University of Baghdad

Al Kindy College of Medicine

Department of Community & Family Medicine

# Community & Family Health Module - Part I

## Identification & Measuring of Health Events in Community

## 3rd Year 2017-2018


**By:**

**The Staff members of Community & Family Medicine Department**

**Al Kindy College of Medicine University of Baghdad**

**Note**

- These lectures worth 50% of the degree of the modules
- The topics will be delivered for the students in lectures, tutorials, seminars and group discussion

- **Learning Objectives:** Upon completion of this theme, students will be able to:
  1) Understand the role of preventive medicine practice in healthcare practice
  2) Recognize the importance and level of prevention in health care setting
  3) Distinguish the natural history of the disease and disease spectrum and distribution in the community
  4) Discuss the importance of population data as a key approach to studying disease and social conditions at individual level.
  5) Understand key terms and basic epidemiology & biostatical measurements;
  6) Use epidemiological data to identify health problems, understand their causes, and develop targeted programs in maternal and child health; and
  7) Calculate and interpret ratios, proportions, incidence rates, mortality rates, prevalence, and years of potential life lost.
  8) Calculate and interpret summery statistics (mean, median, mode, ranges, variance, standard deviation)
  9) Understand the concept of population & sampling & determine whether or not a situation represents random sampling or not.
  10) Distinguish between a parameter and a statistic define sampling error and be able to identify both bias and homogeneity in samples., and normal distribution.
  11) Calculate the appropriate probabilities and z-scores from actual data as an answer to a question about the data, assuming the data is normally distributed.
  12) Prepare and apply tables, graphs, and charts such as arithmetic- scale line, scatter diagram, pie chart, and box plot for data presentation.
  13) Understand the concept of screening and the difference between diagnostic test & screening test. Describe the characteristics of diseases appropriate for screening. Understand validity and reliability of screening test
  14) Describe the process of surveillance & continues monitoring of diseases and how could we evaluate surveillance systems
  15) Recognize the importance of research & descriptive studies in assessing the impact of disease in community and formulate the hypothesis for exposure-disease association. Know the advantage and limitation of case report, case series, correlational study, and cross sectional study
  16) Recognize the importance of analytic research in measuring the association between suspected exposure and interesting outcome. calculate the risk of association in each design OR in case control, AR RR in cohort & interventional designs
  17) Understand the statistical inference and calculation used to reach a valid inference
  18) Distinguish between estimation in general and statistical estimation using the concept of p-values.
  19) Calculate a confidence interval for the true mean of a population given the true mean and standard deviation.
  20) Understand the role of medical research in Evidence based medical practice and the levels of evidence.

- **Introduction to Community Health:**
  - Each year, millions of people die of preventable deaths. Leading causes included cardiovascular disease, chronic respiratory disease, unintentional injuries, diabetes, and infectious diseases. According to estimates made by the World Health Organization (WHO), about 57 million people died worldwide in 2015, 2/3 from non-communicable diseases, including stroke, cardiovascular diseases, lung diseases cancer, and diabetes.
  - In 2015, 7.8 million children died before reaching the age of 5. Child mortality is caused by a variety of factors including poverty, infections & other environmental hazards, and lack of maternal education.
  - Every day, approximately 830 women die from preventable causes related to pregnancy and childbirth.
  - Other emerging risk factors and conditions that affect the health of community: Obesity, sexually transmitted diseases, smoking, violence, addiction, accidents.
  - **What is a Community?** A social unit (a group of people) who have something in common, such as norms, values, or identity. Often - but not always - communities share a sense of place that is situated in a given geographical area (e.g. a country, village, town, or neighborhood).
  - **What is Health** can be defined: Complete physical, mental, social, spiritual wellbeing and not merely the absence of disease or infirmity with the ability to lead a socially and economically productive life.
  - **Community Health:** A subject of study within the medical and clinical sciences which focuses on population groups and communities as opposed to individual patients. It is a discipline which concerns itself with the study & improvement of the health characteristics of families & communities.
  - Community health focuses on prevention (preventive medicine). This can be offered through three levels:

1) Primary healthcare: Interventions that focuses on the individual or family and usually provided at primary health care centers (PHC center). It mainly focuses on health equity, accessibility & acceptability of health services. Its main aim is to provide local care to a patient because professionals related to primary care are normal generalists or family medicine specialists.

Essential 8 **Elements** of Primary Health Care (PHC):

- ✓ E– Education concerning prevailing health problems and the methods of identifying, preventing and controlling them.
- ✓ L– Locally endemic disease prevention and control.
- ✓ E–Expanded programme of immunization against major infectious diseases.
- ✓ M– Maternal and child health care including family planning.
- ✓ E– Essential drugs provision.
- ✓ N– Nutritional food supplement, an adequate supply of safe and basic nutrition.
- ✓ T– Team working to screen & control communicable and non-communicable disease and promotion of mental health.
- ✓ S– Safe water and environment sanitation (sanitary collection and disposal of waste, spraying insecticides to control vectors like mosquitoes & rodents).

2) 2sdry an intermediate level of health care that includes diagnosis and treatment of common diseases, performed in a PHC center having specialized equipment and laboratory facilities.

3) Tertiary healthcare: Interventions that take place in a hospital setting or specialized centers, such as specialized medical or surgery intervention or advanced medical investigation and treatment. This healthcare (specialized consultative healthcare) is provided by the medical specialists like cardiologists, urologists, dermatologists, ophthalmologist etc.

- Comprehensive Preventive Healthcare is especially important given the worldwide rise in morbidity & mortality rates and the cost in management of most common diseases.
- There is general consensus as preventive healthcare measures are cost-effective, & increase the quality of life dramatically.
- Preventive medicine specialists are trained in both clinical medicine and public health, and are equipped to understand and reduce the risks of disease, disability, and death in individuals and in population groups.

- Primary Health Care is concerned with establishing a system of health which meets the essential needs of most of the population.
  - ✓ It is "first-contact" care, serving as a point-of-entry for the patient in to the health care system (usually in PHC center)
  - ✓ It includes continuity by virtue of caring for patients over a period of time, both in sickness and in health
  - ✓ It is comprehensive care, drawing from all the traditional major disciplines for its functional content
  - ✓ It serves a coordinative function for all the health-care needs of the patient
  - ✓ It assumes continuing responsibility for individual patient follow-up and community health problems
  - ✓ It aims to promote health of family and community.

**Public Health:** **T**he science and art of preventing disease, prolonging life and promoting health through organized efforts and informed choices of society, organizations, public and private, communities and individuals."

Public health practice requires multidisciplinary teams of public health workers and professionals including physicians specializing preventive/community medicine/infectious disease, psychologists, medical assistants, nurses, midwives, medical microbiologists, environmental health inspectors, pharmacists, dentists, dietitians and nutritionists, veterinarians, public health engineers, public health lawyers, sociologists, community development workers, communications experts, bioethicists, computer informaticians, and others. Public Health has become an important specialty in developed countries in the early of 19th century and it is now the leader of health all over the ward.

**Health is the responsibility of all**

Individual responsibility, Community responsibility, State responsibility & International responsibility

## Community Medicine (Preventive or Social Medicine):

The science that concerns with the promotion of health, prevention, control, and management of diseases, disabilities, and other health problems in the community. It is the branch of medicine (only physicians) that deals with community rather than individual.

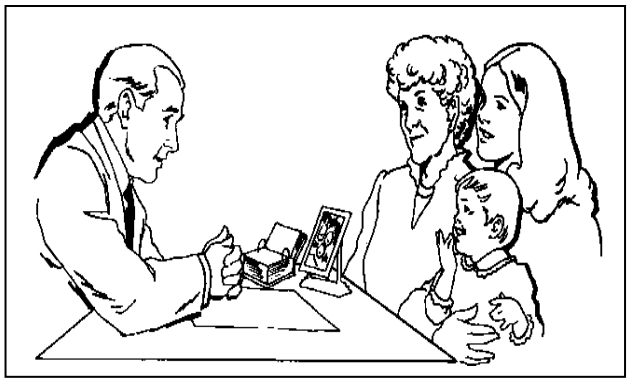
The emphasis in community medicine is on the early diagnosis of disease, the recognition of environmental and occupational hazards to good health, and the prevention of disease in the community through collaborative community action & research.

## Family Medicine (FM):

A specialty devoted to comprehensive health care (preventive & curative) for people of all ages. Unlike other specialties that are limited to a particular organ or disease, family medicine integrates care for patients of all genders and every age, and advocates for the patient in a complex health care system.

## Community & Family Physicians are able to :

- ✓ Provide Evidence based health care (manage acute health problems, chronic health problem, & emergency services).
- ✓ Provide health promotion services.
- ✓ Communicate efficiently.
- ✓ Carry out studies and research.
- ✓ Plan and implement action plan.
- ✓ Educate and train his/her colleagues.
- ✓ Refer patients when it is needed.
- ✓ Activate community participation.
- ✓ Coordinate with other community sectors.
- ✓ Update himself regularly



☞ **Community Medicine physician** focuses on the health of communities, and defined populations (workers, soldiers, sport players…. etc.). Its goal is to protect, promote, and maintain health and well-being and to prevent disease, disability, and death. He/she possesses core competencies in biostatistics, epidemiology, environmental and occupational medicine, planning and evaluation of health services, management of health care organizations, research into causes of disease and injury in population groups, and the practice of prevention in clinical medicine. They apply knowledge and skills gained from the medical, social, economic, and behavioral sciences.

☞ **Family physician** is the first contact and gate of health care system in primary health care centers. The scope of family medicine encompasses all ages, sexes, each organ system to diagnose and treat common diseases and health problems in his/her community. Advance cases of diseases which need a specialized personnel and center would be refereed accordingly.

## Aims of Preventive Medicine Specialty

1. Promote and preserve health and longevity in individuals and community by adoption of healthy life style and health education.
2. Prevent and limit diseases, injuries, and other ill health effects.

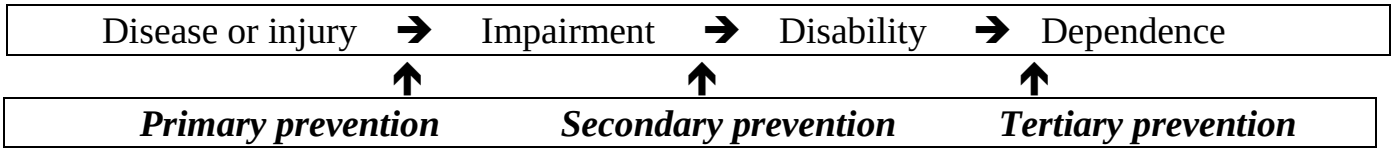
**3.** Enhance quality of health care system and assure that all populations have access to appropriate and cost effective care.

**4.** Assess and monitor the health of communities and populations at high risk to identify health problems.

- **<u>Concept of Prevention</u>:**

➢ Prevention is the anticipatory action taken to reduce the possibility of an event or condition occurring or developing, or to minimize the damage that may result from the event or condition if it does occur.

➢ Prevention may take place at any point along the spectrum of the disease, from the prevention of the disease or injury to the prevention of impairment, disability or dependency.




➢ Categories of disease prevention:

1- ***Primordial prevention:*** (Acting on risk factor) consists of actions and measures that inhibit the emergence of risk factors in the form of environmental, economic, social, and behavioral conditions and cultural patterns of living etc.

2- ***Primary prevention:*** (Acting before disease occurrence) Activities designed to prevent onset of disease. We act before the development of the sign and symptoms of the disease. Ex: Immunization, ban on smoking, speed limit, seat belts.

Primary prevention may be accomplished by measures of:

- Health promotion: It is a process of enabling people to increase control over & to improve health. This could be done by health education, environmental modification, nutritional interventions, and lifestyle and behavioral changes.
- Specific protection: Specific measures for certain disease such as immunization and seroprophylaxis, chemoprophylaxis, use of specific nutrients or supplementations, Protection against occupational hazards

3- ***Secondary prevention:*** (Acting after disease occurrence) Early identification of health problems to reduce the risk of progression or transmission. Ex: Early diagnosis of HT, DM, breast CA, STD.

4- ***Tertiary prevention:*** (Acting after complications occurrence) Focused on rehabilitation to reduce the impairment. Ex: Learning to walk after stroke, adjusting diet and life style after MI, learning to live with DM.

5- Quaternary prevention: Prevention from iatrogenic (medical) harm. This action taken to identify patients at risk of (over) medication, to protect him from new medical invasion or to suggest to him interventions, which are ethically acceptable.

So, prevention is any intervention that seeks to reduce or eliminate diagnosable conditions and may be applied at individual level, as in immunization, or the community level, as in the chlorination of the water supply.

**Control:** The term disease control describes ongoing operations aimed at reducing:

- Incidence of disease
- Duration of disease
- Complications (physical & Psychological)
- Financial burden

## Disease Elimination:

Interruption of transmission of disease in the community, EX: elimination of measles, polio and diphtheria

## Disease Eradication:

It is the process of "Termination of all transmission of infection by extermination of the infectious agent through surveillance and containment". To-date, only one disease has been eradicated, that is smallpox.

## • History of Preventive Medicine

**"An ounce of prevention is worth than a pound of cure"**

Henry De Bracton, 1240

The concept of preventive health care is not modern. Before the 20th-century biomedical revolution, important advances had been made in the understanding and prevention of infectious diseases and nutritional deficiency. Long before the advent of scientific investigation into disease processes, men were explaining disease and attempting to avoid illness. In surveying historical efforts to prevent communicable diseases and nutritional deficiency diseases three categories of prevention are apparent; (1) individual control over personal health through adherence to dietary and hygiene codes; (2) social control over health by means of isolating diseased individuals or protecting large groups of people from environmental dangers; (3) application of increased scientific understanding of disease.

The "explosion" of knowledge during the 20th century has made medicine more complex, and treatment more costly, but the benefits of modern medicine have not yet penetrated the social periphery in many countries.

The important goals which have emerged are prevention of disease, promotion of health and improvement of the quality of life of individuals and groups or communities.

Family Medicine as specialty was recognized in Europe and America in the 1950s. It resembles the general medical practitioner that provides comprehensive primary health care.

## Issues of Preventive Medicine

1. General Epidemiology & biostatistics: Measures the occurrence and distribution of diseases in population.
2. Research Methodology
3. Monitoring and screening of health problems in community
4. Primary Health Care, Includes:
    - ❖ Health education.

- ❖ Management of common diseases in community.
- ❖ Maternal and child health care including family planning.
- ❖ Mental health.
- ❖ Accidents and injuries.
- ❖ Geriatric Health

5. Nutritional Health Nutritional Disorders (under nutrition, over nutrition & obesity).
6. Infectious Diseases (Communicable disease) Control.
7. Non-Communicable Diseases management e.g. HT, DM, Cancer.
8. Environmental &Occupational Health.
9. Health Services Management (planning, leading, organization & evaluation of health services).

**Public Health Problems** ***((Problems associated with poverty and overcrowding))***

1. In Developing Countries:
   - ❖ Infectious diseases e.g. TB, malaria….etc.
   - ❖ Malnutrition.
   - ❖ Poor health education.
   - ❖ Limit access to health services.

2. In Developed Countries: ***((Problems associated with industrialization, affluence, aging, violence, and medical intervention))***
   - ❖ Chronic diseases e.g. IHD, HT, DM…etc.
   - ❖ Over nutrition and obesity.
   - ❖ Violence and drug addiction.
   - ❖ Sexually transmitted diseases (STD)

## The Concept of Health and Disease

- ➢ The ultimate goal of preventive medicine is to keep the community healthy or promote and protect the health of community
- ➢ Better health is central to human happiness and well-being. It also makes an important contribution to economic progress and development, as healthy populations live longer, are more productive, and save more.
- ➢ Multidisciplinary approach to health: Many factors influence health status and a country's ability to provide quality health services for its people. Ministries of health are important actors, but so are other government departments, donor organizations, civil society groups and communities themselves. For example: investments in roads and electricity, can improve access to health services.
- ➢ **Health** is defined as" Complete physical, mental and social wellbeing and not merely the absence of disease or infirmity-WHO 1948". Then add spiritual, and in recent years the statement is amplified to include the ability to lead a socially and economically productive life.
- ➢ **Disease**: It is the converse of Health. It refers to any change from a normal state of health or an abnormal state in which part or all of the body is not properly adjusted or is not capable of carrying on its normal functions. Literary, "DIS-EASE", is the opposite of ease, when something wrong in the body function, or any deviation from normal. The

words "disease", "illness", "sickness", are loosely interchangeable, but are better regarded as not wholly synonymous.

- ☞ **Disease:** A cluster of signs, symptoms and laboratory findings linked by a common patho-physiologic sequence.
- ☞ **Illness:** The subjective state of the individual who feels aware of not wellbeing (The ill individual may or may not be suffering from disease).

  Note: Disease is an objectively measurable pathological condition of the body. Tooth decay, measles, or a broken bone, are examples. In contrast, illness is a feeling of not being normal and healthy. Illness may, in fact, be due to a disease. However, it may also be due to a feeling of psychological or spiritual imbalance.
- ☞ **Sickness:** the social role assumed by an individual suffering from an illness
- ☞ **Syndrome:** When the signs and symptoms have not yet clearly been placed in a common patho-physiologic sequence.

- **<u>Causes of Disease:</u>** some diseases have a well understood etiology, others have a partially understood etiology, and others have an undetermined etiology. The main categories of disease include:
  **1)** Infectious disease – caused by disease producing microorganisms
  **2)** Nutritional deficiency disease – caused by the lack of a particular, necessary nutrient
  **3)** Congenital disease* – is present at birth and is the result of some condition that occurred in utero (maternal infection, use of drugs or alcohol, etc.)
  **4)** Genetic diseases (Inherited disease): are passed to the child via the parent's reproductive cells
  **5)** Degenerative diseases: This occurs when there is a wearing down of part of the body leading to loss of function. This may be due to aging, excessive caloric intake, radiation, errors in gene function, etc.
  **6)** Neoplastic diseases – these are tumors which are new growth of cells or tissues. Tumors may be benign or malignant.
  **7)** Immunologic diseases – this occurs when some of our immunologic defenses attack our own bodies. Are also called autoimmune diseases.
  **8)** Iatrogenic disease: caused by health care personnel during the delivery of health care
    - Could be due to use of contaminated equipment
    - Could be caused by the administration of drugs

  **9)** Psychogenic diseases – are caused, at least in part, by emotional factors
  **10)** Idiopathic diseases – diseases that have an undetermined cause SLE

## ➢ <u>Signals of Disease?</u>

- ☞ Symptoms – subjective changes in body function such as pain or malaise
- ☞ Signs – objective changes that can be observed and measured such as fever, swelling, or a rash

❖ Clinical disease: characterized by signs & symptoms.
❖ Non-clinical (Inapperant, Asymptomatic) disease➔Preclinical state, sub clinical disease, chronic disease, latent disease, or carrier state.

❖ What are the stages of disease?

- Period of incubation: the time been acquiring the infection and the appearance of the first signs or symptoms. This may be a constant time for every individual who acquires the infection or a variable time depending upon the disease.
- Communicable period (infectious diseases): the time interval from infection to development of infectiousness

  e.g : Chicken pox an infectious disease caused by the varicella-zoster virus the communicable period for chicken pox is shorter than the incubation period, so a child with chicken pox becomes infectious to others before developing symptoms
- Latent Period of Chronic Disease: Interval between exposure to a disease-causing agent and the appearance of manifestations of the disease"
- Prodromal period – when the first signs and symptoms appear.
- Period of illness: When the disease is most acute & the overt signs and symptoms of the disease occur.
- Period of decline – this is where the signs and symptoms subside. If the decline occurs quickly, it is said to occur by crisis. If the decline occurs over a longer period of time, it is said to occur by lysis.
- Period of convalescence - this is where the person regains strength and the body returns to its pre-diseased state.

➔It is important to recognize the board spectrum of disease severity in order to identify the level of prevention.

