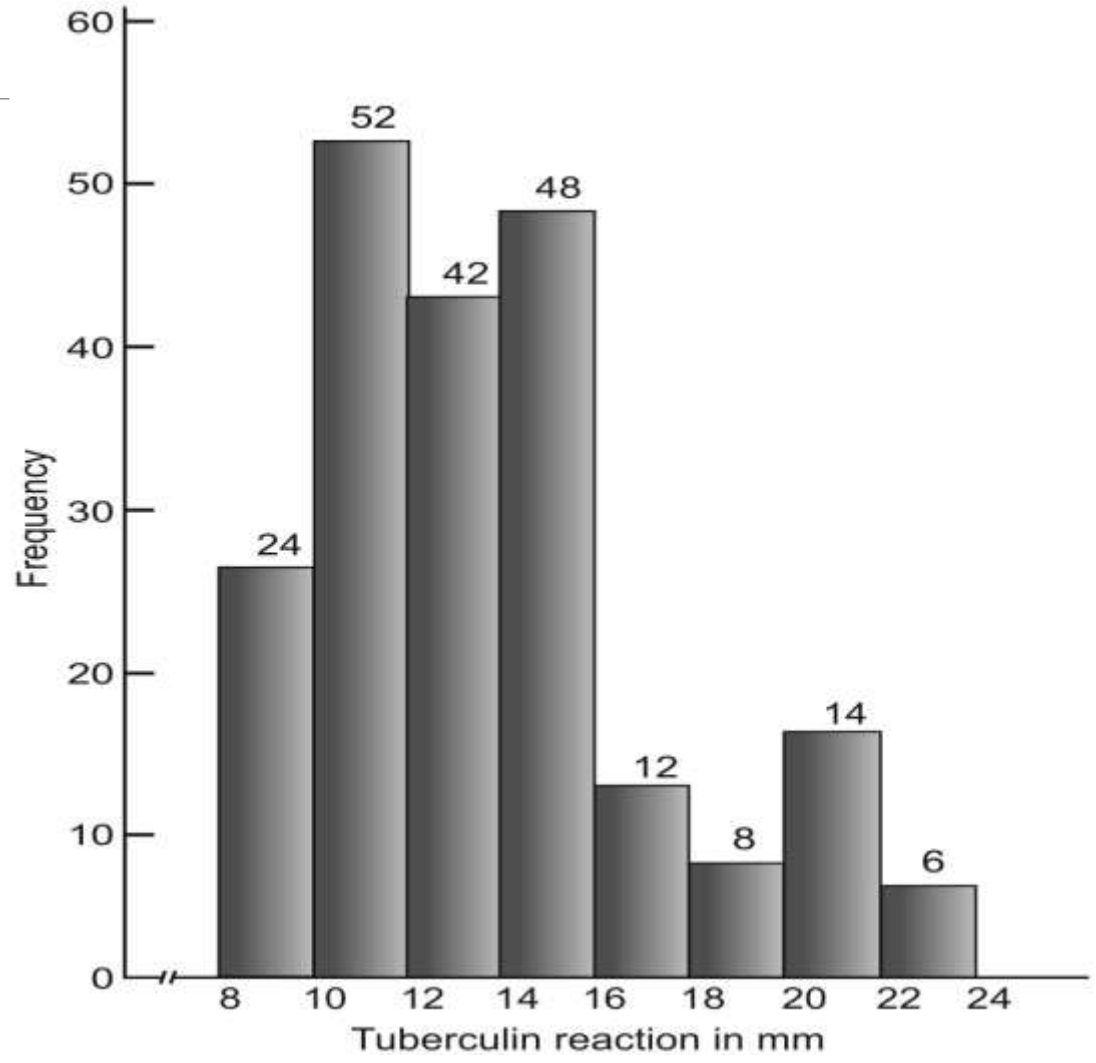
Pie diagram

- Consist of a circle whose area represents the total frequency (100%) which is divided **into segments**.
- Each segment represents a proportional composition of the total frequency



