Biostatistics I

By
Prof. Khalil Mohamed Abbas Ayad

Professor of Public Health & Community Medicine

Faculty of Medicine – Tanta University

Second Year

2018-2019

Acknowledgments

This two-year curriculum was developed through a participatory and collaborative approach between the Academic faculty staff affiliated to Egyptian Universities as Alexandria University, Ain Shams University, Cairo University , Mansoura University, Al-Azhar University, Tanta University, Beni Souef University , Port Said University, Suez Canal University and MTI University and the Ministry of Health and Population(General Directorate of Technical Health Education (THE). The design of this course draws on rich discussions through workshops. The outcome of the workshop was course specification with Indented learning outcomes and the course contents, which served as a guide to the initial design.

We would like to thank **Prof.Sabah Al- Sharkawi** the General Coordinator of General Directorate of Technical Health Education, **Dr. Azza Dosoky** the Head of Central Administration of HR Development, **Dr. Seada Farghly** the General Director of THE and all share persons working at General Administration of the THE for their time and critical feedback during the development of this course.

Special thanks to the **Minister of Health and Population Dr. Hala Zayed and Former Minister of Health Dr. Ahmed Emad Edin Rady** for their decision to recognize and professionalize health education by issuing a decree to develop and strengthen the technical health education curriculum for pre-service training within the technical health institutes.

Contents

Overall course aim and intended learning outcomes 8

Chapter (1) : **(uncertainty, research, and statistics)** ---------- 10

Chapter (2) : **Variables and Data** ---------- 22

Chapter (3) : **Scales of Measurement of variables** ---------- 28

Chapter (4) : **Organization and Presentation of Data** ---------- 45

Chapter (5) : **Population and Samples** ---------- 110

Chapter (6) : **The Normal Distribution** ---------- 124

Chapter (7) : **Demography** ---------- 150

Chapter (8) : **Vital Statistics** ---------- 164

Chapter (9) : **Hospital statistics** ---------- 173

References 178

توصيف مقرر دراسى

1- بيانات المقرر		
الرمز الكودى :	اسم المقرر : الإحصاء الحيوي	الفرقة / المستوى : (
التخصص :	عدد الوحدات الدراسية : نظرى 2 عملى	

2- هدف المقرر:	**Provide the students with theoretical and practical contents on statistics to be able to handle efficiently and interpret different types of data relevant to their future career on medical records and disease registries.**
3- المستهدف من تدريس المقرر :	
ا. المعلومات والمفاهيم :	**By the end of the course, the student should be able to :** - Define the basic concept of biostatistics and their relevance to biomedical field. - Define data sources, collection techniques, and measurement scales. - Describe types of variables. - Identify different methods of data summarization, organization and presentation. - Define populations and samples, and random samples selection techniques. - Describe the normal distribution curve and related characteristics (AUC, examining data for normality). - Define demography, census, population pyramid, population growth and some vital statistics.
ب- المهارات الذهنية :	**By the end of the course, the student should be able to :** - Differentiate between methods of data presentation. - Calculate measures of central tendency and dispersion. - Construct frequency distribution table for individual and grouped data. - Apply methods of graphical presentation.

By the end of the course, the student should be able to : - Use computer tools to make relevant library search and be familiar with some statistical packages. - Handle raw data performing calculations on numerical summarization, designing tables and making graphs. - Interpret areas under normal curve (AUC). - Criticize population pyramids and compare between different shapes of them. - Handle different types of hospital statistics.	**ج- المهارات المهنية :**
By the end of the course, the student should be able to : - Develop skills needed for data presentation and making reports. - Work independently or as a part of team demonstrating creativity. - Enhancing team work sprit.	**د- المهارات العامة :**

4- محتوى المقرر :	**Definitions and importance of statistics.** **Sources and techniques of data collection.** **Types and measurements of variables :** - Classifications of variables. - Scales of measurements of variables. - Errors of measurements. **Presentation of data :** - Numerical presentation. - Tabular presentation. - Graphical presentation. **Population and samples.** **Normal distribution curve.** **Demography and vital statistics :** - Census - Population pyramid. - Population growth - Vital indices (morbidity - mortality - fertility) **Hospital statistics :** - Inpatient census. - Measures of efficiency of bed utilization.
5- أساليب التعليم والتعلم :	**Lectures** **Practical activities including :** - **Computer lab. Activities.** - **Assignments (class and home).**
6- أساليب التعليم والتعلم للطلاب ذوى القدرات المحدودة :	• Individual guidance • Individual feedback • Remedial programs

7- تقويم الطلاب :	
أ- الأساليب المستخدمة	MCQ – Problem solving – oral – practical
ب- التوقيت	Every month (Quiz) (at 5th, 9th, and 12th weeks) Mid-term exam End-term exam
ج- توزيع الدرجات	Three quizzes 20% Mid-term exam 20% End-term exam 60%
8- قائمة الكتب الدراسية والمراجع :	
أ- مذكرات	Handouts by the nominated lecturer.
ب- كتب ملزمة	As nominated by the MOH dedicated experts.
ج- كتب مقترحة	Text book of Biostatstic e.g Encyclopedia of Biostatistics (2010), Peter Armitage 2nd edition.
د- دوريات علمية أو نشرات .... الخ	Not recommended at this level.

.

1- Overall Course Aims :

Provide the students with theoretical and practical contents on statistics to make them able to handle efficiently and interpret different types of data relevant to their future career on medical records and disease registries.

2- Intended learning outcomes (ILOs) :

A. Knowledge and Understanding :

By the end of the course, the student should be able to :

a1 : Define the basic concept of biostatistics and their relevance to biomedical field.

a2 : Define data sources, collection techniques, and measurement scales.

a3 : Describe types of variables.

a4 : Identify different methods of data summarization, organization and presentation.

a5 : Define populations and samples, and random samples selection techniques.

a6 : Describe the normal distribution curve and related characteristics (AUC, examining data for normality).

a7 : Define demography, census, population pyramid, population growth and some vital statistics.

a8 : Define hospital statistics and how to measure the efficiency of bed utilization.

B. Intellectual Skills:

By the end of the course, the student should be able to :

b1 : Differentiate between methods of data presentation.

b2 : Calculate measures of central tendency, dispersion and relative standing.

b3 : Construct frequency distribution table for individual and grouped data.

b4 : Apply methods of graphical presentation.

C. Professional and Practical Skills :

By the end of the course, the student should be able to :

c1 : Use computer tools to make relevant library search and be familiar with some statistical packages.

c2 : Handle raw data and performing calculations on numerical summarization, designing tables and making graphs.

c3 : Interpret areas under normal curve (AUC).

c4 : Criticize population pyramids and compare between different shapes of them.

c5 : Handle different types of hospital statistics.

D. General and Transferable Skills :

By the end of the course, the student should be able to :

d1 : Develop skills needed for data presentation and making reports.

d2 : Work independently or as a part of team demonstrating creativity.

d3 : Enhancing team work sprit.

Chapter (1)

Uncertainty, Research, and Statistics

By the end of this chapter, the student should be able to :

- Define statistics, biostatistics.
- Be oriented with the importance of statistics in medical field.
- Recognize the different types of statistical analysis techniques.
- Understand the different sources of data and types of data collection techniques.

How do we go to uncover the truth about things?

We have two tools to pursue scientific inquiry :

1- We have our *senses*, through which we experience the world and make *observations*.

2- We have the *ability to reason*, which enables us to make *logical inferences*.

Clearly, we need both tools ;

- All the individual observations in the world would not in themselves create a theory, and ;
- All the logic in the world is not going to create an observation.

Background :

Prior to the twentieth century, medical research was primarily based on *trial* and *empirical evidence* ;

- *Diseases* and the *risk factors* associated with it were not well understood.
- *Drugs* and *treatments* for diseases were generally untested.

As medicine has moved to become more *evidence based*, biostatistics has become more important and relevant to its practice. It has also become increasingly evident that the interpretation of much of the research in health sciences depends, to a large extent, on biostatistical principles and methods.

What is Biostatistics ?

Biostatistics is a branch of applied statistics that is concerned with the application of statistical methods to *medicine and other biological fields*. It deals with development and application of the most appropriate statistical methods for :

- *Collection* of data.
- *Organiztion* and *Summarization* of collected data.
- *Presentation* and *Analysis* of the summarized data.
- *Interpretation* and **Decision Making** on the basis of analyzed data.

Statistics is the science of :

Gaining **information** from numerical and categorical **data**

Statistical Methods :

They are objective methods by which *group trends* are abstracted from measurements or observations on many *separate individuals*. Statistical methods help us make *scientific* and *intelligent* decisions.

Statistical methods can be used to find answers to the questions like :

- What kind and how much data need to be collected?
- How can we organize and summarize the data?
- How can we analyse the data and draw conclusions from it?
- How can we assess the strength of the conclusions and evaluate their uncertainty?

Statistical Analysis :

Statistical analysis is concerned specifically with ***making sense of data*** and permitting valid conclusions or inferences to be drawn from them. So that, we can obtain *valid* and *defensible* answers to the questions that prompted the research. Statistical analysis enables us *making wise decisions in the face of uncertainty*.

There are two distinct phases of the statistical analysis, the *descriptive* and the *inferential* (analytic) phases.

❑ The Descriptive Phase of Statistics :

After we have run a study, we usually get masses of *raw data*. Usually, we are unable to make any sense of the data in such rough and crude form. A data set in its crude original form is usually very large, consequently, such a data set is not very helpful in drawing conclusions or making decisions.

It is easier to draw conclusions from summary tables and graphs than from the original version of a data set. So, we reduce data size by constructing tables, drawing graphs, or calculating summary measures. The portion of statistics that helps us do this type of statistical analysis is called descriptive statistics, which make up a *small part* of the field of statistics. It provides procedures for organizing, summarizing, and presenting the data in ways that can *easily be communicated* to others.

❑ The Inferential (Analytic) Phase of Statistics :

In statistics, the collection of all elements of interest is called a *population*. The selection of a subset of elements from this population is called a *sample*. A major part of statistics deals with making decisions, inferences and predictions about populations based on results obtained from samples. The area of statistics that deals with such decision-making procedures is referred to as inferential statistics.

Because it is often impractical to survey the *entire population*, we rarely know population values. In inferential statistics, we try to estimate various population values on the basis of the corresponding sample values (i.e generalizing beyond the data).

The population values, such as the population mean and standard deviation, and population range, are referred to as parameters. For every statistic we calculate from sample data, there is a corresponding parameter. We use the known (*statistics derived from a sample*) to reveal the secrets of the unknown (*population parameters*).

Probability, which gives a measurement of the likelihood that a certain outcome will occur, acts as a link between descriptive and inferential statistics. As inferential statistics is based on the probability theory, this should alert us to the fact that inferential statistics never really *proves* anything.

Statistics cannot prove anything, it just put limits to uncertainty.

Estimating parameters from statistics is no more mysterious than judging the state of a person's health from a collection of vital signs. Both reduce uncertainty but both also have the possibility of error. The beauty of statistics is that, it offers procedures for putting known boundaries on the expected error.

- ***A Parameter*** is a value or characteristic associated with a population. Parameters are denoted using Greek letters e.g the population mean (μ) and population standard deviation (δ).
- ***A Statistic*** is a summary numerical value or characteristic associated with a sample. Statistics are denoted using Roman letters e.g the sample mean ($\bar{x}$) and sample standard deviation (s).

Sample statistics are estimates of the corresponding population parameters (e.g $\bar{x}$ is an estimate of μ)

Collection of Data

(Sources – Types - Techniques)

Sources of data :

1- Routinely kept records : *Hospital medical records* for example contain a large amount of data on patients.

2- Surveys : If the data needed to answer a question are not available from routinely kept records, the logical alternative source may be a *survey* or *a census*.

3- Experiments : Frequently the data needed to answer a question are available only as a result of an *experiment* or *clinical trials*.

4- External sources : The data needed to answer a question may already exist in the form of *published reports*, commercially available *data banks* or the *research literature*.

Types of data :

a- Primary data : gathered by the researcher and usually through a survey or research experiment or clinical trial.

b- Secondary data : the data that have been already collected and recorded by somebody else and readily available for others.

Advantages and disadvantages of secondary data :

Advantages

- Faster.
- Less expensive.
- Less activities and efforts (Field trip, Survey etc.).

Disadvantages

- May be not adequate.
- May not meet the specific needs of the researcher.
- Outdated information.
- Variation in definitions.
- Inaccurate or biased.

Data collection techniques :

Data-collection techniques allow us to *systematically* collect data about our elements of a study (people, objects, events) and about the settings in which they occur. We have the following data collection techniques :

(1) Using available records :

Usually there is a large amount of data that has already been collected by others, although it may not necessarily have been analyzed or published.

(2) Observing :

Observation is a technique that involves systematically selecting, watching and recording behavior of the target elements in a study.

(3) Interviewing (face-to-face) :

An interview is a data-collection technique that involves oral questioning of respondents, either individually or as a group. Answers to the questions posed during an interview can be recorded by writing them down or by tape-recording the responses, or by a combination of both. *Checklist* or *questionnaires* are the usual tools.

(4) Administering written questionnaires :

A written questionnaire (also referred to as self-administered questionnaire) is a data collection tool in which written questions are presented that are to be answered by the respondents in written form. A written questionnaire can be administered in different ways, such as by :

- Hand-delivering questionnaires to respondents and collecting them later.
- Sending questionnaires by mail with clear instructions on how to answer the questions.
- Gathering all or part of the respondents in one place at one time, giving oral or written instructions, and letting the respondents fill out the questionnaires ; or

(5) Focus group discussions :

A focus group discussion allows a group of 8 – 12 participants to freely discuss a certain subject or task with the guidance of a facilitator or reporter. The interviewer creates a supportive environment, asking focused questions to encourage discussion and the expression of differing opinions and points of view.

Exercise [1]

1. **The science of collecting, organizing, presenting, analyzing and interpreting data to assist in making more effective decisions is called :**
   (a) Statistic
   (b) Parameter
   (c) Population
   (d) Statistics
2. **Methods of organizing, summarizing, and presenting data in an informative way are called :**
   (a) Descriptive statistics
   (b) Inferential statistics
   (c) Mathematical statistics
   (d) Analytic statistics
3. **The methods used to determine something about a population on the basis of a sample is called :**
   (a) Inferential statistics
   (b) Descriptive statistics
   (c) Applied statistics
   (d) Theoretical statistics
4. **A specific numerical value of characteristic of a population is called :**
   (a) Statistic
   (b) Parameter
   (c) Variable
   (d) Sample

5. **A set of all units of interest in a study is called :**
   (a) Sample
   (b) Population
   (c) Parameter
   (d) Statistic

6. **A part of the population selected for study is called a :**
   (a) Variable
   (b) Data
   (c) Sample
   (d) Parameter

7. **Listings of the data in the form in which these are collected are known as :**
   (a) Secondary data
   (b) Raw data
   (c) Quantitative data
   (d) Qualitative data

8. **Data that are collected by any body for some specific purpose and use are called :**
   (a) Qualitative data
   (b) Primary data
   (c) Secondary data
   (d) Continuous data

9. **The data obtained by conducting a survey is called :**
   (a) Primary data
   (b) Secondary data
   (c) Continuous data
   (d) Qualitative data

10. Routine registration is the source of :

(a) Primary data
(b) Secondary data
(c) Qualitative data
(d) Continuous data

11. Questionnaire method is used in collecting :

(a) Primary data
(b) Secondary data
(c) Published data
(d) True data

12. In inferential statistics, we study :

(a) The methods to make decisions about population based on sample results
(b) How to make decisions about mean, median, or mode
(c) How a sample is obtained from a population
(d) None of the above

13. In descriptive statistics, we study

(a) The description of decision making process.
(b) The methods for organizing, displaying, and describing data.
(c) How to describe the probability distribution.
(d) None of the above.

14. Data in the Population Census Report is :

(a) Grouped data
(b) True data
(c) Secondary data
(d) Primary data

15. Statistic is a numerical quantity, which is calculated from :

(a) Population
(b) Sample
(c) Data
(d) Observations

16. **Which branch of statistics deals with the techniques that are used to organize, summarize, and present the data :**
    (a) Advanced statistics
    (b) Probability statistics
    (c) Inferential statistics
    (d) Descriptive statistics

17. **A parameter is a measure which is computed from :**
    (a) Population data
    (b) Sample data
    (c) Test statistics
    (d) None of the above

18. **You asked five of your classmates about their height. On the basis of this information, you stated that the average height of all students in your university or college is 67 inches. This is an example of :**
    (a) Descriptive statistics
    (b) Inferential statistics
    (c) Parameter
    (d) Population

Chapter (2)

Variables and Data

By the end of this chapter, the student should be able to :

- Know different ways of classifications of variables.
- Differentiate between variables and observations.
- Construct a data matrix.

Variable :

A variable is any characteristic under study or investigation, related to different elements (subjects, objects or events), can be *observed* or *measured* , and is liable to *variation* or *change* (assumes different values for different elements or within elements at different occasions). For example, age, sex, weight, marital status and blood group are variables.

Data :

Are the raw materials and the basic building blocks of statistics. A single observation or measurement is called a datum or a data point. Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

Measurement or Observation :

The data value of a variable for an element (sampling unit) is called measurement or an observation.

Data Set (Database) :

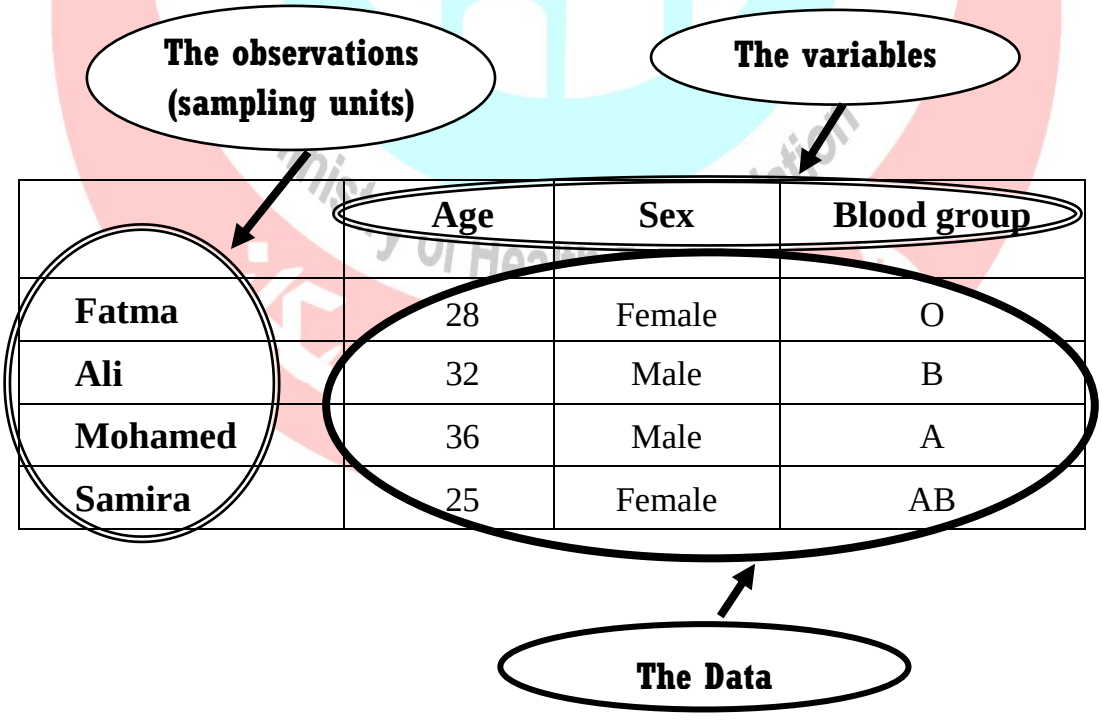
A data set is the collection of observations or measurements on one or more variables for a set of sampling units (elements) from an investigation or survey.

Data Matrix :

Sample data can be presented in a table, which is often called data matrix. In a data matrix, rows usually correspond to observations and columns to variables.

Explain the meaning of : an element, a variable, an observation, data set, data matrix ?

An example of data matrix, 32 years (for the variable age), or female (for the variable sex).



	Age	Sex	Blood group
Fatma	28	Female	O
Ali	32	Male	B
Mohamed	36	Male	A
Samira	25	Female	AB

Classification of Variables :

The first step in any statistical analysis is to identify the type of data (variables) you have. The type of data will determine what kinds of statistics you will be able to use.

Two ways of classification :

- Quantitative versus Qualitative.
- Continuous versus Discrete.

1- Quantitative versus Qualitative

❑ *Quantitative (Numerical) Variables :*

They are variables that yield *measurements* for which the value has numerical meaning (countable or noncountable) i.e numbers represent counts or measurements. For example, we can count the number of cars owned by a family, but we cannot count the height of a family member.

If we want to know someone's weight, we can use a weighing machine, we don't have to look at him and make a guess (which would be approximate), or ask how heavy they are (very unreliable). Similarly, if we want to know the diastolic blood pressure we can use a sphygmomanometer.

Examples of quantitative data include height measured in inches or centimeters, blood pressure (mmHg), age (years), weight (Kgm), heart rate (beats/min), …etc.

Suppose we collect information on the ages (in years) of 50 students selected from a university. The data values, in the order they are collected, are recorded in in the following table. For instance, the first student's age is 21, the second student's age is 19 (second number in the first row), and so forth. The data in such table are *quantitative raw data*.

21	19	24	25	29	34	26	27	37	33
18	20	19	22	19	19	25	22	25	23
25	19	31	19	23	18	23	19	23	26
22	28	21	20	22	22	21	20	19	21
25	23	18	37	27	23	21	25	21	24

❑ *Qualitative (Categorical) Variables :*

Variables that cannot be measured numerically but can be assigned into different categories are called *qualitative* or *categorical* variables. They are variables that yield *observations* on which individuals can be categorized according to some characteristic or quality i.e the value indicates different groupings that are distinguished by some non-numeric characteristics.

Examples include gender (male – female), educational level (illiterate – read and write – high education), marital status (single – married – divorced – widowed), religion, occupation, nationality...etc.

Suppose we ask the same 50 students about their marital status. The responses of the students are recorded in a table. In this table, S, M, D, and W are the abbreviations for single, married, divorced, and widow, respectively. This is an example of *qualitative (or categorical) raw data*.

S	S	M	M	D	W	M	M	D	M
S	S	S	M	S	S	M	S	M	S
M	S	D	S	M	S	S	S	M	D
S	M	M	S	S	M	S	S	S	S
M	S	S	W	D	M	S	W	S	M

2- Continuous versus Discrete

- *Continuous Variables :*

These are variables that can assume any numerical value over a certain interval or intervals. A continuous variable is one with potentially an infinite (unlimited) number of possible values in any interval e.g. height (1.83, 1.74…Cm), weight (48.72, 65.83…Kgm)…etc. Notice that all of these variables can be properly *measured* and have *units of measurement* attached to them. This is a characteristic of all continuous variables.

Contrasted with discrete variables, there are *no gaps* in the real values that a continuous variable may assume. A value of a continuous variable can occur at any point along the scale of values. Weight, height, temperature, and time are commonly used continuous variables.

Note that, if we have any two values of a continuous variable, we can be assured that there are other real values between these two. Values of continuous variables are often expressed as whole numbers, possibly conveying the false impression that they are discrete. We may say, "I am 70 inches tall and weigh 185 pounds as of 5:30 p.m. At this moment, the outside temperature is 28 degrees. All of these

variables are continuous, even though their values are stated in terms of integers. This use of integers is strictly for convenience.

In contrast to *ordinal* values, the difference between any pair of adjacent values is *exactly the same*. The difference between birth weights of 4000 g and 4001 g is the same as the difference between 4001 g and 4002 g, and so on. This property of *real numbers* is known as the interval property. Moreover, a blood cholesterol level, for example, of 8.4 µg/ml is exactly twice a blood cholesterol of 4.2 µg/ml. This property is known as the ratio property.

- *Discrete Variables :*

These are variables whose values are *countable*. In other words, a discrete variable can assume only certain values with no intermediate values. They are variables for which data have distinct categories and a *limited* (finite) number of possible values in any interval.

Counting is the mathematical operation most often used with discrete variables where the data produced are real numbers. The values are usually *whole numbers* (integers) e.g. number of children in a family (two or three children but not 2.5) and number of beds in a hospital. They have the same *interval* and *ratio* properties as continuous data.

Discrete variables *can take values only at specific points along the scale*, such variables *leave gaps* where no real values of the variable are found. Several examples of discrete variables are heart rate (the number of pulses per unit of time), white blood cells in a given sample of blood, and number of chromosomes in a given cell.

All qualitative data are discrete.
Quantitative data may be continuous or discrete.

Chapter (3)

Scales of Measurement

(Scales used to measure variables)

By the end of this chapter, the student should be able to :

- Understand the concept of measurement.
- Be oriented with the different scales of variable measurement.
- Understand and differentiate between the types of errors in data measurement.

Measurement is a way of refining our ordinary observations so that we can *assign numerical values* to it. Measurement allows us to go beyond simply describing the presence or absence of an event or thing to specifying *how much, how long, or how intense* it is. With measurement, our observations become more accurate and more reliable.

Most of us think of measurement in terms of *numerical values* that represent *a quantity* of some sort (such as height, weight, or temperature). Moreover, the numbers representing these quantities can be subjected to *arithmetic manipulations* such as adding, subtracting, multiplying, and dividing.

In reality, however, much of our measuring in daily life does not involve the use of ***precise quantitative*** scales. Think of the ways in which we describe the state of a person's health ;

- At the *most primitive level*, we might distinguish between being alive and being dead.

- At *a more abstract level*, we might say that the person is in better health today than yesterday and is continuing to improve.
- At an *even more abstract level*, we may say that various indicators suggest that the person is in better health than most of the members of his or her cohort.

In each of these examples, we are using *words* to describe an individual's state of health. However, *words are limited in ways that numbers are not*. They are often *subjective* and *imprecise* ;

- If Fatma says she "*often*" has headaches, and Ali complains that he is "*frequently*" has headaches, in which individual do headaches occur with greater frequency?.
- We cannot say, because both words (often, frequently) are imprecise.
- However, if by actual count Fatma experiences 150 headaches in a year compared to Ali 70, we can state, unambiguously that Fatma has headaches more frequently than Ali.

Measurement scales permit us to *substitute numbers for words*, to *perform arithmetic operations* on these numbers, and to *calculate various statistics*. These, in turn, allow us to describe phenomena with great *objectivity* and *precision*.