- **<u>Natural History of Disease</u>**

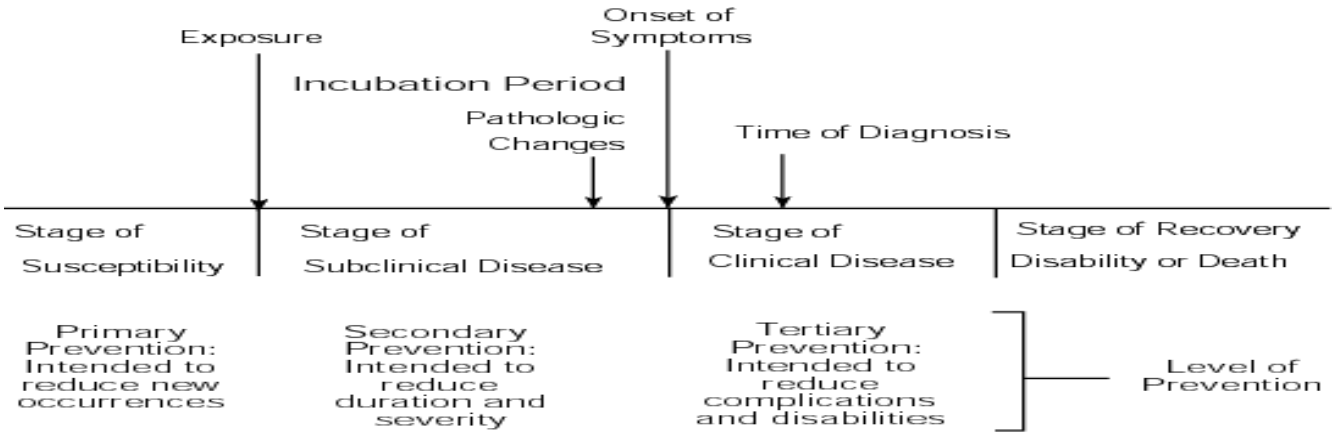
The progress of a disease process in an individual over time, in the absence of intervention. It describes the course of the disease in an individual starting from the moment of exposure to the causal agents till one of the possible outcomes occurs.

➢ Importance of natural history

✓ Natural history is as important as causal understanding for the prevention and control of disease.

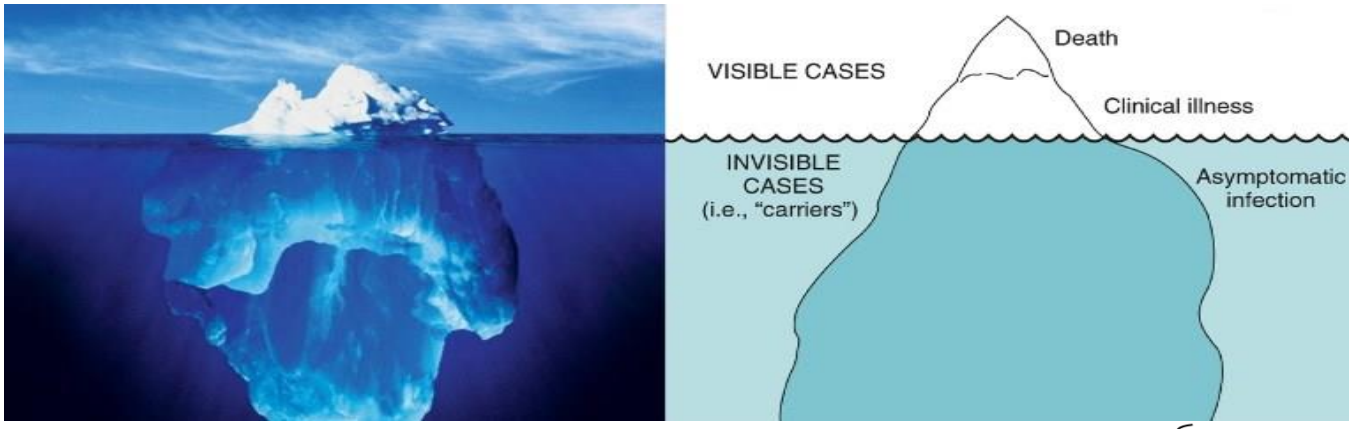
✓ The earlier you can become aware of the attack the more likely you will be able to intervene and save lives.

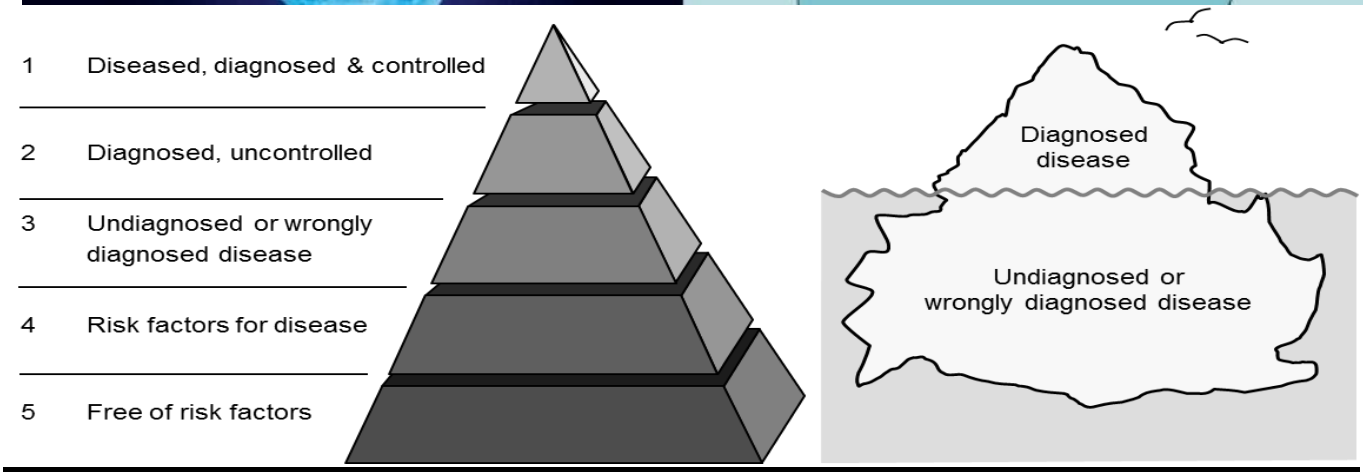
✓ The outcome will depend on the interactions of host, agent and environmental factors.



➢ Iceberg phenomena of disease

- ✓ In the community/society a far larger proportion of disease (e.g., diabetes, hypertension) is hidden from view of the general public or physician.
- ✓ In this context the analogy of an iceberg is widely used to describe the disease pattern in the community. For example, HIV infection has broad clinical spectrum, from inapparent* to severe and fulminating**.
- ✓ The diversity in presentation makes the presentation of disease in community as an **ice berg (Small apparent tip and wide hidden base).**
- ✓ The concept of the "iceberg phenomenon of disease" gives an idea of the progress of a disease from its sub-clinical stages to overt or apparent disease state.
- ✓ The submerged portion of the iceberg represents the hidden mass of the disease (e.g., subclinical cases, carriers, undiagnosed cases).
- ✓ The floating tip represents what the physician sees in his practice/chamber/hospital etc. The remaining Large Hidden part of the iceberg is what constitutes the mass of unrecognized disease in the community.
- ✓ The concept of the "iceberg phenomenon of disease "gives an idea of the progress of a disease from its sub-clinical stages to overt or apparent disease state.



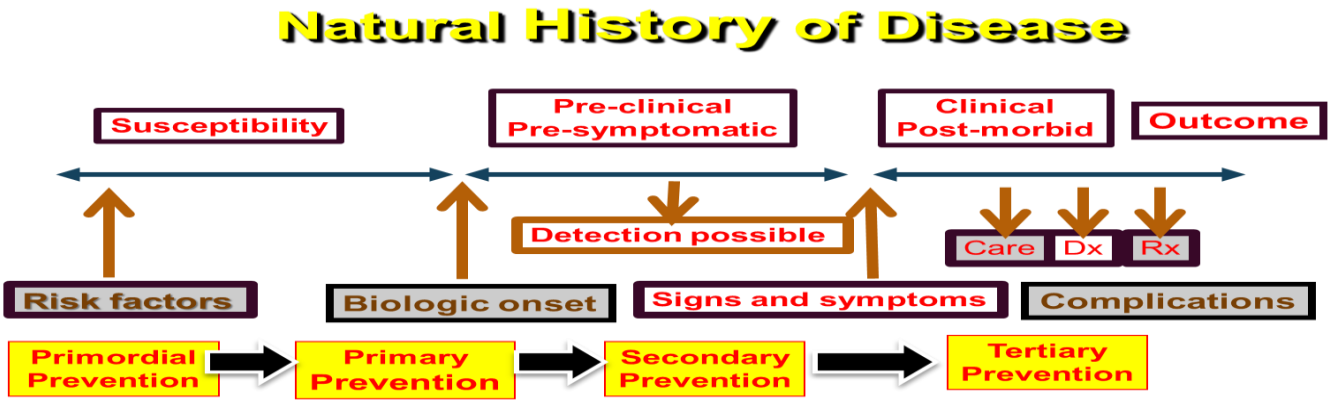

➢ **<u>Consequences or effects of a disease:</u>**

**(Disease→ Impairment →Disability → Handicap)**

1. Impairment: Functional loss in a part of the body.
2. Disability: Functional loss plus psychological upset.
3. Handicap: Impairment of the social role played.

| Ex: Disease | Impairment | Disability | Handicap |
|---|---|---|---|
| Polio | Paralyzed legs | Inability to walk | Unemployed |
| Brain injury | Mild mental retardation | Difficulty learning | Social isolation |

## Link between Natural History of Disease & prevention level




- **How can we identify disease and risk factors?**

☞ Epidemiology and biostatistics are the sciences that quantify diseases and health events in community

☞ **Epidemiology:** is the study of the distribution and determinant of disease or health related status or events (all health outcomes) in specified population, and the application of this study to control of health problems'

☞ It is basically the study of health and disease in population and how we can improve the health and prevent the disease at community level.

☞ **Epidemiology** is a Greek word (epi = upon, demos= people, district, logos= word, discourse) ➔ Study what fall on human population. It is the basic science of preventive Medicine. Epidemiology can be:

**1)** Descriptive Epidemiology studies the Distribution (Frequency of health events) by person, time and place. It is the starting point to **formulate the hypothesis** for better understanding any disease or health event. It asks: **what** is the problem and its frequency, **who** is involved, **where**, and w**hen**?

☒ **Person:** Who is getting the disease and who is not? ➔Age, Sex, Race, Marital status, socioeconomic status [usually measured by education, occupation and income].

☒ **Place:** Where is the rate of the disease highest or lowest? ➔ Geographical distribution of the disease.

☒ **Time:** When dose the disease occur commonly or rarely? ➔ Epidemic, seasonal changes, secular trend (changes over years).

**2)** Analytic Epidemiology studies the Determinants (Search for causes or risk factors) by Agent, Host, and Environment. It aids in **testing a hypothesis** about the cause of disease by studying how exposure of interest relates to the disease of interest. It observes the Agent, Host and the Environment. It asks: **how** and **why** the disease has occurred.

☒ **Agents: could be:**

- Biological (micro-organisms)
- Physical (temperature, radiation, trauma, others)
- Chemical (acids, alkalis, poisons, tobacco, medications / drugs, others)
- Environmental (nutrients in diet, allergens, others)
- Nutritional (under- or over-nutrition)

☒ **Host factors:** are intrinsic factors that influence an individual's exposure, susceptibility, or response to a causative agent. These include: Genetic, Immunologic state, Personal behavior (life-style factors): diet, tobacco use, exercise, etc and Personal characteristics (described before, under "person"), including: age, gender, socio-economic status, etc.

☒ **Environmental factors** are **extrinsic** factors which affect the agent and the opportunity for exposure. These include:

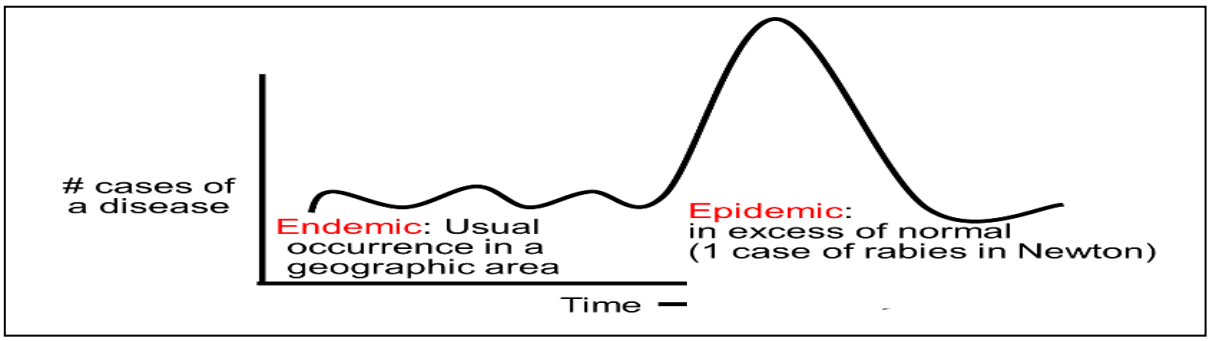
- Physical factors: e.g. geology, climate (temperature, humidity, rain, etc)
- Biological factors: e.g. insects that transmit an agent
- Socioeconomic factors: e.g. crowding, sanitation, and the availability of health services

## Some Epidemiological terms in disease determination:

- Sporadic: Refers to a disease that occurs infrequently and irregularly
- Endemic: Disease that is usually & habitually present in a community.
- Epidemic Refers to an increase, often sudden, in the number of cases of a disease above what is normally expected in that population in that area.
- Outbreak: Carries the same definition of epidemic, but is often used for a more limited in time & geographic area.
- Pandemic: Refers to an epidemic that has spread over several countries or continents, usually affecting a large number of people.
- Cluster: Refers to an aggregation of cases grouped in place and time that are suspected to be greater than the number expected.
- Hyperendemic: Refers to persistent, high levels of disease occurrence.
- Examples:
  - ✓ 22 cases of legionellosis occurred within 3 weeks among residents of a particular neighborhood (usually 0 or 1 per year) : **EPIDEMIC (OUTBREAK)**
  - ✓ Average annual incidence was 364 cases of pulmonary tuberculosis per 100,000 population in one area, compared with national average of 134 cases per 100,000 population. **HYPERENDEMIC**
  - ✓ Over 20 million people worldwide died from influenza in 1918–1919: **PANDEMIC**
  - ✓ Single case of histoplasmosis was diagnosed in a community: **SPORADIC**
  - ✓ About 60 cases of gonorrhea are usually reported in this region per week, slightly less than the national average: **ENDEMIC**

- Epidemic curve: Graph that depicts the relation between disease cases against time (time could be hours e.g. food poisoning, days e.g. cholera, weeks e.g. typhoid fever,

or months e.g. kala-azar). An "epidemic curve" shows the frequency of new cases over time based on the date of onset of disease. The shape of the curve in relation to the incubation period for a particular disease can give clues about the source.




- Surveillance: The ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event for use in public health action to reduce morbidity and mortality and to improve health.
- Surveillance: Information for Action, The word surveillance comes from the French word for "watching over." It is the monitoring of people behavior, activities, or other changing information.

☞**Biostatistics:** is the science of collecting, organizing, summarizing, analyzing, and making inference about data collected from community. is the field of study which can be classified into;

1- ***Descriptive:*** Concern with collection, classification, organization and summarization (reduction) of the data. (They are merely descriptive & used to describe the basic features of the data in a study. They provide simple summaries about the measures make no attempt to draw conclusion)). They can be:

a- Tabular (Tables).
b- Diagrammatic (Figures).
c- Numerical (Numbers).

2-***Inferential:*** Concern with drawing conclusions about population from the data that collected from representative sample (making inference, hypothesis testing, determining relationship, and making prediction). The conclusion drawn will influence sub- sequent decision.

♦ Epidemiology & Biostatistics is an indispensable tool in preventive medicine.
♦ All medical studies rely on the quantification of health and disease events in populations, which combine epidemiology and biostatistics (If we measure…We know better).
♦ Epidemiology is about the understanding of disease development and the methods used to uncover the etiology, progression, and treatment of the disease
♦ The methods and tools of biostatistics are used to analyze the data collected to investigate a question and to aid decision making

## Uses of Epidemiology &Biostatistics:

1- Measure & analyze the health status and health problems in the community.
2- Compare the health status of the community with others.

3- Planning for the health services.
4- Evaluation of the health services & estimation the future needs.
5- For research purposes. "**Statistics is vital & central to most medical research".**
6- Evaluation of published paper.

**Examples**

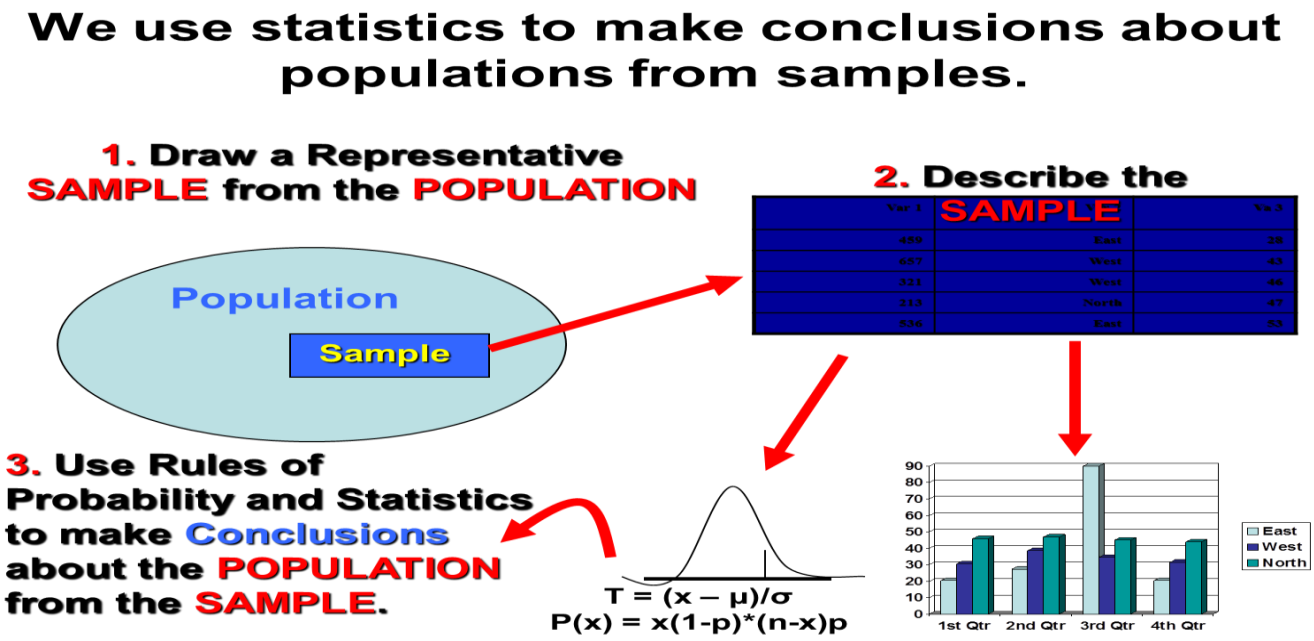
- ✓ Computing age-adjusted cancer incidence rates (lung cancer. Leukemia…..) to determine trends over time and locality in Iraq
- ✓ Calculating the risk of developing brain tumors following cell phone use
- ✓ Quantifying the relationship between use of Cox-2 inhibitors and myocardial infarction

## • Recognition of Data & Variables in Population & Sample

- Preventive Medicine focuses on health issues in populations. Change the size and characteristics of the population is important determinant of disease.
- The relationship between health and population dynamics guides the need for changes in medical practice. Population data are essential to defining and measuring public health problems and the groups of people in which they occur.
- Changes in health influence **vital events**, including births, deaths, and divorce, in turn leading to population changes. Migration, the movement of people from place to place, is another demographic force that leads to new health issues and problems.
- The target of preventive medicine is a population with certain characteristic of interest: for example, hypertensive patients or cholesterol level in elderly individuals, or diabetic patient responds to their different line of treatment.

- **Population:** largest collection from which we have an interest at particular time**. (Collection of entire people you want to understand).** If a population of value consists of fixed number of these value→ "Finite population", but it consists of an endless values → "Infinite population"
- Sometimes data from the entire population is difficult or even impossible to obtain or the population does not even exist. For example: How can we find hypertensive or diabetic patients in Baghdad? So we rely on data obtain from sample from that population.

- **Sample**; A limit number of values drawn from the population (part of population intended to represent the population).
- *Parameters:* A descriptive measure computed from the data of population.
- *Statistics:* A descriptive measure computed from the data of a sample.
- Our aim is to get valid information & generalize our statistic from the sample to the parameter of the population.