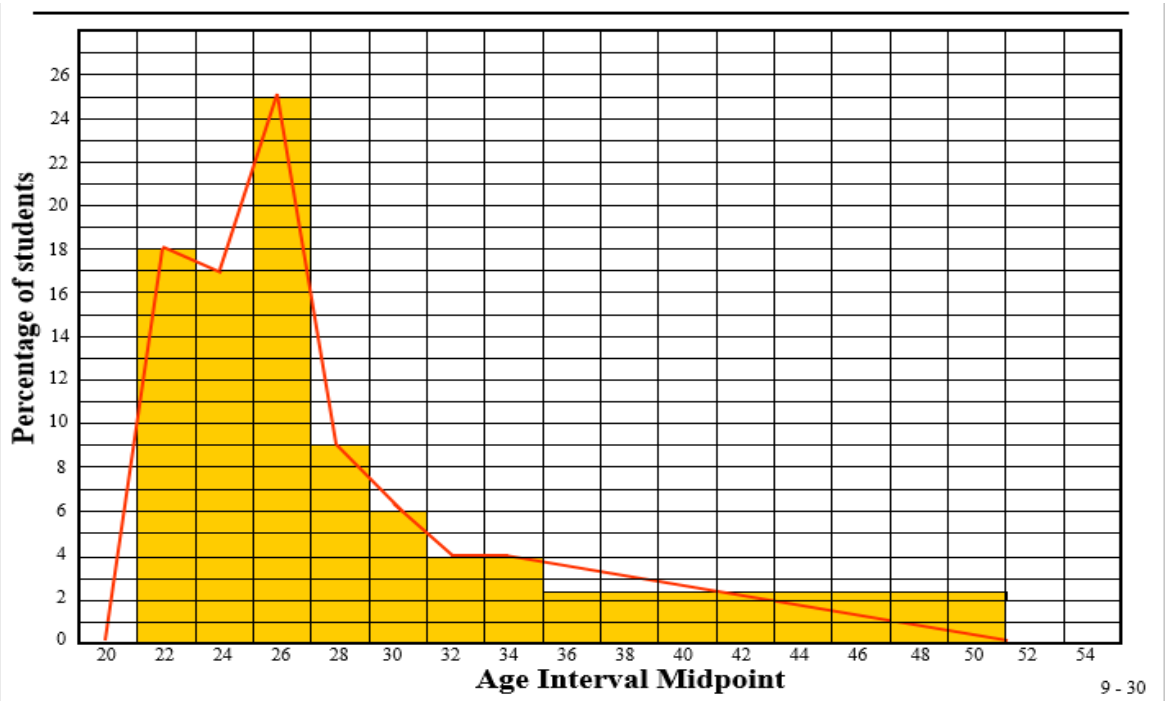
Histogram

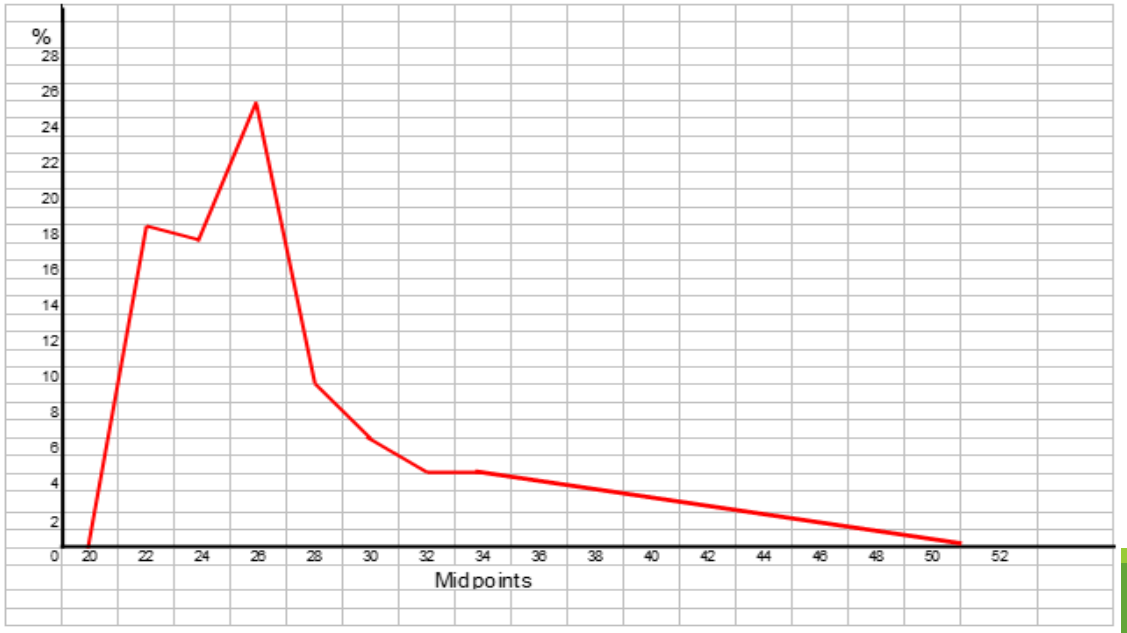
- it is very similar to the bar chart with the difference that the rectangles or bars are **adherent (without gaps).**
- It is used for presenting continuous quantitative data.
- Each bar represents a class and its height represents the frequency (number of cases), its width represent the class interval.



Frequency polygon

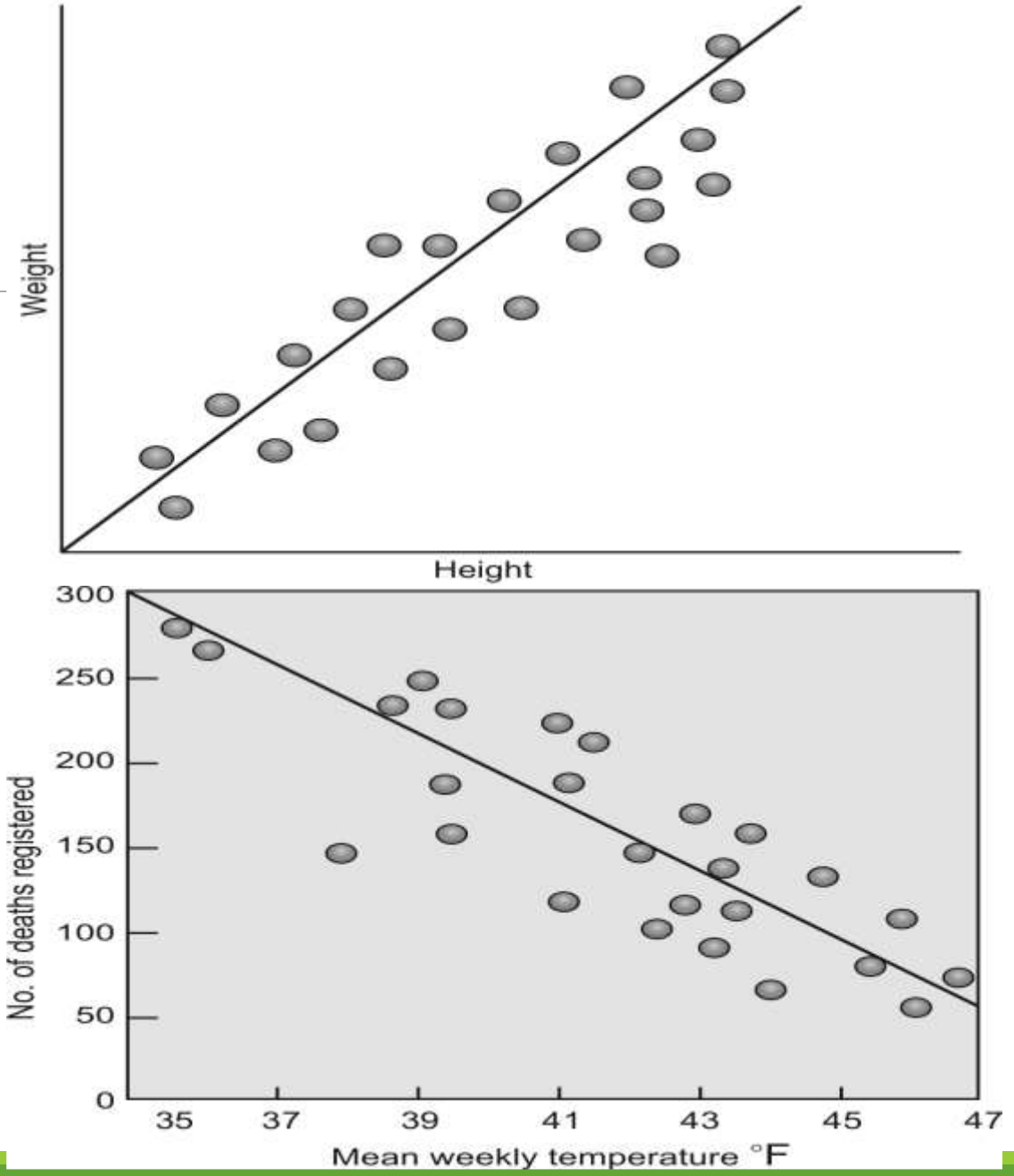
- Derived from a histogram by connecting the **mid points** of the tops of the rectangles in the histogram.
- The line connecting the centers of histogram rectangles is called frequency polygon.
- We can draw polygon without rectangles so we will get simpler form of line graph