There are four measurement scales (nominal, ordinal, interval, ratio).

(1) Nominal scale :

The word nominal is derived from the Latin word “name”. There has been some disagreement among experts whether a nominal scale should even be described as a scale. Most would agree that it should. The reason is that we do *name things*, and this *naming* permits us to do other things as a result.

Labeling or naming allows us to make *qualitative distinctions* or to *categorize* and then *count* the frequency of persons, objects, or things in each category. As you will see later, a chi-square statistic is appropriate for data derived from a categorical (nominal) scale.

Nominal scale uses names, numbers or other symbols to distinguish one measurement from another that *can not be ordered* one above the other. Examples include : eye colour, sex, race, blood group, religion, nationality, residence…etc.

With a nominal scale, **numbers** are assigned to objects or events simply for identification purposes. Performing arithmetic operations on these numbers, such as addition, subtraction, multiplication, or division, would not make any sense. The numbers do not indicate more or less of any quantity. A baseball player with the number 7 on his back does not necessarily have more of something than a player identified by the number 1. Other examples include your social security number, your driver’s license number, or your credit card number.

The variable "blood group" is a *nominal* variable. Notice two things about this variable, which is typical of all nominal variables :

- The data do not have any *units of measurement*.
- The categories *cannot be ordered* in any meaningful way.
- *Counting* is the only mathematical procedure used.

Blood groups of 100 patients.

Blood group	Number of patients (Frequency)
O	65
A	15
B	12
AB	8

This table shows how the number, or frequency, of the different blood groups is distributed across the four categories. So, 65 patients have a blood type O, 15 blood type A, and so on. We can't say that being in any particular category is better, or shorter, or quicker, or longer, than being in any other category.

Nominal scales are frequently used in the health sciences. A few examples are scales designed to classify gender, blood group, medical diagnosis, type of treatment (surgery, radiation, chemotherapy), and cause of death.

Typically, counting is the mathematical procedure used with nominal scales. We count the number of patients of each sex, the number with each blood group, the number classified in each diagnostic category, and so on. Because we count or enumerate instances of people, objects, or events that share the characterisics that define the classes, nominal data are often referred to as *enumerative* data and *attribute* data.

Important points about nominal scale :

- A variable measured on a nominal scale may have one, two or more subcategories depending upon the extent of variation. For example, the variable "gender" can have only in two values : male and female (*Alternative*, *Dichotomous* variable). "blood group" can be classified in many sub-categories as A, B, AB, O (*Non-Alternative*, *Polychotomous* variable).
- Meaningful sample statistics include : frequency and relative frequency distribution (how many observations in each class) ; mode (class with most frequent observations).

(2) Ordinal scale :

Has the characteristics of a nominal scale except that measurements *can be arranged in some order*.

- Ordering is *an intrinsic part* of the class definition and *not arbitrary*.
- However, the differences among the ordered categories are *not necessarily equal* (important).
- Examples include : scores, patient status (unimproved, stable, improved), cancer staging, severity of a disease (mild, moderate, severe) .. etc.

The Glasgow Coma Scale (GCS) measures the degree of brain injury following head trauma. A patient's Glasgow Coma Scale score is judged by their responsiveness, *as observed* by a clinician, in three areas : eye opening response, verbal response and motor response.

The GCS score can vary from 3 (death or severe injury) to 15 (mild or no injury). In other words, there are 13 possible values or categories of brain injury.

The Glasgow Coma Scale is an *ordinal* variable. Notice two things about this variable, which is typical of all ordinal variables :

- The data *do not have any units* of measurement (so the same as for nominal variables).
- The ordering of the categories *is not arbitrary* as it was with nominal variables. It is now possible to order the categories in a meaningful way.

A frequency table showing the (hypothetical) distribution of 90 Glasgow Coma Scale scores.

Glasgow Coma Scale Score	Number of patients
3	8
4	1
5	6
6	5
7	5
8	7
9	6
10	8
11	8
12	10
13	12
14	9
15	5

We can say that a patient in the category '15' has less brain injury than a patient in category '14'. Similarly, a patient in the category '14' has less brain injury than a patient in category '13', and so on.

However, there is one additional and very important feature of these scores, (or any other set of ordinal scores). Namely, the difference between any pair of adjacent scores is *not necessarily the same* as the difference between any other pair of adjacent scores.

For example, the difference in the degree of brain injury between Glasgow Coma Scale scores of 5 and 6, and scores of 6 and 7, is not necessarily the same. Nor can we say that a patient with a score of say 6 has *exactly* twice the degree of brain injury as a patient with a score of 12.

Because ordinal data are not real numbers, it is not appropriate to apply any of the rules of basic arithmetic to this sort of data. You should not add, subtract, multiply or divide ordinal values.

Important points about ordinal scale :

- It is not possible to apply any *mathematical operations* to measurements of an ordinal scale.
- The *inappropriate statistical manipulation* and *interpretation* of data generated from ordinal scales can produce a great deal of confusion because of this limitation.
- Meaningful sample statistics include : frequency and relative frequency distribution, median, mode.

(3) Interval scale :

Values are measured on a continuous scale with well-defined units of measurement but *no natural origin* of the scale i.e. the zero is *arbitrary* (not true), so that *differences* between values are meaningful but *not ratios*.

It uses numerical units of measurement where any two successive points are *equally spaced*. This scale has *no true zero point* (i.e zero point on the scale does not represent the true or theoretical absence of the variable being measured). An example of this scale is temperature (0^o is the point at which water freezes and does not represent absence of temperature).

When we can specify both the *order* of events and the *distance* between events, **we have an interval scale**. The distance between any two intervals on this type of scale is *equal* throughout the scale. The central shortcoming of an interval scale is its lack of an absolute (true) zero point, a location where the user can say that there is a complete absence of the variable being measured.

An example may make clear the difference between an arbitrary zero point and an absolute zero point. **Scores on intelligence** tests are considered to be on an interval scale. With intelligence test scores, the anchor point is set at a mean IQ value of 100 with a standard deviation (SD) of 15. A score of 115 is just as far above the mean (one SD) as a score of 85 is below the mean (one SD). Because we have a relative zero point and not an absolute one, it is meaningless to

say that a person with an IQ of 120 is twice as intelligent as a person with an IQ of 60. Some additional examples of interval scales include Calendar date. Is the year 2000 twice as old as the year 1000? The answer is no. Why?

(4) Ratio scale :

Has the characteristics of interval scale but measurements begin at *a true zero point*. Therefor, the *differences and ratio* between values on the scale are meaningful. For example weight (50 Kgm is twice 25 Kgm), height, .etc.

Measurement reaches its highest levels with interval scales and ratio scales because they are truly quantitative. The numbers can be meaningfully added, subtracted, multiplied, and divided. Moreover, with both ratio and interval scales we can state that equal differences in scale values are themselves equal.

The only differences between interval and ratio scales is the fact that the interval scales have an arbitrary zero point, whereas the zero in a ratio scale denotes the absence of the quantity being measured. Indeed, people in northern climes are accustomed to hearing readings below zero during the winter months. However, if we are measuring the blood pressure or pulse rate of a person, a reading of zero means a complete absence of these vital signs. Clearly, a person cannot be below zero in blood pressure or pulse rate.

Because the zero point in a ratio scale is real, we can state meaningful ratios between and among various measurements. Indeed, the scale derives its name from this fact. A pulse rate of 120 is twice as great as a pulse rate of 60, and a white blood cell count of 16,000 is 4 times as great as a count of 4000. We cannot make similar statements with interval scales. A temperature of 98 degrees is not twice as great as a temperature of 49 degrees.

Summary :

- **Nominal -** categories only.
- **Ordinal -** categories with some order.
- **Interval -** differences but not ratios (no natural starting point).
- **Ratio -** differences and ratios (a natural starting point).

Another way for classification of variables :

(1) Qualititative (categorical, classificatory)

- **Binary – dichotomous** □□

  Gender : male or female.

  Employment status : employed or not emploed.

- **Nominal** □□□

  Residency place : Center, North, South, East, West.

  Marital status : single, married, widowed, divorced.

- **Ordinal** ■■□

  Socioeconomic level : high, medium, low.

  Blood glucose level (Nil, +, ++)

(2) Quantitative (Numerical)

- **Discrete**

  Number of offspring : 1, 2, 3, 4.

- **Continuous**

  Glucose in blood level : 110 mg/dl, 145 mg/dl.

Meaning of the measurement scales.

Scale	Characteristic question	Examples
Nominal	**Is A different than B?**	Marital status Eye colour Gender Religion
Ordinal	**Is A bigger than B?**	Stage of disease Severity of pain Level of satisfaction
Interval	**By how many units A and B differ?**	Temperature Calender date IQ test
Ratio	**How many times bigger than A is B?**	Distance Height Weight

Operations that make sense for variables of different scales.

	Operations that make sense			
	Counting	**Ranking**	**Addition / Subtraction**	**Multiplication / Division**
Nominal	+			
Ordinal	+	+		
Interval	+	+	+	
Ratio	+	+	+	+

Errors of Measurement of Variables

So far, we have looked at four types of measurements (nominal, ordinal, interval, and ratio) and at two types of quantitative variables (discrete and continuous). Different types of errors are commonly associated with the use of each of the various measurements.

1- Classification errors :

It is the most frequently errors made with *nominal scales*. This is because nominal scales involve placement of persons, objects, or events into classes or categories.

2- Counting errors :

Because the data for *nominal scales* are collected by counting, counting errors may also be obtained. However, in discrete quantitative variables (i.e interval or ratio scaled) counting errors may also be made.

3- Judgmental errors :

The use of *ordinal scales* typically involves judgment, which is often subjective. Consequently, judgmental errors are most frequently associated with ordinal measurement.

4- Measurement errors :

The variables likely susceptible to this measurement errors are the *continuous variables*. Our measuring instrument may be inaccurate or out of calibration.

5- Rounding errors :

Recall that, with continuous variables, there are no gaps between real values of the variable. At some point, *we must round the value* that we obtain. The last digit in the number reflects this rounding error.

Exercise [2]

1. **When the characteristic being studied is non-numeric, it is called a :**
   (a) Quantitative variable
   (b) Qualitative variable
   (c) Discrete variable
   (d) Continuous variable
2. **When the variable being studied can be reported numerically, the variable is called a :**
   (a) Quantitative variable
   (b) Qualitative variable
   (c) Independent variable
   (d) Dependent variable
3. **The weights of students (Kgm) in a college/school is a :**
   (a) Discrete Variable
   (b) Continuous Variable
   (c) Qualitative Variable
   (d) None of the above
4. **The number of accidents in a city during 2017 is :**
   (a) Discrete variable
   (b) Continuous variable
   (c) Qualitative variable
   (d) Constant
5. **The first hand and unorganized form of data is called :**
   (a) Secondary data
   (b) Organized data
   (c) Primary data
   (d) None of the above

6. **Questionnaire survey method is used to collect :**
   (a) Secondary data
   (b) Qualitative data
   (c) Primary data
   (d) None of the above

7. **A variable that assumes only limited values in a range is called :**
   (a) Continuous variable
   (b) Quantitative variable
   (c) Discrete variable
   (d) Qualitative variable

8. **A variable that assumes any value within a range is called :**
   (a) Discrete variable
   (b) Continuous variable
   (c) Independent variable
   (d) Dependent variable

9. **Colours of flowers are an example of :**
   (a) Quantitative variable
   (b) Qualitative variable
   (c) Skewed variable
   (d) Symmetric variable

10. **Number of family members in different families in a town is an example of a :**
    (a) Discrete variable
    (b) Continuous variable
    (c) Dependent variable
    (d) Qualitative variable

11. Which of the following is an example of nominal data?

(a) Number of people on a course
(c) List of different species of bird visiting a garden over the past week
(d) Popularity rating of UK top ten television programs
(e) Heart rate

12. Which of the following are examples of ratio data?

(a) Number of people on a course
(b) Cancer staging scale
(c) List of different species of bird visiting a garden over the past week
(d) Popularity rating of UK top ten television programs

13. Which of the following are examples of Ordinal data?

(a) Number of people on a course
(b) List of different species of bird visiting a garden over the past week
(c) Popularity rating of UK top ten television programs
(d) Heart rate

14. Select from the following, an example of the categorical variables :

(a) Number of episodes of disease in a patient over a year.
(b) Serum bilirubin level (mg/dL).
(c) Severity of haemophilia (mild /moderate /severe).
(d) Reduction in blood pressure following antihypertensive treatment (mmHg).

15. Select from the following, the variable which can be measured on a nominal scale :

(a) Height in cm.
(b) Ethnic group.
(c) Age categorized as young, middle-aged or old.
(d) Age in years.

16. **Select from the following, the statement which you believe to be true :**
    (a) A nominal variable has categories that can be ordered in some way.
    (b) Quantitative variable takes true numerical values.
    (c) A binary categorical variable can be either nominal or ordinal.
    (d) Nominal data are usually measurements of some type.

17. **For which type of data is the mode the most appropriate descriptive statistic? :**
    (a) Ordinal
    (b) Interval/ Ratio
    (c) Nominal
    (d) Quantitative

18. **A sample of 400 Regina households is selected and several variables are recorded. Which of the following statements is correct? :**
    (a) Total household income (in $) is interval level data.
    (b) Socioeconomic status (recorded as "low income", "middle income", or "high income" is nominal level data.
    (c) The number of people living in a household is a discrete variable.
    (d) The primary language spoken in the household is ordinal level data.

19. **A _____ is a numerical characteristic of a sample and a _____ is a numerical characteristic of a population :**
    (a) Sample, population
    (b) Population, sample
    (c) Statistic, parameter
    (d) Parameter, statistic

Chapter (4)

Organization and Presentation of Data

By the end of this chapter, the student should be able to :

- Be oriented with the different general ways of organization and presentation of data.
- Calcule manually the different measures of numerical data presentation.
- Construct different types of tables using raw data sets.
- Be orientated with uses of different types of graphs according to the type of data.
- Draw different types of graphs.

Useful information is not immediately evident from a set of raw data. Collected data need to be organized, reduced and presented in such a way so that real information may be extracted from them.

There are three general ways of organizing and presenting data :

1- ***Numerical*** (Summary statistics).
   - Measures of Central Tendency (Location).
   - Measures of Variability (Dispersion).
   - Measures of Relative Standing (Quantiles and Outliers).
   - Measures of Shape (Skewness and Kurtosis).

2- ***Tabular*** (Tables).
   - Reference tables.
   - Frequency tables.

- Relative frequency tables.
- Cummulative frequency tables.
- Contingency tables.

3- *Graphical* (Graphs or Charts).

- Pie chart.
- Bar chart.
- Histogram.
- Frequency polygon.
- Scatter diagram.

I- Numerical Methods

(Summary Statistics)

Why do we need both "central tendency" and "dispersion to describe a numerical variable. Example (age)

11 12 13 14 15 16 17 18 19	Mean = 15.0 SD = 2.7	A
7 9 11 13 15 17 19 21 23	Mean = 15.0 SD = 5.5	B

(a) Measures of Central Tendency

These are values around which the measurements of a set of data tend to concentrate. They are used as summary measures for the data series. They include :

- Mean
- Median
- Mode
- Midrange

The Mean ($\overline{X}$) (Arithmetic Mean - Average) :

The mean is the center of gravity (rather than the physical center) of a distribution. It is computed by dividing the sum of all observations in the series (ΣX) by their number (n). The mean of a whole *population* is usually denoted by μ, while the mean of a *sample* is usually denoted by $\overline{X}$.

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

$\overline{X}$ Is pronounced "x-bar" and denotes the mean of a set of sample values.

Σ Which is the capital Greek letter "sigma". It is the sign of summation frequently used in statistics.

x Is the the individual data values of the variable represented by the data set.

n Represents the number of values in a sample.

μ Is pronounced "mu" and denotes the mean of all values in a population.

$$\mu = \frac{\Sigma X}{N}$$

N Represents the number of values in a population.

Characteristics of the Mean :

- The mean is the most commonly used measure of central tendency.
- It is not necessarily equal to one of the sample values.
- The sum of deviations of the values from the mean is equal to 0.

$$\text{i.e } \Sigma (X - \bar{X}) = 0$$

- As calculation of the mean uses every value in the data, the mean is *sensitive* to extreme values where it is pulled to the direction of the extreme values. The median and mode are not so affected? i.e the mean may become very misleading and worrisome as a measure of centrality in a series having extreme values?.

The Median (M_d) :

The median is a measure of *central-ness* (the physical center rather than the center of gravity) of a distribution. It is the value that divides the series into two equal groups after all values have been ordered so that half of the values are greater than and half are less than the median.

***Calculation* :**

- Arrange the values.
- When the number of values (N) is odd, the value of the median is easily determined because there will always be a middle value. The median is the number located in the exact middle of the list.

  The order of the value of the median in this case = $\frac{N+1}{2}$

- When the number of values (N) is even, there will be two middle values. In this case, the mean of these two values is the median.

  The order of the first middle value = $\frac{N}{2}$

  The order of the second middle value = $\frac{N}{2} + 1$

NB :

1- You should differentiate between the order of the value and the value itself because the median is the value itself and not its order.

2- The mean and median will have the same value when a distribution is symmetrical. When a distribution has some extremely high scores (positive skew), the mean will have greater numerical value than the median. If the a distribution has some extremely low scores (negative skew), the mean will be lower in value than the median.

The Mode (M_o) :

The mode is that value or category in the data that has the highest frequency (most frequently occurring i.e. the peak of a distribution). In this sense, the mode is a measure of *typical-ness* or *common-ness*. The mode is completely unaffected by extreme data values. The data series may have no mode or it may have several modes.

The mode is useful in practical epidemiology such as determining the ***peak of disease occurrence*** in the investigation of a disease outbreak.

NB :

1. If two or more measures appear the same number of times, and the frequency they appear is greater than any other measures, then each of these values is a mode ;
2. If every measure appears the same number of times, then the set of data has no mode.

Midrange :

Is a rough estimate of the midpoint for the data set irrespective of the number of values in each half (compared to the median??). It is calculated by adding the lowest and highest data values and dividing by 2. It is affected by extremely high or low data values.

MR = highest value + lowest value / 2

Example (1) :

Suppose that the length of stay for a sample of 10 inpatients were as follows :

1 – 2 – 5 – 5 – 7 – 10 – 14 – 14 – 20 – 30

- We can see that two observations (5 and 14) occurred twice and no other observation occurred more than once. So, this distribution has two modes, 5 and 14.
- Because the number of cases is even, the median will be the average of the two middle cases after all cases have been ranked in order. With 10 cases, the first middle case will be the (N/2) or (10/2) or fifth case. The second middle case is the (N/2) + 1 or sixth case. The median will be the value halfway between the fifth and sixth cases. The value of the fifth case is 7 and the value of the sixth case is 10. The median for these data is (7 + 10) / 2 = 8.5
- The mean is found by first adding up all the observations and then dividing by their number. The sum of the scores is 108, so the mean is 108 / 10 = 10.8

Note that : The mean is a higher value than the median. This indicates a positive skew in the distribution (a few extremely high scores). By inspection we can see that the positive skew is caused by the two patients who stayed more days (20 and 30) in the hospital than the other 8 patients.

Example (2) :

It has been suggested that the body's natural analgesics - the endorphins - might be implicated in the infant apnea syndrome or "near-miss SIDS (sudden infant death syndrome)". In one aspect of this study, B-endorphin levels were obtained from the cerebrospinal fluid of 8 infants who had stopped breathing for a period of at least 20 seconds.

The endorphin levels are shown below in an array, which is a display of a data set arranged in order of magnitude, from low to high.

47 50 52 52 54 66 66 90

The **mean** is calculated as follows : 477/8 = 59.62

To find the **median**, count down to the (n + 1) / 2 th case. In the present example, we count down 9/2 = 4.5. The 4.5th case is halfway between the 4th and the 5th values. The median is therefore 53.

In the present example, there are **two modes** ; a value of 52 and a value of 66. Both occur with a frequency of 2.

N:B

If a data set more than one number occuring more frequent than one, the numbers with equal frequencies are the modes. However, in case of unequal frequencies, the most frequent one is the mode of the data set.

Comparison of Mean, Median, Mode, and Midrange.

Measure of Center	Definition	How Common?	Existence	Takes Every Value into Account?	Affected by Extreme Values?	Advantages and Disadvantages
Mean	$\overline{X} = \frac{\Sigma X}{n}$	Most familiar "average"	Always exists	Yes	Yes	Works well with many statistical methods
Median	Middle value	Commonly used	Always exists	No	No	Often a good choice if there are some extreme values
Mode	Most frequent value	Sometimes used	Might not exist, may be more than one mode	No	No	Appropriate for data at the nominal level
Midrange	$\frac{high + low}{2}$	Rarely used	Always exists	No	Yes	Very sensitive to extreme values

Important considerations in numerical presentation of data :

- When the variable is quantitative with symmetric distribution, then the *mean* is proper measure of center.
- In a case of quantitative variable with skewed distribution, the *median* is good choice for the measure of center.
- The *mode* should be used when calculating measure of center for the qualitative variable.

(b) Measures of Dispersion

Measures of central tendency alone are incomplete summaries of data. Therefore, for full and meaningful description of a distribution, measures of central tendency *should be paired* with measures of dispersion.

Measures of dispersion are characteristics that are used to describe the spread, variation and scatter of a series of values.

- A more formal way of thinking about dispersion is that measures of dispersion *complement* measures of central tendency by telling you something about *how well* a measure of central tendency represents all the scores in a distribution.
- When the dispersion or variability in a set of scores is low, the mean of a set of scores *does a great job* of describing most of the scores in the sample.
- When the dispersion or the variability in a set of scores is high, however, the mean of a set of scores *does not do such a great job* of describing most of the scores in the sample.

The commonly used measures of dispersion include :

- Range.
- Variance (V).
- Standard deviation (SD).
- Coefficient of variation (COV).
- Interquartile range (IQR).

Range (R) :

It is the distance (difference) between the highest and lowest values in a series. It can be calculated by subtracting the lowest value in the series from the highest one.

The range is *a simple and quick* measure of variability. However, the range may be *misleading*. Because it is based on only two observations in the distribution, the highest and the lowest, it provides no information concerning the *scatter within the series*. Occasionally one or both of these measures is so deviant from the rest of the values in a data set that they provide little insight into the overall variability of the values within that set.

Variance (V) :

As dispersion can be considered as the differences between the data values and their average (mean), an appropriate measure of dispersion might be the mean of these differences i.e how far, on the average, each value is from the overall mean. However, this mean difference *is not very useful* since positive differences cancel out negative differences.

> The mean difference = The sum of differences between the data values and their mean / number of observations.

$$\text{The mean difference} = \frac{\sum (X - \overline{X})}{n}$$

One way to overcome the problem of canceling of positive and negative values is use the absolute deviation from the mean |-----| i.e ignore the sign of the difference and treat all negative differences as if they are positive. The resulting measure of dispersion is called the mean deviation.

Mean deviation = The sum of absolute deviations of the data values from the mean / number of observations

$$\textbf{The mean deviation} = \frac{\sum |X - \overline{X}|}{n}$$

Another way to overcome the *problem of canceling* of positive and negative values is to square all the differences. The average of theses squared differences is called the variance.

The variance = The sum of squared deviations of the data values from the mean / number of observations – 1.

$$\textbf{The variance (V)} = \frac{\sum (X - \overline{X})^2}{n-1} = \frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}$$

Unfortunately, the process of squaring the data also squares the units of measurement, and therefore, the variance is measured in square units. To bring the units of measurement back to those of the original data, the square root is taken and the result is called the standard deviation i.e the variance is used *mainly* in the calculation of standard deviation.

Standard deviation (SD) :

It is the positive square root of the variance i.e

$$SD = + \sqrt{V}$$

Standard deviation is the most useful measure of dispersion. When the SD of any sample is small, the sample mean is close to each individual value. The SD decreases when the sample size increases.

Loosely speaking, it's the average ("**standard**") amount by which all the scores in a distribution differ (**deviate**) from the mean of that same set of scores.

$$\mathbf{SD} = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

SD can be considered *as a kind of average of the absolute deviations of observed values from the mean* of the variable in question. However, the standard deviation have its drawbacks, as its values can be *strongly affected* by a few extreme observations.

Coefficient of Variation (COV) :

The coefficient of variation (COV) or *relative standard deviation* (RSD) is the sample standard deviation expressed as a percentage of the mean. It is the ratio of the SD of a series to its mean. It is unitless and is expressed as percentage.

$$COV = \frac{SD}{\overline{X}} \times 100$$

COV is useful when comparing the *relative variation* or spread of the distribution of *different sets of data* specially when *the unit of measurement* used in one set differs from that used in the other.

COV is most useful in comparing the variability of several different samples, each *with different means* as it relates the mean and standard deviation together for each sample. For example, a standard deviation of 10 would reflect something different conceptually if the arithmetic mean were 10 than if it were 1000 !?

Caution must be exercised when using standard deviation as a comparative index of dispersion.

Weights of newborn elephants (Kg)	
929	853
878	939
895	972
937	841
801	826
n = 10, $\overline{X}$ = 887.1, sd = 56.50	

Weights of newborn mice (Kg)	
0.72	0.42
0.63	0.31
0.59	0.38
0.79	0.96
1.06	0.89
n = 10, $\overline{X}$ = 0.68, sd = 0.255	

Incorrect to say that elephants show greater variation for birth-weights than mice because of higher standard deviation.

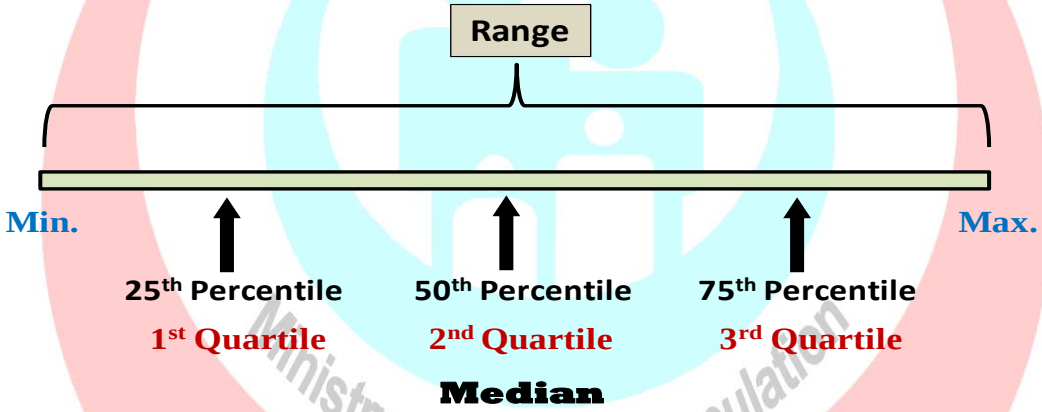
Elephants COV = 0.0637 Mice COV = 0.375

Mice show greater birthweight variation.