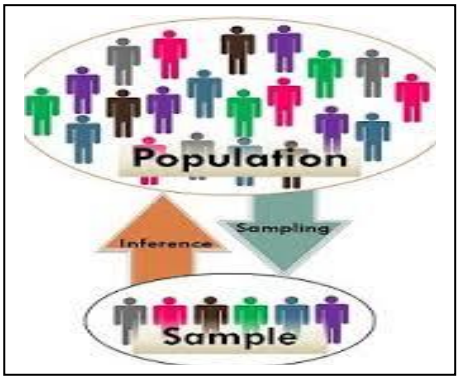

- The available raw information can be called data. Any information in record, descriptive report or symbolic representation of an event, or process may constitute a data point. But these data which are constituted for a set of variables should be arranged in a manner suitable for statistical analysis.
- Sources of data for medical and health studies include:

**1-** Population -level data:

**a)** Vital statistics (birth rates, death rates, pregnancy rates, abortion rates, low birth weight)
**b)** Demographic, economic, housing, geographical, and other data from the Census and other government data-gathering activities
**c)** Summaries of disease and injury reporting systems and registries
**d)** Workplace monitoring systems(e.g.: accident and other work disease and injury)
**e)** Environmental monitoring systems (e.g., air pollution measurements)
**f)** Health programs and campaigns data

**2-** Individual-level data

**a)** Vital events registration (births, deaths, marriages)
**b)** Disease and injury reporting systems and registries
**c)** National surveys
**d)** Medical records
**e)** Questionnaires - in person, by telephone, mailed
**f)** Biological specimens (routinely or specially collected)




- As we can't study the whole population, this information is usually based on a sample from certain population in a certain point of time, and we usually want to use this information to make inductive inferences about the corresponding population.
- Target Population: The whole group of interest
- Study (sampled) Population: The subset of the target population that actually sampled
- Sample: The individuals who were actually measured and comprise the available data.

- **Sampling Methods:**

In order to make a valid inference about population, we need a scientific sample from that population. We have 2 methods of sampling:

***1*- Probability sample**; It's the sample that drawn from a population in such a way that every member of population has the same chance of being included in the sample. The results of this sample are amenable for generalization (valid inference). We have:

a- Simple random sample. b- Systematic sample.
c- Stratified sample. d- Cluster sample.
e- Multistage sample.

2- **Non- Probability sample;** the results of this sample are not amenable for generalization. We have: a- Convenience sample.
b- Quota sample.
c- Snowball sample

- **Simple random sample**

If a sample of size "n" is drawn from a population of size "N" in such a way that every member in the population (N) has the same chance of being selected in this sample (n). The mechanism of drawing is called "Random sampling", can be done by:

1- Lottery method. 2- Computer program. 3- Random number table.

Simple random sample requires: 1- Sample frame (we know all population).
& 2- Sample fraction.

- **Systematic Sample:** Done by:

1- Assign a unique identification number to each member of the population (N).
2- Locate at random starting point.
3- Selection of the sample at regular interval (every 3, every 4 ...etc.).

- **Stratified sample**

Simple random sample & systematic sample can't ensure that the structure of the sample will be similar to the structure of the population regarding certain characteristic. To overcome such problem, we use **stratified sample** by dividing the population into strata regarding certain character (age, sex….etc.), then a random or systematic sampling will be applied to each stratum.

- **Cluster sample.**

The selection of group (cluster) of study instead of individuals. This is usually done in big studies, and the clusters are often geographical unit like villages, districts, schools…..etc. But these clusters contain similar person ((high interclass correlation)), then the generalization of the result may be affected.

- **Multistage sample**

Sampling procedures that carry out in phases (stages) and usually involve more than one sampling method. This is usually done in two or more stages in very large diverse population.

- **Convenience sample**

This sample is representative to the site and time of data collection (e.g., in surgical ward in certain time), but the drawback that the sample is not representative to the total population. Convenience sampling is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher and when probability sample can't be obtained. It is the most common type of sample in medical research. The disadvantages are the risk that the sample might not represent the population as a whole, and it might be biased by volunteers.

**Quota sample**

The composition of the sample as in term of age, sex, social class…etc., is decided and all that require is to find the right number of population to full these quota.

- We measure the **(weight, height, blood pressure…..etc.)** of interest in each member of the sample

## • Variables Types & Classification:

- **Variables:** quantities which vary (take different values in different person, place, &/or time e.g. WT, HT, BP..etc. These variables usually collected from individuals and need statistical analysis( Statistics deals with variables ) to draw conclusion regarding those individuals.
- **Random variable;** is the variable that arise as a result of chance factors, so can't be exactly predicted in advance. e.g. HT, WT, when a child born, we can't predict exactly his/her HT or WT at maturity. Most of human variables are random variables.
- ***Variables*** are subdivided into;

1- Quantitative (Numerical); can be measured by the usual sense e.g. age, WT, HT….etc. This can be either:

a- ***Discrete***, taking some value in discontinues set of value (charect. by a gap or interruption ,have no fraction) e.g. No. of teeth, No. of admissions, Number of children, Number of attacks of asthma per week….etc..

b- ***Continuous***, taking some value in an infinity divisible range of value (doesn't posses a gap or interruption, have a fraction, we can find another value somewhere in between) e.g. WT, HT, Serum cholesterol, Blood sugar….etc.

2- Qualitative (Categorical); can't be measured by the usual sense but must describe in category. This can be either:

a- Nominal, defined by un ordered categories. E.g., color of the eye, blood group, sex, marital state….etc. (Nominal variable with only two probabilities is called **"dichotomous variable"** e.g., life or death, sick or not ….etc.

b- Ordinal, defined by ordered categories. E.g., educational state categorized into primary, secondary and higher education, Cancer staging I, II, III……etc.

- **Ex: Type of variables:** For each of the following, identify the type of variable.

a- **Gender**: (Qualitative, nominal, dichotomous).

b- **Serum bilirubin**: (Quantitative, continuous).

c- **Severity of hemophilia, mild- moderate-sever**: (Qualitative, ordinal).

d- **Height in cm**: (Quantitative, continuous).
e- **Number of X-ray films taken in week**: (Quantitative, discrete).
f- **Ethnic group**: (Qualitative, nominal).
g- **Age as categorized as young-middle age- old**: (Qualitative, ordinal).
h- **Age in years**: (Quantitative, continuous).

## ➢ Classification of variables

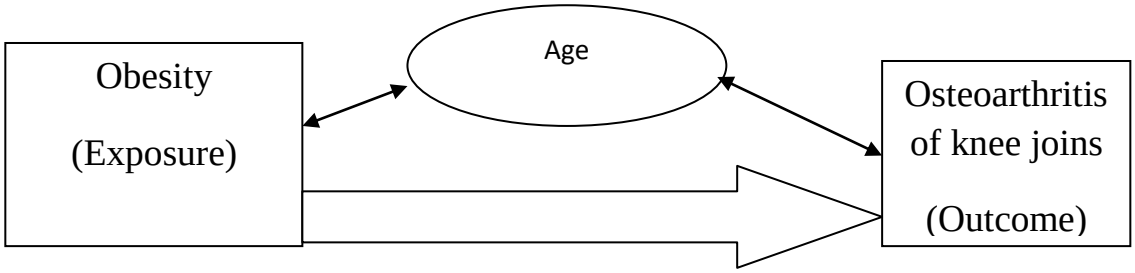
In preventive medicine and in research work we should classify the variables in order to find the relationship. Classification system for variables and how the classification influences the presentation of results are:

**1-** Explanatory variables (Independent variables, Risk factors, Exposure variables, Predictors, first variable): These variables explain, influence, or determine the outcome. Manipulating of these variables will change the outcome.

**2-** Outcome variables (Dependent variables, endpoint variable, Response variables, second variable): The variables that observed as a result of the independent variable being manipulated. These are the counterpart of the explanatory variables (variables that are explained/influenced /determined by explanatory variables).

**3-** Confounding variables **(Third variables)**: The variables that are associated with the exposure variables of interest and have an effect to the outcome variables and as a result, they influence the relationship between the exposure and outcome. Or, alternatively it may mask a real association between them. These variables are called confounders and the ability to adjust or analytically control for the presence of a confounder depends on how well this variable is measured.

**Example1:** in assessing whether obesity can lead to osteoarthritis of knee joins, age is a confounder. Age is associated with increasing of individual weight and it is also a risk factor for osteoarthritis of knee joins. If we don't control age we will be unable to know whether osteoarthritis is caused by obesity or by age. A confounder is a variable other than the one being investigated which is associated with both the exposure and the outcome and can cause bias in a study.




**Example 2:** A study found that People who drink a lot of coffee (exposure) are at risk of Ca lung (outcome). But this spurious relationship is obtained because of confounder. Smoking is usual associated with coffee drinking, people who drik a lot coffee tend to smoke a lot, and smoking (confounder) is the real cause of Ca lung.

- **Quantifying and Measuring Health Events & Diseases**

  - Measuring events {disease events or health events} is at the heart of the preventive medicine. If one can't quantify, he can't do risk assessment, so we can't intervene.
  - Measurements of disease events or health events give us information about the health situation in the community. This type of information is often used in planning of health services, besides, studying the nature of the disease in order to apply effective management and control as well as Evaluation

## What do we measure?

## I- Counting Frequency, Relative frequency and Proportion:

- This can be simply done by counting the number of event. If we classify the event into categories, frequency is the number of individuals falling into each category.
- While proportion is the fraction of values in each category, which is determined by dividing the frequency by total number of sample and usually multiply the result by100 to express the result as **relative frequency or percentage**.
- Example: We have a sample of 112 subjects and 32 are found to be diseased. The number 32 is the frequency of the disease and the number 80 is the frequency of non-disease in our sample. The percentage (relative frequency) of diseased in this sample is (32 ÷ 112)/100 = 28.57% and by the same way the percentage (relative frequency) of non disease is 71.43%.
- Cumulative frequency distribution (CF) and cumulative relative frequency distribution (CRF) are used to facilitate obtaining information when we have several categories and we try to find the frequency and percentage in more than one category.

In the above example, if we classify the sample according to their age and we use five categories, the data will be as follow:

| Age group (years) | Frequency | Relative frequency | CF | CRF |
|---|---|---|---|---|
| < 10 | 18 | 16.07 | 18 | 16.07 |
| 10-20 | 28 | 25.00 | 46 | 41.07 |
| 21-30 | 32 | 28.57 | 78 | 69.64 |
| 31-40 | 22 | 19.64 | 100 | 89.29 |
| >40 | 12 | 10.71 | 112 | 100.00 |
| Total | 112 | 100 | | |

## II- Counting Rate and Ratio:

- It is necessary to convert raw numbers (frequencies) into rates. Frequencies are useful in many ways, but there is a great danger of misleading by just going to compare with frequencies. For valid comparisons we should use the rates.
- **A rate** is a numerical expression of the frequency of a condition in a given population measured in a specified period of time.
- A rate is a fraction—the upper part (the numerator) is the number of people affected by the problem, event, or condition of interest in certain time; the lower part (the

denominator) is the number of persons in the population who are at risk of experiencing the problem, event, or condition in the same time.

- **A ratio** is defined as the relative size of two quantities expressed by dividing one (numerator) by the other (denominator). So, Proportions and rates are ratios when the numerator is part of the denominator. But we usually imply ratio when we describe a two quantities and the numerator not belong to the denominator. For example if we have 100 students and 30 of them get hepatitis A virus infection during one month, the proportion of infection is 30/100 and equal to 30%. The rate of infection is 300 per 1000 students per month, and the ratio of disease/non disease in this sample is 300:700 or 1 to 2.33 students.
- Three types of rates, which are commonly mentioned: crude, specific, and adjusted (or standardized).

1) **Crude rates** are computed for the entire group or population; they disregard factors such as age, gender, and race. For the above example the crude rate for hepatitis A infection is 300 per 1000 students per month. But, crude rate d doesn't take into account differential distribution of underlying factors (i.e. sex) that can influence rates of disease within populations.
2) **Category specific rate**: by dividing the number of people who get the disease or event in a particular category (age, sex, race….) by the total population of the same category (age, sex, race….) during the same period. Specific rates consider these differences among subgroups or categories of diseases. We can calculate specific rate for males or specific rate for females. In the previous example suppose our sample of students contains 500 males and 500 female and the number of infected persons was 200 among male and 100 among females.

**NOTE**: The specific infection rate is 200/500 and equal to 0.4 or 4 per 10 male per month. The quotient is often multiplied by any suitable constant of 10n where n = 2,3,4,5,6. This constant is used to transform the result of the division into a uniform quantity in a power of 10 (100, 1000…etc) and to get rid of fraction. The size of this constant may equal to 100 1000, 10 000 and so on depending upon the value of n. In our previous example and in the same way, the specific infection rate among females is 100/500 = 0.2 or 2 per 10 females per month.

In big studies with large sample size and the observation lasts for year or years, **total midyear population** is used as denominator and the annual crude event rate is defined as the number of event in a calendar year divided by total midyear population.

3) **Adjusted or standardized rates** are used to make valid summary comparisons between two or more populations possessing different category (age, sex, race….) distributions. It is used to answer the question whether the two different populations have had a similar experience (i.e. death rate) if both populations had the same distribution of the underlying factor (age, sex, race….)? The process of age or sex-adjustment gives both populations the same distribution of the underlying factor thus removing or nullifying its effect on the summary measure. Adjustment is nothing more than obtaining a weighted average for each category. The weighting is derived from a standard population. The frequency in each category is multiply by its

corresponding number of standard population to specify the relative weight of the contribution of each category stratum to the summary ('category-standardized') rate

- **Measures of Morbidity (Disease):**

The best example in using rate and ratio in medical data & research is measuring of morbidities as well as measuring of mortalities. Morbidity measures are used to refer to the extent of disease or disease frequency within a defined population. Two important measures of morbidity are prevalence and incidence.

1) **Prevalence** refers to the number of cases of a given disease or health event that exists in a population at a specified time. It can be clarified as the probability of an individual from a population having a disease at a specified point in time and calculated as:

**Prevalence** = No. of all cases / Total population at that specified time X K ($10^n$)

We have two type of prevalence:

a) Point prevalence: How many people in a given population have the disease at this point in time (a snapshot)? It equals the number of disease cases in a population at a single point in time

b) Period prevalence: How many people in a given population ever had the disease during a given period of time (weeks, months, years….)? It equals the point prevalence at the beginning of a study period in addition to the number of new cases that occurred during the remainder of the study period.

**Example**: in a vaccine survey of children below five years at a primary health care center, if the number of full vaccinated children is 800 out of total 1000 children, then the point prevalence of full vaccination among children less than five is 80% at that point of time. If we interview 300 children (their mother or father) and 180 of them reported completeness of their vaccination during the 2 months before the interview, then the period prevalence of completing full vaccine in this population over the last 2 months is calculated as: 180/300 × 100 = 60.0%.

Prevalence should be distinguished from incidence (other measurement of morbidity)

2) **Incidence** of an event can be only calculated in prospective studies when new cases can be measured. An incidence rate (sometimes referred to simply as incidence) is a measure of the frequency that a new case of illness or health problem, occurs in a susceptible population over a specified period of time. The formula for calculating an incidence rate follows:

**Incidence rate** = New cases occurring during a given time period / population at risk during the same time period X $10^n$

**Ex:** A population at risk is composed of 100 obese, above 40 years males (all are free DM). Twenty-five men develop diabetes and are confirmed by laboratory testing during one year. The incidence of DM among that sample is 25% per year.

☒ For better understanding, suppose one was interested in finding out how many

people living in a district had Hypertension. If 100 out of 1000 indwellers examined were positive for HT, what will this proportion (10%) be called incidence or prevalence? It is obviously a prevalence measure because the 10% figure was arrived at by examining people at only one point in time (a cross sectional estimate). We have no idea as to when exactly this 10% actually became HT.

- ☒ But if one wished to know how many people in this district newly develop HT in a certain period of time. Let us say all people were examined at the start of the calendar year (Jan) and 10% of 1000 indwellers are found to be HT. This means that 900 people are non-HT or healthy at the start of the year. Let us this 900 population is again screened for HT at the start of the next year (after 1 year) and 9 people found positive. This 10% *(9/900)* is the one year incidence of HT in this population.
- ☒ Incidence has a longitudinal (follow-up) component in it (as compared to prevalence which a cross sectional component in it). The other difference is that the denominator in incidence contained only population at risk (people liable to have the disease), while in prevalence, the denominator is the total population.
- ☒ In general, incidence is usually used to quantify events with short duration e.g., acute illness, while prevalence is used to quantify chronic illness such as osteoarthritis which have long duration and dates of onset are difficult to pinpoint.
- ☒ Regarding the relation between incidence and prevalence, prevalence is based on both incidence (risk) and duration of disease (prevalence =incidence X disease duration). High prevalence of a disease within a population may reflect high risk, or it may reflect prolonged survival without cure. Conversely, low prevalence may indicate low incidence, a rapidly fatal process, or rapid recovery. We often use prevalence rather than incidence to assess the burden of the disease in a community or planning for health services.

## • Measures of Mortality

- If we wish to address the risk of dying in a population, we must deal with mortality measures. Mortality risk (or rate) is an example of incidence where death is the outcome of interest.
- Crude mortality rate is the annual death rate from all causes, and the crude mortality rate per 1000 population is calculated as: The total number of death from all causes in one year/ Number of persons in the population at midyear.
- We may not always be interested in a rate for the entire population; perhaps we are interested only in a certain age group, in one sex, or in one socioeconomic group. Thus we calculate a category specific rate: The number of death in certain category / Number of persons in that category
- We could also place further restriction by specifying the cause of death (e.g., lung cancer, tuberculosis deaths) and calculate the cause specific death rate as Number of deaths due a specific disease/ Total midyear population. Cause-specific mortality risk is the incidence risk of fatal cases of a particular disease in the population at risk of death from that disease.
- Proportionate mortality ratio which is deaths due to a particular case in relation to deaths from all causes can be also calculated as: Total number of deaths due to a certain disease / Total number of deaths from all causes X 100%

- To understand the above calculations, suppose we have population with below data in year X:

| Age group in year | Total midyear population | No. of deaths |
|---|---|---|
| **<45** | 400000 | 50 |
| **45-55** | 300000 | 100 |
| **56-65** | 200000 | 700 |
| **≥66** | 100000 | 1000 |
| **TOTAL** | 1000000 | 1850 |

  - ✓ The annual crude mortality rate year X per 1000 persons was: 1850/1000000 X 1000 = 1.85 per 1000 population per year X.
  - ✓ The age specific mortality rate (≥66) in year X per 1000 persons was: 10000/1000000X1000=1 per 1000 population per year X.
  - ✓ In the same way, we can find sex, disease specific…..etc. mortality rate
  - ✓ If the causes of those 1850 deaths were 500 pneumonia, 200 leukemia, 200 acute myocardial infarction 300 stroke, , 100 Hodgkin's disease, , and 550 others).
  - ✓ The cause specific mortality rate for acute MI in this population was: 200/1000000X1000=0.2 per 1000 or better 2 per 100000 population per year X.
  - ✓ In the same way, we can find any cause specific rate…..etc. pneumonia, stork...etc.
  - ✓ The proportionate mortality ratio for pneumonia was 500/1850 X 100= 27.03%

- We should distinguish between cause specific mortality rate and a case fatality rate (percent of people diagnosed as having the disease die within a certain lime after diagnosis). **A case fatality rate** (percent) is calculated as follow: The number of individuals dying during a specified period of time after disease diagnosis divided by the number of individuals with that disease X 100.
- In the above example, the cause specific mortality rate for pneumonia in this population was: 500/1000000X1000=0.5 per 1000 or better 5 per 100000 population per year X. While to calculate case fatality rate, we should identify the number of population diagnosed to have pneumonia in that year. Suppose we had 3000 cases of pneumonia, then, the case fatality rate was 500/3000 X 100 = 16.67%. The "survival rate" from pneumonia in that year was 83.33 % (100%-case fatality rate)
- An important use of mortality data is to compare two or more population, or one population in different time period. Such population may differ in regard to many characteristics that affect mortality, of which age distribution is the most important. The mortality rate may more in one country than other as more elderly person present in this country. Therefore, adjustment" or "standardization" methods (not the crude rate, as in morbidity) should be used for comparing mortality in such population as this method effectively holds constant such characteristic as age. .
- The following are common mortality rates used in health statistics:

$$\text{Infant mortality rate} = \frac{\text{Number of infant (< 1 year) deaths}}{\quad}$$

$$\text{Neonate mortality rate} = \frac{\text{Number of neonate } (\leq 28 \text{ days}) \text{ deaths}}{\text{Number of live births}}$$

$$\text{Perinatal mortality rate} = \frac{\text{Number of stillbirths + deaths in 1st week of life}}{\text{Total births}}$$

$$\text{Stillbirth rate} = \frac{\text{Number of intrauterine deaths after 28 weeks}}{\text{Total births}}$$

$$\text{Maternal mortality rate}^{**} \text{ (usually per 100,000)} = \frac{\text{Number of deaths among women due to after 28 week}}{\text{Number of live births}}$$

NB These rates are usually related to one year

** it is actually ratio not rate, but it is internationally accepted and used as rate

- **<u>Measures of location (Central Tendency)</u>**

  - One of the most useful aspects of statistics in medical research is its ability to describe large data sets using only a few numbers (summarization of data).
  - A descriptive measure is a single number that is used to describe a set of data. Numerical data in ratio and interval scales (quantitative, not qualitative) variables can be described by measuring their location (central tendency) and variability (dispersion or spreading). This can be done by calculating mean, median & mode:

    **1) Arithmetic mean** (or **mean):** The most commonly used measure of central tendency.
    - ✓ It is the "average" which is obtained by adding all the values in a sample or population and dividing them by the number of values.
    - ✓ In medical statistics we have population mean (parameters for population) and calculated as $\mu = \Sigma x / N$, where: $\mu$ =population mean, $\Sigma$=sum, and N = No of values in population. And we have sample mean (statistics for sample): and calculated as $X = \Sigma x / n$, where: X= sample mean, n = No of values in the sample.
    - ✓ The mean is the most widely used measure for central tendency because of the following advantages:
      1) Simple to calculate and to interpret.
      2) Unique, single and only single value.
      3) Comprehensive, take all the values in consideration
      4) Most of the important test of statistical significant are based on the mean
    - ✓ But, the mean is affected by extreme values. Large and small values can disrupt the usefulness of this measurement. Suppose four persons receive the following

charges, $95, $95, $90, and $1280. The charge for the five physicians is found to be $390, a value that is not very representative to their charges.