Scattered diagram

- It is useful to represent the relationship between two numeric measurements.
- Each observation being represented by a point corresponding to its value on each axis



Measures of central tendacy

- Variable usually has a point (center) around which the observed values lie.
- The three most commonly used averages are:
  - **The arithmetic mean**
  - **The Median**
  - **The Mode**

Measures of central tendacy

1. Mean:-

- The arithmetic average of the variable *x*.
- It is the preferred measure for **interval or ratio variables** with relatively symmetric observations.
- It has **good sampling stability** (e.g., it varies the least from sample to sample), implying that it is better suited for making inferences about population parameters.
- It is **affected by extreme values**

Measures of central tendacy

2. Median:-

- The middle value (Q_2, the 50^{th} percentile) of the variable.
- It is appropriate for **ordinal measures** and for **skewed interval or ratio measures.**
- It has low sampling stability.
- The rank of median for is **(n + 1)/2** if the number of observation is **odd** and **n/2** if the number is **even**

Measures of central tendacy

3 Mode:-

- The most frequently occurring value in the data set.
- May not exist or may not be uniquely defined.
- It is the only measure of central tendency that can be used with **nominal variables**, but it is also meaningful for quantitative variables that are inherently **discrete.**
- Its **sampling stability** is very low

Measure of dispersion

- Measures of variability depict how similar observations of a variable tend to be.
- Variability of a **nominal** or **ordinal** variable is rarely summarized numerically.
- The measure of dispersion describes the degree of variations or dispersion of the data around its central values: (dispersion = variation = spread = scatter).

Range - **R**

Standard Deviation - **SD**

Coefficient of Variation **-COV**

Measure of dispersion

- **Range:-**
  - It is the difference between the **largest** and **smallest** values.
  - It is the simplest measure of variation.
  - **Disadvantage:-** it is based only on two of the observations and gives no idea of how the other observations are arranged between these two.

Measure of dispersion

- **Standard deviation:-**
  - Represents the average spread of the data around the mean.
  - “Average deviation” from the mean.
  - **SD = √ Σ (mean – x)2 / n – 1**
- **Uses:-**

1.It summarizes the deviations of a large distribution from mean in one figure used as a unit of variation.

2. Indicates whether the variation of difference of an individual from the mean is by chance, i.e. natural or real due to some special reasons.

3.It also helps in finding the suitable size of sample for valid conclusions.

Measure of dispersion

- **Coefficient of variation:-**

- The coefficient of variation expresses the standard deviation as a percentage of the sample mean.