The interquartile range (IQR) :

Four commonly used percentiles are the quartiles. Quartiles divide a data set into 4 approximately equal numbers of observations. The sample values have to be ordered from the smallest to the largest. Then 25% of the cases fall below the first quartile (Q1), 50% fall below the median (Q2), and 75% fall below the third quartile (Q3).

The interquartile range is a frequently used measure based on quartiles. It is the difference between the value at the third quartile (Q3) and the value at the first quartile (Q1). Hence 50% of the observations in a data set fall within the interquartile range.



Q1 = (n + 1) / 4 ranked values.

Q2 = (n + 1) / 2 ranked values.

Q3 = 3(n + 1) /4 ranked values.

Calculation :

- To find the first quartile, compute the value 0.25 (n + 1).
- The second quartile found by computing the value 0.5 (n + 1).
- The third quartile found by computing the value 0.75 (n + 1).

Example :

Find the first and third quartiles of the data.

30	75	79	80	80	105	126	138
149	179	179	191	223	232	236	240
242	245	247	254	274	384	470	495

n = 24

To find the first quartile, compute (n + 1) 0.25 = 6.25

(105 + 126) / 2 = 115.5

To find the third quartile, compute (n+1) 0.75 = 18.75

(245 + 247) / 2 = 246

NB :

Of the measures of center and variation, the sample *mean* and the sample *standard deviation* are the most commonly reported. Since their values *depend on the sample selected*, they vary in value from sample to sample. In this sense, they are called *random variables* to emphasize that *their* values vary according to the sample selected. Their values are unknown before *the sample is chosen*. Once the sample is selected and they are computed, they become known sample statistics.

- If the distribution is normally distributed, use Mean (SD).
- If the distribution is skewed, use Median (IQR).

(c) Measures of Relative Standing :

Quantiles :

Are a set of *cut off points* that divide *ranked* data values into groups containing, as far as possible, *equal numbers of observations*.

Examples of quantiles are the following :

- **Percentile** divides the distribution into *one hundred equal groups*.
- **Deciles** divide data into *10 equal groups*, and are the 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, and 90th percentiles.
- **Quintiles** divide data into *5 equal groups*, and are the 20th, 40th, 60th, and 80th percentiles.
- **Quartiles** divide data into *4 equal groups*, and are the 25th, 50th, and 75th percentiles.
- The **median** is the 50th percentile.

Percentiles :

Because the range is often "victim" of the most extreme values, percentiles are frequently used *to express variability in range*. Percentiles are the values which divide an ordered set of data into 100 equal-sized groups, so that as nearly as possible n% of the observations fall **at** or **below** it.

The 10th percentile and the 90th percentile are more likely to provide a stable measure of variability, because they are not based on the two most extreme values. When we have identified the values of the variable at the 5th and 95th percentiles, we can state that 90% of the cases fall between theses two values.

Definition : In general the *p*th percentile of a data set is the value that p% of the observations fall **at** or **below** it, and the rest of observations fall **at** or **above** it. It essentially splits a data set into two parts : lower *p*% of the data values and the upper (1 – *p*)% of the data values.

Calculating a percentile value :

- Order the sample values from smallest to largest.
- Then compute the quantity (p/100) (n + 1), where n is the sample size.
- If this quantity is an integer, *the sample value in this position* is the pth percentile. Otherwise, *average the two sample values* on either side.

As an illustration, suppose you have birthweights for 1200 infants, which you've put in ascending order. If you identify the birth weight that has 1 per cent (i.e. 12) of the birth weight values below it, and 99 per cent (1188) above it, then this value is the 1st percentile. Similarly, the birth weight which has 2 per cent of the birth weight values below it, and 98 per cent above it is the 2nd percentile. You could repeat this process until you reached the 99th percentile, which would have 99 per cent (1188) of birth weight values below it and only 1 per cent above. Notice that this makes the median the 50th percentile, since it divides the data values into two equal halves, 50 per cent above the median and 50 per cent below.

Note, the first quartile is the 25th percentile, the median is the 50th percentile, and the third quartile is the 75^{th} percentile.

Example :

Take the example of the 30 birthweights :

2860	2994	3193	3266	3287	3303	3388	3399	3400	3421	3447	3508	3541	3594	3613
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
3615	3650	3666	3710	3798	3800	3886	3896	4006	4010	4090	4094	4200	4206	4490
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30

The pth percentile is the value in the p/100(n+1)th position. For example, the 20th percentile is the 20/100(n + 1)th value. With the 30 birth weight values, the 20th percentile is therefore the 20/100(30 + 1)th value = 0.2 × 31st value = 6.2th value. The 6th value is 3303 g and the 7th value is 3388g, a difference of 85g, so the 20th percentile is 3303g plus 0.2 of 85g, which is 3303g + 0.2 × 85g = 3303g + 17g = 3320g.

Calculation of IQR :

- The value which cuts off the bottom 25 per cent of values ; this is known as the first quartile and denoted Q1. Compute the value 0.25(n+1).
- The value which cuts off the top 25 per cent of values, known as the third quartile and denoted Q3. compute the value 0.75(n+1)

The interquartile range is then written as (Q1 to Q3). With the birth weight data : Q1 = 3396.25 g, and Q3 = 3923.50 g. Therefore : interquartile range = (3396.25 to 3923.50) g. This result tells you that the middle 50 per cent of infants (by weight) weighed between 3396.25 g and 3923.50 g.

Outliers :

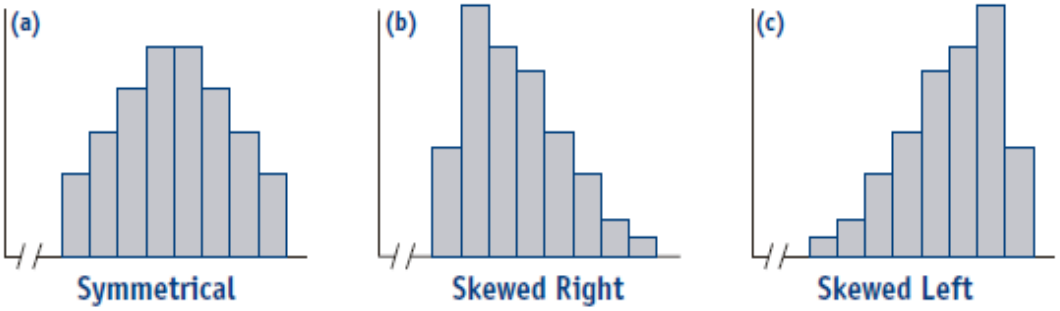
- Sometimes a data set may contain a few points that are *much* larger or smaller than the rest. Such points are called outliers. An outlier is an observation which does not appear to be a part of or belong to the population of interest.
- Outliers can arise because of a measurement or recording error or because of equipment failure during an experiment, etc.
- An outlier might be indicative of a sub-population e.g. an abnormally low or high value in a medical test could indicate presence of an illness in the subject.

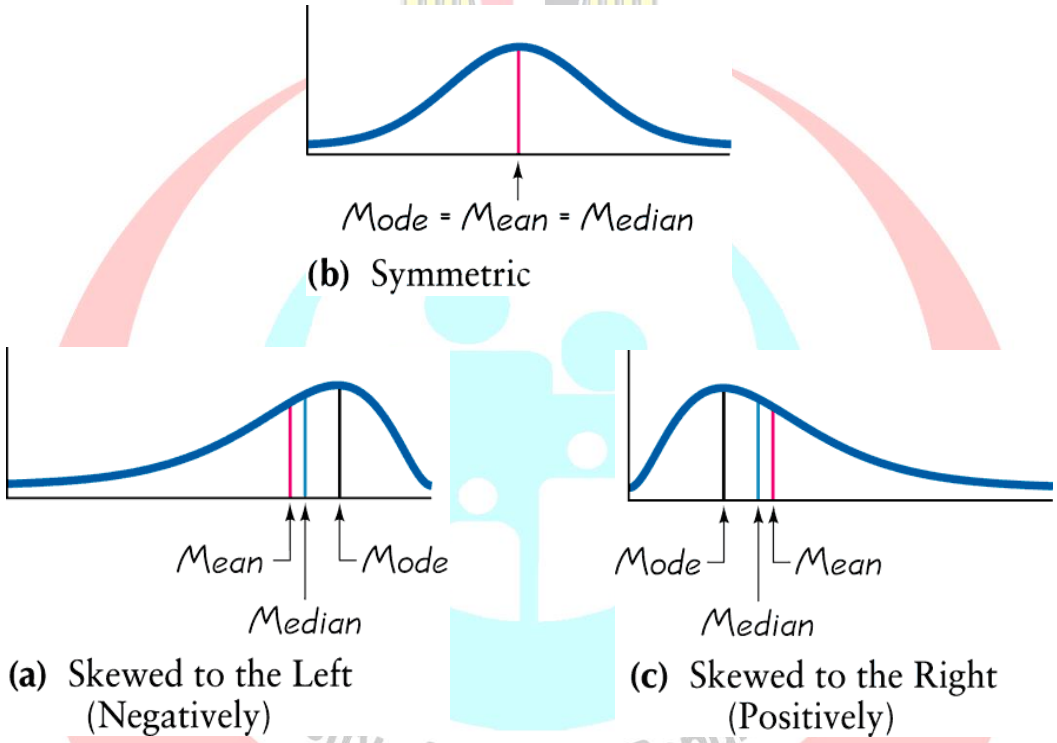
(d) Measures of Shape :

1- Measures of Symmetry (Skewness)

Skewness refers to the degree of asymmetry of the distribution of the variable. The skewness statistic measures the lack of symmetry in a curve. Skewness of 0 implies that the distribution is symmetric.

- When the median and the mean are different, the distribution is skewed. The greater the difference, the greater the skewness.
- Distributions that trail away *to the left* are negatively skewed and those that trail away *to the right* are positively skewed.
- If the skewness is extreme, the researcher should either transform the data to make them better resemble a normal curve or else use a different set of statistics, non-parametric statistics, to carry out the analysis.

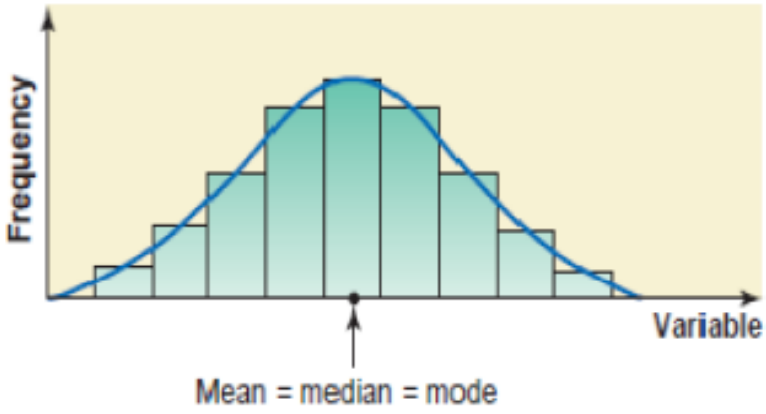




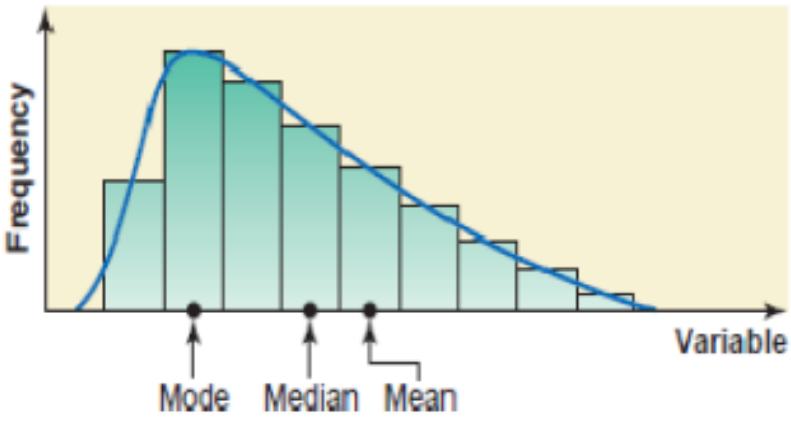
Coefficint of skewness :

Karl Pearson developed the coefficient of skewness to measure the amount and direction of skewness. Computer programs calculate the skewness coefficient for you. This can vary between –1 (strong negative skew), and +1 (strong positive skew). Values of zero or close to it, indicate lower levels of skewness, but do not necessarily mean that the distribution is symmetric.

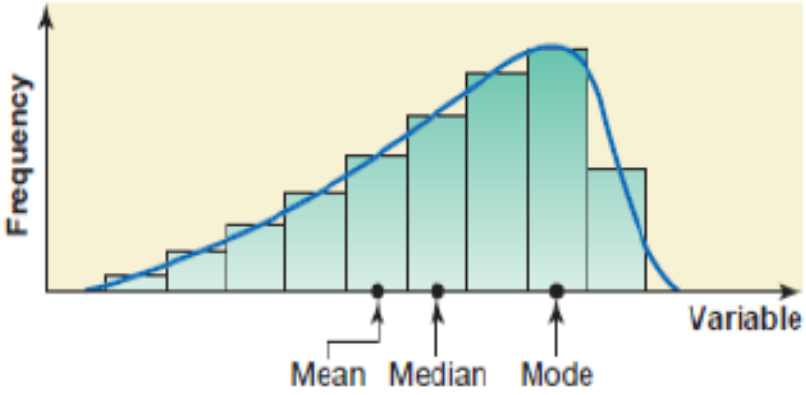
$$\text{Coefficient of skewness} = \frac{3(mean - median)}{SD}$$



Mean, median, and mode for a symetric histogram and frequency distribution curve.



Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.

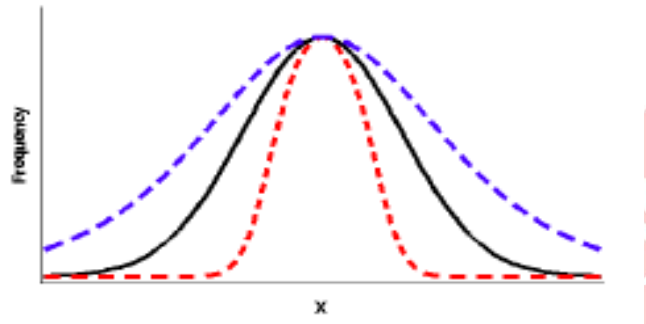


Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.

2- Measures of Peakness/Flatness (Kurtosis) :

Kurtosis refers to the 'peakedness' of the distribution. It is the extent to which scores, are concentrated close to the mean (a "peaked" distribution) or are present in the tails (a "flat" distribution). The standard normal distribution has a kurtosis of 3.

The three curves in the figure below illustrate kurtosis. The curve with the solid line is a normal curve **(Mesokurtotic)**. The curve with the short dashed lines illustrates a peaked distribution known in statistics as a **leptokurtotic** distribution. A leptokurtotic distribution has relative more scores close to the mean and relatively fewer scores in the tails of the distribution. The curve with the longer dashed lines is a **platykurtotic** curve, meaning broad or flat. Here, there are relatively more scores in the tail and relatively few scores close to the mean of the distribution.



Kurtosis - amount of peakedness or flatness of the distribution :

Mesokurtic : normal.

Leptokurtic : peaked, many scores around middle.

Platykurtic : flat, many scores dispersed from middle.

Exercises [3]

1. **If a series of numbers consists of 21 ordered values, the median is :**
   (a) The 11th value in the ordered series.
   (b) The mean between the 10th and 11th values.
   (c) The mean between the 11th and 12th values.
   (d) The 10th value in the ordered series.

2. **The median of a series of numerical values is :**
   (a) Value for which half of the values are higher and half of the values are lower.
   (b) The value located exactly midway between the minimum and maximum of the series.
   (c) The most commonly encountered value among the series.
   (d) A measure of the eccentricity of the series.

3. **The interquartile range includes the following scores? (one correct choice)**
   (a) 50% of the un-ranked scores
   (b) 25% of the ranked scores
   (c) 75% of the ranked scores
   (d) 50% of the ranked scores

4. **Which of the following Greek letters represents the mean of a population? (one correct choice)**
   a. β
   b. α
   c. μ
   d. σ

5. **Sigma squared represents? (one correct choice)**
   (a) Population variance
   (b) Sample standard deviation
   (c) Population standard deviation
   (d) Sample variance

6. **The mean of a distribution is 14 and the standard deviation is 5. What is the value of the coefficient of variation? :**
   (a) 60.4%
   (b) 48.3%
   (c) 35.7%
   (d) 27.8%

7. **The mean of a distribution is 23, the median is 24, and the mode is 25.5. It is most likely that this distribution is :**
   (a) Positively Skewed
   (b) Symmetrical
   (c) Positively skewed
   (d) Negatively skewed

8. **The middle value of an ordered array of numbers is the :**
   (a) Mode
   (b) Mean
   (c) Median
   (d) Mid Point

9. **Which of the following divides a group of data into four subgroups?:**
   (a) Percentiles
   (b) Deciles
   (c) Median
   (d) Quartiles

10. If a distribution is abnormally tall and peaked, then is can be said that the distribution is :

(a) Leptokurtic
(b) Pyrokurtic
(c) Platykurtic
(d) Mesokurtic

11. The measure of location which is the most likely to be influenced by extreme values in the data set is the :

(a) Range
(b) Median
(c) Mode
(d) Mean

12. Which one of the following measurements does not divide a set of observations into equal parts?

(a) Quartiles
(b) Standard Deviations
(c) Percentiles
(d) Deciles

13. Which one is the not measure of dispersion :

(a) The Range
(b) 50th Percentile
(c) Inter-Quartile Range
(d) Variance

14. What is the main difference between the median and mean?

(a) The median uses the ranked values whereas the mean uses the frequencies
(b) The median uses the ranked values whereas the mean uses the actual values
(c) The mean uses the ranked values whereas the median uses the actual values
(d) There is no difference

15. When calculating the median for a dataset consisting of an even number of scores, which of the following is correct? :

(a) Calculate the average value of the three middle ranked scores

(b) Calculate the average value of the two middle ranked scores

(c) Calculate the mode and use instead

(d) Choose either the upper or lower value of the two

16. Which of the following statements concerning the mean is incorrect ? :

(a) The mean is not suitable for nominal data

(b) The mean should always be used as the preferred measure to indicate a typical score

(c) The mean is a more complex descriptive statistic than either the mode or median

(d) The mean provides the most sensible result when the interval/ratio dataset has a symmetrical set of scores.

17. The mean can be interpreted as :

(a) The centre of gravity of a dataset.

(b) The average of the mode and median values of a dataset.

(c) The weight of all the positive deviations.

(d) The relative frequency with the highest value.

II- Tabular Presentation

The purpose of tabulation is to arrange observations that are similar so that *their frequency of occurrence* in the whole group is made clear. A table, if properly constructed, can show *at a glance* many of the properties of the data.

The necessary parts of a table :

1- **The Title :** Which usually appears above the table and give a clear and accurate description of the data. It should be *concise* but *informative* with *a clearly worded statement* declaring what the table shows. The table is usually preceded by an identifying number if two or more tables are presented. It is better to avoid using verbs in the text of the title.

2- **Headings :**

a- Heading of columns (general heading) : which includes all of the column heading and subheadings except the heading of rows and row totals.

b- Heading of rows (heading of stub) : Which is the box at the top of the column at the extreme left of the table (in English system). The heading gives a brief description of the variable which changes in the value from row to row.

3- **The Body :** It is all the rest of the table proper i.e the part which contains the numbers (values).

4- **Totals :** Table may have one or more boxes for totals (row, column and grand totals).

Table (…) : The title ……..

Heading of stub	General heading		Total
	Subheading	Subheading*	
stub	Body		Row total
			Row total
			Row total
Total	Column total	Column total	**Grand total**

*** (Notes)**

Important considerations in the design of tables :

- Column and row headings and subheadings should be clear and comprehensive i.e descriptions of the values or categories taken by the variable concerned.
- The units of measurements should be given for all entries in the table, either in the title or in the row / column headings.
- Notes should be used, where appropriate, to give the source of data, detailed specifications and definitions of terms.

Types of tables :

- Reference tables (data matrix)
- Frequency tables (simple frequency, realtive frequency, cumulative frequency, relative cumulative frequency).

❑ Reference Tables [Data Matrix] :

Table (1) : Data from a Health Center Survey.

Patient	Sex*	Marital status	Satisfaction with services**	Age
A	M	Single	4	18
B	M	Married	2	19
C	F	Single	4	18
D	F	Single	2	19
E	M	Married	1	20
F	M	Single	3	20
G	F	Married	4	18
H	F	Single	3	21
I	M	Single	3	19
J	F	Divorced	3	23
K	F	Single	3	24
L	M	Married	3	18
M	F	Single	1	22
N	F	Married	3	26
O	M	Single	3	18
P	M	Married	4	19
Q	F	Married	2	19
R	M	Divorced	1	19
S	F	Divorced	3	21
T	M	Single	2	20

* (M = Male, F = Female).

** (1) Very dissatisfied (2) Dissatisfied (3) Satisfied (4) Very satisfied.

❑ Frequency tables :

Frequency tables are the most commonly used and most convenient way of summarizing data. In a frequency table, the *sample values* are distributed or classified into groups according to certain properties. Frequency tables allow us to get an overall view of where measurements or observations are concentrated and how spread out they are. A frequency distribution is a tabular summary of data showing the frequency of observations in each of several *non-overlapping* categories or classes.

Frequency Distributions :

A sample of 100 students enrolled at a university were asked what they intended to do after graduation. Forty-four said they wanted to work for private companies/businesses, 16 said they wanted to work for the federal government, 23 wanted to work for state or local governments, and 17 intended to start their own businesses. Table 2.3 lists the types of employment and the number of students who intend to engage in each type of employment. In this table, the variable is the type of employment, which is a qualitative variable. The categories (representing the type of employment) listed in the first column are mutually exclusive. In other words, each of the 100 students belongs to one and only one of these categories. The number of students who belong to a certain category is called the frequency of that category. A frequency distribution exhibits how the frequencies are distributed over various categories. Table 2.3 is called a frequency distribution table or simply a frequency table.

Table 2.3 Type of Employment Students Intend to Engage In

Variable → Type of Employment	Number of Students ← Frequency column
Private companies/businesses	44
Category → Federal government	16 ← Frequency
State/local government	23
Own business	17
	Sum = 100

An example of a simple frequency table (nominal variable) :

Category (hair color)	Frequency (number of children) (n = 95)
Brown	49
Dark	27
Blonde	15
Red	4

As you know, the ordering of nominal categories is arbitrary, and in this example they are shown by the number of children in each - largest first.

An example of a simple frequency table (ordinal variable) :

Satisfaction with nursing care	Frequency (number of patients) (n = 475)
Very satisfied	121
Satisfied	161
Neutral	90
Dissatisfied	51
Very dissatisfied	52

The frequency distributions for the ordinal variable 'level of satisfaction', with nursing care by 475 psychiatric in-patients.

When the variable in question is ordinal, we can allocate the data into ordered categories. 'Level of satisfaction' is clearly an ordinal variable. 'Satisfaction' cannot be properly measured, and has no units. But the categories can be meaningfully ordered, as they have been here. The frequency values indicate that more than half of the patients were happy with their psychiatric nursing care, 282 patients (121+161), out of 475. Much smaller numbers expressed dissatisfaction.

❑ Relative frequency tables :

Often of more use than the actual number of subjects in each category are the percentages. Tables with this information are called relative or percentage frequency tables. A relative frequency table can be constructed as follows :

- Divide each class frequency by its total and multiply the result by 100.
- Check the arithmetic by ensuring that the addition of the percentages of all of all frequencies of class intervals equals 100.

An example of a relative frequency table :

Category (hair color)	Frequency (number of children) (n = 95)	Relative frequency (% of children in each category)
Brown	49	51.6
Dark	27	28.4
Blonde	15	15.8
Red	4	4.2

(49/95) x 100 = 51.6

An example of a relative frequency table (categorized quantitative variable) :

SBP (mm Hg)	Frequency	Relative frequency
100 -	5	16.7
120 -	11	36.7
140 -	12	40.0
160 -	1	3.3
180 – 199	1	3.3
Total	**30**	**100.0**

Percentages are easier to read and comprehend than frequencies. This advantage is particularly obvious in comparing groups of different sizes as shown in table (2).

Table (2) : Relative frequency distribution of patient's satisfaction in two hospitals.

Satisfaction	Hospital A		Hospital B	
	No.	%	No.	%
Very satisfied	103	24.82	312	31.33
Satisfied	82	19.76	279	28.01
Dissatisfied	137	33.01	188	18.88
Very dissatisfied	93	22.41	217	21.78
Total	**415**	**100.00**	**996**	**100.00**

Q– Based on the data presented in table (2), which hospital has the higher relative number of dissatisfied patients?.

A : Because the total numbers are so different, such comparison is difficult from the cell frequencies. To make comparisons easier, cell frequencies are transformed into percentages for both distributions.

It is clear from table (2) : Inspite of the fact that Hospital B has an absolute number of dissatisfied patients (188) which is greater than that of Hospital A (137), yet Hospital A has a much higher percentage of dissatisfied patients (33.01%) than that of Hospital B (18.88%) and about the same percentage of very dissatisfied patients.

Dimensions of frequency tables :

(I) Single variable (one dimension) frequency table :

It describes the distribution of a single characteristic or variable and can be constructed as follows :

A- <u>For Qualitative Variables</u> (*nominal* and *ordinal*) :

For each category of the variable being displayed, the frequencies or occurrences are counted.

Table (3) displays a frequency distribution for the variable "sex" from the health center survey. For purposes of illustration, a column for tallies has been included in this table which <u>would not be included</u> in the final form of the table.

Table (3) : Frequency distribution of patients according to sex.

Sex	Tallies	Frequency
Male	𝍸 𝍸	10
Female	𝍸 𝍸	10
Total		**20**

The meaning of the table is clear. There are 10 males and 10 females in the sample.