2) **Median:** The value that divides the set of data into two equal parts. The number of values equal to or greater than the median equals the number of values less than or equal to the median. It is when the data have extreme values and the mean is considered not suitable measure for central tendency.
   - ✓ Less familiar but also helpful measurements of central tendency when the data is not homogenous because it is not affected by extreme values
   - ✓ Finding the median depends on whether there are an odd or even numbers of values in the list of data. To find the site of the median, first, we must arrange the value in ordered array then, if the No of observations is odd, the site of median is = (n + 1)/2. But, if the No of observations is even, the median is average of (n/2) and (n / 2) + 1.
   - ✓ The most essential advantage of the median, besides simplicity and uniqueness, is that it is not affected by extreme values. So, when we have outliers in our data we shift to the median not the mean.

3) **Mode**: It is the most frequent value in a set of data. It is the only measure that can be used in both qualitative and quantitative variables. In a set of data, it can be no mode, one mode (uninmodal), two modes (bimodal) or three modes (trimodal)…. etc. So the mode is not unique and the least commonly used measure of central tendency.

**Example:** What is the average of serum cholesterol in diabetic patients? If our sample size is 10 and our measurements of serum cholesterol for those 10 patients revealed the following results: 250, 200, 210, 280,360, 220, 250, 190, 240 and 230mg/dl. The sum of these 10 values is 2430, so the mean is 2430/10=243 mg/dl.

- ☞ If we were to exchange the last patient value (230mg/dl) with more extreme value (e.g. 640 mg/dl), the mean would increase from 243 to 284.
- ☞ Unlike the mean, the median is not affected by the extreme values. And to compute the median, we should first put all values in numerical order, and then locate the score in the center of the sample. We order the 10 values above, we would get190, 200, 210, 220, 240, 250, 250, 280, 360, and 640. There are 10 values (even number) and the values in order 5 and 6 represent the halfway point of all values. Since sum of these scores are 490, the median is 490/2= 245.
- ☞ If we have 9 scores, the position of the median (center of the sample) would be order 5 only (240).
- ☞ To determine the mode, The most frequently occurring value is the mode. In our example, the value 250 occurs twice and is the mode. If other value (e.g. 230) also present twice, then there are two values that occur most frequently and the distribution of our data has two modal values (bimodal distribution)

- **<u>Measures of Variation (Dispersion)</u>**

Knowing the central tendency measures only is not enough to describe a total picture of a distribution. We want to know the variation (dispersion) of the values around the central tendency. The common measures of dispersion are:

1) **The range**: The range is simply the highest value minus the lowest value. In our example distribution, the highest (maximum) value is 640 and the lowest (minimum) value is 190, so the range is 640 - 190 = 450.
2) **The standard deviation (SD):** It quantifies the amount of variability or spread about the mean. The SD is usually used with the mean in descriptive statistics to describe the distribution of any set of data, in an interval or ratio scale, adequately. Let's take the set of 250, 200, 210, 280,360, 220, 250, 190, 240 and 230 values.
   - ☞To compute the SD, we first find the distance between each value and the mean. We know from above that the mean is 243. So, the values that are below the mean have negative discrepancies and values above it have positive discrepancies.

| $x$ | 250 | 200 | 210 | 280 | 360 | 220 | 250 | 190 | 240 | 253 |
|---|---|---|---|---|---|---|---|---|---|---|
| $(x-\bar{x})$ | 7 | -43 | -33 | 37 | 117 | -23 | 7 | -53 | -3 | -13 |

   - ☞We find the sum of discrepancies is zero. So, we square each discrepancy to overcome this point.

| $x$ | 250 | 200 | 210 | 280 | 360 | 220 | 250 | 190 | 240 | 253 |
|---|---|---|---|---|---|---|---|---|---|---|
| $(x-\bar{x})$ | 7 | -43 | -33 | 37 | 117 | -23 | 7 | -53 | -3 | -13 |
| $(x-\bar{x})^2$ | 49 | 1849 | 1089 | 1369 | 1368 | 5269 | 49 | 2809 | 9 | 169 |

   - ☞Now, take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 21610. Next, we divide this sum by the number of values minus I. Here, the result is 21610/9 = 2401.11. This value is known as the variance and it is in square unit ($mg/dl^2$).
   - ☞To get the SD with no square unit, we take the square root of the variance, and this would be $\sqrt{2401.11} = \pm$ **49mg/dl**.
   - ☞The mean tells as the single best point for summarization of the entire sample and the SD tells us how much is, on average, other scores deviate from the mean. Hence, the SD is the average of the deviations from the mean, and the smaller the SD is the more homogenous is the sample. For example, if we have another sample with a mean of 243 and SD of 20, then we would immediately know that the second sample is more homogenous.

3) **The Percentiles and quartiles**: Percentiles and quartiles provide another way of looking at variations in distributions.
   - ✓ Just as the median is the 50th percentile or second quartile of a collection of data, the 25th and 75th percentile are the first and third quartiles respectively.
   - ✓ The interquartile range (IQR) is the middle half of the values. i.e. those lying between the first and third quartiles(Q3- Q1). It is the distance between the scores representing the 25th and 75th percentile ranks in a distribution and in our example it is 280-200=80mg/dl.
   - ✓ Not only 25, 50 and $75^{th}$ percentiles can be calculated, the $30^{th}$, $60^{th}$, $90^{th}$ or any $n^{th}$ percentile can be determined and indicates that a particular measurement in a set of data is larger or smaller than these values.

# • Organizing and Displaying Data: Tables and Graphs

- Descriptive statistics serve as device for organization, summarization of data.
- Tables and graphs are the most common method used to depict the variable
- The purpose of tables and graphs is to present information in a concise way so that readers can understand and remember it more easily.

**1) Tabular (Tables):** Table consists of row(S) & column(S), could be 2x2, 2x3….etc.

*Table must be:*

a- As simple as possible (it is better to have 2-3 simple tables than one complicated).

b- Understandable & self explanatory without references to the text. This is done by:

- The title should be clear (placed above the table), and answer the questions of: What? Where? And When?
- Each row and column should be labeled clearly and concisely.
- Specific unit of the measure for the data should be defined.
- Total should be placed.
- Illustrate symbols, code, and abbreviation by putting a footnote below the table.

c- Source of the table (if not original).

d- Avoid too much over ruling.

**Ex:** Table 1: Legitimate total birth (what), England (where), 1969-1970 (where).

| Parity | Mother's age (years) | | Total |
|---|---|---|---|
| | <30 | ≥30 | |
| 0 | 514,108 | 49,895 | 564,003 |
| 1-3 | 583,889 | 234,084 | 817,973 |
| ≥4 | 22,216 | 64,894 | 87,110 |
| Total **(all parity)** | 1,120,213 | 348,873 | 1,469,089 |

**Source.** Osborn J.F., (1975). J.R statistic. 24, 75-84.

**2) Diagrammatic "Graphical","Figuer".** The form of the diagram varies according to the nature of the data:

***The graph should be;***

1- As simple as possible. 2- Self explanatory.

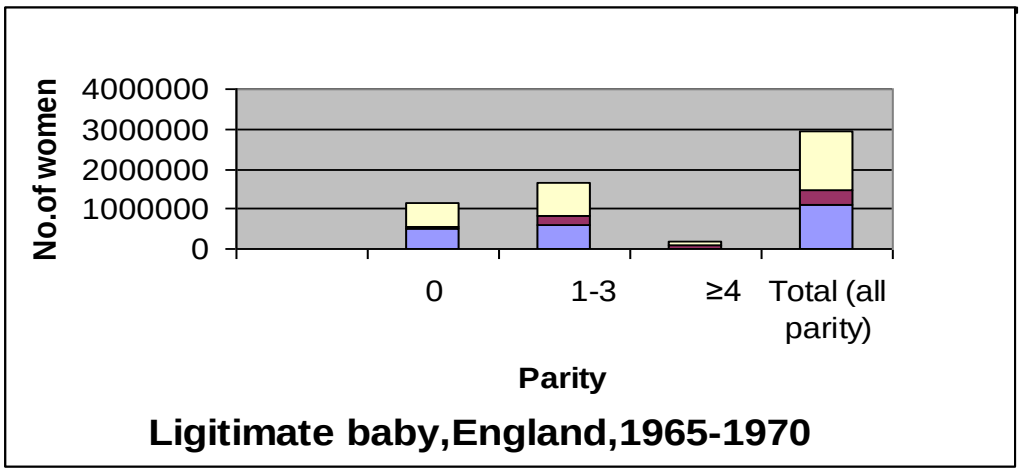
3- The title should answer the questions of what, where and when?

4- Key, if there is more than one line. 5- Source, if it's not original.

6- Scale and unit must be placed. 7- Separate the title from the graph.

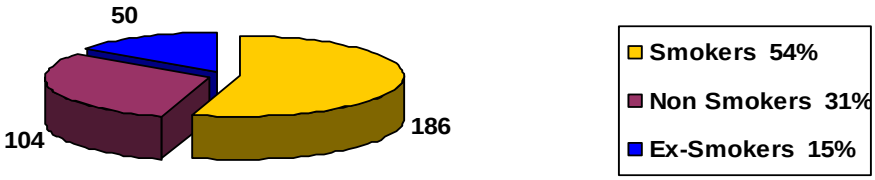
a- For ***categorical data,*** we have "chart", this can be:

- Bar chart; This is a graphical representation of the (relative) frequencies or magnitudes by rectangles of constant width drawn with length proportional to the (relative) frequencies or magnitudes concerned, as the EX. Below:




- Pie chart; This is a graphical representation of the (relative) frequencies or magnitudes by a circle whose area represent the total frequency and which is divided into segments which represent the proportional composition of total frequency, as the EX. Below:

A pie chart represents division of a total quantity into component parts. The total quantity (100%) represents an entire circle. Each wedge represents a proportionate component part of the total.




**Bar chart shows the distribution of the study sample regarding their smoking status**

- Picto-chart ((Picto-gram)). This is a graphical representation of the (relative) frequencies by using symbols (drawing or picture) relevant to the subject matter. Symbols of different size should not be used. A unit value of the data should be represented by standard symbol which may repeat to represent magnitude. As in the Ex.

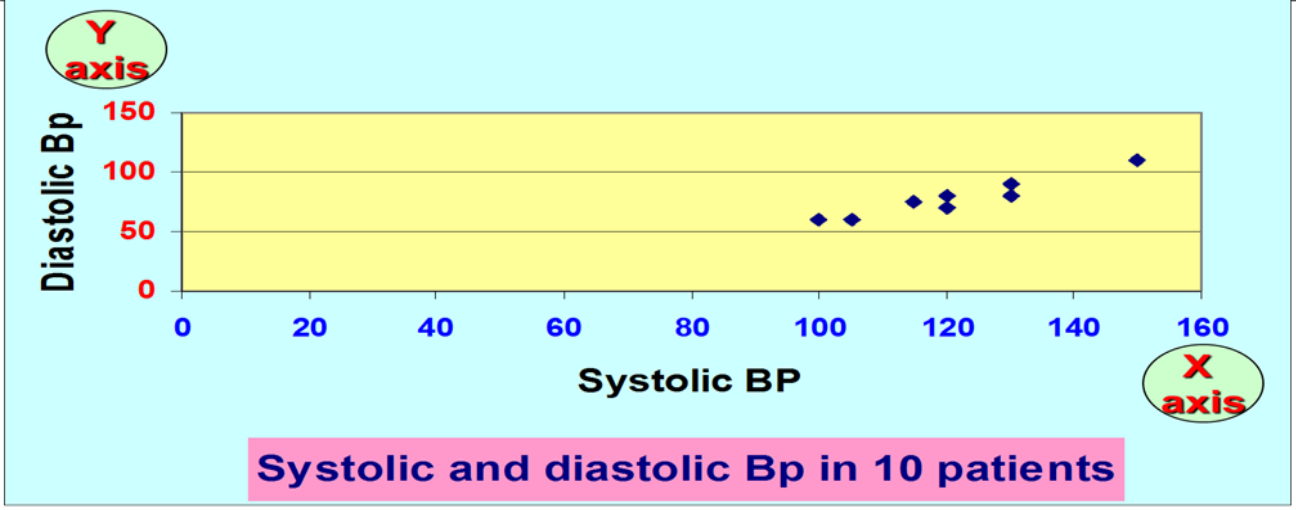
| | | |
|---|---|---|
| Parity | 0 | ♀♀♀♀♀ |
| | 1-3 | ♀♀♀♀♀♀♀ |
| | ≥4 | ♀ |

Fig. Pictogram, No. of mothers (all ages) in 100, 000.

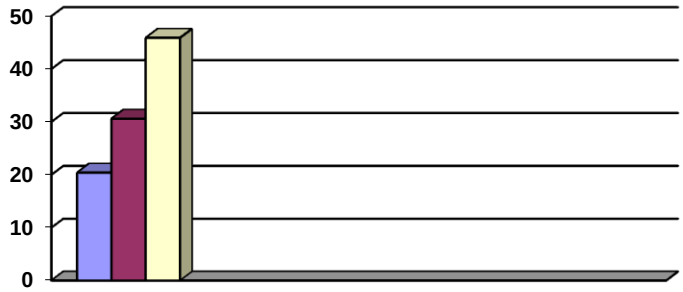
b-***For numerical data***

- Scattered diagram (dot-graph). Each observation is marked as a dote corresponding to its value on each axis (X & Y). The pattern made by the dotes is an indication of relation between the two axes, which may be linear if the follow a straight line or curved if not.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 120 | 115 | 105 | 100 | 130 | 150 | 130 | 120 | 120 | 105 |
| 80 | 75 | 60 | 60 | 90 | 110 | 80 | 70 | 80 | 60 |

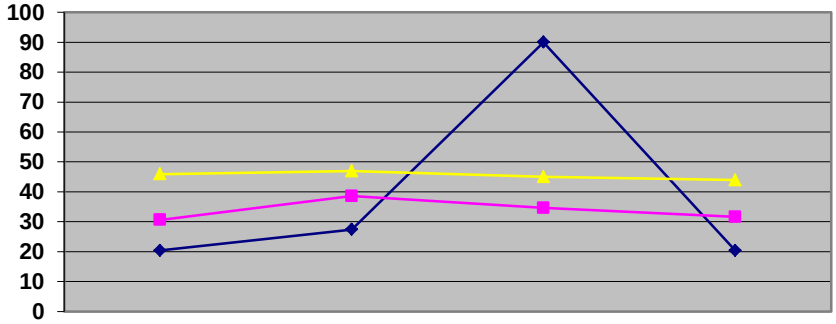



- Histogram: This is a graphical representation of frequency distribution in which rectangle proportional in the area to the frequencies are erected on the horizontal axis. The base lines are continuous (because we are dealing with continues variables). The width of the rectangles should be equal. As in the Ex.




*Disadvantage: can't present more than one set of data.

- Frequency polygon: If we join the tops of the rectangles in the histogram →Polygon (total area of histogram= total area of polygon). It is only appropriate when the variables on the horizontal axis are continues and there is only single value on the vertical axis. As in the Ex. Advantage: can present more than one set of data.



- Stem & Leaf plots

  - It is the combination of a graphic technique and a sorting technique of the data
  - The raw data can still be obtained from the graph for further analysis
  - The stem is the leading digit(s) of the data
  - The leaf is the trailing digit(s) of the d**ata**

Ex: A pediatric registrar is investigating the amount of lead in the urine of children from a nearby housing estate. In a particular street there are 15 children whose ages range from 1 year to under 16, and in a preliminary study the registrar has found the following amounts of urinary lead (nmol/24 hrs): 0.6, 2.6, 0.1, 1.1, 0.4, 2.0, 0.8, 1.3, 1.2, 1.5, 3.2, 1.7, 1.9, 1.9, 2.2

| Stem | Leaf |
|---|---|
| 0 | 1,4,6,8 |
| 1 | 1,2,3,5,7,9,9 |
| 2 | 0,2,6 |
| 3 | 2 |

**Stem and leaf "as they come"**

| Stem | Leaf |
|---|---|
| 0 | 6,1,4, 8 |
| 1 | 1, 3, 2, 5, 7, 9, 9 |
| 2 | 6, 0, 2 |
| 3 | 2 |

**Ordered stem and leaf plot**

## • Recognition the Distribution of variables

- One of the most important things to know about a variable is its distribution. Knowledge of the probability distribution of the variables provides the clinicians and researchers with a powerful tool for summarization and describing a set of data and for reaching conclusion about population on the basis of a sample drawn from that population.
- We have several types of distribution in statistics:

I- For discrete variables we have

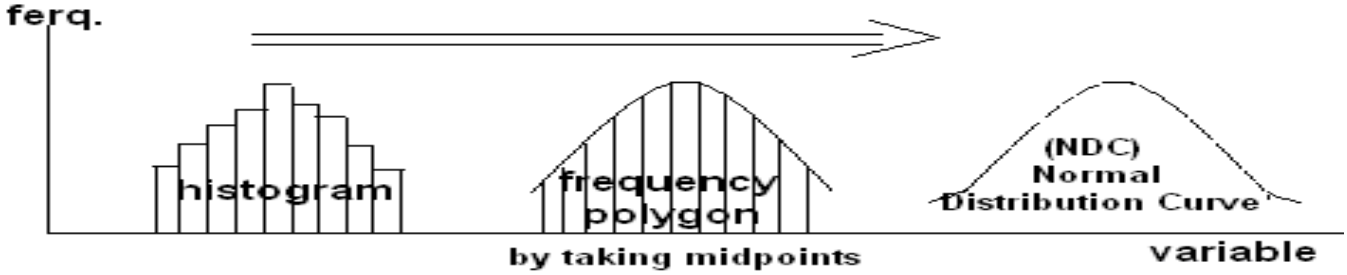
a- Binomial Distribution: dichotomous outcomes (disease, not disease, A-B, heads-tails, yes-no, on-off, right-wrong, etc.)

b- Poisson Distribution Useful for studying rare random events.

II- But the “Normal distribution" "Gaussian distribution", for continues variables is the most important one. This is because:

- Most human variables (age, weight, height, blood pressure, blood sugar…) naturally have a “bell shaped” distribution.
- The distributions are tied to probabilities, and it is the probability which will be of interest to us

**Ex:** If we take large number population and put their age distribution in a histogram, then in a frequency polygon. We get a bell shaped curve represent the distribution.



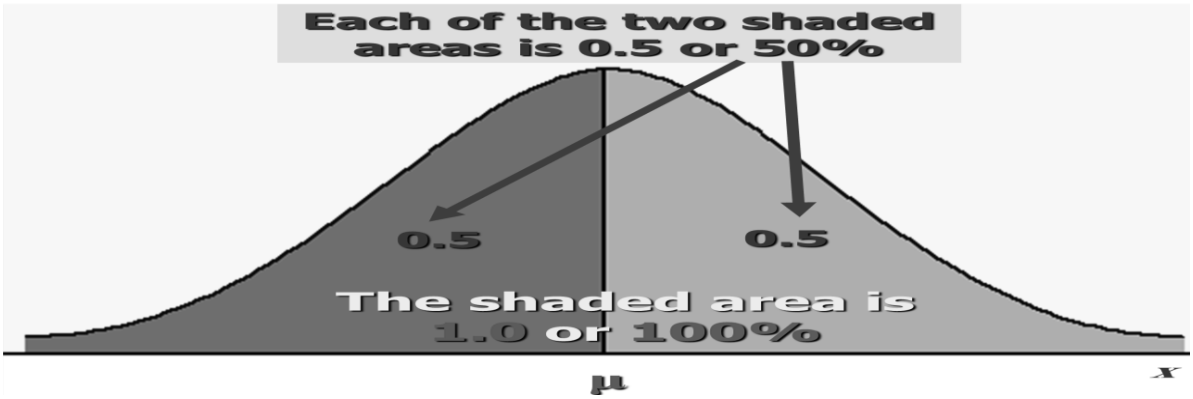
**The normal distribution** "Gaussian distribution", "Bell Shaped distribution" is the most important distribution in the statistics, **the parameters of this distributions are**:

1- The mean (μ) → Measure of location.
2- The standard deviation (∂) →Measure of dispersion.