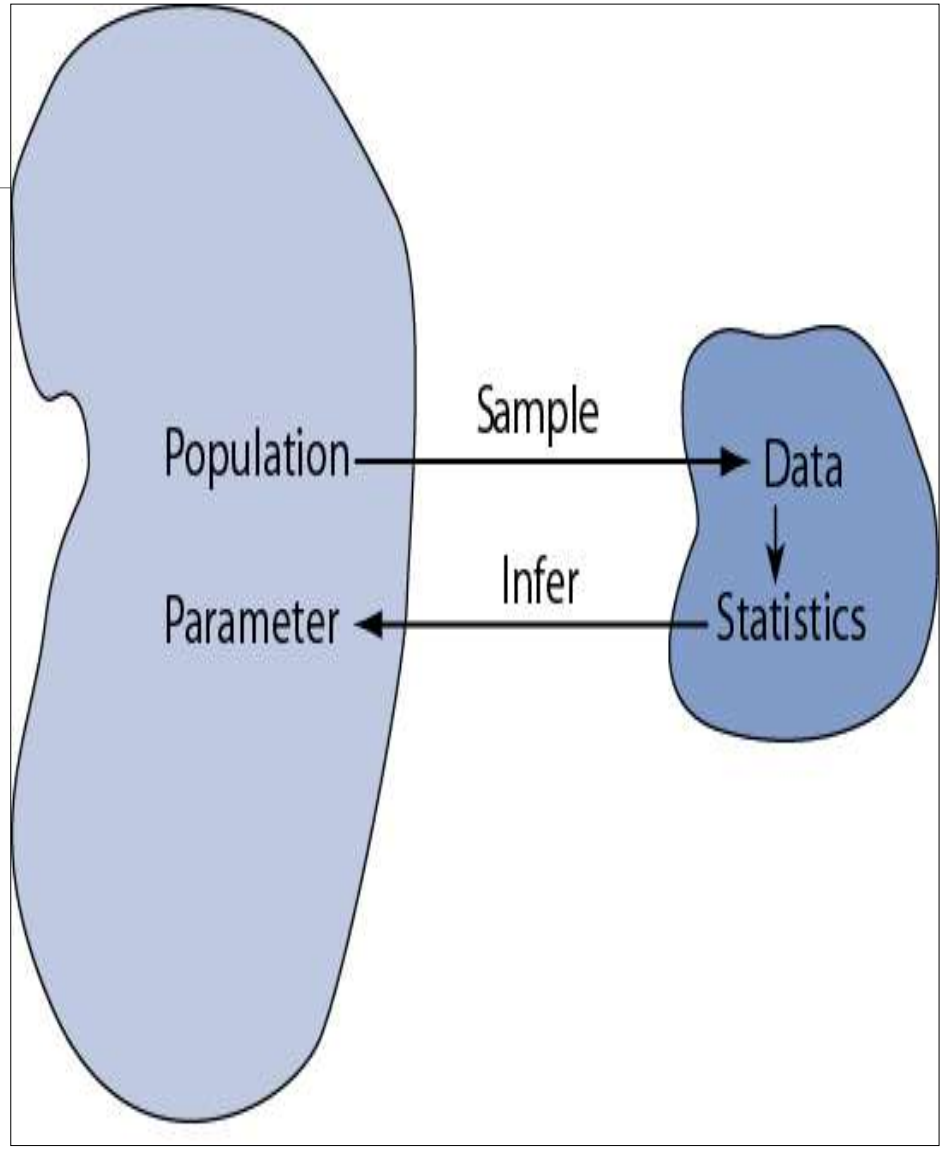
- **C. V = SD / mean * 100**

- C.V is useful when, we are interested in the relative size of the variability in the data.

Inferential statistics

- It refers to the process of selecting and using a sample to draw inference about population from which sample is drawn.

- Two forms of statistical inference
  - **Hypothesis testing**
  - **Estimation**



Hypothesis Testing

- During investigation there is assumption and presumption which subsequently in study must be proved or disproved.
- Hypothesis is a supposition made from observation. On the basis of Hypothesis we collect data.
- Two Hypothesis are made to draw inference from Sample value-

A. Null Hypothesis or hypothesis of no difference.

B. Alternative Hypothesis of significant difference.

The Null Hypothesis and the Alternative Hypothesis are chosen before the sample is drawn.

Hypothesis Testing

- Null Hypothesis states that the observed difference is entirely due to sampling error, that is - it has occurred purely by chance.

  E.g:- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.

- Alternative Hypothesis of significant difference states that the sample result is different that is, greater or smaller than the hypothetical value of population.

- A test of significance such as Z-test, t-test, chisquare test, is performed to accept the Null Hypothesis or to reject it and accept the Alternative Hypothesis.

Error in Hypothesis Testing

	Decision	
	Accept Ho	Reject Ho
Ho true	Correct decision	Type 1 error
Ho false	Type 2 error	Correct decision

P value

- The probability of committing Type 1 Error is called the P-value. Thus p-value is the chance that the presence of difference is concluded when actually there is none.
- When the p value is between 0.05 and 0.01 the result is usually called significant.
- When p value is less than 0.01, result is often called highly significant.
- When p value is less than 0.001 and 0.005, result is taken as very highly significant.

- **Confidence Interval :-**
  - The interval within which a parameter value is expected to lie with a certain confidence level as could be revealed by repeated samples is called confidence interval.
  - A point estimate from a sample population may not reflect the true value from the population so it is often helpful to provide a range.
- **Confidence Level** : The degree of assurance for an interval to contain the value of a parameter

  $(1-\alpha)$.

Test of significance

- Test of significance is a formal procedure for comparing observed data with a claim (also called a hypothesis) whose truth we want to assess.

Parametric	Non parametric
➢ Based on specific distribution such as Gaussian	➢ Not based on any particular parameter such as mean ➢ Used when the underlying distribution is far from Gaussian (applicable to almost all levels of distribution) and when the sample size is small

Test of significance

Parametric test	Nonparametric
▪ Student's t- test(one sample, two sample, and paired) ▪ Z test ▪ ANOVA F-test ▪ Pearson's correlation(r)	▪ Sign test(for paired data) ▪ Wilcoxon Signed-Rank test ▪ Wilcoxon Rank Sum test ▪ Chi-square test ▪ Spearman's Rank Correlation(p) ▪ ANOCOVA ▪ Kruskal-Wallis test

Test of significance

Purpose of application	Parametric test	Non-Parametric test
Comparison of two independent groups.	't'-test for independent samples	Wilcoxon rank sum test
Test the difference between paired observation	't'-test for paired observation	Wilcoxon signed-rank test
Comparison of several groups	ANOVA	Kruskal-Wallis test
Quantify linear relationship between two variables	Pearson's Correlation	Spearman's Rank Correlation
Test the association between two qualitative variables	_	Chi-square test

Sampling

- **Sample size:-**This is the sub-population to be studied in order to make an inference to a reference population(A broader population to which the findings from a study are to be generalized)
- Optimum sample size determination is required for the following reasons:

1.To allow for appropriate analysis

2.To provide the desired level of accuracy

3.To allow validity of significance test

Sampling

Disadvantages of inappropriate sample size	
Small sample	**Large sample**
1 Even a well conducted study may fail to answer it research question 2. It may fail to detect important effect or associations 3. It may associate this effect or association imprecisely	1The study will be difficult and costly 2. Time constraint 3. Loss of accuracy.

Sampling

- **Random error**: error that occur by chance. Sources are sample variability, subject to subject differences & measurement errors. It can be reduce by averaging, increase sample size, repeating the experiment.

- **Systematic error**: deviations not due to chance alone. Several factors, e.g patient selection criteria may contribute. It can be reduce by good study design and conduct of the experiment.

Sampling

- **Precision**: the degree to which a variable has the same value when measured several times. It is a function of random error.
- **Accuracy**: the degree to which a variable actually represent the true value. It is function of systematic error.
- Approaches for estimating sample size and performing power analysis depend primarily on:

1. The study design
2. The main outcome measure of the study

Sampling

- There are four procedures that could be used for calculating sample size:

1. Use of formulae
2. Ready made table
3. Nomograms
4. Computer software

Sampling

- **USE OF FORMULAE FOR SAMPLE SIZE:-**

The appropriate sample size for population-based study is determined largely by 3 factors

1. The estimated prevalence of the variable of interest.

2. The desired level of confidence.

3. The acceptable margin of error

- For population >10,000.