For some nominal variables, the researcher might have to make some choices about the <u>number of categories</u> included in the table. For example, the distribution of the variable "marital status" could be reported using the categories listed in table (4).

Table (4) : Frequency distribution of patients according to marital status.

Marital status	Frequency
Single	10
Married	7
Divorced	3
Total	**20**

It might be wanted to compare between "married" and "non married", where the researcher may not be concerned with the difference between single and divorced and treat both as simply "non married". In that case, these categories could be collapsed and treated as a single entity. However, by collapsing information details have been lost, as in table (5).

Table (5) : Frequency distribution of patients according to marital status.

Marital status	Frequency
Married	7
Non married	13
Total	**20**

Frequency distributions for ordinal variables are constructed following the same steps used for nominal variables. Table (6) reports the frequency distribution of the "Satisfaction" variable from the health center survey.

Table (6) : Frequency distribution of patients according to their satisfaction.

Satisfaction	Frequency
Very satisfied	4
Satisfied	9
Dissatisfied	4
Very dissatisfied	3
Total	**20**

Again, the price paid for collapsing information details is that some information is lost as in table (7).

Table (7) : Frequency distribution of patients according to their satisfaction.

Satisfaction	Frequency
Satisfied	13
Dissatisfied	7
Total	**20**

B- <u>For Quantitative Variables</u> (*interval* and *ratio* variables) :

Quantitative variables usually have a large number of possible values which require some collapsing or grouping of categories to produce reasonably <u>compact frequency distributions</u>.

For construction of a frequency distribution table for a quantitative variable you must :

- Group a set of observations into a set of <u>non overlapping</u> intervals so that each value in a set of observations can only be placed in one of the intervals. These intervals are usually referred to as **class intervals**.
- Too few class intervals are undesirable because many frequencies will be crowded in each class and so much information would be lost. On the other hand, too many intervals will contain a few or no frequencies and the *objective of summarization* would not be achieved.
- Determine how many categories (class intervals) to use and how wide these categories should be. A commonly followed rule of thumb states that there should be no fewer than 4_intervals and no more than 12.
- Class intervals should be of the *same width* to help in *comparison* between the frequencies of any two intervals. The width (W) of the class interval may be determined by dividing the range (R) of the data set by the desired number of class intervals (K).

$$W = \frac{R}{K}$$

- Class interval widths of 5 units or multiplies of 5 tend to make summarization more comprehensive.
- It is preferred to order class intervals from smallest to largest measurements. The lower limit of the first class interval should be equal to or smaller than the smallest observation in the data set, and the upper limit of the last class interval should be greater than the largest observation.

Important considerations :

- Constructing frequency tables for ***discrete*** data is often less problematic than with continuous data, because the number of possible values (categories) which the discrete variable can take is often *limited*.
- Organizing raw ***continuous*** data into a frequency table is usually impractical, because there are theoretical an infinite number of possible values (observations). The corresponding frequency table is likely to have a large, and thus unhelpful, number of rows.
- The most useful approach with continuous data is to group them first, and then construct a frequency distribution of the grouped data.

Example [Reference Table] :

Infant ID (n = 30)	Birthweight (g)	Apgar score	Sex	Mother smoked during pregnancy	Mother's parity
1	3710	8	M	No	1
2	3650	7	F	No	1
3	4490	8	M	No	0
4	3421	6	F	Yes	1
5	3399	6	F	No	2
6	4094	9	M	No	3
7	4006	8	M	No	0
8	3267	5	F	Yes	5
9	3594	7	F	No	2
10	4206	9	M	No	4
11	3508	7	F	No	0
12	4010	8	M	No	2
13	3896	8	M	No	0
14	3800	8	F	No	0
15	2860	4	M	No	6
16	3798	8	F	No	2
17	3666	7	F	No	0
18	4200	9	M	Yes	2
19	3615	7	M	No	1
20	3193	4	F	Yes	1
21	2994	5	F	Yes	1
22	3266	5	M	Yes	1
23	3400	6	F	No	0
24	4090	8	M	No	3
25	3303	6	F	Yes	0
26	3447	6	M	Yes	1
27	3388	6	F	Yes	1
28	3613	7	M	No	1
29	3541	7	M	No	1
30	3886	8	M	Yes	1

Among the 30 infants in the table above, there are none with the same birth weight, and a frequency table with 30 rows and a frequency of 1 in every row would add very little to what you already know from the raw data (apart from telling you what the minimum and maximum birth weights are).

One solution is to group the data into groups of equal width, to produce a *grouped frequency distribution* which would be certainly worthwhile if there is enough data values.

Grouped frequency distribution table for birthweight of 30 infants.

Birthweight (gm)	Birthweight (g)
2700 - 2999	2
3000 – 3299	3
3300 - 3599	9
3600 - 3899	9
3900 - 4199	4
4200 - 4499	3

Open-ended groups :

One problem arises when one or two values are too away from the general mass of the data, either much lower or much higher. These values are called outliers. Their presence can mean having lots of empty or near-empty rows at one or both ends of the frequency table.

For example, one infant with a birthweight of 6050 g would mean having five empty cells before this value appears. One favoured

solution is to use open-ended groups. If you define a new last group as ≥ 5000 g, you can record a frequency of 1 in this row, and thus incorporate all of the intervening empty groups into one.

Frequency diatribution table for discrete quantitative variable :

Number of times inhaler used in the past 24 hours	(Frequency) Number of children (n = 53)
0	6
1	16
2	12
3	8
4	5
≥ 5	6

Constructing frequency tables for quantitative discrete data is often less problematic than with continuous metric data, because the number of possible values which the variable can take is often limited. The above table is a frequency table showing the number of times in the past 24 hours that 53 asthmatic children used their inhaler. We can easily see that most used their inhaler once or twice. Notice the open-ended row showing that six children had used their inhaler five or more times.

Example :

The systolic and diastolic blood pressure as recorded in a sample of 30 patients are presented in table (8) [Reference Table] :

Table (8) : Systolic and Diastolic Blood Pressures of 30 patients.

Patient No.	**SBP** (mmHg)*	**DBP** (mmHg)**
1	125	78
2	104	72
3	133	82
4	153	84
5	117	70
6	142	85
7	194	116
8	134	87
9	119	69
10	120	75
11	144	81
12	128	87
13	135	76
14	118	78
15	137	82
16	151	85
17	148	84
18	121	74
19	134	77
20	151	85
21	146	76
22	141	94
23	169	101
24	152	88
25	138	87
26	140	89
27	156	98
28	152	99
29	124	77
30	114	81

* Systolic blood pressure. ** Diastolic blood pressure.

The frequency table that describes the distribution of systolic blood pressures is constructed as follows :

1- The range is (194 – 104) + 2 = 92.

2- A table of 5 or 6 intervals seems reasonable.

3- Determine the width of each class interval as follows :

92 / 5 = 18.4 , Therefore, a class interval size of 20 is suitable.

Table (9) : Frequency distribution of systolic blood pressure.

SBP (mm Hg)	Tallies	Frequency
100 -	𝍸	5
120 -	𝍸 𝍸 /	11
140 -	𝍸 𝍸 //	12
160 -	/	1
180 – 199	/	1
Total		**30**

Q : Could you construct a frequency distribution table for diastolic blood pressure in the same way as systolic blood pressure in the previous example?

Important considerations in construction of tables for continuous variables :

- A frequency distribution is a distribution of the total number of observations over an arbitrarily defined classes.
- That is, we divide the overall range of values into a number of classes and count the number of observations that fall into each of these classes or intervals.
- The number of observations falling under a class is called class frequency.
- There should not be too few or too many classes.
- As possible, equal class intervals are preferred. But the first and last classes can be open-ended to enclose extreme values.
- Each class should have a class mark to represent the classes. It is also named as the class midpoint. It can be found by taking simple average of the class boundaries or the class limits of the same class.
- In practice, the number of classes and the appropriate class width are determined by trial and error.
- Once a possible number of classes is chosen, the appropriate class width is found.
- The process can be repeated for a different number of classes.
- Ultimately, the analyst uses judgment to determine the combination of the number of classes and class width that provides the best frequency distribution for summarizing data.

(II) Two dimensional frequency tables (contingency tables) :

Contingency tables are designed *to study the relationship between two variables* such as height and weight. The rows correspond to one variable and the columns to the other variable.

For construction of a contingency table proceeds as follows :

- Determine the suitable number of class intervals and the width of each for both variables. The two variables need not have the same number of intervals.
- Enter a tally stroke for each case into the box that corresponds to its value and this step is repeated for each variable separately.
- Add up the number of tally strokes in each box for rows and columns where the sum of all the row totals must equal the sum of all the column totals because each of these should equal the grand total of cases in the contingency table.

Example :

The above simple process can be illustrated by an example using the previously given 30 cases. The relationship between diastolic and systolic blood pressure can be obtained as follows :

- Compute the range for DBP (69 to 116) and SBP (105 to 195).
- For such a small sample, 5 or 6 classes for each variable seems suitable.
- Compute the class interval for each variable by dividing the corresponding range by the proposed number of classes (5 or 6).
- Construct the rows and columns corresponding to the class intervals computed in the previous step.

- Transfer the data to the contingency table using tally strokes (as seen in table 9).

Table (10) : Frequency distribution of Systolic and Diastolic blood pressures using the previously given 30 patients.

SBP (mm Hg)	DBP (mm Hg)						Total
	65-	75-	85-	95-	105-	115-	
100 -	///	//					5
120 -	/	//// //	///				11
140 -		////	//// /	//			12
160 -				/			1
180 – 199						/	1
Total	**4**	**13**	**9**	**3**	**0**	**1**	**30**

Interpreting Contingency Tables :

- If there is a strong association or relationship between the two variables, then the association can be seen from the cell frequencies. Where both variables tend to have values in the same direction (one of them is large and the other tends to be large also).
- When the association is in the opposite direction (one variable has a large value and the other variable tend to have a small value) then the concentration of frequencies will be in the diagonal boxes but in the opposite direction than before.
- If there is no association, or only weak association, between the variables then the cell frequencies are distributed more widely throughout most of the cells with little or no concentration along the diagonal cells.

❑ Cumulative Frequency Distribution :

To do this, we successively add, or cumulate, the frequency values one by one, starting at the top of the column. The cumulative frequency for each category tells us how many subjects there are in that category, and in all the lesser-valued categories in the table. For example, 35 of the total of 154 patients had a GCS score of 7 or less.

The Glasgow Coma Scale scores of 154 road traffic accident patients

GCS score	Frequency (number of patients) (n = 154)
3	10
4	5
5	6
6	2
7	12
8	15
9	18
10	14
11	15
12	21
13	13
14	17
15	6

A cumulative frequency table provides us with a somewhat *different view of the data*. Note that although you can legitimately calculate cumulative frequencies for both metric and ordinal data, it makes no sense to do so for nominal data, because of the *arbitrary* category order.

The Glasgow Coma Scale scores showing the cumulative frequency values.

GCS score	Frequency (number of patients) n = 154	Cumulatative frequency (cumulative number of patients)
3	10	10
4	5	15
5	6	21
6	2	23
7	12	35
8	15	50
9	18	68
10	14	82
11	15	97
12	21	118
13	13	131
14	17	148
15	6	154

Cumulative frequency is found by adding successive frequencies i.e

10 +5 = 15

15 + 6 = 21

and so on, ..

Cumulative and relative cumulative frequency for the categorized birthweight.

Birthweight (g)	No of infants (frequency)	Cumulative frequency	% cumulative frequency
2700–2999	2	2	6.67
3000–3299	3	5	16.67
3300–3599	9	14	46.67
3600–3899	9	23	76.67
3900–4199	4	27	90.00
4200–4499	3	30	100.00

III- Graphical Presentation

A graph or chart provides an easily understood picture of the data and gives the user *a nice* and *impressive* overview of its **essential features**. However, the various graphical techniques should be regarded *as aids* to thinking about data *rather than as substitutes* for the statistical analysis of that data.

There are five main types of diagrams that are commonly used and particularly useful for data presentation (pie chart, bar chart, histogram, frequency polygon, and scatter diagram).

1- Pie chart :

It is suitable for summarizing data *arranged in categories* and on *percentage* basis. It is specially useful in presenting data that consist of a *small number* of categories.

Pie chart is a circle divided into wedges (segments or slices) that correspond to the *percentages or frequencies* of the distribution i.e the size of the slice is proportional to the frequency or percentage of cases belonging to the category it represents.

Because there is 360^0 in the circle, each 1% of the distribution can be represented by a sector of the circle with a central angle of 3.6^0.

A **disadvantage** of a pie chart is that it <u>can only represent one variable</u>. You will therefore need a separate pie chart for each variable you want to chart. Moreover a pie chart can lose clarity if it is used to represent more than four or five categories.

Example :

The method of drawing a pie chart can be demonstrated using the following data.

Table (11) : Distribution of 72 males according to blood group.

Blood group	Frequency	Percentage
A	20	27.8
B	30	41.7
AB	10	13.9
O	12	16.7
Total	**72**	**100.0**

- Draw a circle of a suitable size.
- Show the radius at any position (although the 12 o'clock position is commonly used).
- The angle that belong to each segment is calculated as follows :

corresponding percentage x 360 / 100

Thus, the angle corresponding to each segment for blood group data is as follows :

Blood group A = 360 x 27.8 / 100 = 100

Blood group B = 360 x 41.7 / 100 = 150

Blood group AB = 360 x 13.9 / 100 = 50

Blood group O = 360 x 16.6 / 100 = 60

NB : The sum of all angles must add to 360^0.

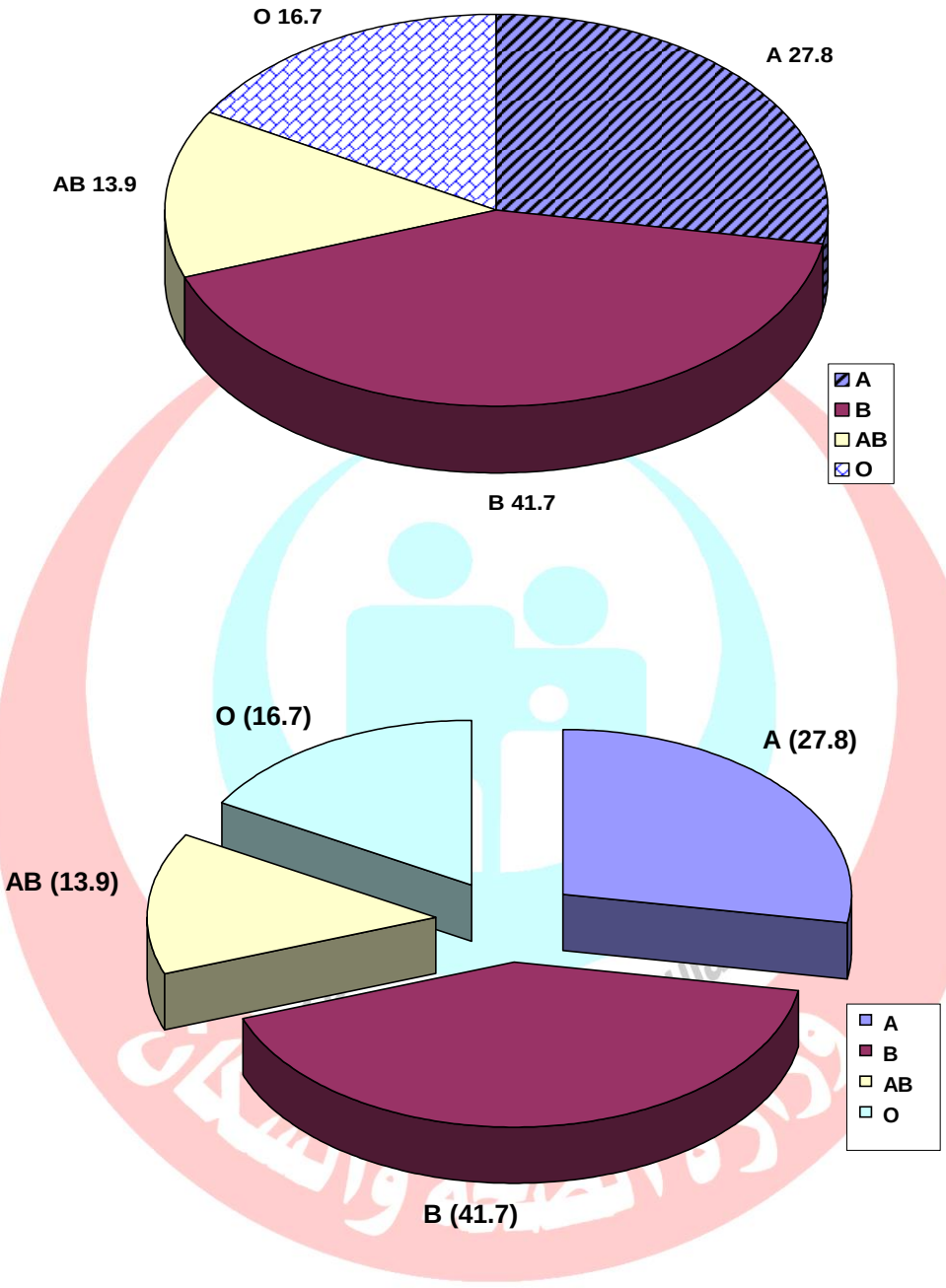


Figure (1) : Pie chart of blood group among males.

2- Bar chart :

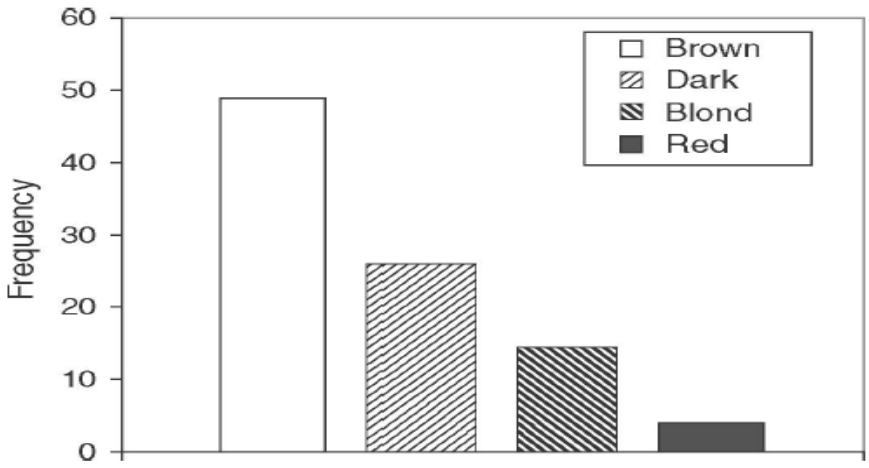
It is a tool for presentation of *categorical variables* where the *various categories* are represented on one axis (usually the horizontal) and *frequencies* or *percentages* of each category along the other axis (usually the vertical).

A vertical bar represents each category, and the height of each bar represents the frequency (or relative frequency) corresponding to each one. The bars should be *separate* and of *equal width* so the total area of all the bars is equal to the sample size, or n.

Bar chart may be **simple** when it represents one variable or it may be **multiple** (clustered or stacked) when it represents the comparison of more than one variable in the different comparison groups.

- **The simple bar chart**

  The simple bar chart is appropriate if only one variable is to be shown. Note that the bars should all be the same width, and there should be (equal) spaces between bars. These spaces emphasise the categorical nature of the data.



Example :

- Figure (2) (a simple bar chart) represents the previous example of distribution of blood groups among males (table 11).

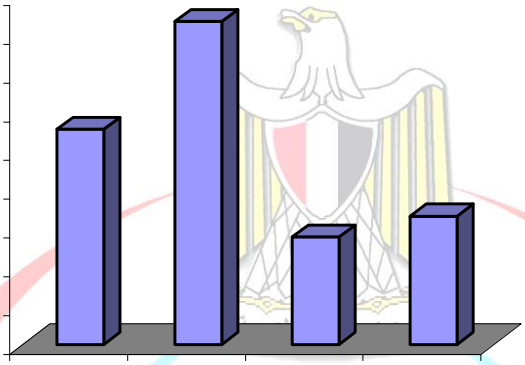
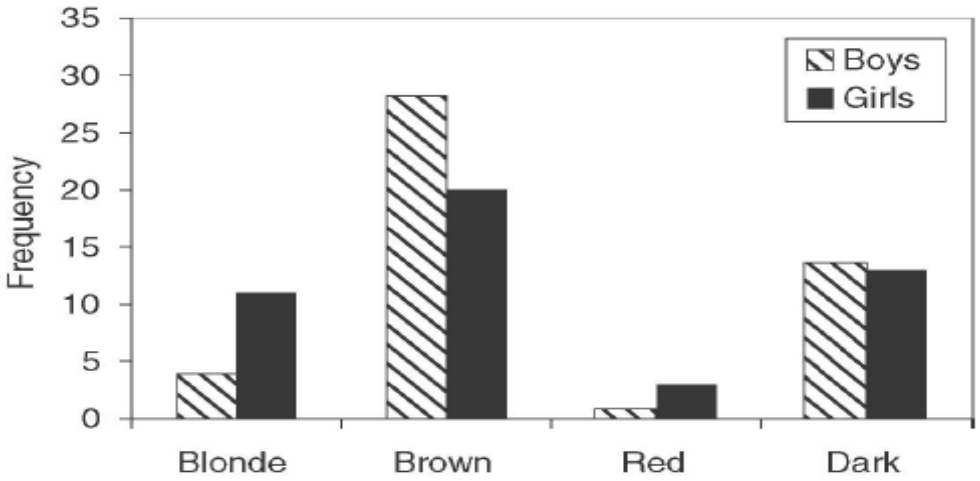


Figure (2) : Bar chart for blood groups among males.

- **The clustered bar chart**

  If you have more than one group you can use the clustered bar chart. There are two ways of presenting a clustered bar chart. This arrangement is helpful if you want to compare the relative sizes of the groups within each category (e.g. redheaded boys versus redheaded girls).

	Frequency	
Hair colour	Boys	Girls
Blonde	4	11
Brown	29	20
Red	1	3
Dark	14	13



Clustered bar chart of hair color by sex

Alternatively, the chart could have been drawn with the categories boys and girls, on the horizontal axis. This format would be more useful if you wanted to compare category sizes within each group. For example, red haired girls compared to dark haired girls. Which chart is more appropriate depends on what aspect of the data you want to examine.

Example :

- Figure (3) (a multiple bar chart) represents the distribution of blood groups and sex as shown in table (12).

Table (12) : Distribution according to blood group and sex.

Blood group	Sex			
	Male		Female	
	No.	%	No.	%
A	20	27.8	15	25.9
B	30	41.7	18	31.0
AB	10	13.9	12	20.7
O	12	16.7	13	22.4
Total	72	100.0	58	100.0

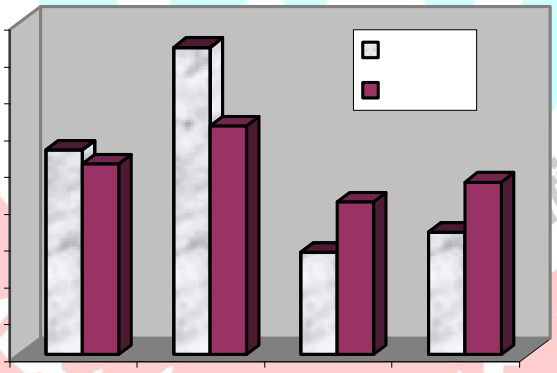
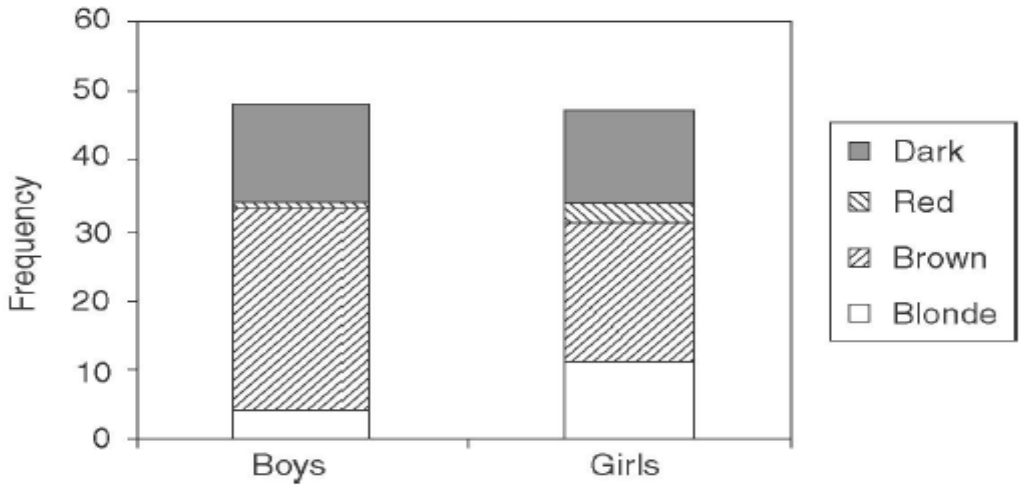


Figure (3) : clustered bar chart for blood group and sex.

- **The stacked bar chart**

  The figure shows a stacked bar chart for the same hair color and sex data. Instead of appearing side by side, as in the clustered bar chart, the bars are now stacked on top of each other. Stacked bar charts are appropriate if you want to compare the total number of subjects in each group (total number of boys and girls for example), but not so good if you want to compare category sizes between groups, e.g. redheaded girls with redheaded boys.



A stacked bar chart of hair color by sex

3- Histogram :

A continuous metric variable can take a very large number of values, so it is usually impractical to plot them without first grouping the values. The grouped data is plotted using a frequency histogram. It is a special form of a bar chart that presents categories (intervals) of a grouped **quantitative** variable. The bars are *not separated* by any space on the X axis. The frequency or percentage of data in each category is depicted on the Y axis.

A histogram looks like a bar chart used with discrete data except that each bar in a histogram represents an interval (category or class) of possible values rather than a *single value* but without any gaps between adjacent bars. This emphasizes the continuous nature of the underlying variable.

The *width of the bar* represents the interval of each category and the total area of each bar is proportional to the corresponding frequency or percentage of each category.

The steps for drawing the histogram can be illustrated using the following example :

Table (13) : Distribution according to age.