Population parameters

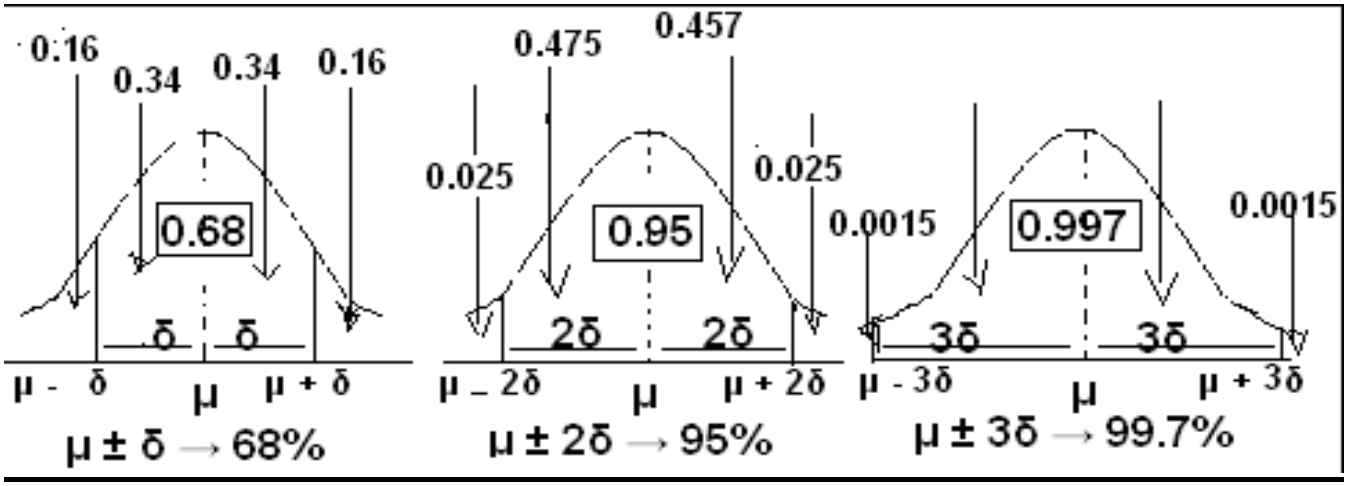
## Characteristic of the Normal Distribution

1- Used for the continuous variables, between 0 and ∞
2- Symmetrical about its mean (μ), ((either side of mean is a mirror image of other side.
3- Mean, median, and mode are equal.
4- The total area under the curve is equal to one, 50% on the left &50% on the right of a perpendicular erected at the mean.



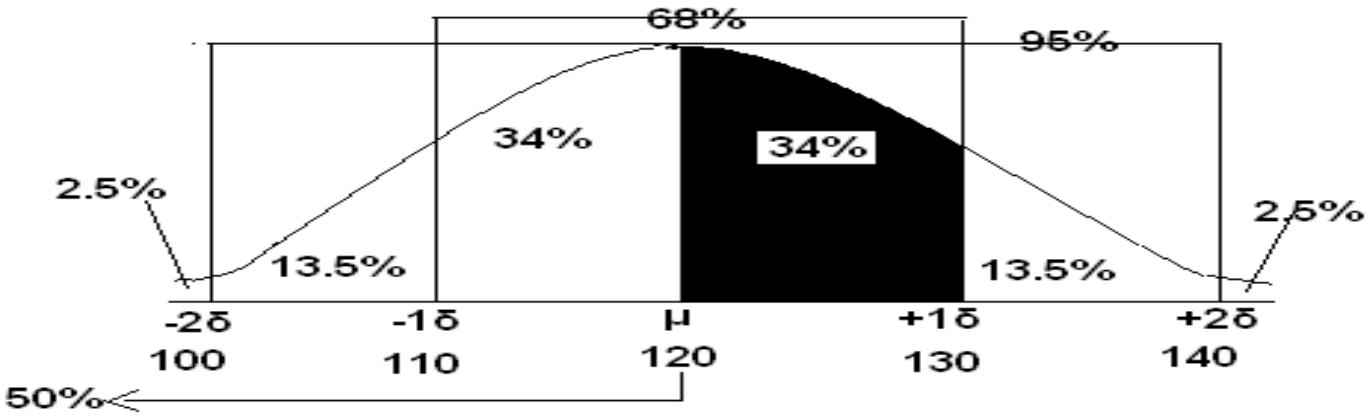

5- The normal distribution is completely determined by the parameters (μ) & (∂). Different values of μ shift the graph along the X-axis, while different values of ∂ shift the graph along the Y-axis (determine the degree of flatness or peakness of the graph).
6- μ ± 1∂ → 68% of the area.
μ ± 2∂ → 95% of the area.
μ ± 3∂ → 99.7% of the area.

- Since we know the shape of the curve, we can (using calculus) calculate the area under the curve




**Ex:** If population mean of systolic blood pressure is 120 mmHg with population standard deviation of 10 mmHg. What is the probability of getting a patient with systolic BP:

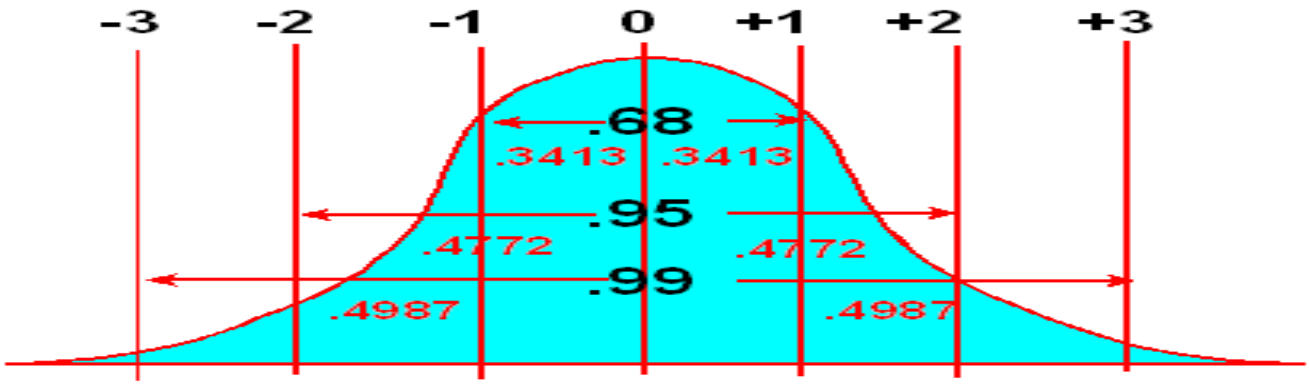
a) between 120 &130 mmHg, b) < 120mmHg, c) < 100 mmHg d) between 120 & 125mmHg?




**Answers:**

a) From 120 to 130 we move one δ, so the probability is 34% (0.34) (i.e. half of 68%).
b) Probability of less than 120 mmHg is 50%.
c) Probability of less than 100 mmHg is 2.5%.
d) To calculate the Probability of SBP between 120 and 125 mmHg or any other probability, we must follow Z scale or standard normal distribution.

- **The Standard Normal Distribution** "Z-distribution".
  - By standardizing normally distributed scores, one could better understand and compare score. Standardizing is a process through which scores are transformed into a common scale (z-scores).
  - Z distribution or Z score is the normal distribution curve which has a mean of zero and a standard deviation of one **(μ=0, & ∂ =1).** → $\boxed{Z = x - \mu / \partial}$
  - If we know the population mean (**μ**) and population standard deviation (**∂**), for any value of X we can compute a z-score by subtracting the population mean and dividing the result by the population standard deviation

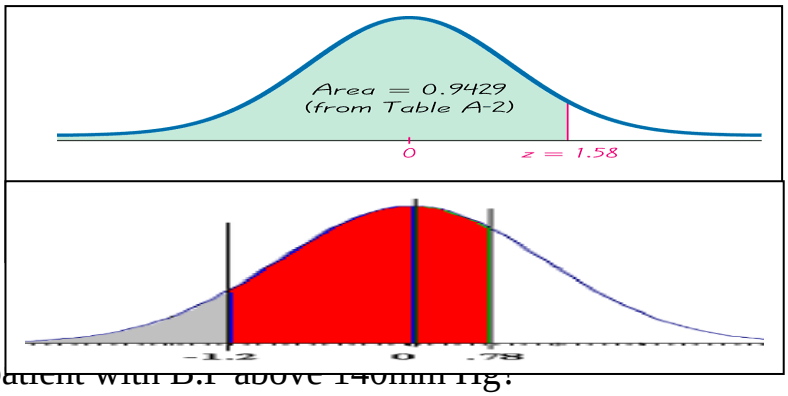



**Properties of Z Distribution (Z-score)**

- 90% of the values of a normal variable lie within ± 1.65 sample standard deviations from the sample mean

- 95% of the values of a normal variable lie within ± 1.96 sample standard deviations from the sample mean
- 99% of the values of a normal variable lie within ± 2.58 sample standard deviations from the sample mean
- From Z table, we can find any probability at Z scale.

**EX:** From Z table: Find P ($z \geq 1.58$)
P($z < 1.58$) = 0.9429
P($z > 1.58$) = 1 - 0.9429 = 0.0571




**Ex:** Calculate p(-1.2<Z<0.78)
P (-1.2<Z<0.78) = 0.7823-0.1151
=0.6672

**Ex:** What is the probability of having a patient with B.P above 140mm Hg?
Z = x - μ/ ∂
= 140-120/10 = +2 P(x ≥ 140) → P ( Z ≥ +2). &From the Z-table, P=0.023.

**Ex:** If the total cholesterol values for a certain target population are approximately normally distributed with a mean of 200 (mg/100 mL) and a standard deviation of 20 (mg/100 mL), what is the probability that a person picked at random from this population will have a cholesterol value greater than 240 (mg/100 mL)?
Z = x - μ/ ∂ = 240-200/ 20 = 2 P (x >240) → P(Z > 2). = 0:0228 or 2.28%

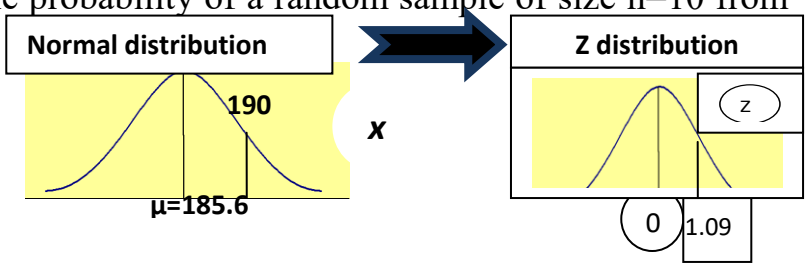
- **Z-value for Sampling Distribution of the Mean**

  - In medical research and calculation we cannot measure everyone in the population that does not mean we cannot study populations or make any conclusions about them.
  - We select proper representative sample to study and then generalize the results
  - It is important to know that we use Z-distribution when the variance (or standard deviation) of the population is known or the sample size more than 30.
  - The formula is:

$$Z = (\bar{x} - \mu) / (\partial/\sqrt{n})$$

  - Standard error (SE): is the measurement of variation between sample and population. SE = Standard deviation (SD)/ ⌋sample size (n). So, if we increase the n, SE will decrease. If we use the all the population, there is no SE.

**Ex:** If the cranial length of certain large human population which is normally distributed μ= 185.5 mm and ∂ = 12.7 mm, what is the probability of a random sample of size n=10 from this population will have x ≥ 190mm?
Z = ($\bar{x}$ - μ) / (∂/√n)
= (190-185)/ (12.7/√10) = 1.09.
P(x ≥ 190) → P ( Z ≥ 1.09).
& From the Z-table, P=0.1379.



- **Distribution of the Sample Proportion**

**Proportion = part/whole** (the numerator is part form the denominator).

When the sample size is large (≥30), the distribution of sample proportion (P) is approximately normally distributed. The mean of the distribution will be equal to the true population proportion, and the variance of the distribution will be equal to P (1-P)/n. We can use Z-distribution, and to calculate Z:

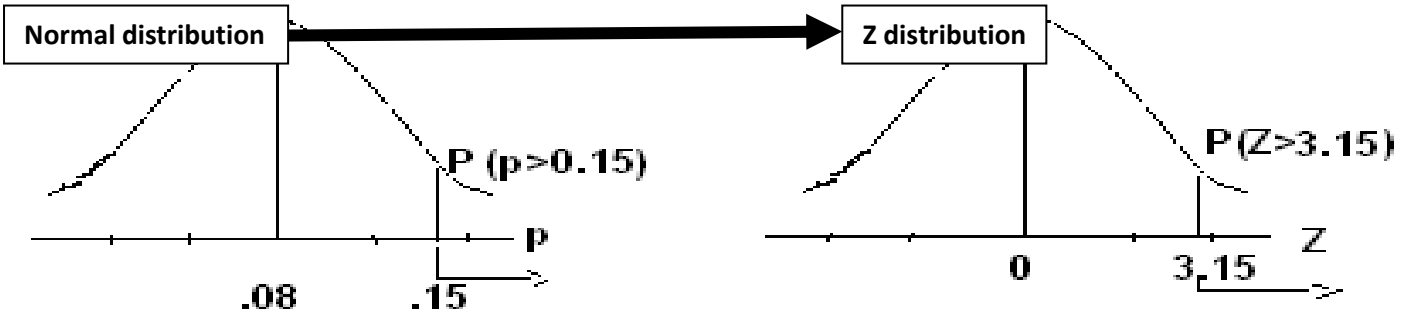
$$Z = [\hat{p} - P] / [\sqrt{P(1-P)/n}]$$

**Ex:** suppose in a certain human population the proportion of color blindness is 8%, if randomly we select 150 individuals from this population, what is the probability that the proportion of color blindness in this sample will be greater than 15%?

$$Z = [\hat{p} - P] / [\sqrt{P(1-P)/n}]$$

$$= [0.15-0.08] / [\sqrt{o.o8\ (1- 0.08)\ /150}] \quad = 3.15$$

$P(P \geq 15\%) \rightarrow P(Z \geq 3.15)$. & From the Z-table, P=0.0008




- **The t-distribution & t-test ((Student's t-test))**
  - In a case when the population variance ($\partial^2$) is unknown and the sample size is small (n≤30), we can use the sample variance ($S^2$) as a best point estimator for population variance ($\partial^2$) **but** in this situation the distribution will not follow the standard normal distribution (Z-distribution) but follow the **t-distribution.**
  - As in standard normal distribution in which we have the Z-table, here we have the t-table which depend on the df= (n-1) → Row of the table & α (probability of error) = 1 - α/2 → column of the table.

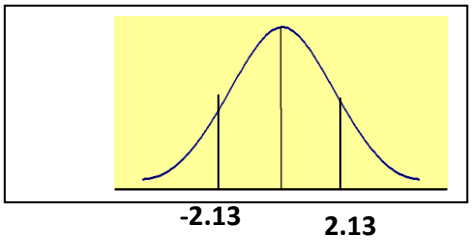
**Ex:** A certain breeds of rats show a mean weight gain of 65gm during the first 3 months of life, a sample 16 0f these rats was taken and feed a new diet from birth until the age of 3 months, the mean weight was 60.75gm with S= 3.84gm. Is this mean differ significantly from population mean?

**$t = (x-\mu) / (S/\sqrt{n})$**

$= (60.75 - 65) / (3.84/\sqrt{16}) \quad = -4.43$

From the t-table (α=0.05): $t_{1-\alpha/2}$, df =15.

→ 2.13 (tabulated t) Since the calculated t located out of the range of tabulated the difference between x & μ is statistically significant.



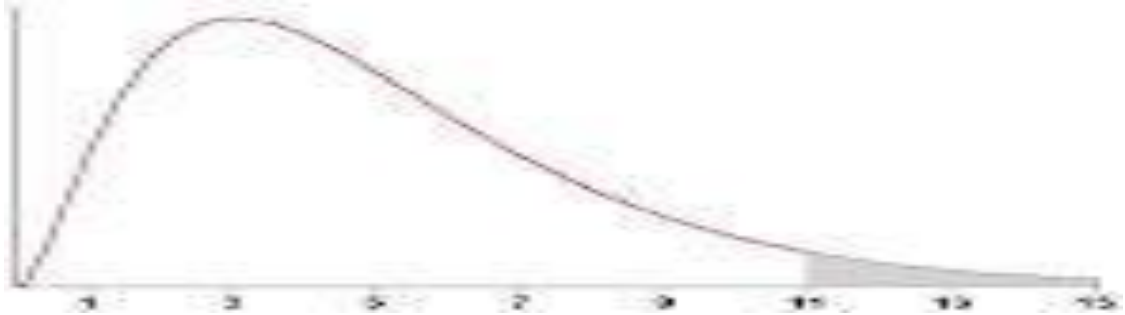
➢ Other uses of t test
- ✓ Comparing the significant difference between two samples means. [Small sample size ($n \leq 30$) and the population variance ($\partial 2$) is a known].
- ✓ Paired difference: Many studies are designed to produce observation in pair e.g., single individual has pair of reading (before & after), Ex measurement of BP before and after treatment

## • The Chi-Square distribution ($X^2$-test)

➢ Most frequently employed statistical technique for analysis of count or frequency data to find the association between two variables.

➢ $X^2$-test statistic is use with qualitative (categorical) variables e.g. marital status (single, married, widowed), or with discrete numerical variable → Used for the frequencies associated with these variables (most accurate when the variable is dichotomous e.g. life or death, disease or not, male or female….etc.).

➢ Chi-Square distribution may be derived from the normal distribution, but it is a skewed distribution (not normal), started from zero and has only one tail (only positive values).



➢ The chi-square test is a non-parametric test It depends on:

1- Observed values (O); number of subjects in our sample that fall into the various categories of the variable of interest (data of the sample).

2- Expected values (E); number of subjects that we would expect to observe in our sample if there is no association between variables. To calculate the expected values: **(E)**= [**Raw margin** X **Column margin**] / **Grand total**

* Always $\sum (O) = \sum (E)$.

$$X^2 = (O-E)^2 / E$$

* and from $X^2$ distribution table **df** =(r-1)(c-1) $\alpha$=0.05 we find p-value.

**Ex:** A group of 350 adults who participated in a health survey were asked whether or not they were on diet, there responses by sex were as in the table.

| | **Male** | **Female** | **Total** |
|---|---|---|---|
| **On diet** | 14 | 25 | 39 |
| **Not on diet** | 159 | 152 | 311 |
| **Total** | 173 | 177 | 350 |

Regarding the above data is the association between sex and being on diet is statistically significant ($\alpha$=0.05)?

☞From the table above (observed values), we calculate the expected values using the formula: **E**= [**Raw margin** X **Column margin**] / **Grand total**.

| | **Male** | **Female** | **Total** |
|---|---|---|---|
| **On diet** | 19.3 | 19.7 | 39 |
| **Not on diet** | 153.7 | 157.3 | 311 |
| **Total** | 173 | 177 | 350 |

☞We calculate the $X^2$-value for each cell using the formula: $X^2 = (O-E)^2 / E$.

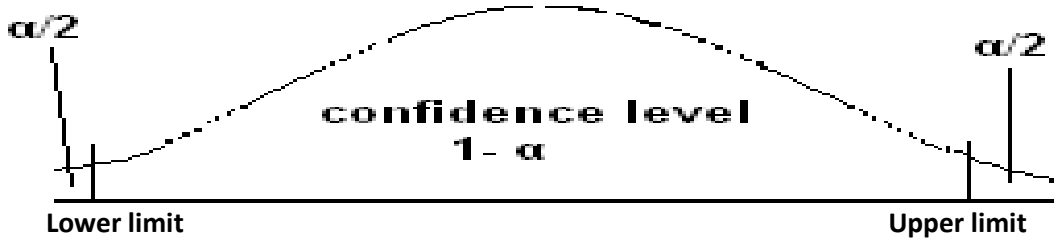
| | **Male** | **Female** | **Total** |
|---|---|---|---|
| **On diet** | 1.465 | 1.426 | 2.881 |
| **Not on diet** | 0.163 | 0.179 | 0.342 |
| **Total** | 1.638 | 1.605 | **3.243** |

☞We calculate the total $X^2$-value for the table (3.243), The df=(r-1)(c-1) ➡ (2-1)(2-1) = 1, $\alpha$=0.05 and from $X^2$-distribution table we find p-value for df=1 and $\alpha$=0.05 (3.841).

Thus, 3.841>3.243➡The association is not significant.