$n=Z^2pq/d^2$

USE OF READYMADE TABLE FOR SAMPLE SIZE

Such table that give ready made sample sizes are available for different designs & situation

Table 3: Estimating an incidence rate with specified relative precision [Formula: $n = (Z_{1-\alpha/2} / \varepsilon)^2$]

Relative precision (ε)	Confidence level		
	99%	95%	90%
0.01	66358	38417	27061
0.02	16590	9605	6766
0.03	7374	4269	3007
0.04	4148	2402	1692
0.05	2655	1537	1083
0.06	1844	1068	752
0.07	1355	785	553
0.08	1037	601	423
0.09	820	475	335
0.10	664	385	271
0.12	461	267	188
0.14	339	197	139
0.16	260	151	106
0.18	205	119	84
0.20	166	97	68
0.22	138	80	56
0.24	116	67	47
0.26	99	57	41
0.28	85	50	35
0.30	74	43	31
0.32	65	38	27
0.34	58	34	24
0.36	52	30	21
0.38	46	27	19
0.40	42	25	17
0.42	38	22	16

Sampling

- **USE OF COMPUTER SOFTWARE FOR SAMPLE SIZE:-**
  - Epi-info
  - nQuerry
  - Power & precision
  - Sample
  - STATA
  - SPSS

STILL CONFUSED……

Smart people don't do it alone……

Call a statistician

- Sample selection
- Sample size determination
- Analysis of data

TERMS USED TO DESCRIBE THE MAGNITUDE OF AN EFFECT

- **Relative risk** — The relative risk (or risk ratio) equals the incidence in exposed individuals divided by the incidence in unexposed individuals. The relative risk can be calculated from studies in which the proportion of patients exposed and unexposed to a risk is known, such as a cohort study

- **Odds ratio** — The odds ratio equals the odds that an individual with a specific condition has been exposed to a risk factor divided by the odds that a control has been exposed. The odds ratio is used in case-control studies

- **Absolute risk** — The relative risk and odds ratio provide an understanding of the magnitude of risk compared with a standard. However, it is more often desirable to know information about the absolute risk.

The "attributable risk" represents the difference in the rate of a disease in an exposed, compared with a nonexposed, population. It reflects the additional incidence of disease related to an exposure taking into account the background rate of the disease

- **Number needed to treat** — The benefit of an intervention can be expressed by the "number needed to treat" (NNT). NNT is the reciprocal of the absolute risk reduction (the absolute adverse event rate for placebo minus the absolute adverse event rate for treated patients).
- NNTs from different studies cannot be compared unless the methods used to determine them are identical.
- When the outcome is a harm rather than a benefit, a number needed to harm (NNH) can be calculated similarly..

TERMS USED TO DESCRIBE THE QUALITY OF MEASUREMENTS

- **Reliability** — Reliability refers to the extent to which repeated measurements of a relatively stable phenomenon fall closely to each other. Several different types of reliability can be measured, such as inter- and intraobserver reliability and test-retest reliability.
- **Validity** — Validity refers to the extent to which an observation reflects the "truth" of the phenomenon being measured

MEASURES OF DIAGNOSTIC TEST PERFORMANCE

- **Sensitivity** — The number of patients with a positive test who have a disease divided by all patients who have the disease. A test with high sensitivity will not miss many patients who have the disease (ie, few false negative results).
- **Specificity** — The number of patients who have a negative test and do not have the disease divided by the number of patients who do not have the disease. A test with high specificity will infrequently identify patients as having a disease when they do not (ie, few false positive results).
- sensitivity and specificity are interdependent. Thus, for a given test, an increase in sensitivity is accompanied by a decrease in specificity and vice versa

Predictive values — The positive predictive value of a test represents the likelihood that a patient with a positive test has the disease. Conversely, the negative predictive value represents the likelihood that a patient who has a negative test is free of the disease

Definitions of sensitivity, specificity, and positive and negative predictive values

	Disease present	Disease absent
Test positive	**A**	**B**
Test negative	**C**	**D**
Sensitivity = A ÷ (A + C)		
Specificity = D ÷ (B + D)		
Positive predictive value = A ÷ (A + B)		
Negative predictive value = D ÷ (C + D)		

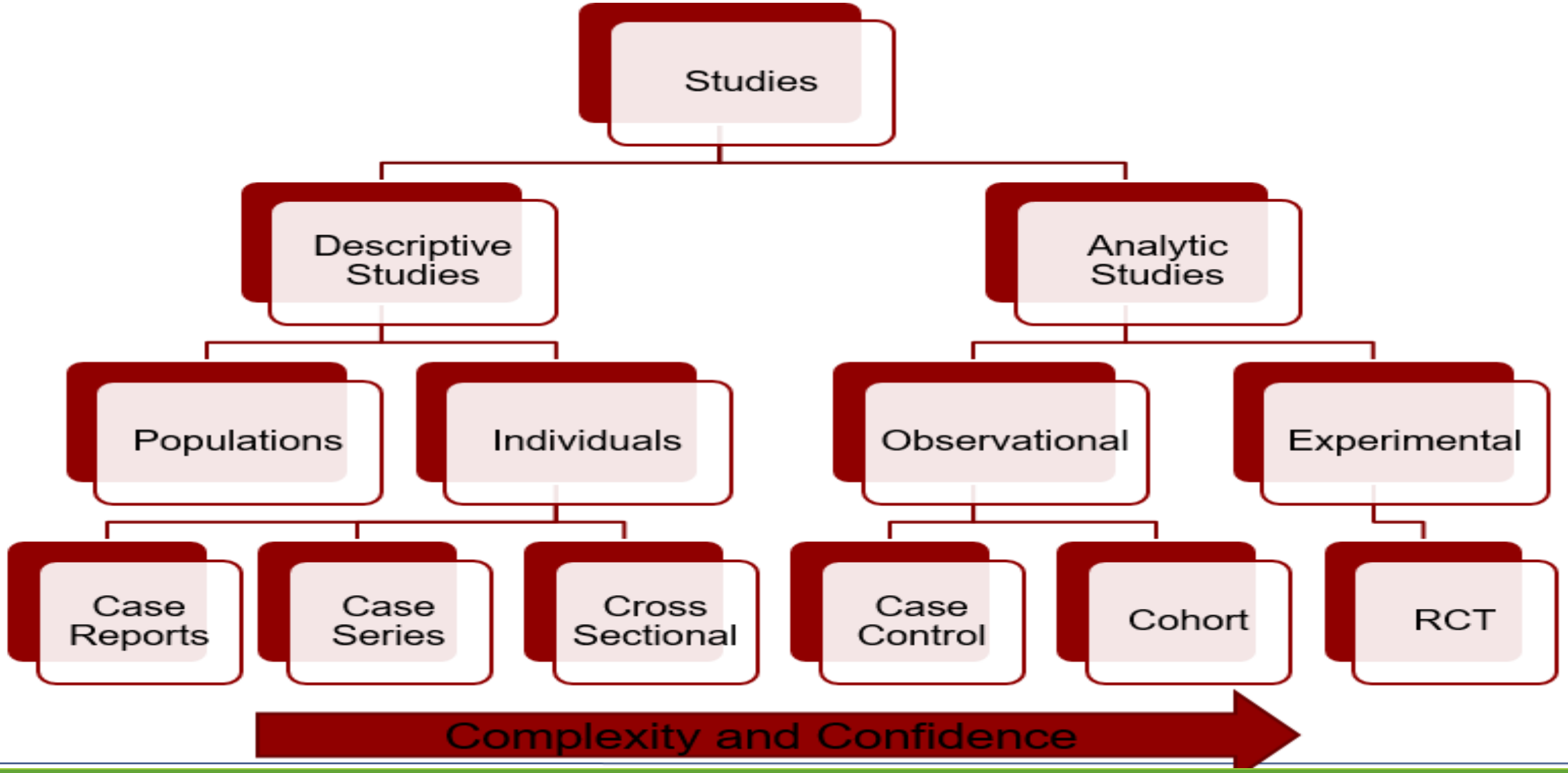
Accuracy — The performance of a diagnostic test is sometimes expressed as accuracy, which refers to the number of true positives and true negatives divided by the total number of observations