Age (years)	Frequency	Relative frequency (percentage)
25 –	3	14.3
30 –	5	23.8
35 –	7	33.3
40 –	4	19.1
45 – 49	2	9.5
Total	**21**	**100.00**

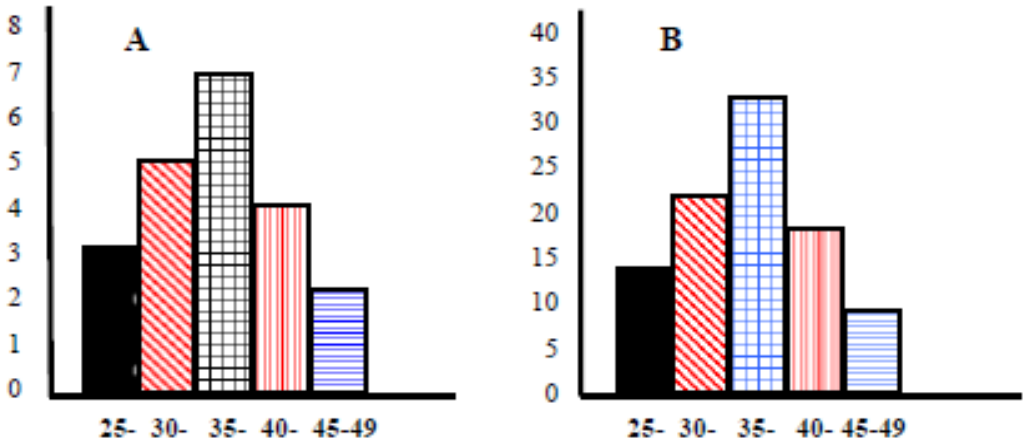


Fig. (4) : Distribution according to age (Histogram).

A : Frequency B : Percentage

Important points about histogram :

- A continuous metric variable can take a very large number of values, so it is usually impractical to plot them without first grouping (categorizing) the values. The grouped data is plotted using a frequency histogram, which has frequency plotted on the vertical axis and group size on the horizontal axis.
- A histogram looks like a bar chart but *without any gaps* between adjacent bars which emphasizes the continuous nature of the underlying variable. If the groups in the frequency table are all of the same width, then the bars in the histogram will also all be of the same width.
- One limitation of the histogram is that it can represent only one variable at a time (like the pie chart), and this can make comparisons between two histograms difficult, because, if you try

to plot more than one histogram on the same axes, invariably parts of one chart will *overlap* the other.

- Histogram is mostly used in the **explorative step** of data analysis. The histogram depicts the frequencies of a categorized scale variable, similarly to a bar chart, but it is not allowed to interchange bars.

One limitation of the histogram is that it can represent only one variable at a time (like the pie chart), and this can make comparisons between two histograms difficult, because, if you try to plot more than one histogram on the same axes, invariably parts of one chart will overlap the other.

4- Frequency Polygon

As in histograms, frequency polygons are used for displaying **quantitative** variables. It uses the same axes as the histogram, however, it is a special type of *line graph*. Frequency polygons are particularly more useful than histograms because *several* frequency distributions can be plotted easily and compared on one graph.

The frequency polygon is particularly useful when we want to compare two or more distributions with comparable n's. When sample n's are not comparable, relative frequency polygons should be used to compare two or more distributions. Otherwise, the differences in n's may distort the visual comparisons. For example, if two distributions are identical in form, but one is based on twice as many observations as the other, the frequencies of the group with the larger n will appear

twice as high as those of the group with the smaller n. By converting to proportions or percentages, we can show all the frequencies relative to 1.00 or 100%

A frequency polygon is a closed geometric figure used to graphically display frequency distributions. To prepare a frequency polygon, you must first add two classes to each end of the distribution. Because the frequency is 0 for these added classes, the curve is brought down to the horizontal axis at the midpoint of the two classes that are added. (Hence the polygon is a "closed" figure). Each of the remaining points in the frequency polygon is positioned over the midpoint of its corresponding class.

To draw a frequency polygon, the following is done :

- Determine (marking) the mid-point of each class interval represented on the horizontal axis of the graph.
- These points are then connected by straight lines.
- At the ends, the points are connected to the mid-points of the previous and succeeding intervals of zero frequency.

NB : The height of a given dot above the horizontal axis corresponds to the frequency or percentage of the relevant class interval.

Examples :

- The frequency polygon for data in table (13) is displayed in Fig. (5).
- The frequency polygon for data in table (14) is displayed in Fig. (6).

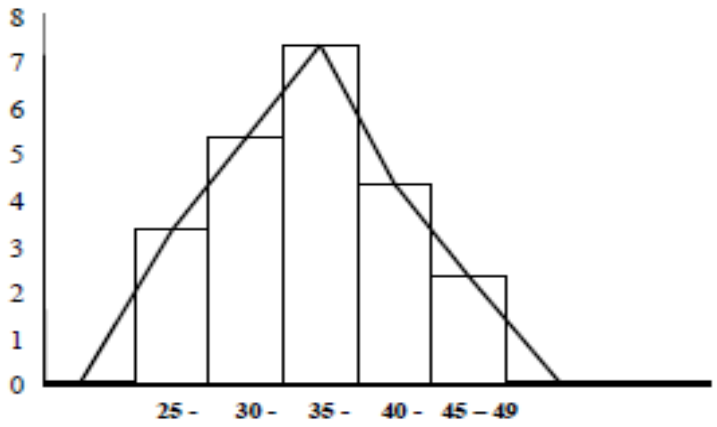


Fig. (5) : Distribution according to age (Frequency polygon)

Table (14) : Distribution of age according to sex.

Age in years	Sex			
	Male		Female	
	No.	%	No.	%
25 –	8	16.0	10	14.3
30 –	12	24.0	15	21.4
35 –	15	30.0	20	28.6
40 –	9	18.0	13	18.6
45 – 49	6	12.0	12	17.1
Total	**50**	**100.0**	**70**	**100.0**

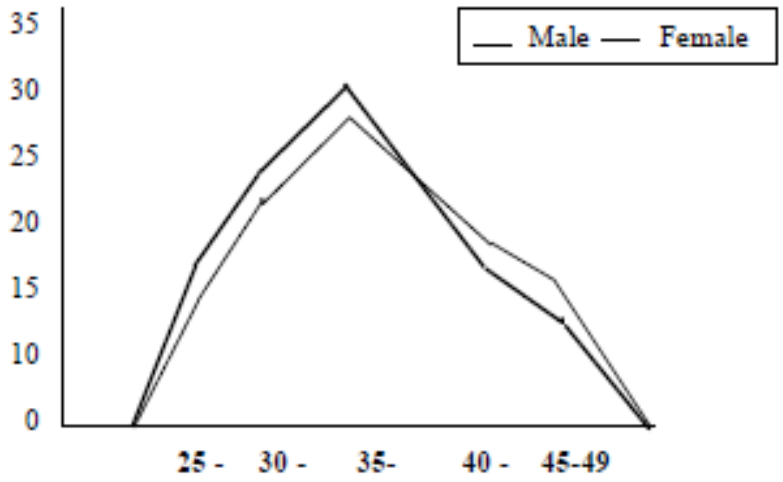


Fig. (6) : Distribution of age according to sex (Frequency polygon).

5- Scatter Diagrams

The relationship between two variables can be shown graphically in a scatter diagram, as shown in Fig. (7). A scatter diagram is a graph in which each individual or unit measured is entered as a point, the position of each point being determined by the values for the two characteristics measured.

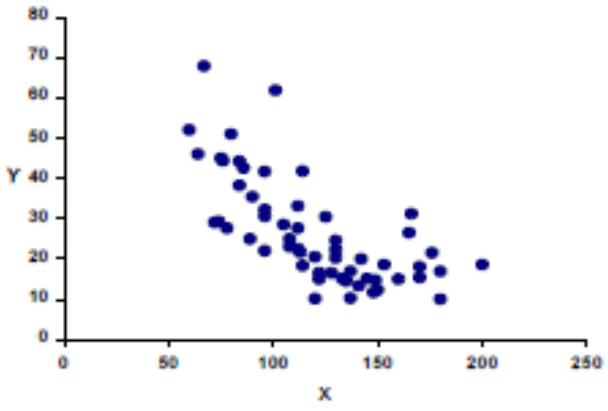


Fig. (7) : Scatter diagram for the relation between X and Y variables

Chapter (5)

Population and Samples

By the end of this chapter, the student should be able to :

- Define sample and population.
- Recognize the importance and reasons of sampling.
- Understand Technique of sampling.
- Know types of random samples.

A population is the term statisticians use to describe a large set of items (subjects, objects, events) that have *common observable characteristics*. It is the group *to which you want to generalize* your findings.

A sample is a group or a subset of the population, that you observe or collect data from, selected in such a way that it is *representative* of the larger population.

NB : A sample that represents the characteristics of the population *as closely* as possible is called a representative sample. You can not generalize your research findings based upon non-representative (biased) samples.

Parameter and Statistic :

- A number that describes a population is called a parameter.
- A number that describes a sample is called a statistic.

If we take a sample and calculate a statistic, we often use that statistic to infer something about the population from which the sample was drawn.

What is Sampling?

- Measuring a small portion of something and then making a general statement about the whole thing.
- Process of selecting a number of units for a study in such a way that the units represent the larger group from which they are selected.

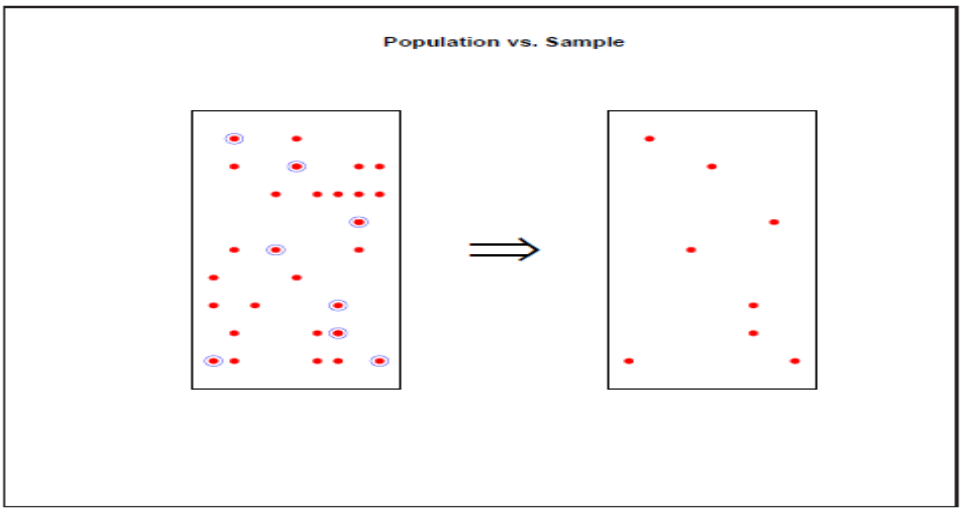


Figure 1: Population and Sample

Reasons for sampling :

1- A study of an entire population is *impossible* in most situations.
2- Samples can be studied *more quickly* than populations. Speed can be an important factor in certain situations.
3- A study of a sample is *less expensive* (cost-effective) than a study of an entire population.
4- Sample results are often *more accurate* than results based on a population.

However, certain occasions necessitate the study of the whole population e.g in tuberculosis survey to treat the diseased and during surveillance to control an epidemic of a communicable disease.

Methods of Sampling (Sampling procedures) :

1- Methods which do not follow probability theory i.e non probability samples :

- Purposive samples.
- Convenience samples.
- Quota samples.
- Snow ball samples.

Characters of non-probability sampling :

- In non-probability sampling, the chance of a member being included in the sample is not known.
- Results of non-probability samples can not be generalized from the sample to the population.
- This procedure also does not allow the researcher to calculate sampling statistics that provide information about the precision of the results.
- Non-probability samples tend to be cheaper, less complicated and less time consuming than probability samples.

2- Methods which follow probability theory i.e probability (random) samples :

- Simple random samples.
- Systematic random samples.
- Stratified random samples.
- Cluster random samples.
- Multisatge random samples.

Characters of probability sampling :

- The researcher knows the exact possibility of selecting each member of the population.
- Probability samples are the only type of samples where the results can be generalized from the sample to the population.
- In addition, probability samples allow the researcher to calculate the precision of the estimates obtained from the sample and to specify the sampling error.
- A probability sample tends to be more difficult and costly to conduct.

(I) Non probability sampling

- ***Purposive (judgment) samples :***

  These are samples chosen according to the *researcher own judgment* about who to include in the sample i.e *not chosen in a random way* and thus results *can not be generalized* to the whole population. They are easier in selection than probability samples.

  Members are chosen for the sample based on known factors so they can only be used in *pilot studies*, or in *pre-testing*. *Prior knowledge* and *research skills* are used in selecting the respondents or elements to be sampled.

  The researcher may lack the information regarding the population from which he has to collect the sample.

  As with all non-probability sampling methods, the degree and direction of error introduced by the researcher cannot be measured, and statistics that measure the precision of the estimates cannot be calculated.

- ***Convenience samples :***

  This type of non-probability sampling involves the sample being drawn from that part of the population which is close to hand. That is, readily available and convenient (i.e easily accessible participants with no randomization). This method saves money, time and effort ; but at the expense of information and credibility.

- ***Quota samples :***

  Selecting participants in number proportionate to their numbers in the larger population, no randomization. It raises analytic confidence and representativeness.

  Quota sampling is often confused with two probability sampling methodologies, stratified and cluster sampling. The primary differences between the methodologies is that with stratified and cluster sampling the classes are mutually exclusive and are isolated prior to sampling. In quota sampling, the classes cannot be isolated prior to sampling and respondents are categorized into the classes as the survey proceeds. As each class fills or reaches its quota, additional respondents that would have fallen into these classes are rejected or excluded from the results.

  It is used in sampling of public opinions. Each enumerator is asked to obtain the required information from a specified number of individuals (classified in different groups). The enumerator choose these individuals in a way which enable him to get the information easily and quickly. It is of no use in public health or clinical practice.

  An example of a quota sample would be a survey in which the researcher desires to obtain a certain number of respondents from various income categories. Generally, researchers do not know the incomes of the persons they are sampling until they ask about income. Therefore, the researcher is unable to subdivide the

population from which the sample is drawn into mutually exclusive income categories **prior to drawing the sample**. Bias can be introduced into this type of sample when the respondents who are rejected, because the class to which they belong has reached its quota, differ from those who are used.

- ***Snow ball samples :***

  It refers to identifying someone who meets the criteria for inclusion in the study. Selection of additional respondents is based on referrals from the initial respondents. It is also called chain referral as group members identify additional information-rich members to be included in the sample.

(II) Probability (random) Sampling

The most important consideration is that any sample should be representative of the population from which it is taken. For example, if your population has equal numbers of male and female babies, but your sample consists of twice as many male babies as female, then any conclusions you draw are likely to be misleading.

The key to building representative samples is randomization. "Randomization" is the process of randomly selecting population members for a given sample, or randomly assigning subjects to one of several experimental groups, or randomly assigning experimental treatments to groups. For a sample to be truly random, the *basic characteristic* of sampling is that all members of the population have an equal and independent chance of being included in the sample.

Random sampling means that only chance determines whether or not a particular unit in the population will be part of the sample. Though the word random implies that a pattern does not exist ; a specific plan or strategy is necessary to insure that a pattern does not exist. *A random sample is a sample that has been selected using a definite plan.*

Unfortunately, this is rarely possible in practice, since this would require a complete and up-to-date list (name and contact details) of the target population. Such a list is called a sampling frame. In practice, compiling an accurate sampling frame for any population is hardly feasible. This problem applies to two close relatives of simple random sampling – systematic random sampling, and stratified random sampling.

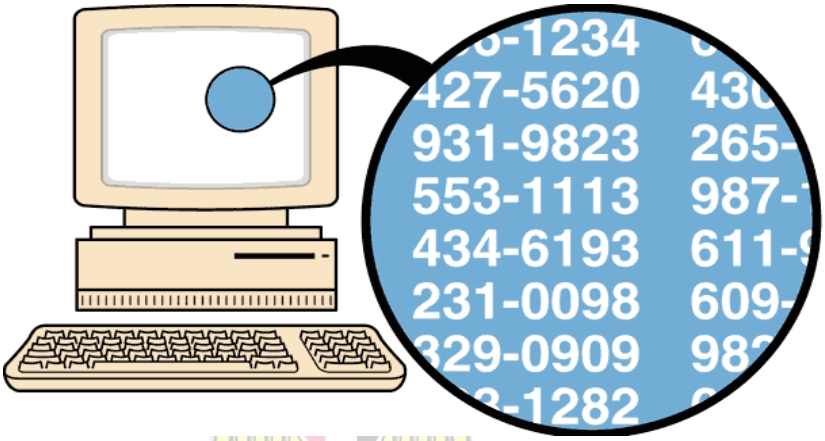
Types of random samples

(1) Simple random samples :

This method assures that each member of the population has an equal probability (chance) of being chosen for the sample. The population from which a simple random sample is drawn should be uniform or homogenous and every unit must have an identification number (ID), and an exhaustive list of all members of the population of interest, called a sampling frame, must be available. The recommended way to select such a sample without introducing researcher bias is to use a table of random numbers or a computer-generated list of random numbers.

Samples may be drawn with or without replacement. In practice, however, most simple random sampling for survey research is done without replacement ; that is, a person or item selected for sampling is removed from the population for all subsequent selections.

An example of a simple random sample would be a survey of County employees. An exhaustive list of all County employees as of a certain date could be obtained from the Department of Human Resources. If 100 names were selected from this list using a random number table or a computerized sampling program, then a simple random sample would be created.



(2) Systematic random samples :

It is one in which every kth item is selected ; k is determined by dividing the number of items in the sampling frame by the desired sample size. It is easy to draw and spreads more evenly over the population from which the sample is drawn. A *random starting point* at the beginning of an ordered population is chosen, and then the remainder of the sample is chosen according to a predetermined schedule.

For example, suppose we have a population where N = 100 and it is desired to have a sample n = 10 : so 100 / 10 = 10. We select a random number between 1 and 10, say 2 and consequently we chose the number, 2, 12, 22,….92.

A researcher may choose to conduct a systematic sample instead of a simple random sample for several reasons :

- Systematic samples tend to be easier to draw and execute.
- The researcher does not have to jump backward and forward through the sampling frame to draw the sample.
- A systematic sample may spread the members selected for measurement more evenly across the entire population.

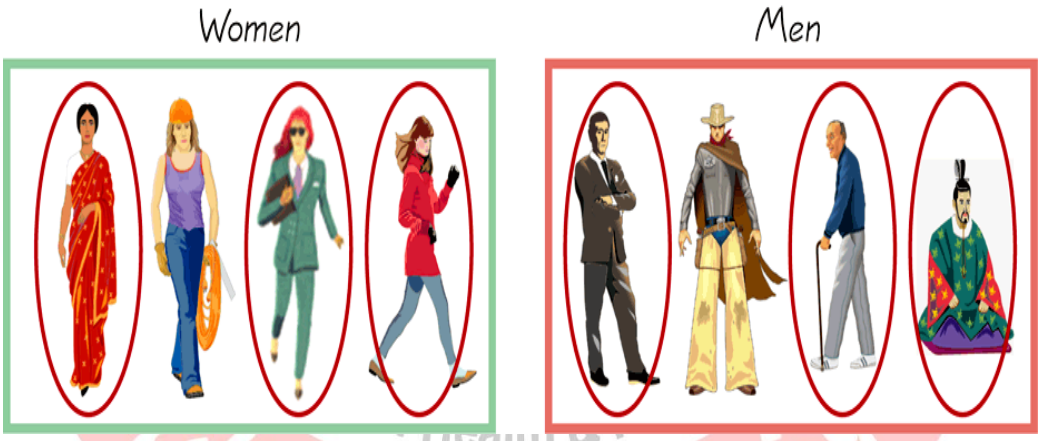
- Systematic sampling may be more representative of the population and more precise.

One of the most attractive aspects of systematic sampling is that this method can allow the researcher to draw a probability sample without complete prior knowledge of the sampling frame. For example, a survey of visitors to the County's publications desk could be conducted by sampling every 10th visitor after randomly selecting the first through 10th visitor as the starting point. By conducting the sample in this manner, it would not be necessary for the researcher to obtain a comprehensive list of visitors prior to drawing the sample.

(3) Stratified random samples :

It is the ***most suitable procedure*** for ensuring representativeness of samples. The value of stratification is that if a heterogeneous population is divided into homogeneous strata (subgroups), the accuracy of the sample can be increased as different groups are properly represented. The sample is drawn through :

- Stratifying the population i.e dividing the population into different homogeneous strata according to certain characteristics that may substantially invalidate the results. The sampling frame is first broken down into strata relevant to the study, for example men and women ; or nonsmokers, ex-smokers and smokers.
- Selecting simple or systematic random samples from each stratum the size of which is proportionate to the size of the corresponding stratum.
- Sum up the total of the samples drawn from the different strata to get the final sample.



(4) Cluster sampling :

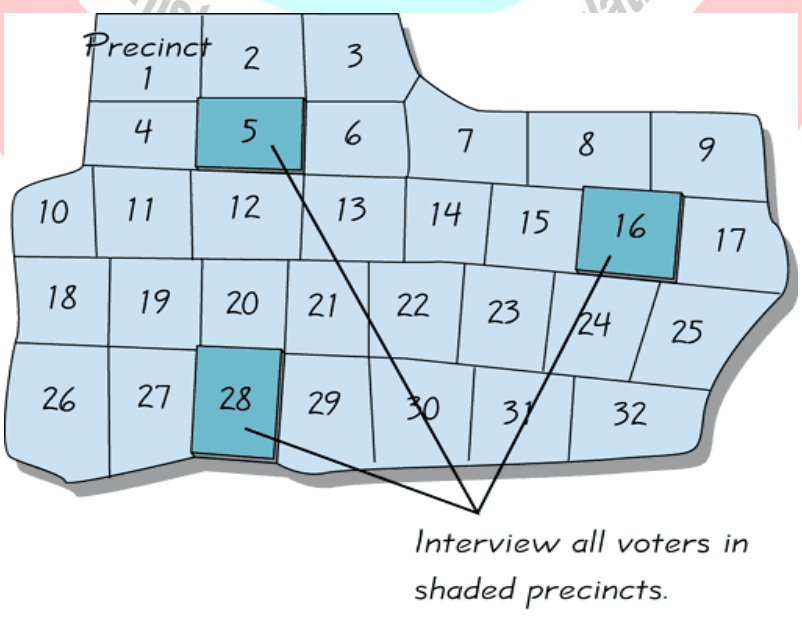
Here, the population is divided into sampling units, or groups and a random sample of groups is chosen i.e the sample is made up of groups (clusters), not individuals. For example, in a city, all the residents who live in randomly selected blocks are chosen.

Cluster sampling is similar to stratified sampling because the population to be sampled is subdivided into mutually exclusive

groups. After the clusters are defined, a simple random sample of the clusters is drawn and the **members of the chosen clusters are sampled**. If all of the elements (members) of the clusters selected are sampled, then the sampling procedure is defined as one-stage cluster sampling. If a random sample of the elements of each selected cluster is drawn, then the sampling procedure is defined as two-stage cluster sampling.

Cluster sampling involves randomly selecting groups, not individuals. Any intact group with similar characteristics is a cluster. Examples of clusters include classrooms, schools and hospitals.

There are drawbacks to cluster sampling. First, a sample made up of clusters may be less representative than one selected through random sampling. A second drawback is that commonly used inferential statistics are not appropriate for analyzing data from a study using cluster sampling.



(5) Multisatge random samples :

In this type, we draw one sample in two or more stages. It is resorted to if the distribution of the population is over a large area and we have no enough funds, or when it is impossible to carry other types of sampling.

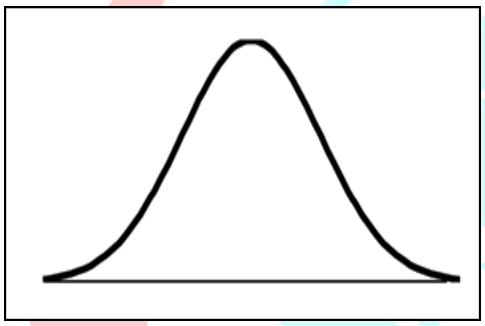
Chapter (6)

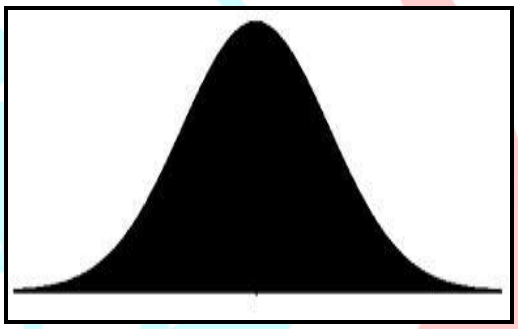
The Normal Distribution

(The Gaussian distribution)

By the end of this chapter, the student should be able to :

- Understand the concept of data normalness.
- Know the properties of Normal curve.
- Be oriented with and calculate areas under the Normal curve (AUC) using z-table.
- Understand the standardized Normal curve and z-score.





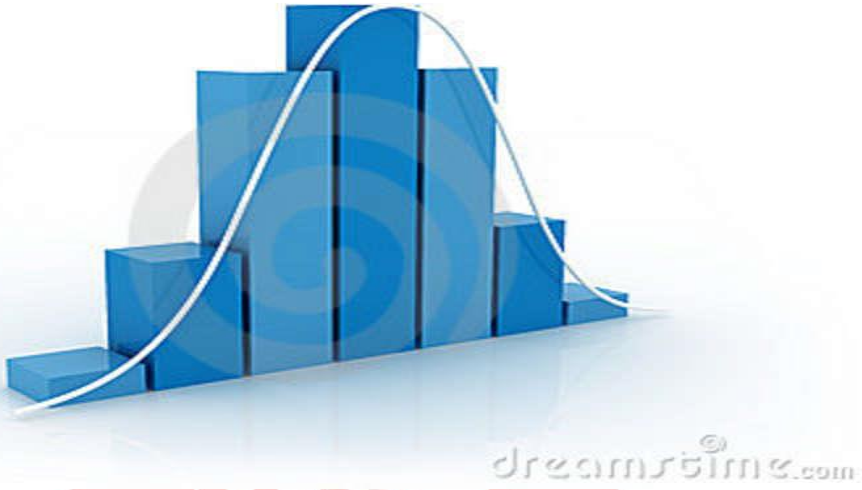
There is one particular symmetric bell-shaped distribution, known as the Normal distribution, which has a special place in the heart of statisticians. The *Normal*, or *Gaussian* distribution (named in honour of the German mathematician C.F.Gauss, 1777 – 1855) is *the most important* probability distribution in statistics.