# • Estimation of the Confidence Interval (CI).

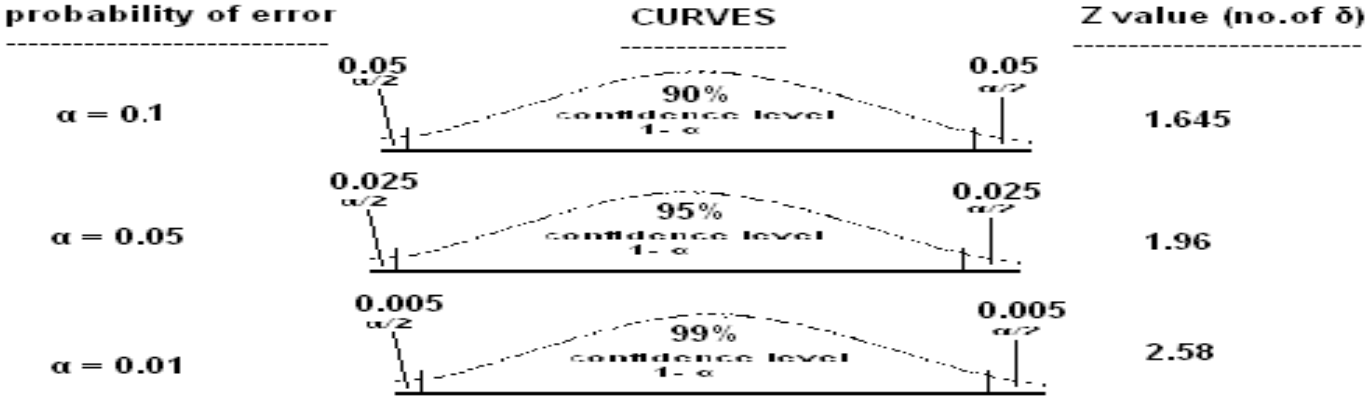
➢ CI: Consist of two numerical values (upper & lower values) defining the interval with which lies the unknown parameter with certain degree of confidence. These values depend on the confidence level which is equal to 1- α, and α is the probability of error.




➢ If α (probability of error) is 10% or 0.1, and α/2 is 0.05 on each side of the curve, then the confidence level 1- α will be 0.9 or 90%, and so for α = 0.05, α/2=0.025.

➢ If α (probability of error) is 5% or 0.05, & and α/2 is 0.025 on each side of the curve, then the confidence level 1- α will be 0.95 or 95%.

➢ If α (probability of error) is 1% or 0.01, & and α/2 is 0.005 on each side of the curve, then the confidence level 1- α will be 0.99 or 99%.



## Calculation of the Confidence Interval (C.I)

**CI = Estimator ± Reliability coefficient (z or t-value) X standard error (SE)**

### 1- Calculation of C.I for the population mean (μ)

The sample mean**(x)** is the best point estimator for the population mean **(μ).**

**a-** ***When the population variance is known or the sample size is large (>30****), the formula is:*

$$\textbf{C.I } (\mu) = \bar{x} \pm Z\,(\partial / \sqrt{n})$$

**Ex.** The mean of serum bilirubin level 16 four days old infants was found to be 5.89mg\dl with $\partial = 3.5$mg\dl. Find:

- 90% CI for the mean?
  90% CI (μ) = 5.98 ±1.64 X 3.5√16 = 5.98 ± 1.43 = (4.54 - 7.42)
- 95% CI for the mean?
  95% CI (μ) = 5.98 ±1.96 X 3.5√16 = 5.98 ± 1.71 = (4.26 - 7.69)
- 99% CI for the mean?
  99% CI (μ) = 5.98 ± 2.58 X 3.5√16 = 5.98 ± 2.25 = (3.72 - 8.24)

**Note:**

- ✓ Form this example we can notice that the width of CI is directly related to the level of confidence; smallest with 90% (but not reliable level of confidence, i.e. high probability of error) and largest with 99% (a highly confident estimation but very wide range) **and that's why 95% level of confidence is the most practical to use.**
- ✓ Another factor can affect the width of confidence level, although it is not illustrated in this example but it is obvious from the rule of confidence interval, that the width of the interval is reciprocally related to the square root of the sample size, i.e. to the sample size, so we can decrease the width of this interval by taking larger samples.

**b-** ***When the population variance is unknown or the sample size is small (≤30****), the formula is:*

$$\text{C.I } (\boldsymbol{\mu}) = \bar{x} \pm t\,(S / \sqrt{n}) \quad \textbf{df=n-1}$$

**Ex.** The mean of WT of 16 rats was found to be 60.75 Kg with S = 3.8 mg\dl. Find the 95% CI for the mean?

95% CI (μ) = 60.75 ± 2.13 X 3.8√16
= 60.75 ± 2.04 = (58.71 – 62.79)

### 2- Calculation of C.I for the population proportion (P).

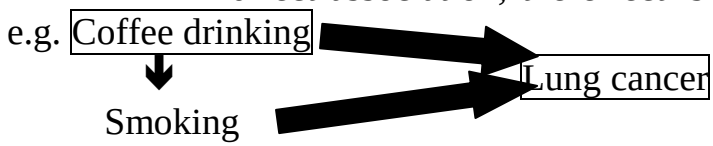
The sample proportion (P) is the best point estimator for population proportion (P). *We use* ***z-test*** *and the formula is*:

$$\textbf{C.I (P)} = p \pm Z\sqrt{[(p(1-P)/ n)]}$$

**Ex:** In a survey, 300 adults were interviewed, 123 said they have yearly medical check up. Find 95% C.I for the proportion having yearly checkup.

95% **C.I (P)** $= p \pm 1.96\sqrt{[(p\,(1-P)/\sqrt{n})]}$
$= 123/300 \pm 1.96\sqrt{[0.41(1-0.41)/ 300]}$
= 0.41± 0.06 = (0.35- 0.47)

- **<u>Causal Association (causal inference)</u>**
  - An understanding for the causes of the disease is important in the health field not only for prevention but also in the diagnosis and application of the treatment.
  - In the first step, If we determine that an exposure is associated with a disease, the next step is whether the observed association reflect a causal one.
  - The initial step may consist of clinical observation; the second step is tried to identify correlation from available data. Then we can conduct new studies to determine whether there is an association between an exposure and outcome.
  - Association is a statistical relationship between variables. While causation is a matter of judgment assessment of the meaning of that observed association.
  - Observation ➡ Correlation ➡ Association ➡ Causation
  - Presence of correlation not necessarily indicates a valid statistical association & the presence of valid statistical association not necessarily indicate there is causal relation.
  - Types of association:
    1- Spurious ((false, factitious)) association ➡No real association between the factor and outcome.
    2- Real ((True)) association, which is either;
       - ❖ Direct ((causal)), e.g. smoking➡lung cancer.
       - ❖ Indirect association; the effect is due to hidden factor(s) "**confounder**",




Or the factor leads to the outcome through intermediate steps, **Factor ➡ step I ➡ step II ➡ Disease**

  - Guide lines "Hill's Criteria" for judging whether an association is causal.
    1. Temporal relationship: Exposure to the factor must have occurred before the disease developed with sufficient time. E.g., Asbestos increases the risk of Ca-lung, but Ca occurs after 20 years of exposure. If Ca occurs after 3 years➡Asbestos not the cause.
    2. Strength of association: Measured by relative risk and odd's ratio; the stronger the association ➡the stronger the suggestion about causal association.
    3. Dose- response relation -ship: As the dose of the exposure increase, the risk of the disease also increase➡ strong evidence of causal association.
    4. Replication of the finding (consistency): replication of the finding; the same finding in different studies and in different population.
    5. Biological plausibility: Coherence with current biological knowledge. E.g., smoking considers strongly a cause of Ca lung as pathological changes in the lung tissue.
    6. Consideration of alternative explanation: The extent to which the investigator have taken other possible explanations into account.
    7. Reversibility: If a factor is a cause of a disease, we would expect reduction in the risk of disease if we decrease the exposure.
    8. Specifity of the association: Specific association between the exposure & the disease.
    9. Coherence with other knowledge: If a relation ship is causal, we would expect the finding to be consistent with other data (biological, chemical ….etc.)

- **Identification of Health Problems & Diseases in Community**

  1- Screening
  2- Surveillance
  3- Research work

## 1) Screening:

- The presumptive identification of unrecognized disease or defect in community by the application of test, examination, or other procedures which can be applied rapidly to a symptomatic population to sort out those who probably have the disease from those who probably not.
- One of the central concerns in clinical medicine is differentiating the normal from the abnormal. In preventive medicine, this is important at group (population) level.
- The basic purpose of screening for disease is to separate from a large group of apparently well persons who have the probability of having the disease under the study. A screening test is not intended to be diagnostic; and those with positive test sent on for further evaluation.
- There is often confusion between screening and diagnosis. This often happens because the same tests can be used for both purposes.
- While screening occurs in individuals who are asymptomatic, or not suspected of having the disease in question, diagnosis is used to confirm whether someone actually has a disease.
- For example, a blood sugar test in someone that is healthy would be a screening test. It is an initial step
- The same test in someone with symptoms of diabetes would be diagnostic.
- The **concept** of screening is that early detection, before the development of symptoms → more favorable prognosis because the treatment began before the treatment began before the disease become clinically manifested → More effected → Reduction in morbidity and mortality. (2ndry prevention)

**Disease appropriate for screening:** Disease should be:

1. Prevalent with detectable preclinical phase.
2. Serious.
3. Treatment giving before symptoms develop should be more beneficial. E.g. HT, TB, Ca breast, PKU….etc. Of course, it is not applicable for Gall stone, IBS, Influenza…etc.

- The interval of time between the point at which the disease can detected by screening and the point at which the individual become symptomatic have been termed "detectable preclinical phase"
- The prevalence of detectable preclinical phase of disease and thus number of cases detected by screening can be increased by screening high risk group.
- Diagnostic and screening tests have major differences due to the reasons for performing them. Diagnostic tests usually done on a patient who has a clinical problem, more likely

to have the disease. While screening tests done on apparently healthy people, before they have any symptoms.

- **Criteria for screening test:** screening test should be:
  1. Valid and accurate, with high sensitivity and specify.
  2. Feasible; easy to perform, low cost, not complicated (not need highly skilled person).
  3. Acceptable to population; painless, noninvasive, no side effects or under risk of complications.
  4. Reliable and reproducible results.
  5. Effective: impact on the course of the disease.

**Unfortunately**, there is no screening test currently available that can meet all these criteria.

- **Gold Standard:** Test or examination that denote the presence of the disease without any doubt. Ex: Endoscope for DU, Histopathology for Ca.

**Results of screening test**

| Screening Test | Gold standard Test Disease | Disease Free | Total |
|---|---|---|---|
| Positive | (a) True Positive | (b) False Positive | (a+b) Total Test +ve |
| Negative | (c) False Negative | (d) True Negative | (c+d) Total Test -ve |
| Total | (a+c)Total Disease +ve | (b+d) Total Disease -ve | (a+b+c+d) Grand Total |

☞ **Validity of screening test:**

The validity of a screening test is measured by how good it is at distinguishing between individuals who have the disease and those don't have. Two components of validity: **Sensitivity** and **Specifity**.

☒ **Sensitivity**: Ability of the test to identify correctly those who have the disease (True positive). In other words, of all those who are truly with disease, how many are picked up by the test?

**Sensitivity = a / a+c X 100%**

☒ **Specifity**: Ability of the test to identify correctly those who do not have the disease (True negative). In other words, of all those who are truly without the disease, how many are picked up by the test?

**Specifity = d / b+d X 100%**

☞ **Predictive values of a screening test**

In clinical setting, a different measure may be important for the physician which is whether or not an individual actually has the disease giving that the result of the test is known (either +ve or –ve). This is called the **predictive value** of the test. We have:

1. **Positive Predictive Value ($PV^+$):** Is the probability that person actually has the disease giving that he or she tests positive.

$$PV^+ = a / a+b \text{ X } 100\%$$

***Positive Predictive Value is also called "Yield" of the test**

2. **Negative Predictive Value ($PV^-$):** Is the probability that person actually has no disease giving that he or she tests negative.

$$PV^- = d / c+d \text{ X } 100\%$$

**Also we have:**

- The false-positive rate is the likelihood of a positive result in patients known to be free of the disease: **b / b+d X 100%** and equals (1-specificity);
- The false-negative rate is the likelihood of a negative result in patients known to have the disease: **c / a+c X 100%** , and equals (1-sensitivity).
- Accuracy of test. (TP+TN) / Total

Thus, sensitivity and the false-negative rate describe how the test performs in patients with disease, whereas specificity and the false-positive rate describe how the test performs in patients without disease.

**Example:** Supposing one is interested in validating the use of CXR for the diagnosis of pulmonary TB. The gold standard for diagnosing TB is the culture of AFB from the sputum. To validate the use of CXR, we would have to select 200 TB suspect's people and perform both CXR and sputum cultures for them. The results given in the table:

| Screening test CXR | Gold standard Test (Culture) Disease | Disease Free | Total |
|---|---|---|---|
| Positive | (a) 80 | (b) 70 | (a+b) 150 |
| Negative | (c) 20 | (d) 30 | (a+b) 50 |
| Total | (a+c) 100 | (b+d) 100 | (a+b+c+d) 200 |

- ☒ Sensitivity = a / a+c X 100% = 80/100 X100 = 80%
- ☒ Specifity = d / b+d X 100% = 30/100 X100 = 30%
- ☒ $PV^+$ = a / a+b X 100% = 80/150 X 100 = 53%
- ☒ $PV^-$ = d / c+d X 100% = 30/50 X 100% = 60%

**Example:** A simple routine test for the presence of HIV was carried on the 300 high risk subjects (IV drug users). A more accurate but expensive test (the gold standard) was also carried out to assess the accuracy of the routine test. The following results were obtained.

| Routine test | Gold standard Test Disease | Disease Free | Total |
|---|---|---|---|
| Positive | 92 | 10 | 102 |
| Negative | 2 | 196 | 198 |
| Total | 94 | 206 | 300 |

- ☒ Sensitivity = a / a+c X 100% = 92/94 X100 = 97.9%
- ☒ Specifity = d / b+d X 100% = 196/206 X100 = 95.1%
- ☒ $PV^+$ = a / a+b X 100% = 92/102 X 100 = 90.2%
- ☒ $PV^-$ = d / c+d X 100% = 196/198 X 100% = 99%
- ☒ The estimated prevalence = 94/300 X 100= 31.3%

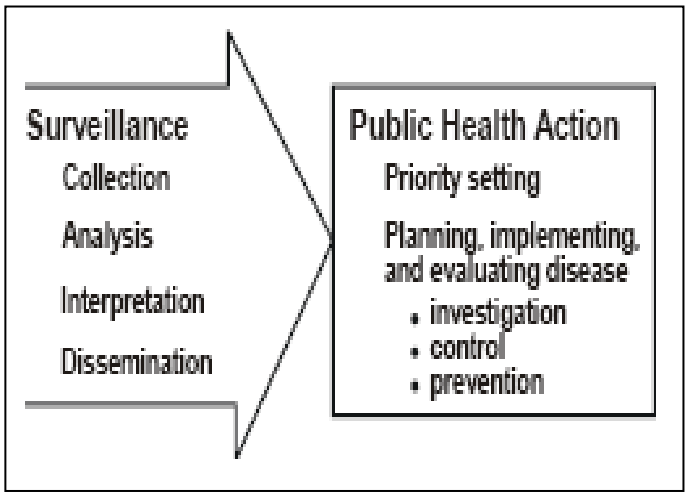
➢ **Relation of $PV^+$ to disease prevalence**

The $PV^+$ of the screening test (yield of the test) is increased as the prevalence of the disease is increased as shown in the below table:

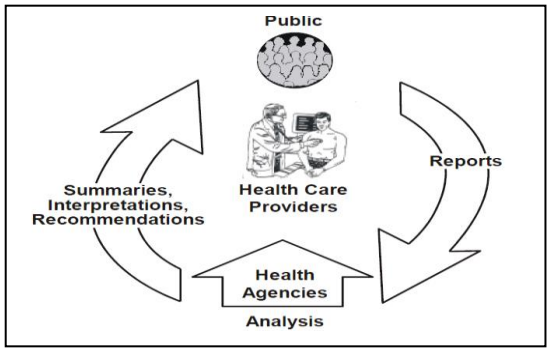
| Disease Prevalence | Test Result | Gold standard Test Disease | Disease Free | Total | $PV^+$ |
|---|---|---|---|---|---|
| 1% | Positive | 99 | 495 | 594 | 17% |
| | Negative | 1 | 9405 | 9406 | |
| | Total | 100 | 9900 | 10000 | |
| 5% | Positive | 495 | 475 | 970 | 51% |
| | Negative | 5 | 9025 | 9030 | |
| | Total | 500 | 9500 | 10000 | |

## 2) Surveillance:

- ➢ The ongoing, systematic collection, analysis, interpretation, and dissemination of data regarding a health-related event for use in public health action to reduce morbidity and mortality and to improve health.
- ➢ Surveillance: Information for Action
- ➢ It is the monitoring of people behavior, activities, or other changing information.
- ➢ Purposes of Public Health Surveillance
  - **1)** Assess public health status
  - **2)** Define public health priorities
  - **3)** Identify risk factors
  - **4)** Identify high-risk population or groups
  - **5)** Evaluate programs
  - **6)** Stimulate research




- ➢ Data Sources and Methods for Surveillance
  - **1)** Notifiable diseases
  - **2)** Laboratory specimens
  - **3)** Vital records (birth, death, ….).
  - **4)** Surveys
  - **5)** Administrative data systems
- ➢ Steps in surveillance:
  - 1) Assess the population
  - 2) Select the outcome or process for surveillance
  - 3) Collect surveillance data



4) Calculate, analyze and compare rates
5) Apply risk stratification methodology
6) Report and disseminate surveillance information

- Surveillance of diseases trends includes :
  a. Identifying , investigate and control outbreaks or epidemics;
  b. Identifying specific population groups at high risk of sickness and death from priority diseases;
  c. Confirming current priorities among disease control activities;
  d. Evaluating the impact of preventive and curative activities on the incidence and prevalence of priority diseases in the community;
  e. Monitoring disease trends so as to adjust plans to meet current needs.

- Surveillance can be divided into:

**1) Passive Surveillance:** Reporting of cases by health workers at their discretion. It uses already existing systems in your districts. It relies on the periodic reports which you prepare and submit to your district team.

**2) Active Surveillance:** This is the regular or periodic collection of case reports from health care providers or facilities. In other words it is done all the time by collecting data in health unit or laboratory. Data collected by active surveillance is more accurate than other types of surveillance.

**3) Sentinel Surveillance:** Because underreporting is common for certain diseases, an alternative to traditional reporting is sentinel reporting, Relies on a prearranged sample of health-care providers who agree to report all cases of certain conditions. These sentinel providers are clinics, hospitals, or physicians who are likely to observe cases of the condition of interest.