Definition of accuracy

	Disease present	Disease absent
Test positive	A	B
Test negative	C	D

Accuracy = The number of true positives plus the number of true negatives expressed as a percentage (ie, [A + D]/[A + B + C + D]).

Experimental statistics



Case Study Design

- Often a description of a individual case's condition or response to an intervention.
  - data may be qualitative, quantitative, or both
  - Case series: observations of several similar cases are reported
- **Strengths:-**
  - Enables understanding of the totality of an individual's (or organization, community) experience
  - The in-depth examination of a situation or 'case' can lead to discovery of relationships that were not obvious before

Case Study Design

- Useful for generating new hypotheses or for describing new phenomena

- **Weaknesses:-**
  - No control group
  - Prone to selection bias and confounding
    - The interaction of environmental and personal characteristics make it weak in internal validity
  - Limited generalizability

Cross-sectional Study

- Researcher studies a stratified group of subjects at one point in time
- **Strengths:-**
  - Fast and inexpensive
  - No loss to follow-up (no follow-up)
  - Ideal for studying prevalence
  - Data is useful for planning of health services and medical programs
- **Weaknesses:-**
  - Difficult to establish a causal relationship from data collected in a cross-sectional time-frame
  - Not practical for studying rare phenomena

Cohort Study

- A group of individuals who do not yet have the outcome of interest are followed together over time to see who develops the condition
- May identify risk by comparing the incidence of specific outcomes in exposed and not exposed participants
- Types of cohort study..
  - Prospective (concurrent)
  - Retrospective (historical)
  - Restricted (restricted exposures)

Cohort Study

- **Strengths:-**
  - Powerful strategy for defining incidence and investigating potential causes of an outcome before it occurs
  - Time sequence strengthens inference that the factor may cause the outcome
- **Weaknesses:-**
  - Expensive – many subjects must be studied to observe outcome of interest
  - Potential confounders: eg, cigarette smoking might confound the association between exercise and CHD

Case-Control Study

- Identify groups with or without the condition
- Look backward in time to find differences in predictor variables that may explain why the cases got the condition and the controls did not

MATCHING

• **CHARACTERISTICS OFTEN USED**

–age

–gender

–body mass index (weight / height2)

–smoking status

–marital status

Case-Control Study

	Case	control
Exposed	A	B
Unexposed	C	D
Total	A+C	B+D

Odds ratio:- A*D/B*C

OR = 1 then exposure is NOT related to disease

OR>1 then exposure is POSITIVELY related to disease

OR<1 then exposure NEGATIVELY related to disease

Case-Control Study

Strengths

- Useful for studying rare conditions
- Short duration & relatively inexpensive
- High yield of information from relatively few participants
- Useful for generating hypotheses

Weaknesses

- Increased susceptibility to bias:
  - Separate sampling of cases and controls
  - Retrospective measurement of predictor variables
- No way to estimate the excess risk of exposure
- Only one outcome can be studied

Experimental study

- It is a collection of research designs which use manipulation and controlled testing to understand causal processes.
- Generally, one or more variables are manipulated to determine their effect on a dependent variable
- **Characteristics or Features of Experimental Design:-**