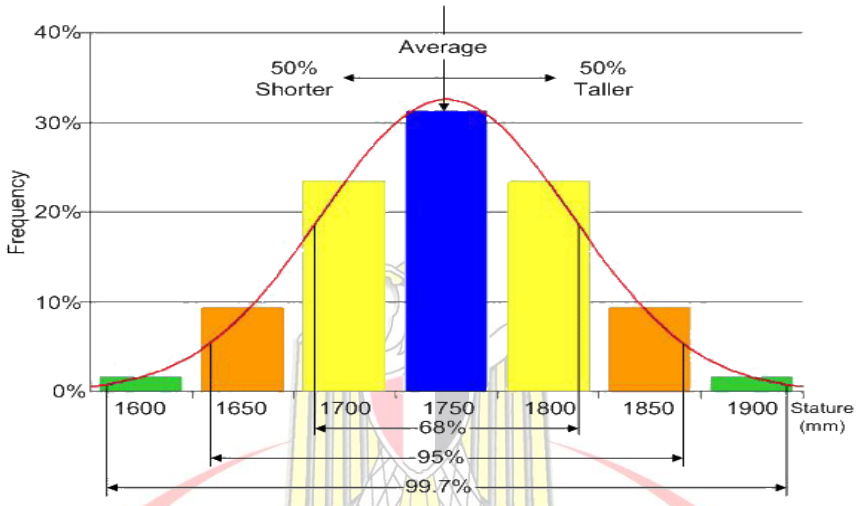
Normal-ness

It is important to stress that, in this context, the word “Normal" is a statistical term and is not used in the dictionary or clinical sense. Thus, in order to distinguish between the two, statistical and dictionary “normal", it is conventional to use a capital letter when referring to the **N**ormal distribution.

In many circumstances, the normal distribution, adequately describes the distribution of natural phenomena. For example, the IQ score of a group of children selected at random are normally distributed. The distribution of heights, weights, blood sugar, heart rate, and cholesterol levels are a few variables for which the normal distribution can be used to describe their biologic variation.

The normal distribution is the *most widely known and used* of all distributions. It is the **cornerstone** distribution of statistical inference, represnting the distribution of the possible estimates of a population parameter that may arise from different samples. *Parametric statistics* are based on the assumption that the variables are normally distributed.





Properties of the Normal distribution curve :

1- It is bell shaped.

2- Symmetrical around the mean (either side is mirror image of the other side).

3- The mean, median and mode are equal (all coincide).

4- The total area under the normal curve above the X- axis is equal to 1.

5- Normal distributions are uniquely defined by *two parameters*, the mean (μ) and the standard deviation (σ). There is a *limitless* number of normal distributions because there is an infinite number of possible pairs of means and standard deviations.

6- It has inflection points at $\mu - \sigma$ and $\mu + \sigma$.

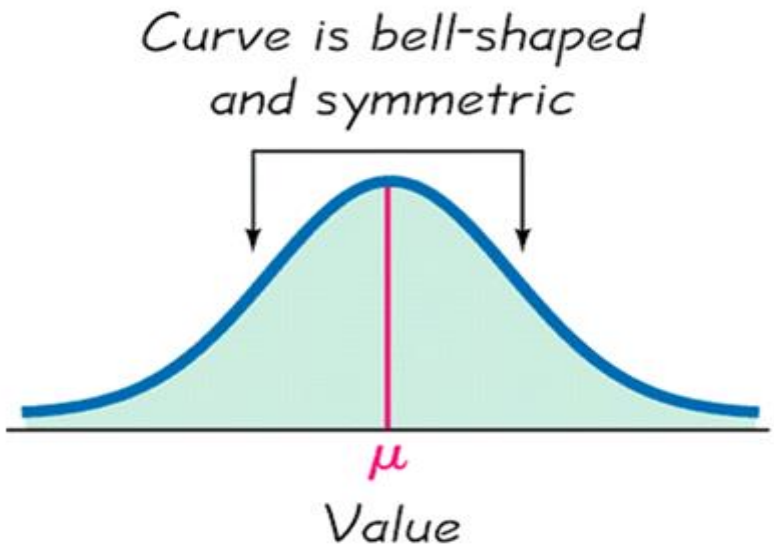
7- About 2/3 (68.26%) of the area of a normal distribution is within one standard deviation of the mean. $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6826$

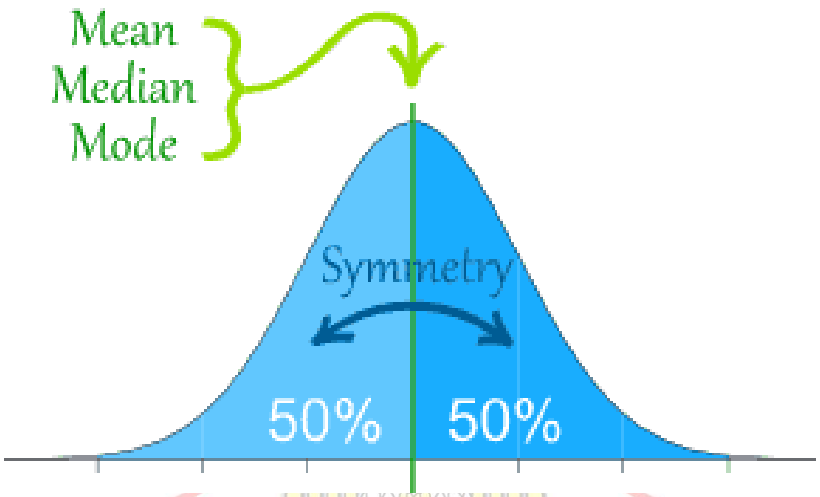
8- Approximately 95% of the area of a normal distribution is within two standard deviations of the mean. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9544$

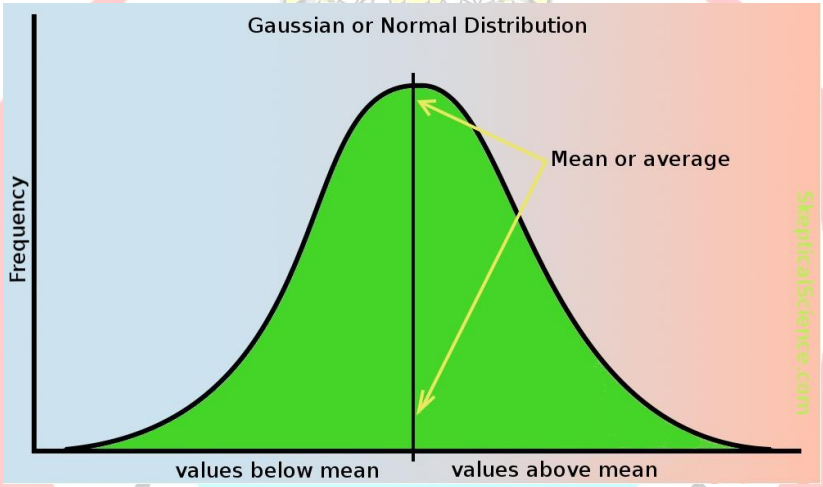
9- Approximately 99% of the area of a normal distribution is within three standard deviations of the mean. P ($\mu - 3\sigma \leq X \leq \mu + 3\sigma$) = 0.9974

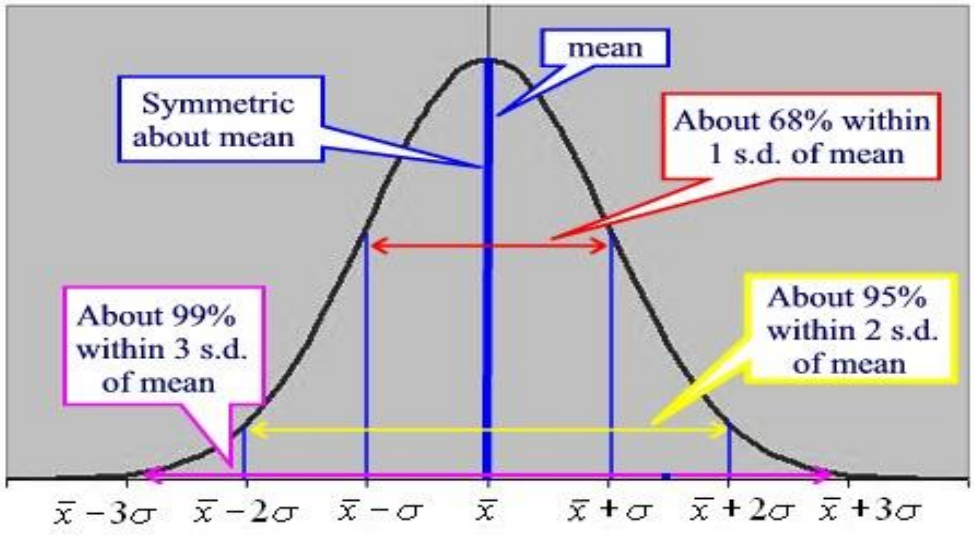
As mentioned earlier, the Normal distribution is described completely by two parameters, the mean (μ) and the standard deviation (σ). This means that for any Normally distributed variable, once the mean and variance (σ^2) are known (or estimated), it is possible to calculate the probability distribution for that population.

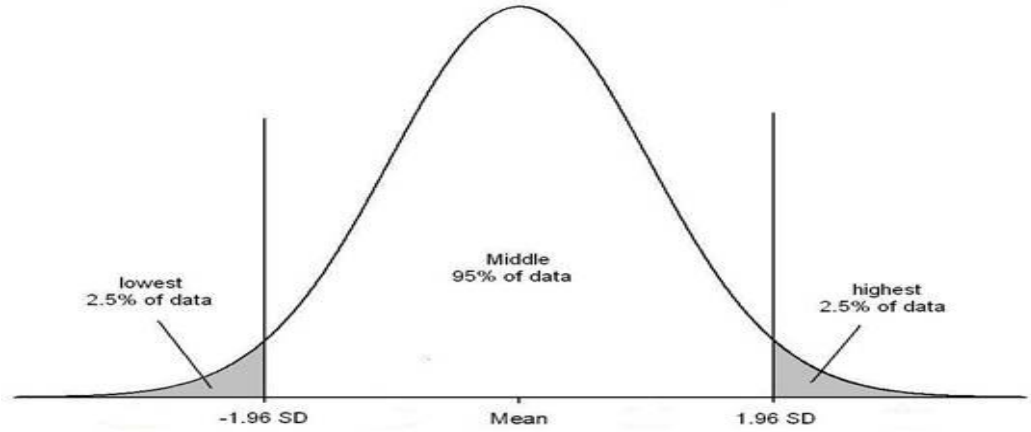
This distribution is the most important one in statistics. It is important partly because it approximates well the distributions of many variables. The main reason for its prominence, however, is that most inferential statistical methods make use of properties of the normal distribution **even** when the sample data are not bell-shaped.

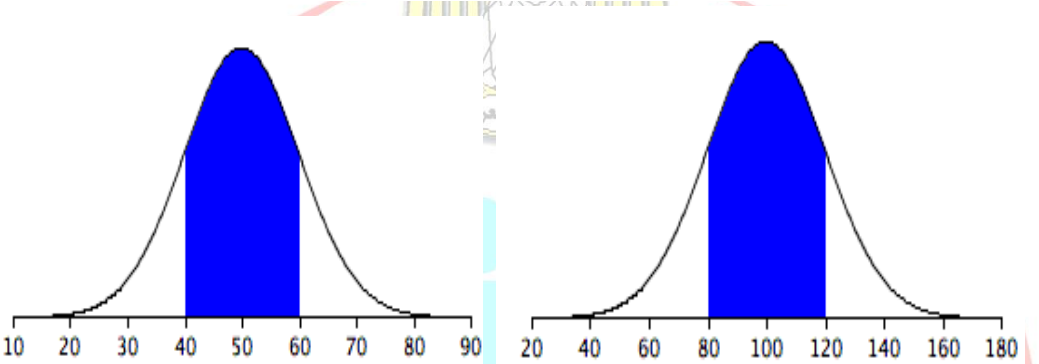



Mean
Median
Mode
Symmetry
50%
50%


Gaussian or Normal Distribution
Mean or average
Frequency
values below mean
values above mean
SkepticalScience.com

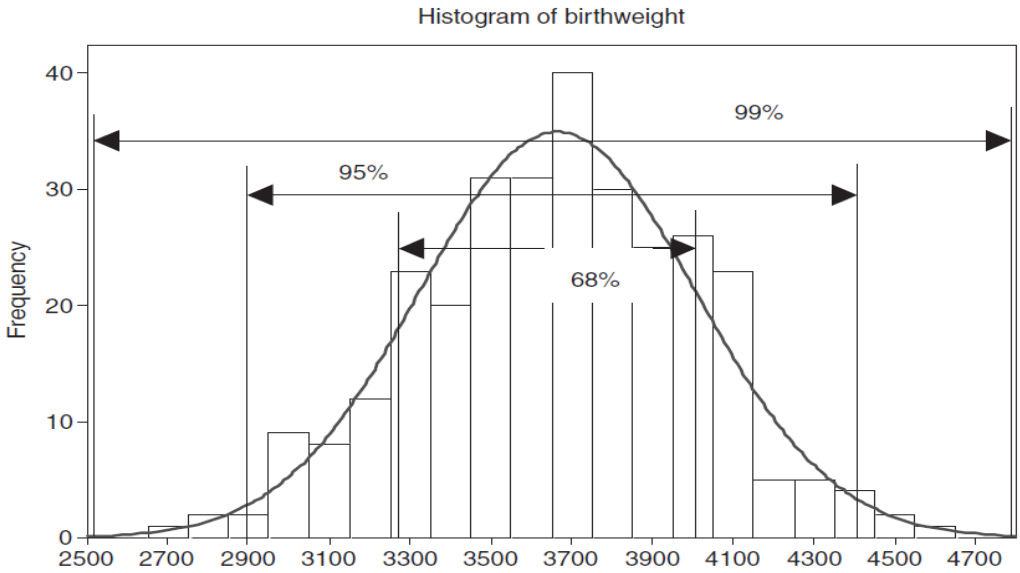

mean
Symmetric about mean
About 68% within 1 s.d. of mean
About 99% within 3 s.d. of mean
About 95% within 2 s.d. of mean
$\bar{x}-3\sigma$ $\bar{x}-2\sigma$ $\bar{x}-\sigma$ $\bar{x}$ $\bar{x}+\sigma$ $\bar{x}+2\sigma$ $\bar{x}+3\sigma$





Normal distribution with a mean of 50 and standard deviation of 10. 68% of the area is within one standard deviation (10) of the mean (50).

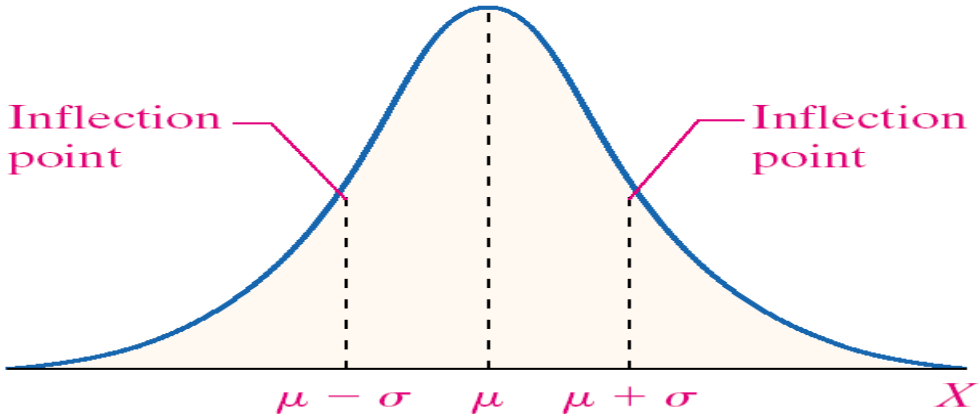
Normal distribution with a mean of 100 and standard deviation of 20. 68% of the area is within one standard deviation (20) of the mean (100).



Minitab calculates these birthweights to have a mean of 3644 g, and a standard deviation of 377 g. In words, the area properties are as follows :

- About 68 per cent of the birthweights will lie within one standard deviation either side of the mean. That is, from 3644 g − 377 g to 3644 g + 377 g, or from 3267 g to 4021 g.
- About 95 per cent of the birthweights will lie within two standard deviations either side of the mean. That is, from 3644 g − 754 g to 3644 g + 754 g, or from 2890 g to 4398 g.
- About 99 per cent of the birthweights will lie within three standard deviations either side of the mean. That is, from 3644 g − 1131 g to 3644 g + 1131 g, or from 2513 g to 4775 g.

So, if you have some data that you know is Normally distributed, and you also know the values of the mean and standard deviation, then you can make statements such as, 'I know that 95 per cent of the values must lie between so-and-so and so-and-so.'



Example (1) :

Normal curves for two populations with different means :

Population #1	Population #2
$\mu_1 = 50$	$\mu_2 = 70$
$\sigma_1 = 4$	$\sigma_2 = 4$

Draw the normal curves for both populations.

Summary : The two curves are exactly the same, except one curve is to the right of the other curve.

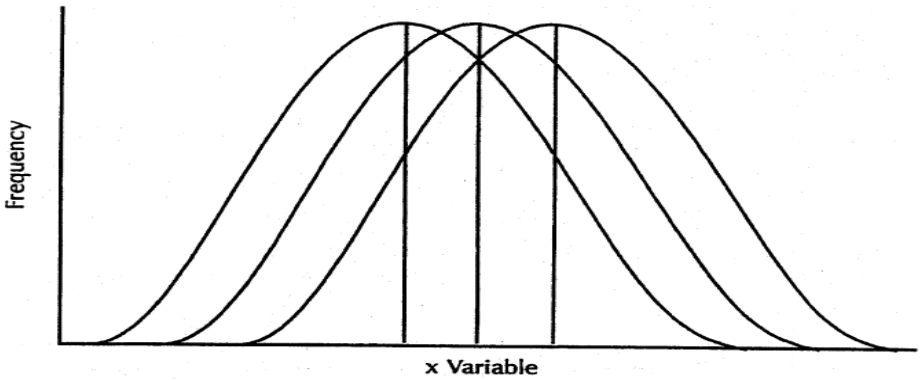
Example (2) :

Normal curves for two populations with different standard deviations.

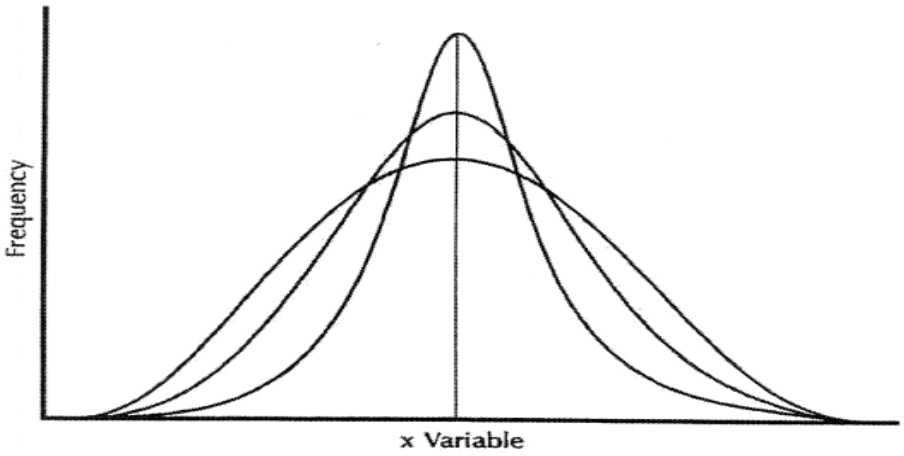
Population #1	Population #2
$\mu_1 = 50$	$\mu_2 = 50$
$\sigma_1 = 4$	$\sigma_2 = 7$

Draw the normal curves for both populations.

Summary : Increasing the standard deviation causes the curve for population #2 to become flatter and more spread out.



Same standard deviations and different means.



Same mean and different standard deviations.

Meaning of the Standard Deviation :

Standard deviation is a measure of the dispersion or spread of the data. If we are estimating from a sample and if there are a large number of observations, the standard deviation can be estimated from the range of the data, that is, the difference between the smallest and the highest value. Dividing the range by 6 provides a *rough estimate* of the standard deviation if the distribution is normal, because 6 standard deviations (3 on either side of the mean) encompass 99%, or virtually all, of the data.

On an individual clinical level, knowledge of the standard deviation is very useful in deciding whether a laboratory finding is normal, in the sense of "healthy." Generally a value that is *more than 2 standard deviations* away from the mean is suspect, and perhaps further tests need to be carried out.

For instance, suppose as a physician you are faced with an adult male who has a hematocrit reading of 39. Hematocrit is a measure of the

amount of packed red cells in a measured amount of blood. A low hematocrit may imply anemia, which in turn may imply a more serious condition. You also know that the average hematocrit reading for adult males is 47. Do you know whether the patient with a reading of 39 is normal (in the sense of healthy) or abnormal?. You need to know the standard deviation of the distribution of hematocrits in people before you can determine whether 39 is a normal value. In point of fact, the standard deviation is approximately 3.5 ; thus, plus or minus 2 standard deviations around the mean results in the range of from 40 to 54 so that 39 would be slightly low. For adult females, the mean hematocrit is 42 with a standard deviation of 2.5, so that the range of plus or minus 2 standard deviations away from the mean is from 37 to 47. Thus, if an adult female came to you with a hernatocrit reading of 39, she would be considered in the "normal" range.

Determining the Area Under the Normal Curve :

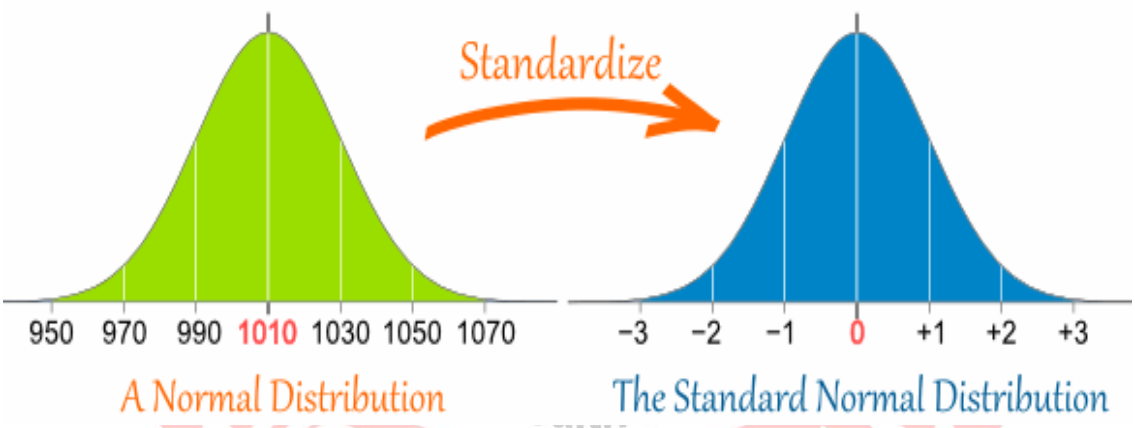
(Determining probabilities using the Normal distribution)

If the total area under the curve is 1.00 (or 100 percent), then it is possible to use the curve in a probabilistic manner. For example, we can obtain the probability that an individual's heart rate will be between 68 and 72.

To obtain the probability, the actual measurements (heart rate) must be converted to a *standard scale* for which all the areas under the curve have already been calculated. The corresponding table in the appendix gives the tabulated values_for area under the standard normal curve.

The standard normal distribution, commonly denoted by Z, has a mean equal to 0 and standard deviation equal to 1. Table in the appendix is constructed so that for each value on the standard normal scale, Z, the area under the curve from minus infinity to z and the area from z to plus infinity are given. These areas are referred to as left-hand and right-hand areas.

To calculate an area under the normal distribution, the points on the x-variable scale (heart rate) are converted to the **z-scale** and then Z-table (A2) is used. If the variable (heart rate) has mean μ and standard deviation δ, then to convert a point x (a specific heart rate measurement) to the corresponding point z on the z-scale.



The standardized normal distribution is one whose mean = 0, standard deviation = 1, and the total area under the curve = 1.

Standardized Variable :

A variable is said to be standardized if it has been adjusted (or transformed) so that its mean equals 0 and its standard deviation equals 1.

Standardization can be accomplished using the formula for a z-score :

$$Z = \frac{x - \mu}{\sigma}$$

The z-score represents the number of standard deviations that a data value is away from the mean.

❑ Finding area under the standard normal curve (AUC) :

1. Find the area under the standard normal curve to the left of Z = -1.40.
2. Find the area under the standard normal curve to the right of Z = 1.85.
3. Find the area under the standard normal curve between Z = 0.50 and Z = 2.25.

Notation for the probability of a standard normal random variable :

- $P\ (Z < a)$ represents the probability a standard normal random variable is less than a.
- $P\ (Z > a)$ represents the probability a standard normal random variable is greater than a.
- $P\ (a < Z < b)$ represents the probability a standard normal random variable is between a and b

Examples written using probability notation :

1. Find P (Z < -1.40).
2. Find P (Z > 1.85).
3. Find P (0.50 < Z < 2.25).

❑ Finding Z-Scores for given areas :

1. Find the 85th percentile for the Z distribution, i.e. find z_0 such that P ($Z < z_0$) = 0.85.
2. Find z_0 such that P ($Z > z_0$) = 0.25.
3. Find the two values of Z (z_1 and z_2) that include the middle 95% of Z values.

Technique :

Finding the area under any normal curve :

Step 1 : Draw the normal curve with the desired area shaded.

Step 2 : Convert the values of X to Z-values using :

$$Z = \frac{x - \mu}{\sigma}$$

Step 3 : Draw a Z-axis under the X-axis on the normal curve in Step 1, and place the Z-values under the corresponding X-values.

Step 4 : Find the area under the normal curve using the Z-values and the standard normal curve in Appendix.

Examples :

(1) The random variable X is normally distributed with $\mu = 500$ and $\sigma = 100$. Find $P(X < 400)$. Use a graph with labels to illustrate your answer.

(2) The random variable X is normally distributed with $\mu = 500$ and $\sigma = 100$. Find $P(X > 620)$. Use a graph with labels to illustrate your answer.