- **EX on surveillance:**
  a) Evaluating impact of national vaccination campaigns
  b) Estimating impact of TB on health care system)
  c) Identifying outbreaks of rubella and congenital rubella
  d) AFP monitoring

- **Note regarding data**
  - ✓ When available, demographic characteristics such as gender, age, occupation, education level, socio-economic status, immunization status can reveal disease trends
  - ✓ Many ways to display surveillance data :
    - Line graphs for displaying data by time
    - Maps for presenting data in geographic context ( Spot or rate maps)
    - Single/multivariable tables

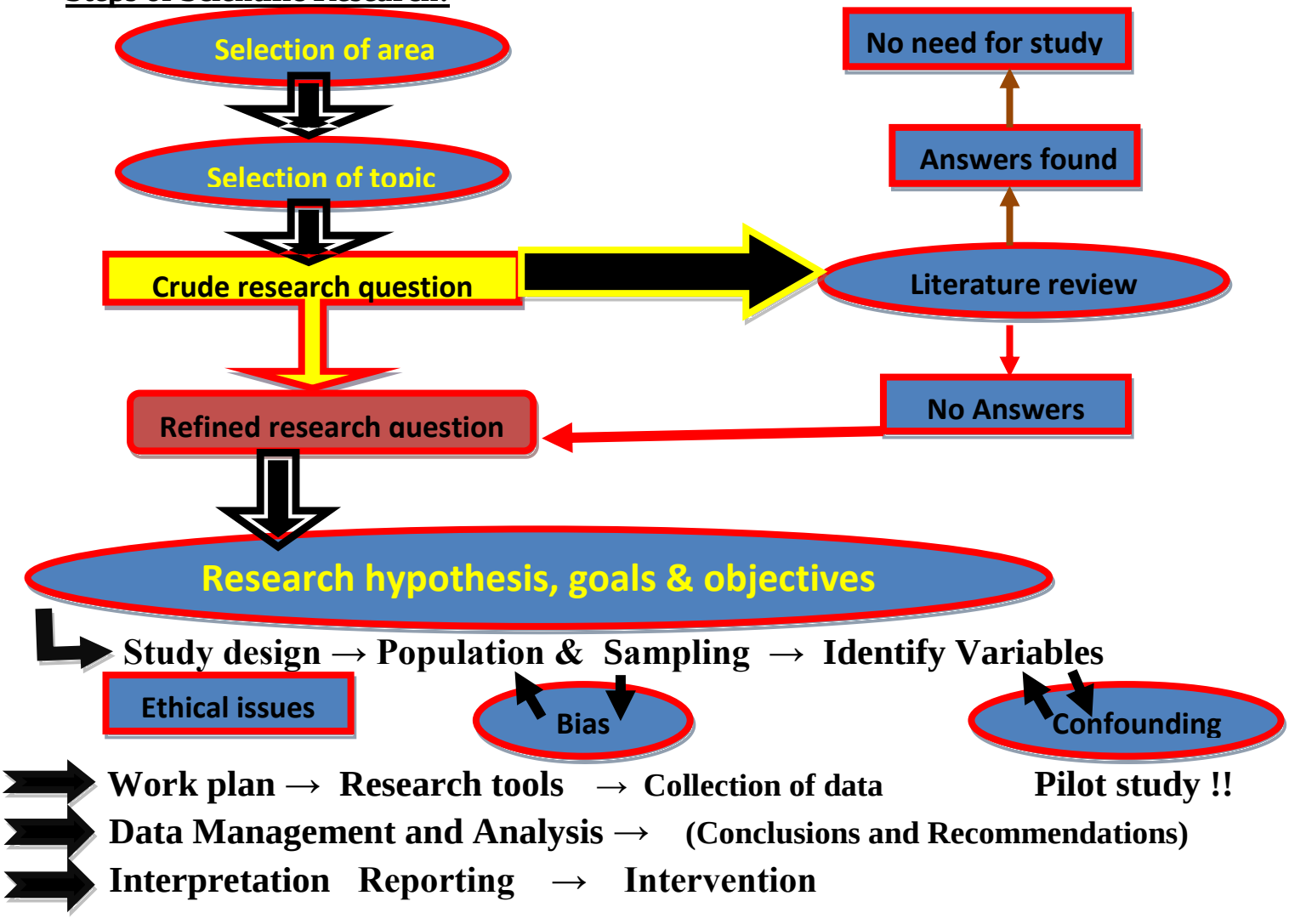
## 3) Medical Research

- ¤ Research can be defined as "the systemic and organized method in searching for information for answering questions to solve a problem & find new knowledge ".
- ¤ The daily practice of medicine need the knowledge about who is likely to develop a particular condition and under what circumstances, what is the best treatment, what is the prognosis, and who we can prevent such illness. This type of knowledge emerges in many cases from epidemiological researches.
- ¤ Research consists of the prefix "Re" which means (again , or a new ) And Search (verb) means ( to examine, or to try )

### Aims of medical research:

- Establishing causes of disease or factors associated with death or disease
- Detecting disease
- Preventing death or disease
- Evaluating treatments for disease.
- Evaluating health services & programs.

### Steps of Scientific Research:



➢ **Selection of Research Area for the researcher is based on the following:**

1-Specialty 2- Interest 3- Scientific background 4- Experience

➢ **Prioritizing and selecting a research topic:** Criteria for selecting a research topic:

1. Relevance: How large or widespread is the problem? Who is affected? How severe is the problem?
2. Avoidance of duplication: If the topic has been researched, the results should be reviewed to explore whether major questions that deserve further investigation remain unanswered. If not, another topic should be chosen.
3. The characteristics of the proposed study: Feasibility, Cost-effectiveness, Applicability of the results, Available interventions

➢ **Research question: The investigator must make sure that:**

1) He has a research question
2) The question is clear and specific
3) It reflects the objectives of the study
4) It has no answer by common sense
5) It has no answer in the LITERATURE
6) Finding an answer to the question will help in solving the problem to be studied.

***It is important to:***

1-Determine the exposure & outcome precisely.
2- Determine the confounding variables
3- Remember that the choice of any study design based on:

- Features of exposure &outcome.
- Time & resources available.
- Results from previous studies.
- Gaps in the knowlege that remain to be filled.

➢ **Components of Research:**

Research papers are organized in systematic way so that the information flow in logic sequence. The sequence of research component is:

a) Title of the research
b) Abstract
c) Introduction: Background information and Statement of the research problem (Scientific justification for the study) + Research objectives
d) Literature review
e) Methods
f) Results
g) Discussion
h) Conclusions and recommendations
i) References
j) Annexes

**a) Title:** Should provide a brief, informative summary that will attract the target audience. A good title should be accurate, and concise. It should make the central objectives of the study clear to the reader.

**b) Abstract: :** Probably the most important part of the paper, since many persons will only read the abstract.

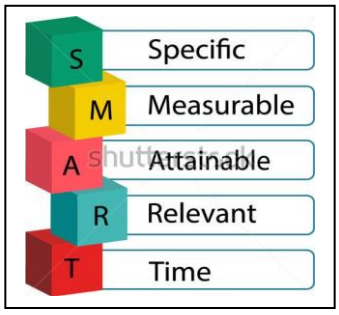
- The abstract/summary should be written only *after* the final draft of the report has been completed. Usually **250** words or less.
- Contains a brief summary of each major section of the paper (introduction, methods, results, conclusions).
- Structured abstract is written with sub section paragraph. (Background, Aim, Method, Results. Conclusions, Recommendation).

**c) Introduction:** Typically limited to a few paragraphs contain some relevant background data related to the problem. It should

- Convince the reader about the relevance of the study (magnitude, severity of the problem).
- Frames the purpose and public health significance of the research.
- The research hypothesis(es) to be investigated / tested should be clearly stated. It should contain a paragraph on what you hoped to achieve.
- Identify a gap in knowledge (Study problem). Provide key background (scope/nature/magnitude of the gap). Be clear that filling this gap will be useful.
- Describe the relevant limitations of previous studies. Emphasize that your approach addresses the limitations of previous studies.
- **Goals and Objectives** The goal (aim) and objectives must be stated at the very beginning of the study, since they will guide the investigator during the process of formulating research questions and hypothesis. They will also help in the prioritization process.
- They will enable the reader or consumer of the work to judge whether the investigator had achieved these objectives or not.

**Goal:** Describes the aim of the work in broad terms

**Objectives:** More specific and relate directly to research question. They may be divided into two types:




***The research objectives should be: "SMART"***

- Closely related to the research question
- Covering all aspects of the problem
- Ordered in a logical sequence
- Stated in action verbs that could be evaluated e.g. to describe, to identify, to measure, to compare, etc.
- Achievable, taking into consideration the available resources and time

**S**→Specific, **M**→Measurable, **A**→ Achievable
**R**→Relevant, **T** → Time-bound

**d) <u>Literature Review:</u>**

surveys scientific articles, books, medical journals, dissertations and other sources relevant to a particular issue, area of research, or theory, providing a description, summary, and critical evaluation of each work."

***Purpose of Literature Review***

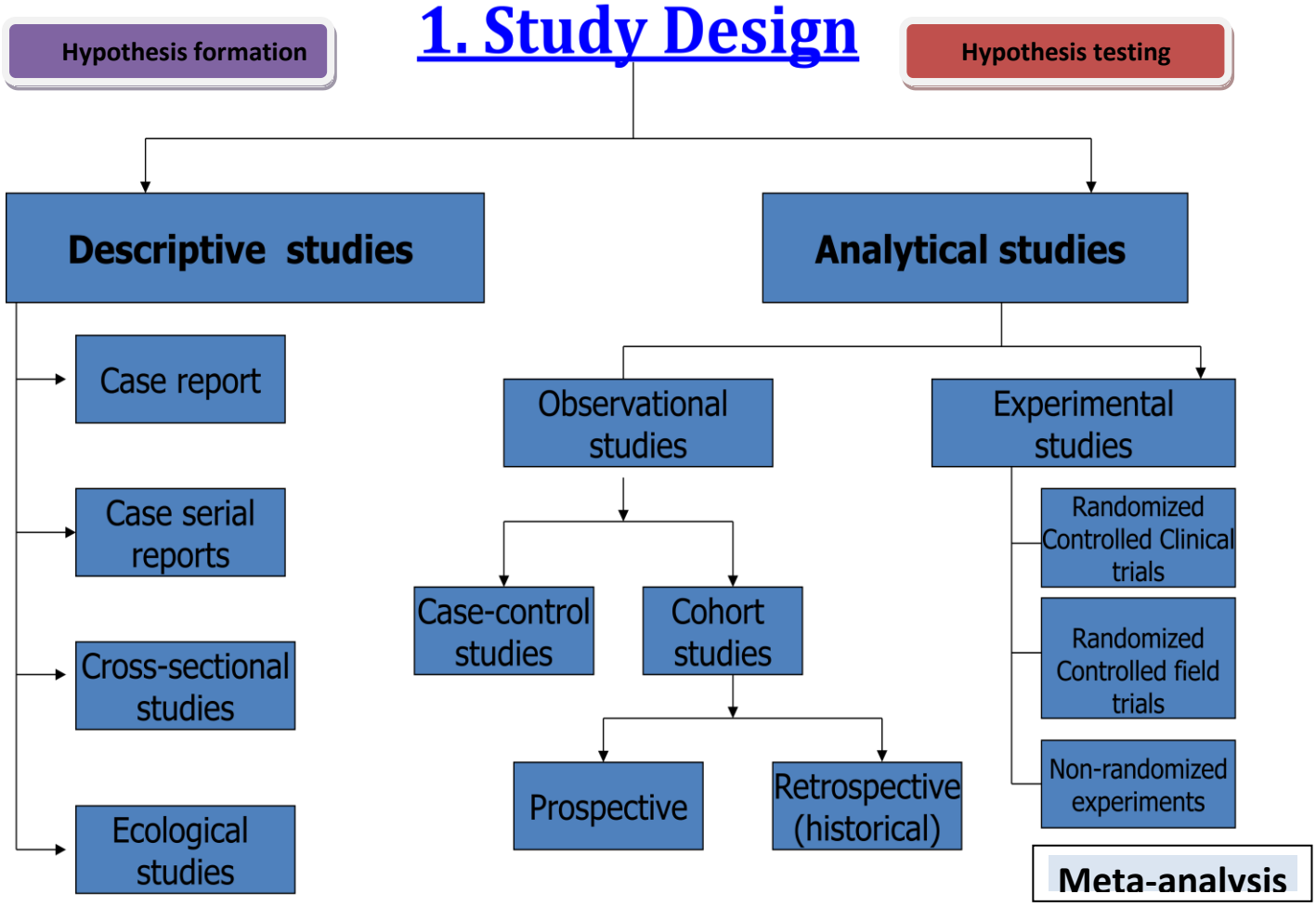
- Understanding of the subject under review
- Describe the relationship of each work to the others under consideration
- Identify new ways to interpret, & shed light on gaps in previous research
- Resolve conflicts amongst seemingly contradictory previous studies
- Identify areas of prior research to prevent duplication of effort
- Place one's original work (in the case of theses or dissertations) in the context of existing literature

**<u>Sources of Literature:</u>**

1) **Books:** Especially on line to find more detailed and up-to-date sources of information.
2) **Journal articles**: these are good sources, especially for up-to-date information. They are frequently used in literature reviews because they offer a relatively concise, up-to-date format for research.
3) **Theses and dissertations**: But can be difficult to obtain since they are not published, but are generally only available from the library or interlibrary systems
4) **Scientific websites**: Avoid commercial websites .
5) **Conference proceedings**: Useful in providing the latest research, or research that has not been published. They are also helpful in providing information about people in different research areas.
6) **Government/corporate reports:** Many government departments and corporations' commission carry out research. Their published findings can provide a useful source of information, depending on your field of study.

**e) <u>Methods: (patients, subjects, materials)</u>** to describe how, when, and where you collected, organized and analyzed the data (that are relevant to the purpose of the study). This section should include all the steps of work:

- ✓ Setting: Place & time of study
- ✓ Study design(Descriptive or analytic)
- ✓ Study population, sample & sampling methods (random or not)
- ✓ How many subjects were eligible (eligibility criteria)
- ✓ How many dropped out (follow up in prospective studies)
- ✓ Variables definition (exposure, outcome , confounders, associated variables)
- ✓ Ethical Considerations
- ✓ Laboratory methods
- ✓ Epidemiologic investigation
- ✓ Diagnostic Evaluation
- ✓ Statistical analysis (Descriptive and inferential)




☒ Important to remember:

- ☞ The scientific value and level of evidence for any medical study are determined to a major extent by the study design.
- ☞ The study design is the path that governs how the data are to be collected & analyzed
- ☞ Errors in study design cannot be corrected afterwards.

**f) Results:**

- ➢ Describe what you found, not what you did (Methods).
- ➢ Should correspond directly with the stated research hypothesis(es).
- ➢ Start with text (prose). Write the tables and figures later.
- ➢ Text descriptions should not be largely redundant with data in tables and figures.
- ➢ Be brief: a picture (table / figure) is worth a thousand words. Avoid reiterate
- ➢ Use tables to highlight individual values.
- ➢ Use figures to highlight trends/relationships.
- ➢ Provide consistent row or column summation.
- ➢ Present results in a logical sequence.
- ➢ Make sure text is consistent with tables / figures

**g) Discussion:** to interpret your results and justify your interpretation

- Describes what was learned from the study and public health implications of the findings.
- Should not be a large re-statement of text from the Results section.
- Should not include presentation of "new" findings not presented in Results section.
- Emphasize strengths of study and what is new /useful.
- Focus on the main results and conclusion.
- Be clear about why results support this conclusion.
- Maintain connection with purpose of the study.
- Contrast the results with similar previous studies, including possible explanations for differences.
- Should state to whom the results most likely apply (generalize).
- State limitations / difficulties (frankly, without apology). (all studies have some limitations)

**h) Conclusions:** In logic sequence from the study.

**i) Recommendations:** In logic sequence from the study.

**j) References:**

- Referencing the research correctly is essential when gathering information from books, journals and Web Pages
- The references can be numbered in the sequence in which they appear in the research and then listed in order in the list of references (Vancouver refer. system).
- DO NOT include references in your abstract.
- Keep it accurate and provide all the relevant details.
- Use a consistent format for your references.

**K) Annexes or Appendices: May include:**

- Interview schedule/ questionnaires and/or other data collection tools).
- Informed consent form
- Institutional/Ethical approval for the study

# • Types of Medical Research

## I- Descriptive Studies

- The investigators merely describe the health status of a population or characteristic of a number of patients.

☒ Describe the pattern of disease occurrence in relation to variables such as person, place and time.

**a-** Person: who is getting the disease? & who is not? Person characteristics are: Age, sex, race, marital status & socio-economic status (education, occupation & income).

**b-** Place: where are the rates of disease highest or lowest?

**c-** Time: When does the disease occur commonly or rarely?

☒ Data provided by descriptive studies are weak because they make no attempt to link cause and effect and therefore no causal association can be determined.

☒ But, these data are often the first steps to a well-designed epidemiological study. They allow the investigator to define a good hypothesis which can then be tested using a better design. They are essential for;

a- Public health administrators {which group(s) or subgroup(s) are more or less affected by the disease}.

b- Epidemiologists {identification of risk factor(s)}

☒ Important to keep in mind:

1) Exposure means risk factors or independent variable
2) Outcome means disease or dependent variable

## ✓ Descriptive studies could be:

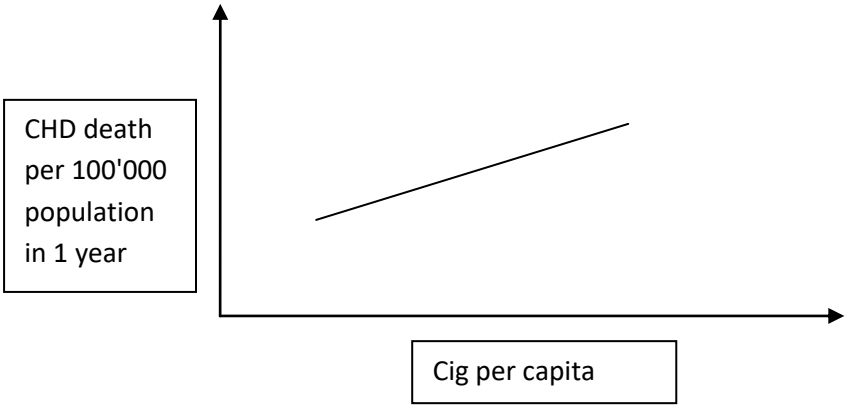
**A- Popular studies** ➡ Correlational studies.

**B- Individual studies**➡ Case report, Case series, & Cross -sectional.

### 1- Correlational study (Ecological study);

Describe the disease in the entire population (the units of study are populations rather than individuals) in relation to the factor of interest such as age, sex, calendar time, utilization of health services, consumption of certain food or medication…….etc.

Ex: To describe mortality from coronary heart disease (CHD) in 5 countries (population) with per capita cigarette consumption in one year.




The ***descriptive measure*** of association in correlational study is "correlation coefficient" (r) which ranges from (-1) to(+1).

If r = -ve → inverse association (may be preventive)
If r = +ve → +ve association (may be causal)
If r = 0 → No association.

***Advantages;**

a- Quick & inexpensive. b- Use already available data. c- Usually used as a first step in investigation a possible exposure- out come relation-ship.

***Limitations;**

a- In ability to determine the temporal relation-ship between exposure & out come.
b- Lack the ability to control for the effect of confounder.
c- Represent average exposure level rather actual individual level (Ecologic Fallacy).

☞**The Ecologic Fallacy**: Observations made at the group level may not represent the exposure-disease relationship at the individual level.

**EX**: An ecologic study finds that 70% of children have TB although 60% were vaccinated.The media report that BCG vaccine will not protect children from TB.

d- Formulate the hypothesis but can't test it.

**Remember: # The presence of correlation doesn't necessarily imply the presence of a valid statistical association.**

**2- Case report study.**

Describe the experience of a single patient. A condition develops in single individual and draws the attention of the clinician or the researcher. It is the first step in disease recognition.

Ex: Kaposi sarcoma in healthy homosexual adult.

Ex: Advanced proliferative diabetic retinopathy in newly diagnosed DM

Generally report a new or unique finding

- ✓ previous undescribed disease
- ✓ unexpected link between diseases
- ✓ unexpected new therapeutic effect

**3- Case series study.**

Describe the event of a group of patients with similar diagnosis. (Collection of case reports)

***Advantages (case report & case series) ;**

a- Recognition of new disease e.g. AIDS.

b- Formulation of hypothesis concerning possible risk factor.

***Limitations;**

a- Based on experience of one or few patients only.

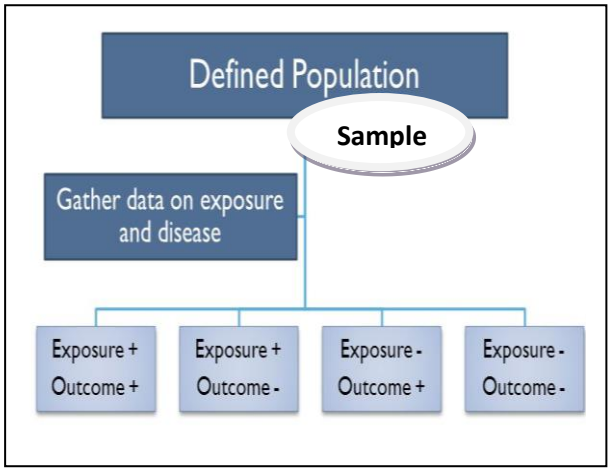
b- Lack of comparison group.

c- Formulate the hypothesis but can't test it.

**4- Cross-Sectional Study (Prevalence Survey)**;

The exposure & the outcome are assessed simultaneously among individual in well-defined population. It is as if we were taking a photo-graph or a slice through the population at a point in time.

- Also called survey or prevalence study
- Usually involves random sampling and questionnaire measurement
- Participants (study sample) are selected to represent study population
- Cannot distinguish whether hypothesized cause preceded the outcome
- used to generate hypotheses



**Ex:** A researches wishes to investigate a possible association between cigarette smoking & CHD in a certain population, 1460 young adults were randomly selected, smoking history was taken & an ECG performed for evidence of CHD.

| | | Out-come(CHD) | | Total |
|---|---|---|---|---|
| | | +ve | -ve | |
| Exposure (Smoking) | +ve | 130 | 730 | 860 |
| | -ve | 20 | 580 | 600 |
| Total | | 150 | 1310 | 1460 |

- ☞ We can calculate the prevalence of disease (CHD) in person with exposure (Smoking) (130/860) and compare it with the prevalence of disease in person without exposure (150/600).
- ☞ Or compare the prevalence of exposure in person with disease (130/150) to the prevalence of exposure in person without the disease (730/1310).