1. Manipulation(M)

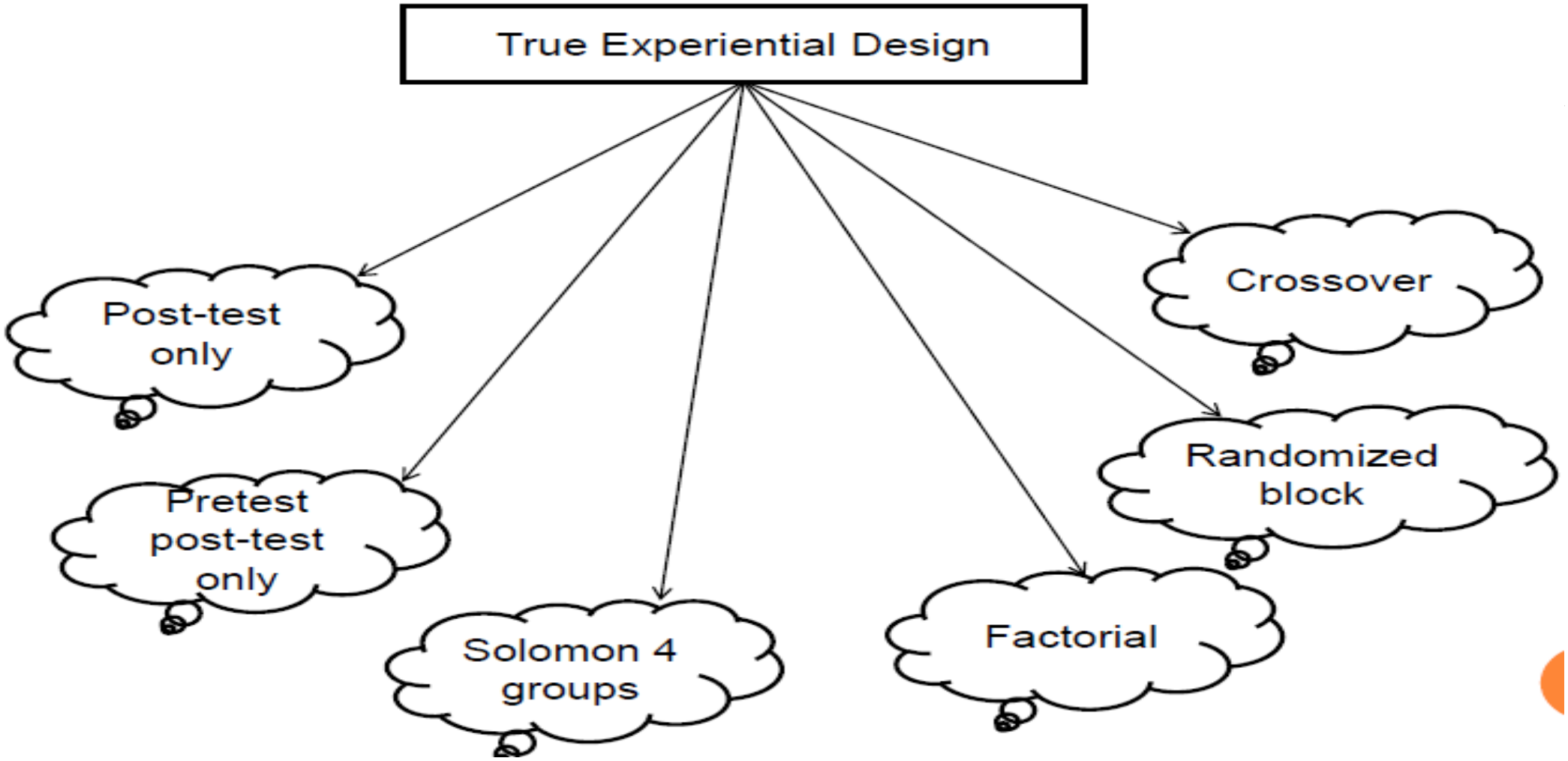
2. Control(C)

3. Randomization(R)

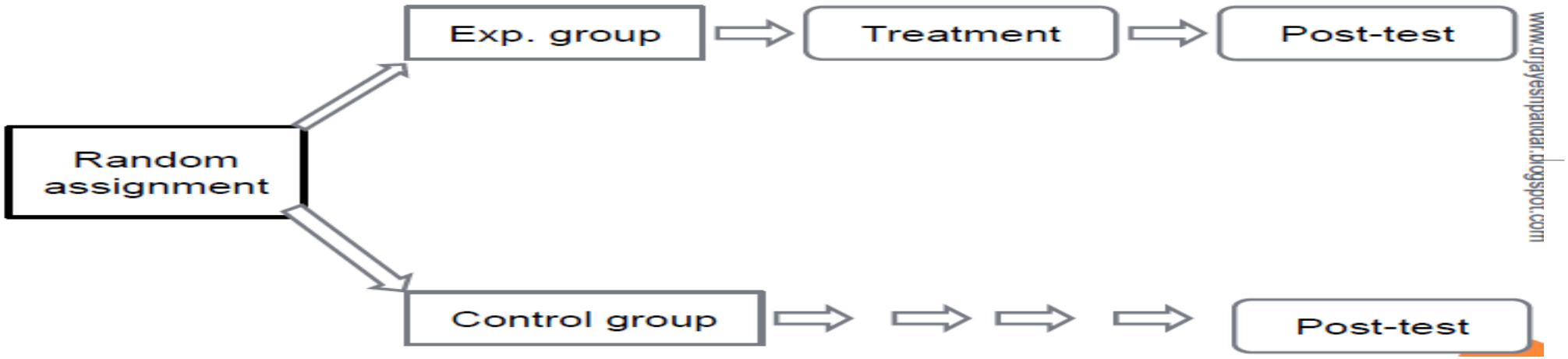
Experimental study

- **Types of Experimental Designs:-**

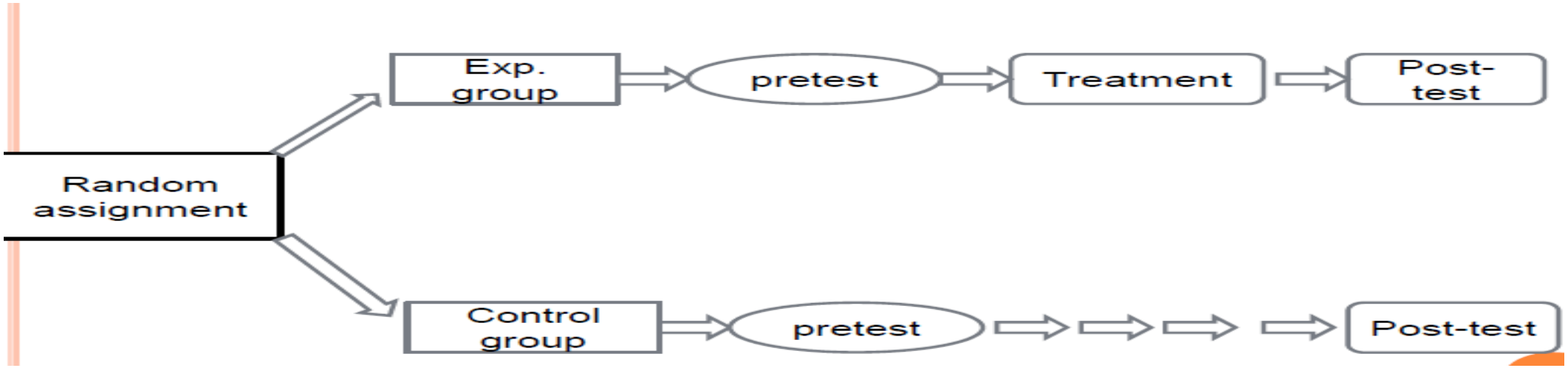
- True-Experimental (Simple):- M+R+C
- Quasi-Experimental:-M+R or C
- Pre-Experimental:- M, no R or C