(3) Scores on the SAT test are normally distributed with $\mu = 500$ and $\sigma = 100$. What score must a student make on the test to be at the 90th percentile? Use a graph with labels to illustrate your answer.

(4) Scores on the SAT test are normally distributed with $\mu = 500$ and $\sigma = 100$. What range in SAT scores (x1 and x2) includes the middle 50% of scores? Use a graph with labels to illustrate your answer.

N:B

- The normal curve is not a single curve but a family of curves, each of which is determined by its mean and standard deviation.
- In order to work with a variety of normal curves, we cannot have a table for every possible combination of means and standard deviations.
- What we need is a standardized normal curve which can be used for any normally distributed variable. Such a curve is called the Standard Normal Curve.

- The Standard Normal Curve (Z-distribution) is the distribution of normally distributed standard scores with mean equal to zero and a standard deviation of one.
- A z score is nothing more than a figure, which represents how many standard deviation units a raw score is away from the mean.
- For scores above the mean, the z score has a positive sign. Example

  + 1.5z. Below the mean, the z score has a minus sign. Example -0.5z.
- Calculate Z score for blood pressure of 140 if the sample mean is 110 and the standard deviation is 10
- Z = 140 – 110 / 10 = 3

Comparing Scores from Different Distributions :

(1) When the population mean and standard deviation is known :

Z is derived from *X* by the following :

$$Z = \frac{x - \mu}{\delta}$$

Thus, the Z-score really tells you *how many standard deviations from the mean* a particular *x*-score is. The whole process is very much like dividing feet by yards in order to express distance or height in terms of yards.

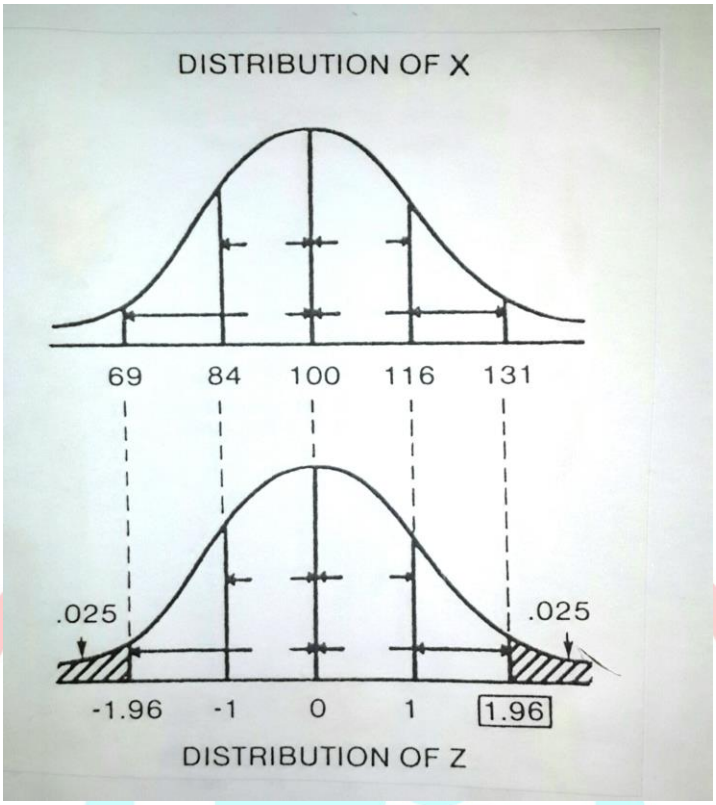


Fig. (12)

Thus, an I.Q. score of 131 is equivalent to a Z score of 1.96 (i.e., it is 1.96, or nearly 2, standard deviations above the mean I.Q.).

$$Z = \frac{131 - 100}{16} = 1.96$$

One of the nice things about the Z distribution is that the probability of a value being anywhere between two points is equal to the area under the curve between those two points. It happens that the area to the right of 1.96 corresponds to a probability of 0.025, or 2.5% of the total curve. Since the curve is symmetrical, the probability of Z being to the left of -1.96 is also 0.025. Invoking the additive law of probability, the probability of a Z being *either* to the left of -1.96 *or* to the right of +1.96 is 0.025 + 0.025 = 0.05.

Transforming back up to *X*, we can say that the probability of someone having an I.Q. outside of 1.96 standard deviations away from the mean (i.e., above 131 or below 69) is 0.05, or only 5% of the population have values that extreme. (Commonly, the Z value of 1.96 is rounded off to ± 2 standard deviations from the mean as corresponding to the cutoff points beyond which, lies 5% of the curve, but the accurate value is 1.96.).

(2) When the population mean and the standard deviation are not known :

Z is derived from *X* using the sample estimates of μ and δ as follows :

$$Z = \frac{X - \bar{X}}{s}$$

All we need know is the sample mean and its standard deviation to transform any distribution into *z*-scores.

A very important use of Z derives from the fact that we can also convert a sample mean (rather than just a single individual value) to a Z score.

The *numerator now* is the distance of the sample mean from the population mean and *the denominator* is the standard deviation of the distribution of means, which is the *standard error of the mean.* This is illustrated in Figure 6, where we are considering means based on 25 cases each.

$$Z = \frac{\bar{X} - \mu}{s.e_{\bar{x}}}$$

The s.e. is $16/\sqrt{25} = 16/5 = 3.2$.

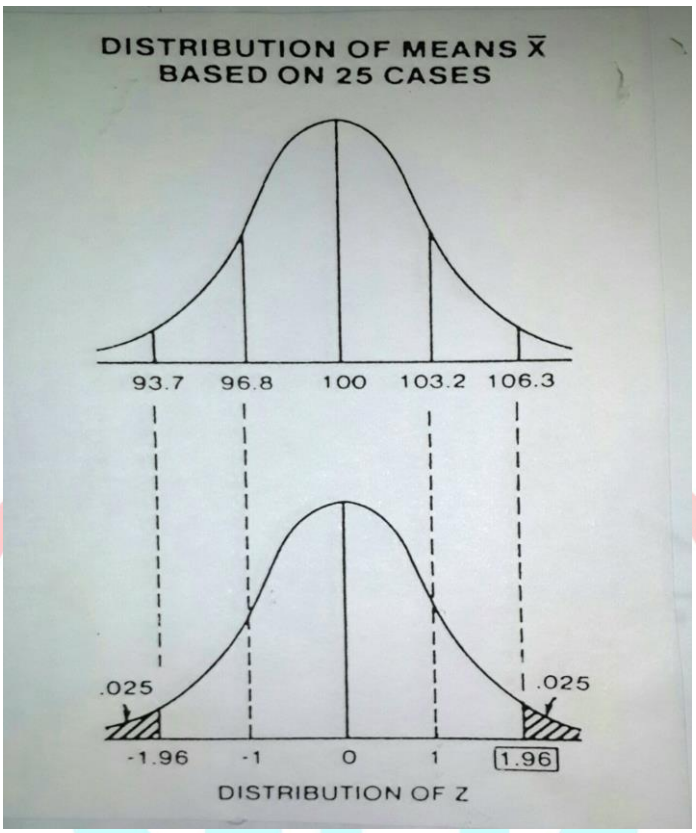


Fig. (13)

Now we can see that a sample mean of 106.3 corresponds to a Z score of 1.96.

$$Z = \frac{106.3 - 100}{3.2} = 1.96$$

We can now say that the probability that the *mean I Q. of a group of 25 people* greater than 106.3 is 0.025. The probability that such a mean is less than 93.7 is also 0.025.

Examples :

(1) 250 medical students were enrolled in an examination marked out of 100. The mean mark obtained was 61, standard deviation 10. Since the pass mark was 50, how many students failed?

A :

$$Z = \frac{50 - 61}{10} = \frac{-11}{10} = -1.1$$

From table A2 in the appendix we find that (0.1357) 13.57% of students will have gained marks of less than 50. Since 250 students sat the exam, 34 will have failed.

(2) It was felt to be appropriate to pass 75% of candidates in a final examination. From previous years the average mark obtained in this exam was known to be 68% with standard deviation of 14. What must the pass mark be set at?

A :

If 75% of candidates pass, 25% will fail. In the appendix (table A2), it is shown that z must be – 0.674 for this to happen.

$$\frac{?-68}{14} = -0.674$$

$$? - 68 = 14 \times -0.674$$

$$? = 68 + (14 \times -0.674) = 58.6\%$$

The pass mark should therefore be set at 59%

When the population mean and standard deviation is known :

$$Z = \frac{x - \mu}{\delta}$$

When the population mean and the standard deviation are not known :

$$Z = \frac{X - \bar{X}}{s}$$

We can also convert a sample mean (rather than just a single individual value) to a Z score :

$$Z = \frac{\bar{X} - \mu}{s.e_{\bar{x}}}$$

Exercise [5]

1. **Marks on a Chemistry test follow a normal distribution with a mean of 65 and a standard deviation of 12. Approximately what percentage of the students have scores below 50?**

   (a) 11%
   (b) 89%
   (c) 15%
   (d) 18%
   (e) 39%

2. **Refer to the preceding question. What is the approximate 90th percentile of the mark distribution?**

   (a) 80
   (b) 90
   (c) 85
   (d) 75
   (e) 95

3. **The marks on a statistics test are normally distributed with a mean of 62 and a variance of 225. If the instructor wishes to assign B's or higher to the top 30% of the students in the class, what mark is required to get a B or higher?**

   (a) 68.7
   (b) 71.5
   (c) 73.2
   (d) 74.6
   (e) 69.9

4. **The grade point averages of students at the University of Cairo are approximately normally distributed with mean equal to 2.4 and standard deviation equal to 0.8. What fraction of the students will possess a grade point average in excess of 3.0 ?**

   (a) 0.7500
   (b) 0.6000
   (c) 0.2734
   (d) 0.2500
   (e) 0.2266

5. **The diameters of steel disks produced in a plant are normally distributed with a mean of 2.5 cm and standard deviation of .02 cm. The probability that a disk picked at random has a diameter greater than 2.54 cm is about :**

   (a) 0.5080
   (b) 0.2000
   (c) 0.1587
   (d) 0.0228
   (e) 0.4920

6. **Suppose the test scores of 600 students are normally distributed with a mean of 76 and standard deviation of 8. The number of students scoring between 70 and 82 is :**

   (a) 272
   (b) 164
   (c) 260
   (d) 136
   (e) 328

7. **The cost of treatment per patient for a certain medical problem was modeled by one insurance company as a normal random variable with mean $775 and standard deviation $150. What is the probability that the treatment cost of a patient is less than $1,000, based on this model?**

   (a) 0.5000
   (b) 0.6826
   (c) 0.8531
   (d) 0.9332
   (e) Cannot be computed without knowledge of additional parameters

8. **The time that a skier takes on a downhill course has a normal distribution with a mean of 12.3 minutes and standard deviation of 0.4 minutes. The probability that on a random run the skier takes between 12.1 and 12.5 minutes is :**

   (a) 0.1915
   (b) 0.3830
   (c) 0.3085
   (d) 0.6170
   (e) 0.6826

9. **The time required to assemble an electronic component is normally distributed with a mean of 12 minutes and a standard deviation of 1.5 min. Find the probability that a particular assembly takes more than 14.25 minutes.**

   (a) 0.9332
   (b) 0.0668
   (c) 0.3413
   (d) 0.4332
   (e) 0.1587

10. Heights of males are approximately normally distributed with a mean of 170 cm and a standard deviation of 8 cm. What fraction of males are taller than 176 cm?

(a) 0.7500
(b) 0.6000
(c) 0.2734
(d) 0.2500
(e) 0.2266

11. The height of an adult male is known to be normally distributed with mean of 175 cm and standard deviation 6 cm. The 20th percentile of the distribution of heights is :

(a) 175
(b) 179
(c) 170
(d) 172
(e) 174

12. The heights of students at a college are normally distributed with a mean of 175 cm and a standard deviation of 6 cm. One might expect in a sample of 1000 students that the number with heights less than 163 cm is :

(a) 997
(b) 23
(c) 477
(d) 228
(e) 456

13. The distribution of weights in a large group is approximately normally distributed. The mean is 80 kg. and approximately 68% of the weights are between 70 and 90 kg. The standard deviation of the distribution of weights is equal to :

(a) 20
(b) 5
(c) 40
(d) 50
(e) 10

14. The daily milk production of Guernsey cows is approximately normally distributed with a mean of 35 kg/day and a std. deviation of 6 kg/day. The probability that a days production for a single animal will be less than 28 kg. is approximately :

(a) 0.41
(b) 0.09
(c) 0.38
(d) 0.12
(e) 0.62

15. Refer to the previous question. The producer is concerned when the milk production of a cow falls below the 5th percentile since the animal may be ill. The 5th percentile (in kg) of the daily milk production is approximately :

(a) 1.645
(b) -1.645
(c) 33.36
(d) 25.13
(e) 44.87

16. Which of the following is NOT CORRECT about a standard normal distribution?

(a) P (0 < Z < 1.50) = .4332
(b) P (Z < -1.0) = .1587
(c) P (Z > 2.0) = .0228
(d) P (Z < 1.5) = .9332
(e) P (Z > -2.5) = .4938

17. **Which value is closest to the 90th percentile for the standard normal distribution?**

(a) -1.3
(b) -0.7
(c) 0.7
(d) 1.3
(e) 1.4

18. **For a set of data that follow a normal distribution how many scores can one expect to find within one standard deviation on each side of the mean, that is two standard deviations in total ? (one correct choice)**

(a) 54%
(b) 99%
(c) 50%
(d) 88%
(e) 68%.

19. **We wish to estimate the cholesterol content in duck eggs. How large a sample should be selected if we can assume that F=15 mg also holds for duck eggs, and we wish our estimate to be correct within 5 mg with 99% confidence?**

(a) 8
(b) 99
(c) 35
(d) 75
(e) 60

20. **If P(-2<Z<k) = 0.6 , where Z is a standard normal random variable, then k is ..**

(a) 0.5773
(b) 0.195
(c) 0.73
(d) 0.55
(e) -0.40

21. The Vitamin C content of a particular brand of vitamin supplement pills is normally distributed with mean 490 mg and standard deviation 12 mg. What is the probability that a randomly selected pill contains at least 500 mg of Vitamin C?

(a) 0.7967
(b) 0.8333
(c) 0.0525
(d) 0.1123
(e) 0.2033

22. A wholesale distributor has found that the amount of a customer's order is a normal random variable with a mean of $200 and a standard deviation of $50. What is the probability that the total amount in a random sample of 20 orders is greater than $4500 ?

(a) 0.1915
(b) 0.0125
(c) 0.3085
(d) 0.0228
(e) 0.4875

23. The time required to assemble an electronic component is normally distributed with a mean of 12 minutes and a standard deviation of 1.5 min. Find the probability that the time required to assemble all nine components (i.e. the total assembly time) is greater than 117 minutes ?

(a) 0.2514
(b) 0.2486
(c) 0.4772
(d) 0.0228
(e) 0.0013

Chapter (7)

Demography

By the end of this chapter, the student should be able to :

- Understand census and methods of estimation of the population.
- Understand and interpret different patterns of population pyramid.
- Calcule growth rate and know factors affecting this rate.
- Know factors affecting fertility rates.
- Be oriented with overpopulation prolem in Egypt, hazards and management.

Demography is the scientific study of human population dynamics. It includes the study of the size, structure and distribution of populations, and how populations change over time due to births, deaths, migration and ageing.

Within this science, the following main subjects will be discussed :

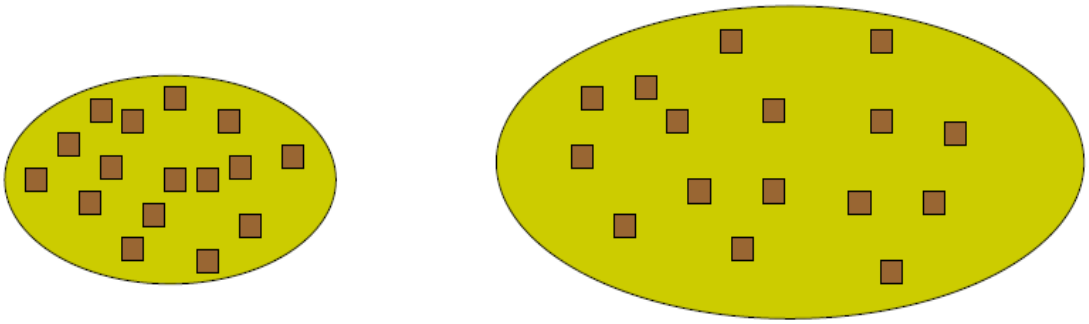
- Census.
- Fertility.
- Population growth.

Population :

Group of individuals of same species living in the same geographic area at the same time.

Population Density :

The number of individuals per unit area at a given time.



Same population with different densities

I- Census

Census is the process concerned with :

1. Enumeration of individuals all over the country at certain time.
2. Collection of demographic and socio-economic data of the population, which include age, sex, nationality, religion, education, occupation, income, marital status, family composition and other data. This data is collected by filling in a special form "census sheet" by head or a member of the family or trained persons. Census information is collected repeatedly at periods usually every 10 years.

Uses of census :

- To give characteristic features of the population.
- To provide data needed to calculate statistical rates and planning of different community programs (educational, health, socio-economic and others).

Census population and estimated population :

- *Census population :* It means the number of the population during the year in which the census was conducted.
- *Estimated population :* It means the number of the population calculated in any of the years between two censuses. It is obtained by application of different methods of estimation based on census data to estimate the number of population.

Whether census or estimated, the number of population is usually referred to as midyear population which means the number of the population in the middle of a given year (the first of July).

Methods of estimation of population :

1- Natural increase method :

Natural increase of population is the difference between number of live births and deaths in the years, which follow the census year. This natural increase value is added to the census population to get estimated population of a given inter-census year. This method can be used for rough estimation of the inter-census population of a given area having no or limited migration.

2- Arithmetic method :

Two consecutive census populations of a given area are taken, e.g. 200,000 in 1970 and 250,000 in 1980. To calculate the estimated population of 1984 the following should be calculated :

- Mean annual increase = (250,000 - 200,000) / (1980 - 1970) = 50,000 /10 = 5000.

- This means that the number of population increased in this area by an average of 5000 persons per year.
- So, the estimated population of 1984 = (250,000) + (5000 X 4) = 270,000.

This method, however, is rough, not accurate, since increase of population is geometric, not simple.

3- Geometric method :

It is the most accurate method of estimation. Special formula is used to find out the annual rate of population growth, to be applied to the last census population to get the estimated population of a given year.

4- Graphic method :

A number of successive census populations are plotted on a graph and joined together by straight line which is then extended over future years. Obtained graph can be used to find out :

- Inter-census population of a given year within the inter-census period.
- Expected populations of the future years, on assumption that growth rate of the population will not change.

Population Pyramid

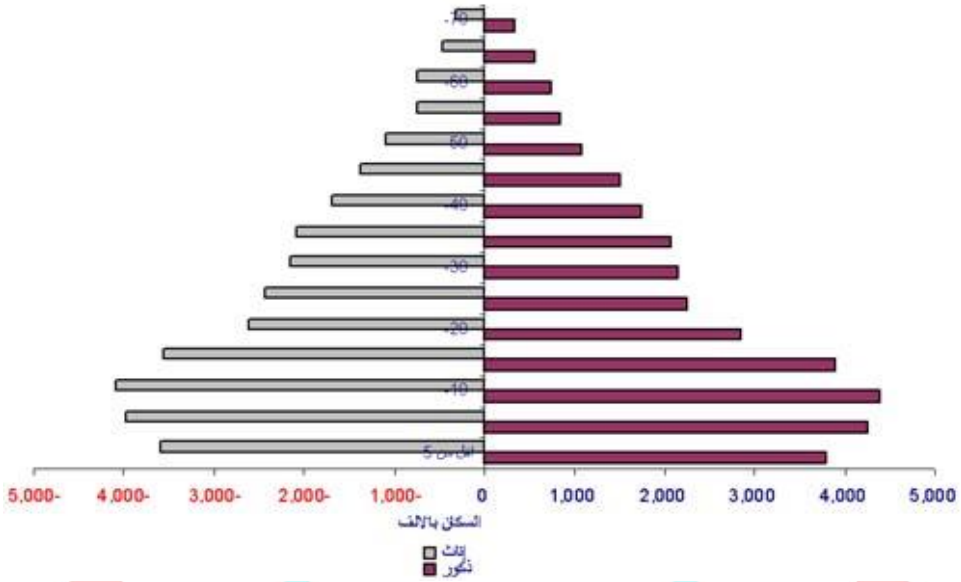
It is graphical presentation of age and sex composition of the population of any country.

- A population pyramid is a special type of bar chart.
- Number (or proportion) of males and females in each age group are plotted on the opposite sides.
- Number (or proportion) of males and females in each age group is directly proportional to the length of the horizontal bar.
- Younger ages (< 15 years) form the base.
- The elderly, of 65 years and over, form the apex.
- In-between, those of 15 – 64 years, form the body of the pyramid.

Shape of population pyramids varies in different countries according to age and sex distribution of the population (percent proportion of each age of both sexes to total population) :

- Young population : larger proportion of people in the younger age groups (<15 years) - in most less developed countries.
- Old or aging population : relatively large proportion of people in the older age groups (> 64 years) - in the more developed countries.

Characteristics of Population Pyramid of Egypt :



Population pyramid of Egypt ; October 2003

1- *Base layers :* Representing age groups below 15 years are broad due to high birth rate. They form around 42% of the population compared to narrower base layers of developed countries, which form 23% only of the population,

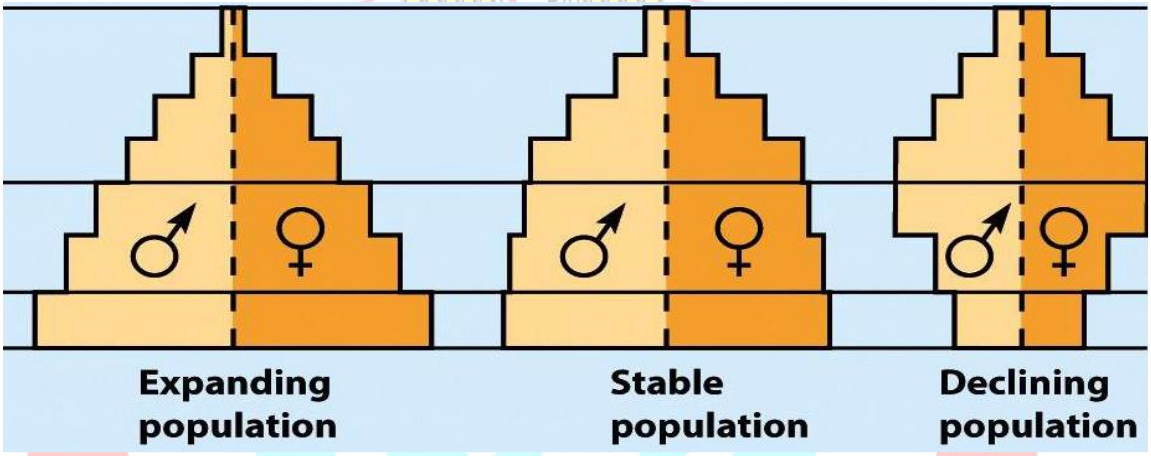
2- *Top of pyramid* is narrow due to smaller proportion of the elderly aged 65 years and over (nearly 7% at present). In developed countries, top of pyramid is flat due to higher proportion of the elderly.

3- *Body of the pyramid* made of strata in-between the base and top, which are smaller than those of developed countries due to less mean life expectancy.

4- *Sex distribution* : Male / female ratio of the population shows insignificant differences with slightly more percent of females.

N:B

- Each year a new cohort is born and added to the bottom of the pyramid, while the older cohorts move up as they age. The pyramid keeps narrowing with loss of members due to death (assuming no migration in or out).
- Rapidly increasing death rates after age 45 result in a narrowing peak in all population pyramids.



Life Expectancy :

It is the average number of years an individual is expected to live at a given age e.g. life expectancy at birth is the average number of years the newborn is expected to live. In Egypt, it is progressively increasing. It was 41.4 years for males compared to 47.0 years for females in 1947 and increased in 1998 to be 62 years for males and 67 years for females.

Median age : Age at which exactly half the population is older and half is younger. In 1999, the median age of Uganda (world's youngest population) was 17.5 years, while that in Italy (world's oldest population) was 39.9 years.

Age Dependency Ratio :

Age Dependency Ratio = Ratio of persons in the 'dependent ages' (under 15 and over 65) to those in the 'economically productive' ages.

$$\frac{P_{0-14} + P_{65+}}{P_{15-64}} \times 100$$

- *Child dependency ratio :* ratio of population under 15 to population 15 to 64.
- *Elderly dependency ratio :* ratio of population 65 and older to population 15 to 64.

Regions	Age <15 %	Age 15-64 %	Age 65+ %	Dependency ratio
Africa	44.9	52.0	3.0	92.2
Latin America	35.7	59.5	4.8	68.0
Europe	20.5	66.8	12.7	49.8

Sex Ratio :

Sex ratio is affected by :

- Sex ratio at birth.
- Differential patterns of mortality for males and females.
- Differential patterns of migration for males and females in population.

II- Fertility

Fertility, in demography, refers to the ability of females to produce healthy offspring. Fertility of a given population is measured by the birth rate, and other fertility indices (will be discussed later). Fertility is determined by fertility motives, fertility attitude and fertility behaviors which may be low or high.

High fertility motives :

1. High infant and child mortality : especially in rural and less developed urban areas of developing countries. Mothers like to have many births so as to ensure that some will continue to live to older ages.
2. Economic motive : with unsatisfactory socioeconomic level, children are employed to support family income.
3. Motives related to traditional concept of family and community welfare, where traditional societies believe that :
   - Marriage is happier and more stable with the many children of big family.
   - Sons continue the family name.
4. The many children of family are God's will.
5. Husbands find demonstration of vitality and manliness in high fertility and having many children.
6. People dislike using contraception as they believe that contraception interferes with sex, and they fear side effects.

Low Fertility Motives :

1. To preserve health of mother and children.
2. For better rearing of children who get more chance of survival, and better lifestyle.
3. For family welfare, to help having better standard of living, with less tension, overwork of parents, and avoid worsening poor economic conditions.
4. To avoid the overpopulation problem and support community development. Also, to help meeting the needs of the population, specially housing, food, education and public services.