☒ **Planning for the Cross-Sectional Study: Determine**

- How you will acquire a registry of your target population
- Develop a valid survey instrument e.g. Questionnaire
  - How will you collect the information? Mail out questionnaires, Telephone interview, Personal interview
  - Physical exam
  - Laboratory information
- How valid are the measurement instruments you are using? Subjective or Objective
- Have you calculated the sample size required to test the hypothesis?
- What statistical methods will you use?

***Advantages**.

1- Quick, inexpensive and less time consuming.
2- Provide information about the frequency & characteristics of the disease.
3- Provide information about the prevalence of the disease.

***Limitations:**

1- Not determine the temporal relation-ship between exposure & out-come.
2- Not determine prognostic factor from risk factor.
3- Liable for information bias (recall or interviewer bias).
4- Formulate the hypothesis but can't test it.

- **KAP (knowledge, attitudes, and practices) study:**

KAP studies are purely descriptive and help to build up a better understanding of the behavior of the population, without necessarily relating this to any disease or health outcome.

- **Population census:**

A cross-sectional study of an entire population. It provides the denominator data for many purposes (e.g., estimation of rates, assessing generalizability, projecting from smaller studies)

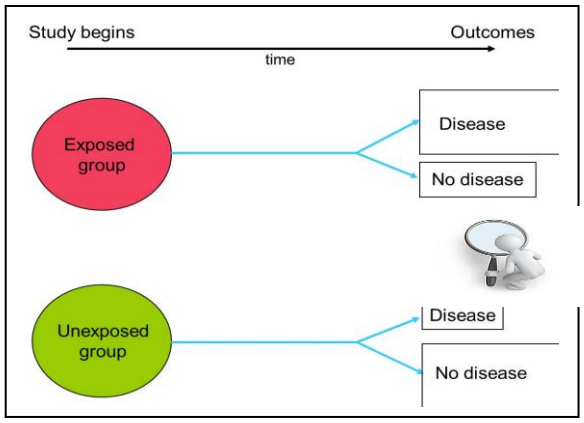
# II- Analytic studies

These studies describe the association between exposure and outcome (test the hypothesis). Can be classified into:

1- **Observational studies**. The investigator simply observe the nature al course of the events, noting who is exposed & non exposed and who has & has not developed the outcome of interest(passive investigator). We have;
   - **a- Case-control study.**
   - **b- Cohort study.**
2- **Interventional studies (clinical trial).** The investigator himself will allocate the exposure (active investigator).

## *Cohort Study.

The group(s) is defined on the basis of presence or absence of the exposure to a suspected risk factor for a disease\outcome. All subjects then followed over a period of time to assess the occurrence of that outcome.




### Design of cohort studies

| ((Starting point)) | | Out-come +ve | Out-come -ve | Total |
|---|---|---|---|---|
| **Exposure** | **+ve** | a | b | **Exposed** |
| | **-ve** | c | d | **Not-exposed** |
| **Total** | | **Present** | **Free** | **N** |

### #Strength;

1- Establish the temporal relationship between exposure and outcome.
2- Allow direct measurement of incidence.
3- Examine multiple effects of single exposure.
4- Suitable for rare exposure.
5- Provide information on confounders.

### *Limitations;

1- Expensive and time consuming.
2- Validity of the results can be seriously affected by losses to follow up (bias).
3- Insufficient for the evaluation of rare disease.

The source of exposure data can be obtained from:

1- Pre-existing data.
2- From the study subject.
3- Direct physical examination.
4- Direct investigation.
5- Direct measurement of environment.

**Estimation of risk ((is there an association?))**

In cohort study, we can measure the **incidence** (absolute risk) directly. The incidence among exposed ($I_e$) and incidence among non-exposed ($I_0$).

**Relative risk (RR)** is the measurement of association in cohort study and can be calculated as follow: **(RR) =** ($I_e$)/ ($I_0$).

If RR=1 →No association, If RR>1 → Positive association

If RR<1 →Negative (inverse association, possibly protective).

**Attributable risk (AR)** provides information about absolute effect of the exposure: **(AR)** = ($I_e$) - ($I_0$). **AR** indicate the number of cases among the exposed that can be attributed to the exposure it self.

**Attributable risk percent (AR%):** **(AR)**/ ($I_e$) x 100% , estimate the proportion of the disease among the exposed that can be attributed to the exposure it self.

**Ex:** To determine the relationship between cigarette smoking and CHD, 8000 healthy individual with age > 45 years were enrolled in a study. 3000 of them were smoker. Within 10 years, 84 of the smoker and 87 0f non-smokers develop CHD.

**1- What is the design of the study?** (Cohort study).

**2- Draw 2x2 table**.

| | | **Out-come (CHD)** | | |
|---|---|---|---|---|
| | | **+ve** | **-ve** | **Total** |
| **Exposure (smoking)** | **+ve** | 84 | 2916 | 3000 |
| | **-ve** | 87 | 4913 | 5000 |
| **Total** | | 171 | 7829 | **8000** |

**3- Is there any relation between smoking and CHD?**

We measure the association by estimating the RR = ($I_e$)/ ($I_0$).

$I_e$(CHD in smokers)=(84/3000)x 0.1 = 0.0028 per year

$I_0$(CHD in non- smokers)=(87/5000)x 0.1 = 0.0017 per year

RR=0.0028/0.0017 = 1.6

AR = Ie – I0 = "Risk Difference"

AR = (28.0 – 17.4) / 1000 = 10.6 / 1000

Among SMOKERS, 10.6 of the 28/1000 incident cases of CHD are attributed to the fact that these people smoke…

Among SMOKERS, 10.6 of the 28/1000 incident cases of CHD that occur could be prevented if smoking were eliminated.

AR% = (AR) / I exposed = "Etiologic fraction"

AR% = (28.0 – 17.4) / 28.0 = 37.9%

Among SMOKERS, 38% of the morbidity from CHD may be attributed to smoking…

Among SMOKERS, 38% of the morbidity from CHD could be prevented if smoking were eliminated.

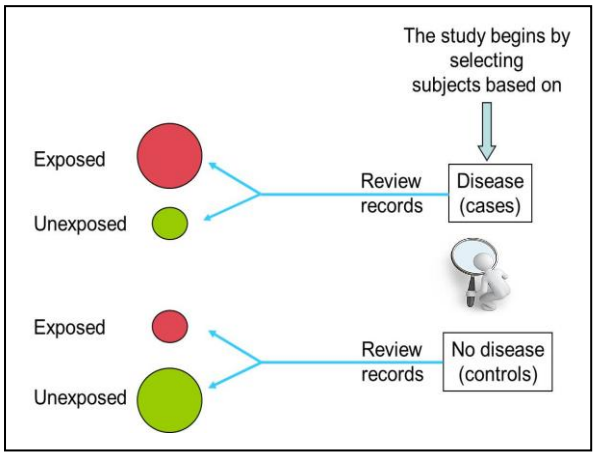
## *Case control study.

Subject are selected on the basis of whether they do (cases) or don't (control) have a particular disease / outcome under study, the groups are then compared with respect to the proportion having a history of an exposure or characteristic of interest.

### Design of case control studies.

**((Starting point))**

| | | Out-come | | Total |
|---|---|---|---|---|
| | | +ve | -ve | |
| **Exposure** | **+ve** | a | b | **Exposed** |
| | **-ve** | c | d | **Not-exposed** |
| **Total** | | **Present** | **Free** | **N** |




### #Strength;

**1-** Quick & inexpensive regarding other analytic studies.
**2-** Depend on already available data.
**3-** Suitable for rare diseases &diseases of long latency period.
**4-** Can examine multiple etiological factors for a single disease e.g. CHD with smoking, diet, physical exercise…..etc.

### *Limitations;

1- Not establish the temporal relationship between exposure and outcome (both exposure and outcome have already occurred.
2- Can't compute the incidence.
3- More prone for bias compared with other analytic studies.
4- In efficient for the evaluation of rare exposure.

### Definition of cases

1- Homogenous disease entity (clear definition of the disease).
2- Depend on strict diagnostic criteria (e.g. MI diagnosed by chest pain more than 30 min. ECG changes and enzyme changes).

### Sources of selection of cases

1- Hospital-based case control study.
- *Advantages*: common, easy and inexpensive.
- *Limitations*: only sever cases of the disease enter the hospital (selection bias).

2- Population based case control.
- *Advantages*: avoid selection bias & describe the disease in the population.
- *Limitations*: difficult, costly and not routinely done.

### Selection of control

It is important to select appropriate control group. The control group should be selected to be comparable to the cases and there is no control group fit for all situations. We have the following types of control:

1- Hospital control;

- *Advantages*: common, easy, more willing to cooperate and inexpensive.
- *Limitations*: They are diseased.

2- General population control.

- *Advantages*: represent healthy population.
- *Limitations*: difficult, costly and not routinely done as it is difficult to contact with healthy individuals.

3- Special (Other) Types of Controls: Friends, Spouses, Siblings/twins

**How many controls? For each control group, how many controls per case?**

✓ The optimal case-control ratio is 1:1

✓ When the number of cases is small, the sample size for the study can be increased by using more than one control e.g. 1:2 1:3 1:4

Matching is a comparative technique of neutralizing all other variables present in cases and controls, except the variable (disease) under study, to eliminate the systematic errors (biases) while conducting the study.

**Estimation of risk ((is there an association?)).**

In case control study we can't calculate the incidence (absolute risk) because we start already disease population (cases) and non disease (control) people. Hence, we use the" **Odd's Ratio"** (OR) which is measure using the following formula: **OR= ad/bc**

**Ex:** To determine the relationship between smoking and CHD, patients with CHD had been compared to patients from orthopedic department; the two groups were matched for age and sex, and were asked about smoking history, the following had been found:

| | | **Out-come (CHD)** | | |
|---|---|---|---|---|
| | | **+ve** | **-ve** | **Total** |
| **Exposure (smoking)** | **+ve** | 112 | 176 | 288 |
| | **-ve** | 88 | 224 | 312 |
| **Total** | | 200 | 400 | **600** |

**1- What is the design of the study? (Case control).**

**2- Is there any relation between smoking and CHD?**

We measure the association by estimating the OR= ad/bc

OR= (112 x224) /(176 x 88) = 1.62 (those who have CHD were 1.62 time more exposed to smoking than those free of disease).

**Note:** OR Approximates RR…

When disease is rare, the proportion of cases in exposed and unexposed groups is low in total (source) population

$a+b \approx b$ and $c+d \approx d \rightarrow$ $RR = a/(a+b)/c/(c+d) \approx a/b / c/d = ad/bc$

## Interventional studies (Clinical trail).

- A design that can provide data of such high quality that it closely resemble the controlled experiments done by basics science researches.
- Similar to cohort study, individual are included on the basis of their exposure status, but the difference is that the investigator himself will allocate the exposure **(active investigator)** while in cohort **(passive investigator).**
- It is the "gold standard" of research designs. Provide the most convincing evidence of relationship between exposure and effect.
- ***Design of randomize trail:***

**Defined population** (Acceptable and eligible group)

↓

**Randomization** (Each individual has the same chance to be included in group 1 or 2).

Group 1: Exposure (New medical or surgical treatment, or preventive measures).

Group 2: No exposure (No or traditional treatment, or placebo)

↓

**Follow up** for a period → Improvement or not → "Compare"

**#Strength:**

1) The temporal relationship between exposure & outcome can be demonstrated with confidence.
2) Many factors and confounders under study can be controlled.
3) Small- moderate differences (10%-20%) which can't be demonstrated by observational studies can be demonstrated in interventional studies.
4) It's the strongest and the most direct epidemiological evidence to judge a causal association.

***Limitations;**

1- The major disadvantage of this study is that this study doesn't represent the real life situation because in real life we can't allow one factor to operate while keeping other factors constant.

2- Expensive and time consuming.

3- Ethical issue: for certain factors, there is some doubt about the benefit or harm to the exposed individuals.

**Selection of study group**

The experimental group should be:

1- Sufficient sample size.
2- Sufficient number of out come or end point (use high risk group e.g. study CHD, we chose male above 40 years old).
3- Need to have accurate and compete follow-up information.

☞ The experimental group should inform about the aim of the study, about the possible side effects or benefits of the agent or may have only placebo→ **Acceptable group**→ screening for eligibility (exclusion of unfit individuals), Causes of exclusion:

1- Definite history of end point under study e.g. MI.

2- Definite need for study treatment e.g. Aspirin for patient with Rheumatoid arthritis.
3- Contra indication for the study treatment e.g. Aspirin in DU.

→ So we have **Acceptable** and **eligible** group → We do "Randomization" to control known and unknown confounders. → Follow up (need compliance)

**Maintenance of Compliance.**

Non-compliance is the major problem in the interventional studies which is related to the ***complexity*** and ***length*** of the follow up. This could be due to:

1- Development of side effect(s).
2- Forgetting to take the treatment.
3- Withdrawal from the trail.
4- The intervention become contra indicated.

**Enhancement of Compliance**

1- Inclusion of interested and high reliable group (high risk group).
2- Frequent contact with participants.
3- Provision of incentive.
4- Implementation of "run in" or "wash out" period before randomization.

**Ascertainment of out come**

It's important to have complete, accurate and similar method of collection of information from both study groups and also a complete follow up of study participants over the duration of the trail. Some end points require short period follow up (e.g. mortality rate after acute MI) → Better contact with all participants during the entire study period. While other require long period of follow up (risk of dying from chronic disease) and may need many years of follow up →Difficult to maintain complete ascertainment.

The ascertainment of outcome can be affected by knowing the treatment status of participants especially if the outcome is subjective (headache, nausea…etc.), this will lead to **"observational bias".** This observational bias can be decreased by doing **blinding** of the exposure status, we have:

1. Single blind design. The investigator alone is aware about the allocation of the intervention.
2. Double blind design. Neither the participants nor the investigator (responsible for the ascertainment of outcome) know to which intervention is allocated.

But blinded trails are usually complex and difficult to conduct and sometime even impossible (evaluation of programs involving change in life style as exercise, cigarette smoking…etc.), so we do un blind trail.

In order to ensure that all the aspects of the program are identical except for actual treatment (experiment), the comparison group is assigned to receive "**Placebo**". Placebo is inert substance in distinguish from active substance.

The use of placebo will minimize the bias in the ascertainment of both subjective and side effect of treatment. If a study doesn't use placebo control, it is impossible to tell whether subjective outcome are due to actual, to extra attention participants receive or merely to their believe that the treatment will help e.g. gastric freezing study in treatment of DU.

**Measurement of association**

Relative risk (RR) "as in cohort" is the measurement of association between exposure and outcome of interest in treated and comparison groups.

**Statistical power of clinical trail**

The statistical power of clinical trail to detect a difference between treatments groups depend on:

1. Sample size; should be sufficiently large.
2. Total number of end points, and to achieve sufficient number of end points we do:
   - Selection of high risk population. E.g. study mortality rate from CHD, we select old age group.
   - Length of follow up period.
3- The difference in compliance between the groups.

**Ex:** An investigator wished to investigate whether giving aspirin in low doses reduces the risk of myocardial infarction (MI). The participants in this study were over 22000 healthy males' physicians who were randomly and equally assigned to receive aspirin or placebo. The study samples were followed over an average period of 60 months. Blinding was done for both patients and clinicians ascertained the outcome. The investigator found that 8 physicians in the aspirin group experienced MI attack during the course of the study comparing with 50 physicians in the placebo group. Regarding the above information answer the following questions:

1. Determine the study design used.
2. Construct 2x2 table for the results showing the exposure and the outcome.
3. Calculate the appropriate measures for risk (per year), and for association.
4. Can we measure the preventable fraction among physicians who take aspirin due to exposure itself?

**Solution:**

1. Randomize double blind clinical trail
2.

| Group | | Outcome | | Total |
|---|---|---|---|---|
| | | MI | No MI | |
| Expos | Aspirin | 8 | 10992 | 11000 |
| | Placebo | 50 | 10950 | 11000 |
| Total | | 58 | 21942 | 22000 |

3- $RR=I_e / I_{-e}$

(8/11000) / (50/11000) = 0.00072/ 0.0042 = 0.17 (protection)

For better understanding: (50/11000) / (8/11000) = 5.8 (those not received aspirin have 5.8 times risk to develop MI than those received)

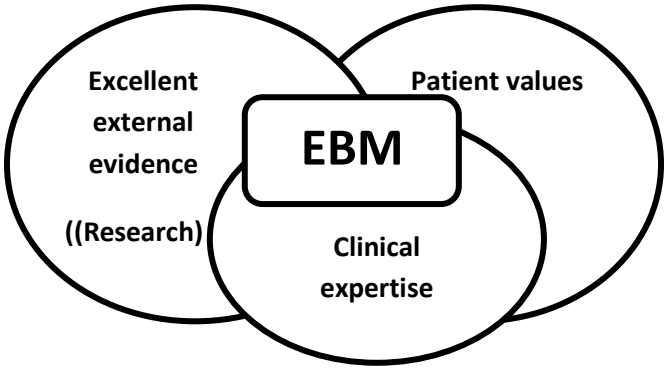
4- AR & AR%

AR = 0.0042 – 0.00072 = 0.00348

AR% = 0.00348 / 0.0042 = 82.85%

- # Evidence Based Medicine (EBM)

  - A Journey from opinion based practice into evidence based practice in medical setting to improve patient care and cure, begins and ends with the patient
  - Physicians must evaluate or critically appraise the evidence; its validity and applicability to the patients
  - We need a scientific tool that helps our experience in induce positive changes in clinical practice according to the will and sake of the patient.
  - Evidence-Based Medicine: The conscientious, explicit and judicious use of current best evidence in making decisions about the care of the individual patient (and population) David Sackett, 1996
  - The integration of best research evidence with clinical expertise and patient values. Sackett et al 2000
  - Evidence based health (medical) practice is the process of systematically reviewing, appraising and using health and medical research findings to aid the delivery of optimum health care in all level primary, secondary and tertiary.
  - EBM aims to optimize decision-making by emphasizing the use of evidence from well designed and conducted research.
  - Weighs three factors to assist with medical decision making
  - Best available clinical evidence
  - Experience of clinician
    1) Patient needs/desires/resources




## Classification for grading evidence

- A++ High-quality meta-analyses, systematic reviews of RCTs, or RCTs with a very low risk of bias
- A+Well-conducted meta-analyses, systematic reviews, or RCTs with a low risk of bias
- B++ High-quality systematic reviews of case-control or cohort studies High-quality case-control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
- B+ Well-conducted case-control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
- C Non-analytic studies; for example, case reports, case series
- D Expert opinion

EBM is concerned with every day clinical issues and questions

- **International Classification Of Diseases 10th Revision (ICD10)**

➢ It is the international standard for defining and reporting diseases and health conditions. It allows the world to compare and share health information using a common language.

➢ The ICD defines the universe of diseases, disorders, injuries and other related health conditions. These entities are listed in a comprehensive way so that everything is covered. It organizes information into standard groupings of diseases, which allows for:

1) easy storage, retrieval and analysis of health information for evidenced-based decision-making;
2) sharing and comparing health information between hospitals, regions, settings and countries; and
3) data comparisons in the same location across different time periods.

➢ It is the diagnostic classification standard for all clinical and research purposes. These include monitoring of the incidence and prevalence of diseases, observing reimbursements and resource allocation trends, and keeping track of safety and quality guidelines.

➢ ICD-10 allows the counting of deaths as well as diseases, injuries, symptoms, reasons for encounter, factors that influence health status, and external causes of disease.

➢ It promotes international comparability in the collection, classification, processing, and presentation of mortality statistics.

➢ Users include physicians, nurses, health workers, researchers, health information managers, policy-makers, insurers and national health programme managers, among others.

## Reverences:


1) Maxey-Rosenau-Last Public Health and Preventive Medicine: 15th Edition by Robert Wallace
2) Epidemiology 3rd Edition) - By Leon Gordis
3) Clinical epidemiology how to do clinical practice research third edition R. brian Haynes
4) A Practical Guide for Health Researchers Mahmoud F. Fathalla WHO Regional Publications Eastern Mediterranean Series 30
5) Hennekens CH, Buring JE. Epidemiology in Medicine. Little, Brown & Co.
6) Barker DJP, Cooper C & Rose CP. Epidemiology in medical practice. Churchill Livingstone
7) Epidemiology and public health; a text and reference book for physicians, medical students and health workers Kindle Editionby Victor Clarence Vaughan (Author)
8) Biostatistics and Epidemiology: a Primer for Health Professionals. S.Wasserthell-Smoller. Springer-Verlag,
9) Clinical Epidemiology and Biostatistics. A primer for Clinical Investigations and Decision Makers. M. Kramer. Springer-Verlag
10) Statistics in Medicine. Colton. Little Brown & Company, Boston