True Experiential Design
Post-test only
Pretest post-test only
Solomon 4 groups
Factorial
Randomized block
Crossover

True experimental design



POST-TEST-ONLY CONTROL DESIGN



PRETEST-POST-TEST-ONLY DESIGN

True experimental design

RANDOMIZED BLOCK DESIGN

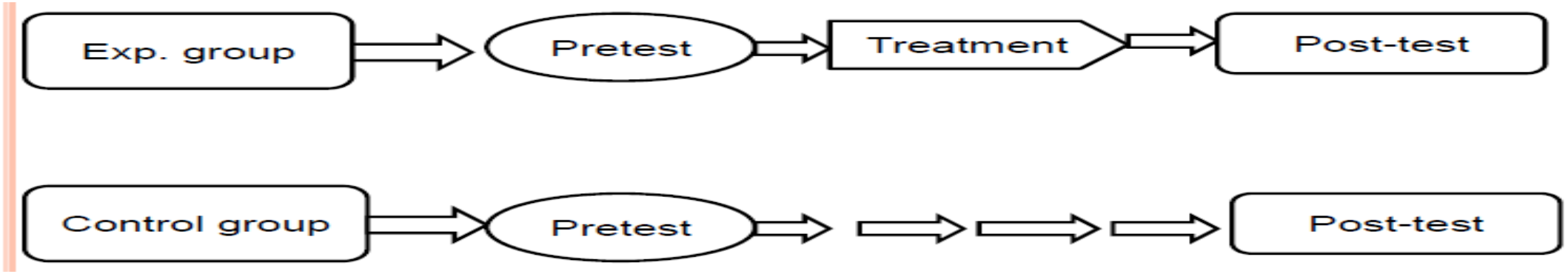
Types of antihypertensive drugs	Blocks		
	Patients with primary hypertension (I)	Diabetic patients with hyper tension (II)	Renal patients with hypertension (III)
A	A, I	A, II	A, III
B	B, I	B, II	B, III
C	C, I	C, II	C, III

CROSSOVER DESIGN

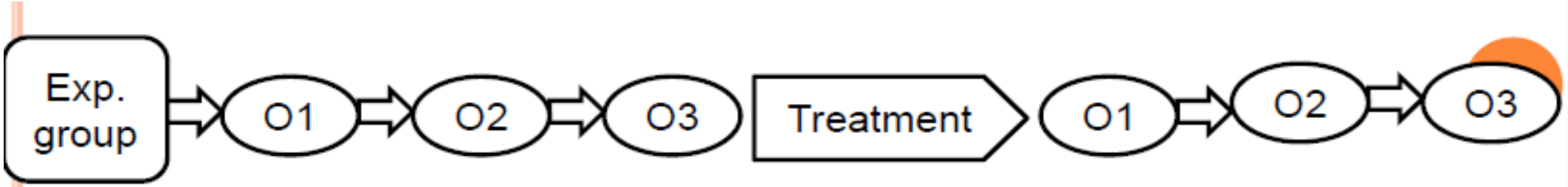
Groups	Protocols of the mouth care	
Group I	Chlorhexidine (α1)	Saline (α2)
Group II	Saline (α2)	Chlorhexidine (α1)

TYPES OF QUASI-EXPERIMENTAL DESIGN

NONRANDOMIZED CONTROL GROUP DESIGN

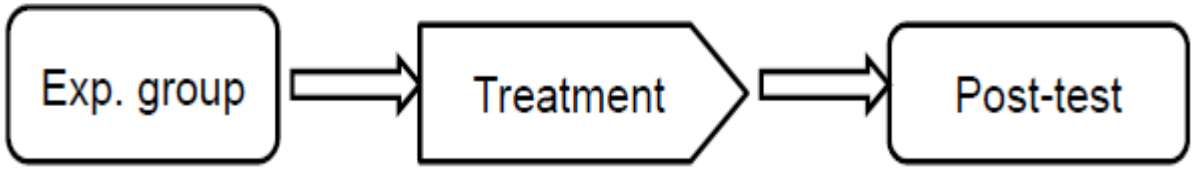


TIME-SERIES DESIGN

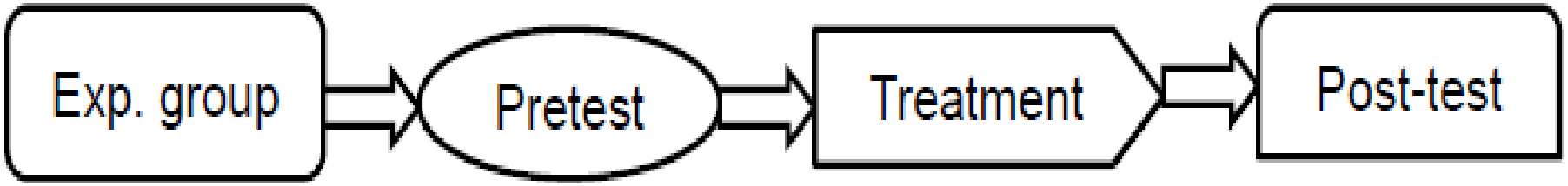


TYPES OF PRE-EXPERIMENTAL RESEARCH DESIGN

ONE-SHOT CASE DESIGN



ONE-GROUP PRETEST-POSTTEST DESIGN



Experimental study

- **Strengths:-**
  - Controls the influence of confounding variables, providing more conclusive answers
  - Randomization eliminates bias due to pre-randomization confounding variables
  - Blinding the interventions eliminates bias due to unintended interventions

Experimental study

- **Weaknesses:-**

- Costly in time and money
- Many research questions are not suitable for experimental designs
- Usually reserved for more mature research questions that have already been examined by descriptive studies
- Experiments tend to restrict the scope and narrow the study question

Meta Analysis

- Statistically combines results of existing research to estimate overall size of relation between variables.
- Helps in
  - Developing theory
  - Identifying research needs,
  - Establishing validity
- Can replace large-scale research studies
- Better than literature reviews