Definitions :

- Parity = Number of children born alive to a woman.
- Gravidity = Number of pregnancies a woman has had whether or not they produce a live birth.
- Fecundity = Physiological capacity to conceive (reproductive potential), probability that a woman will conceive during a menstrual cycle.

III- Population Growth

Two factors influence growth of the population :

1- Natural increase : (The basic factor) :

It is the difference between live births and deaths.

2- Migration : (Minor factor, if any) :

It is the difference between immigration and emigration. It may increase or decrease the population, according to the extent of each.

Rate of natural increase of population (RNI) :

It is the rate of the difference between births and deaths.

$$\frac{No.of\ live\ births - No.of\ deaths\ (in\ certain\ locality\ and\ year)}{Midyear\ population\ of\ same\ locality\ and\ year}\ X\ 1000$$

Rate of natural increase of the population is high in developing countries, and eventually causes overpopulation problem (population crisis).

Overpopulation problem in Egypt

Overpopulation has become a prominent problem in Egypt and other developing countries, due to high RNI which arises from:

- High birth rate (though declined in Egypt at present, yet is still high 23.32 births/1000 population).
- Progressively declining death rate : It is the main factor for RNI becoming so high in Egypt (1.81%) and other countries, to give overpopulation.

Hazards of overpopulation :

The overpopulation problem threatens prosperity of mankind in developing countries which expend most of national income for needs of public services, especially food, thus interferes with socioeconomic development of the country. Hazards of overpopulation involve individuals, family and the nation.

1- Hazards for individuals :

- Infant and young children are exposed to risk of high morbidity and mortality (serious communicable diseases in dense populations).
- Mothers are exposed to hazards of repeated unspaced pregnancies, and also stress, worry and burden of big family size.
- Fathers suffer continues worry of big family, and face financial problems.

2- Hazards for family :

They arise from big family and low percapita income.

- Poor housing and unsanitary living conditions.
- Inadequate feeding and nutrition, especially of vulnerable groups.
- Employment of children who become deprived of school education and joyful social life.
- Some families may neglect care and culture of children who become exposed to maladjustment and delinquency.
- Deficient medical care.

3- National hazards :

Overpopulation is an obstacle to community development, being associated with socioeconomic problems and unsatisfactory standard of living, due to :

- Increase in national income cannot parallel the progressive geometric growth of population, and so the gap between the two lines gets continually wider.
- High proportion of dependent nonproductive group below 15 years (around 42% of total population).
- National hazards are repercussions of family hazards, giving rise to the following problems:
  - Housing : Problem of housing (with increased crowdness index) and slum areas are progressively increasing.
  - Food supply : Food imports are progressively increasing and prices rising, especially of animal protein foods.
  - School education : Available school buildings are cumulatively short of the increasing numbers of children, thus associated with crowdness of schools, and unsatisfactory educational process.
  - Public services : Are increasingly burdened to become inconvenient and not satisfying needs of the public.
  - Employment : Job vacancies and work opportunities cannot cope with the continually increasing young generation. Problem is still accentuated with development of mechanization and automation.

Management of overpopulation problem :

Long term and short term policies are needed. They interact and support each other, and must thus go together side by side.

1- Long-term policy :

National socioeconomic development is needed, to promote national, family and percapita income, and also manage the majority of high fertility motives :

- Better chance of child survival, and lowered infant and child mortality.
- Better standard of living: the family is not in need of employing children for economic support, and feels responsible for child welfare, and so prefers small family size.
- Better chance of education and culture and social change, with ambition for better living conditions.

2- Short-term solution :

- Family planning program, for fertility regulation and birth control.
- Promoting national productivity and investment, to meet needs of the public and verify more exports and less import.

Chapter (8)

Vital Statistics

By the end of this chapter, the student should be able to :

- Know different types of morbidity, mortality and fertility indices.
- Calculate and interpret different vital indices.
- Be oriented with raw and standardized rates.

Vital statistics is concerned with vital events of human life (births, deaths and morbidity). Registered data are used for calculation and presentation of vital rates which are indices (indicators) of :

- Health status of the population, especially vulnerable groups.
- Community development including socioeconomic features, which influences health (e.g., education, culture, nutritional status, and environment).
- Effectiveness (efficiency and utilization) of health services.

I- Birth (fertility) statistics :

They are presented as "Birth Rates" and "Fertility Rates".

1- Birth Rate :

Definition : Birth rate is the number of live births per 1000 population of a certain locality (or country) and year.

$$\text{Birth rate} = \frac{\text{No. of live births in a certain locality and year}}{\text{Midyear population of the same locality and year}} \times 1000$$

Birth rate is usually high in developing countries, due to :

- High fertility motives and behaviors, giving repeated un-spaced pregnancies throughout childbearing period, and no practice of birth control.
- Marriage of girls of traditional communities at young age, where fertility is higher through a long childbearing period.
- *Birth rate in Egypt* : Used to be high (40 - 45) then declined to become 37.5 in 1988, reached 29/1000 in 1996 and 26 / 1000 in 1999, and further lowering is expected (the rate is 15 or less in developed countries). What is the current birth rate in egypt?

2- Fertility Rates :

A woman is considered fertile when she has ever born a baby. A number of fertility rates (indices) are calculated each having particular significance.

a) General fertility rate (GFR) :

Definition : it is the number of live births per 1000 females of childbearing period in a certain locality (or country) and year. Childbearing period is 15 – 44, or 15 – 49 years old.

$$GRF = \frac{\text{No. of live births in certain locality and year}}{\text{No. of females in childbearing period of same locality and year}} \times 1000$$

It is about 3.6/1000 in Egypt in 1996. What is current fertility rate in Egypt?

b) Age-specific Fertility Rate : (Age-specific Birth Rate)

Definition : it is "total fertility of a particular age group" (of the seven 5-year groups) to get the average number of live births born to 1000 females in each of the 5 years of the age groups in a given locality and year.

Childbearing period (15 – 49) includes seven 5-year age groups. The total fertility of the such seven age groups is the number of live births born to1000 females of this group in a given locality and year. Total fertility of all females in childbearing period = sum of total fertility of the seven 5-year age groups.

Advantage : Age-specific fertility rate is a better index of fertility than the GFR as it considers differences in age distribution of females in populations of different countries.

c) Fecundity Rate :

Definition : It is the number of live births per 1000 married women of childbearing age in a certain locality and year. It is a valuable index of fertility being calculated for married not all females in childbearing period.

II- Morbidity Measures :

Calculation of disease frequency is based on the calculation of prevalence and incidence. Before calculation we should know the meanings of population at risk ratio, proportion and rate.

Population at risk :

An important factor in calculating disease frequency is the correct estimate of the numbers of people under study (population at risk). Ideally these numbers should only include people who are potentially susceptible to the diseases being studied; e.g., men should not be included when calculating the frequency of cervical cancer.

Ratio :

The simplest relation between numbers is a ratio and is expressed as X:Y (part : part) e.g. there is 20 male students and 10 female students in a classrooms, the ratio of male to female in this classroom is 20:10 or 2:1

Proportion : (part / total)

The relation between 2 numbers where one of them (the numerator) is always included in the other (the denominator). It is expressed as x / (x + y) x k (k is a constant value usually = 100) and in this case the proportion is called percent. In the above example the proportion of males is 20 / 30 x 100 = 66.7%

Rate :

It is a measure of the change of a quantity per unit time. For example if there is 1000 infants are vaccinated with BCG vaccine during 2011 and 200 failed to form scar, the proportion of failure of vaccination will be 200/1000 and when we express this proportion per time it becomes a rate ; the failure rate will be 200 / 1000/ year.

Measuring disease in a population :

1- Prevalence rate :

Prevalence is the frequency of existing cases in a defined population at a given point in time. Prevalence rate (P) of a disease is calculated as follows :

$$P = \frac{\text{Number of people with the disease or condition at a specified time} \times 10^n}{\text{Number of people in the population at risk at the specified time}}$$

Factors influencing prevalence of a disease :

Increased by	Decreased by
Longer duration of the disease	Shorter duration of the disease
Prolongation of life of patients without cure	High case-fatality rate from disease
Increase in new cases (increase in incidence)	Decrease in new cases (decrease in incidence)
In-migration of cases	In-migration of healthy people
Out-migration of healthy people	Out-migration of cases
Improved diagnostic facilities (better reporting)	Improved cure rate of cases

2- Incidence rate :

The incidence of disease represents the rate of occurrence of new cases arising in a given period in a specified population. Incidence rate (I) is calculated as follows :

$$I = \frac{\text{Number of new events in a specified period } (\times 10^n)}{\text{Number of persons exposed to risk during this period}}$$

Differences between prevalence and incidence.

	Prevalence	Incidence
Numerator	Number of **existing** cases of disease at a given point of time	Number of **new** cases of disease
Denominator	Population at risk	Population at risk
Focus	- Presence or absence of a disease - Time period is arbitrary ; rather a "snapshot" in time	- Whether the event is a new case - Time of onset of the disease
Uses	- Estimates the probability of the population being ill at the period of time being studied. - Useful in the study of the burden of chronic diseases and implication for health services	- Expresses the risk of becoming ill - The main measure of acute diseases or conditions, but also used for chronic diseases - More useful for studies of causation

III- Mortality Measures :

Causes of death are recorded on a standard death certificate, which carries information on age, sex, and place of residence. The International Statistical Classification of Diseases and Related Health Problems (ICD) provide guidelines on classifying deaths. The procedures are revised periodically to account for new diseases and changes in case-definitions, and are used for coding causes of death. The International Classification of Diseases is now in its 10th revision.

Limitations of death certificates :

Data derived from death statistics are prone to various sources of error and provide invaluable information on trends in a population's health status. The usefulness of the data from death certificate depends on :

- The completeness of records.
- The accuracy in assigning the underlying causes of death, Mortality (death) statistics is important for assessing the burden of disease, as well as for studying changes in diseases over time. So, the provision of accurate cause-of-death information is a priority for health services.

i. Mortality of Diseases :

Deaths of a particular disease can be presented in the following mortality rates :

1) Case-fatality Rate :

It is the number of deaths of a particular disease per 100 cases in a certain locality (or country) and year.

$$\text{Case fatality rate} = \frac{\text{No. of deaths of a particular disease in certain year and locality}}{\text{No. of reported cases of this particular disease in same year and locality}} \times 100$$

In outbreaks of diseases in a confined community, case-fatality rate is the number of deaths per 100 diagnosed cases during the period of outbreak. Case-fatality rate of a particular disease varies with severity of disease and whether complicated, early diagnosed, treatment is available, and health status of cases. For example, in case of meningococcal meningitis, case- fatality rate was high (more than 50%) in the past, and then declined with early diagnosis and availability of chemotherapy to below 5%. Invariably fatal infectious diseases (case-fatality is 100%) :

- Rabies, pneumonic plague and pneumonic anthrax, which are acute rapidly fatal diseases.
- ADIS : chronic infection, which is fatal after some varied period of time (months or years).

2) Mortality Rate of a particular disease : "cause-specific rate"

It is used for chronic disease (e-g pulmonary tuberculosis) in developing countries, where accurate number of cases occurring within a year is not available, due to deficient case finding and reporting.

Definition : Mortality rate of a particular disease is the number of deaths of disease per 100,000 population in a certain locality (or country) and year. Mortality rate of pulmonary tuberculosis in Egypt, for example, was 47/100,000 in 1947 and declined to below 2/100,000 at present (why?).

3) Proportionate Death Rate :

It is the percent proportion of the number of deaths of a particular disease to total deaths in a certain locality (or group of population, or country) and year. It shows the relative mortality role of each disease for a particular group (e.g., infants and preschool children), or population. Therefore, causes of death can be arranged by their magnitude and leading causes (major causes) of death can be found.

ii. Death Statistics :

1) General or Crude Death Rate :

- General death rate : As it represents all deaths, for all ages and causes, and not for specific group or cause.
- Crude death rate : Being not suitable for comparison with other countries and needs adjustment first.

Definition : The general death rate is the total number of deaths per 1000 population of a certain locality (or country) and year. It is about 8/1000 in Egypt in 1996 and 7/1000 in 1999. What is current crude death rate in Egypt?

$$\text{Crude death rate} = \frac{\text{Total number of deaths of a certain locality and year}}{\text{Midyear population of same locality and year}} \times 1000$$

Value of death rate :

1. Death rate is influenced by certain specific and general factors and can thus be used for comparison of these factors :
   - For different years in a particular country.
   - In-between different countries provided they have more or less similar age and sex distribution of the population.
2. Death rate is a direct index of specific factors related to morbidity.
   - Health status of the population, and health problems of the community.
   - Effectiveness (efficiency and utilization) of health services.
3. Death rate is indirect index of genera! factors influencing exposure to morbidity and mortality :
   - Socioeconomic and community development. Unsatisfactory development is characterized by unsanitary environment, poor living conditions, malnutrition and others.

- Education, culture, traditions and health awareness and behavior of the population. Illiteracy, and faulty traditional beliefs, habits and lifestyle predispose to morbidity.

How to explain decline of death rate of Egypt :

General and specific factors contribute to prevention and control of morbidity, and thus lower mortality :

1- Progressive community development, with better socioeconomic and environmental circumstances.
2- Upgraded primary health care, which provide preventive and curative health services for urban and rural population.
3- Prevention and control of communicable diseases that have significantly lowered deaths caused by such diseases.

2) Specific Death Rates :

The general death rate represents total deaths per 1000 population, while specific death rates represent deaths of a particular group of the population per 1000 individuals of the group. The group may be represented by age, sex, social, occupation, or some other variable.

Age-Specific death Rate

Deaths notified to the health office are registered in the "death record" under the following groups, in years :

- Below one year ; Infant Mortality Rate (IMR).
- Children 1 – 5 year's mortality.
- 5 – 14 years old.
- 15 – 44 years old.
- 45 – 59 years old
- 60 years old and over.

Chapter (9)

Hospital statistics

By the end of this chapter, the student should be able to :

- Know and calculate inpatient census.
- Understand and calculate measures of efficiency of bed utilization.

Hospital statistics use three sources of data :

1- The number of patients disposed (discharged, died or transferred to another hospital) during the period.
2- The number of beds available in the hospital.
3- The number of patients occupying a hospital bed each night.

Inpatient census :

It is the number of inpatients present at any time. It may be taken at any daily fixed and convenient time (census-taking time). However, it is usually taken at midnight.

Daily inpatient census :

It is calculated as follows :

The patients remaining in the hospital at the census-taking time for a specific day + (the admissions – the discharges, including deaths for the following day) ⟶ the patients remaining at the next census-taking time.

Inpatient service day (inpatient day) :

It refers to the services received by one inpatient in one 24-hour period. The "24-hour period" is the time between the census-taking hours of two successive days. For example : when the census-taking

time is midnight , the 24-hour period will be 12:01 A.M. through 12:00 P.M. (the same as the calendar day).

N:B

- One inpatient day must be counted for each inpatient admitted and discharged during the same day (between two successive census-taking hours i.e inpatient service day should never be reported as a fraction of a day.
- Inpatient census is used to calculate the inpatient service days, as every inpatient receives one inpatient service day each day he is hospitalized. Many inpatient service days are usually provided by a hospital on any one day.

Q- Why the number inpatient service days for a specific day is usually more than the corresponding daily inpatient census for that day?

Example :

At a certain hospital, the census-taking time is midnight, the number of patients in that hospital at midnight June 19 was 475, number of patients admitted on June 20 was 42, number of patients discharged (including deaths) on June 20 was 16. The number of patients both admitted and discharged (including deaths) on June 20 was 6.

- Calculate the daily inpatient census for June 20?
- Calculate the inpatient service days for June 20?

Answer :

Daily inpatient census for June 20 = 475 + (42 – 16) = 475 + 26 = 501

Inpatient service days for June 20 = 501 + 6 = 507

Average daily inpatient service days :

The formula to obtain the average daily inpatient service days for a hospital during a certain period is

$$\frac{\text{Total inpatient service days for the period}}{\text{Number of days of the same period}}$$

Example :

A hospital provided 960 service days to patients during November. Calculate the average daily inpatient service days for during this month.

Answer :

According to the formula, it is 960 / 30 = 32

The average daily inpatient service days during November was 32

How to measure the extent and efficiency of hospital beds utilization ?

There are many summary statistics which are commonly used to measure the degree of bed utilization in a hospital including :

(1) The number of bed-days used :

It is the sum of the occupied bed counts during a certain period.

(2) The bed-days available :

It is the number of beds available in the hospital. We can calculate the number of bed-days available during a certain period by :

The number of beds available X number of days covered by the period.

(3) The bed occupancy rate :

It is the percentage of available bed-days that were actually used during a certain period. That is

$$\text{Bed occupancy rate} = \frac{\text{Number of bed-days used}}{\text{Number of available bed-days}} \times 100$$

Example :

If there are 50 beds in a hospital, over a 30 day period ⟶ 1500 bed-days available. If the number of bed-days used over that period is 1200, then :

The bed occupancy rate is $\frac{1200}{1500} \times 100 = 80\%$

N:B

Bed occupancy rate over 100% can occur indicates that an extra number of beds have to be provided.

(4) The daily bed occupancy rate :

It is the mean number of patients occupying a bed per day. It is calculated by dividing the average daily inpatient service days by the number of beds available.

(5) The average length of stay :

It is the mean time per patient that a hospital bed is occupied during a period of time. It is calculated by dividing the number of bed-days used during a certain period by the number of disposals during the same period.

Importance :

- It is a measure of the efficiency of hospital care (how quickly patients are dealt with)
- When comparing length of stay between hospitals, it is very important to consider the medical specialty, type of patients and facilities available. (different illnesses require different periods of hospitalization).

(6) The average turnover :

It is the mean number of patients that have occupied any one bed during a period. It is calculated by dividing the number of disposals (discharges) by the average number of available beds.

(7) The average turnover interval :

This is the mean length of time that a hospital bed is left empty between two successive patients.

$$\text{It is} = \frac{\text{Number of bed-days available} - \text{number of bed-days used}}{\text{Number of disposals during a period}}$$

This rate is important to indicate the efficiency of hospital scheduling procedures for non-emergency admissions. Where it is important to avoid delay in replacing departing patients, assuming that the demand for places in hospital is always greater than the supply.

Exercise :

1- Define :

- Inpatient census.
- Daily inpatient census.

2- During 1995 a hospital ward with 20 beds had complete bed availability each day and had no occasion to borrow beds from other wards. The midnight occupied bed counts over the year produced the following distribution :

Number of occupied beds	12	13	14	15	16	17	18	19	20
Number of days	2	14	31	64	116	78	24	18	18

Calculate the following :

a- Daily bed occupancy rate.

b- The average length of stay, given that 850 patients were discharged, transferred or died in 1995.

c- The average turnover.

d- The average turnover interval.

REFERENCES

- Adrian Cook, Gopalakrishnan Netuveli & Aziz Sheikh. In : Basic Skills in Statistic, A Guide for Healthcare Professionals. Class Publishing (London) Ltd, 2004.
- Albert Jim and Jay Bennett. *Curve Ball : Baseball, Statistics, and the Role of Chance in the Game*. New York : Springer – Verlag, 2001, Chap. 2.
- Altman, D.G. & Bland, J.M. Presentation of numerical data. *British Medical Journal* (1996) ; 312 : 572.
- Altman, D.G. & Bland, J.M. Treatment allocation in controlled trials : why randomise. *British Medical Journal* (1999) ; 318 : 1209.
- Altman, D.G. ; Machin, D. ; Bryant, T. & Gardner, M.J. *Statistics with Confidence*, 2nd ed. (2000), London, BMJ Books.
- Altman, D.G. *Practical Statistics for Medical Research* London, Chapman & Hall. (1991).
- Armitage, P. ; Berry, P.J. & Matthews, J.N.S. *Statistical Methods in Medical Research*, 4th ed. (2002), Oxford, Blackwells.
- Bigwood, S. & Spore, M. *Presenting Numbers, Tables and Chart*. Oxford University Press (2003).
- Bradford Hill, A. Memories of the British streptomycin trial : the first randomised clinical trial. *Controlled Clinical Trials*, (1990) ; 11 : 77 – 79.
- Briscoe M.H. Preparing Scientific Illustrations : A Guide to Better Posters, Presentations, and Publications. 2nd edition. New York : Springer – Verlag, 1996.
- Campbell, M.J. ; Machin, D. & Walters, S.J. *Medical Statistics : a textbook for the health sciences*, 4th ed. (2007), Chichester, Wiley.

- Cleveland, W.S. *The Elements of Graphing Data*. New York : Van Nostrand Reinhold Co. 1994.
- Cleveland, W.S. *Visualizing Data*. Summit, NJ : Hobart Press, 1993.
- Edwards, A. ; Elwyn, G. & Mulley, A. Explaining risks : turning numerical data into meaningful pictures. *British Medical Journal* (2002) ; 324 : 827 – 830.
- Ehrenberg, A.S.C. 2000. *A primer in data reduction* Chichester, John Wiley & Sons.
- Freeman, J.V. & Campbell, M.J. 2006. Basic test for continuous data: Mann-Whitney U and Wilcoxon signed rank sum tests. *Scope*, 15, (4).
- Freeman, J.V. & Julious, S.A. 2005a. Describing and summarising data. *Scope*, 14, (3).
- Freeman, J.V. & Julious, S.A. 2005b. The Normal Distribution. *Scope*, 14, (4).
- Freeman, J.V. & Julious, S.A. 2005c. The visual display of quantitative information. *Scope*, 14, (2) : 11 – 15.
- Freeman, J.V. & Julious, S.A. 2006a. Basic tests for continuous Normally distributed data. *Scope*, 15, (3).
- Freeman, J.V. & Julious, S.A. 2006b. Hypothesis testing and estimation. *Scope*, 15, (1).
- Freeman, J.V. & Julious, S.A. 2007. The analysis of categorical data. *Scope*, 16, (1) 18 – 21.
- Freeman, J.V. & Young, T.A. 2009. Correlation coefficient : association between two continuous variables. *Scope (18) :* 31 – 33.

- Harris R.L. Information Graphics : A Comprehensive Illustrated Reference. Oxford : Oxford University Press, 1999.
- Harris, Robert. *Information Graphics : A Comprehensive Illustrated Reference*. New York : Oxford University Press, 2000.
- Huff, D. (1991). *How to lie with statistics* London, Penguin Books.
- Julious, S.A. & Mullee, M.A. (1994). Confounding and Simpson's paradox. *British Medical Journal ;* 308 : 1408 – 1481.
- Lang, T.A. & Secic, M. (1997). *How to report statistics in medicine* Philadelphia, American College of Physicians.
- Leroy Thacker, PhD VCU Department of Biostatistics and the VCU Center for Clinical and Translational Research. August, 2013 "Basic Statistical Methods for Residents".
- Medical Statistics from A to Z, A Guide for Clinicians and Medical Students. B.S. Everitt, Cambridge University Press, Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, Sao Paulo, 2006. Second Edition.
- Mosteller, F. and Kruskal, W. et al. *Statistics by Example : Exploring Data*. Reading, MA : Addison – Wesley, 1973.
- Stacey B. Plichta (Professor of Urban Public Health, Hunter College, The City University of New York (CUNY) New York, New York) & Laurel S. Garzon (Director, Graduate Nursing Programs, School of Nursing, Old Dominion University, Norfolk, Virginia) "Statistics for Nursing and Allied Health". Wolters Kluwer Health, Lippincott Williams & Wilkins, 2009.
- Swinscow, T.D.V. & Campbell, M.J. (2002). *Statistics at square one*, 10th ed. London, BMJ Books.

- Tufte, E.R. (1983). *The visual display of quantitative information* Cheshire, Connecticut, Graphics Press.
- Tufte, Edward, *The Visual Display of Quantitative Information*. Cheshire,CT: Graphics Press, 1983.
- Tukey, John W. *Exploratory Data Analysis*. Reading, MA : Addison – Wesley, 1977.
- Utts, Jessica. *Seeing Through Statistics*. Belmont, CA : Wadsworth Publishing Co. 1996.
- Wainer H. Understanding graphs and tables. Ed Researcher (1992) ; 21 : 14 – 23.
- Wainer, H. How to Display Data Badly. *The American Statistician* (1984) ; 38 : 137 – 47.
- White J. Using Charts and Graphs : 1000 Ideas for Visual Persuasion. New York : R.R. Bowker Company, 1984.
- Wildbur, Peter. *Information Graphics*. New York : Van Nostrand Reinhold Co. 1989.

Table 1: *Percentiles of the standard normal distribution.*

Standard score (z)	Percentile (c)	Standard score (z)	Percentile (c)	Standard score (z)	Percentile (c)
-3.4	0.03	-1.1	13.57	1.2	88.49
-3.3	0.05	-1.0	15.87	1.3	90.32
-3.2	0.07	-0.9	18.41	1.4	91.92
-3.1	0.10	-0.8	21.19	1.5	93.32
-3.0	0.13	-0.7	24.20	1.6	94.52
-2.9	0.19	-0.6	27.42	1.7	95.54
-2.8	0.26	-0.5	30.85	1.8	96.41
-2.7	0.35	-0.4	34.46	1.9	97.13
-2.6	0.47	-0.3	38.21	2.0	97.73
-2.5	0.62	-0.2	42.07	2.1	98.21
-2.4	0.82	-0.1	46.02	2.2	98.61
-2.3	1.07	0.0	50.00	2.3	98.93
-2.2	1.39	0.1	53.98	2.4	99.18
-2.1	1.79	0.2	57.93	2.5	99.38
-2.0	2.27	0.3	61.79	2.6	99.53
-1.9	2.87	0.4	65.54	2.7	99.65
-1.8	3.59	0.5	69.15	2.8	99.74
-1.7	4.46	0.6	72.58	2.9	99.81
-1.6	5.48	0.7	75.80	3.0	99.87
-1.5	6.68	0.8	78.81	3.1	99.90
-1.4	8.08	0.9	81.59	3.2	99.93
-1.3	9.68	1.0	84.13	3.3	99.95
-1.2	11.51	1.1	86.43	3.4	99.97

Percent of area under the normal curve between the mean and z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00.00	00.40	00.80	01.20	01.60	01.99	02.39	02.79	03.19	03.59
0.1	03.98	04.38	04.78	05.17	05.57	05.96	06.36	06.75	07.14	07.53
0.2	07.93	08.32	08.71	09.10	09.48	09.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.90	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.83	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	48.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
4.0	49.997									

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9031	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Book Coordinator ; Mostafa Fathallah

General Directorate of Technical Education for Health