Addis Ababa Science and Technology University

School of Interdisciplinary program directorate

Department of Statistics

Probability and statistics

By Ashebir Feyisa. (BSc, MSc in Biostatistics)

Email: asheber.feyisa@gmail.com

2015/2016

Introduction to Statistics

Objectives:

At the end of this session, students should be able to:

- understand statistics and basic terminologies
- understand scales of measurement in statistics
- understand the basic methods of data collection

Definition of Statistics

- The word statistics has several meanings.
  - We can define statistics either in plural or singular sense.
- ***In plural sense:*** statistics is defined as the collection of numerical facts or figures (or the raw data themselves).
- In this sense the word 'statistics' is usually understood by a layman.

- Eg. 1. Vital statistics (numerical data on marriage, births, deaths, etc).
- 2. The average mark of statistics course for students is 70% would be considered as a statistics whereas Abebe has got 90% in statistics course is not statistics.
- **Remark:** statistics are aggregate of facts. Single and isolated figures are not statistics as they cannot be compared and are unrelated.

Definition of Statistics

- **In its singular sense**:- Statistics is the science that deals with the methods of collecting, organizing, presenting, analyzing and interpreting statistical data.

Classification of Statistics

- Statistics may be divided into two main br anches:

  I. Descriptive Statistics
  II. Inferential Statistics

Classification of Statistics cont…

Descriptive statistics:

- Includes statistical methods involving the collection, presentation, and characterization of a set of data in order to describe the various features of the data.
- Methods of descriptive statistics include graphic methods (bar chart, pie chart, e t c) and numeric measures (mean, median, variance e t c).
- Descriptive statistics do not allow us to make conclusions beyond the data we have analyzed.

Classification of Statistics cont…

- Meaningful and pertinent information cannot be realized from raw data unless summarized by the tools of descriptive statistics.
- Descriptive statistics, therefore, allow us to present the data in a more meaningful way which allows interpretation of the data easily.

Classification of Statistics cont…

Inferential statistics:

- Includes statistical methods which facilitate estimation the characteristics of a population or making decisions concerning a population on the basis of sample results.

- In this regard, methods like estimation and hypothesis testing are examples of inferential statistics.

Classification of Statistics cont…

- For example, a biologist collected blood samples of 10 students from biology department to study blood types. Accordingly, the following data is obtained:

- O A O AB A A O O B O

- Summary measures, for example, the proportion of students with blood type O in the sample is 50% is an example of descriptive statistics. We can also describe the data using bar or pie charts.

- However, if he/she wants to get information on the proportion of students with blood type O in the entire class, he/she may use the sample proportion (50%) as an estimate of the corresponding value of the entire class. This is an example of inferential statistics.

Stages in statistical investigation

- A statistical study might involve the following stages: collection of data, organizing and presenting the collected data, analyzing and interpreting the result.
- **Stage 1:** Data collection: this stage involves acquiring data related with the problem at hand.
- **Stage 2:** Organizing: this stage involves the classification or sorting the collected data based on some characteristics or attributes such as age, sex, marital status e t c.
- **Stage 3:** presenting data: Further we may use tables, graphs, charts so on to present the data.

Stages in statistical investigation

- **Stage 4:** Data analysis: a thorough scrutiny or analysis of the data is necessary in order to reach conclusions or provide answers to a problem. The analysis might require simple or sophisticated statistical tools depending on the type of answers that may have to be provided.

- **Stage 5:** Interpretation of the result: logically a statistical analysis has to be followed by conclusions in order to be able to make a decision. The technical terminology used to describe this last process of a statistical study is referred to as interpretation.

Definition of some terms

- **A population:** Consists of all elements, individuals, items or objects whose characteristics are being studied. The population that is being studied is called *target population.*
- **Sample:** A portion of the population selected for study.
- **Sample survey**: The technique of collecting information from a portion of the population.
- **Census survey**: A survey that includes every member of the population.
- **Variable**: is a characteristic under study that assumes different values for different element.
- **Quantitative variable**: A variable that can be measured numerically. The data collected on quantitative variable are called *quantitative data.* Examples include weight, height, number of students in a class, number of car accidents, e t c.

Definition of some terms cont…

- **Qualitative variable:** A variable that cannot assume a numerical value but can be classified into two or more non numerical categories. The data collected on such a variable are called q*ualitative or categorical data.* Examples include sex, blood type, marital status, religion e t c.
- **Discrete variable:** a variable whose values are countable. Examples include number patients in a hospital, number of white blood cells in a droplet of blood sample, number of rodents per plot of farmland e t c.
- **Continuous variable:** a variable that can assume any numerical value over a certain interval or intervals. Examples include weight of new born babies, height of seedlings, temperature measurements e t c.

Definition of some terms cont.....

- **Parameter:** A statistical measure obtained from a population data. Examples include population mean, proportion, variance and so on.
- **Statistic:** A statistical measure obtained from a sample data. Examples include sample mean, proportion, variance and so on.
- **Unit of analysis:** The type of thing being measured in the data, such as persons, families, households, states, nations, etc.

Limitation of statistics

- Statistics deals with only those subjects of inquiry which are capable of being quantitatively measured and numerically expressed.
- Statistics deals only with aggregates of facts and no importance is attached to individual items
- Statistical data is only approximately and not mathematically correct
- Statistics is liable to be misused. Hence expertise in the subject is very essential. Besides, honesty is very important in the use of statistics.

Scales of measurement

- Formally, we distinguish among four levels of measurement scales.

Scales of measurements cont…

Nominal scale:

- It is the simplest measurement scale.
- There is no natural ordering of the levels or values of the scale in nominal scale.
- For example, sex of an individual may be male or female. There is no natural ordering of the two sexes. Others examples include religion, blood type, eye colour, marital status e t c.
- The values of nominal scale can be coded using numerical values;
- However, we cannot perform any mathematical operations on the numbers used to code.

Scales of measurements cont..

Ordinal scale:

- This measurement scale is similar to the nominal scale but the levels or categories can be ranked or order.
- That is, we can compare levels or categories of the scale.
- Therefore, this scale of measurement gives better information on the quantities being measured as compared to nominal scale. For example, living standard of a family can be poor, medium or higher.
- These categories can be ordered as poor is less than medium and medium is less than higher class.
- However, the distance or magnitude between the levels, say between poor and medium, is not clearly known.

Scales of measurements cont…

Interval scale:

- This measurement scale shares the ordering or ranking and labeling properties of ordinal scale of measurement. Besides, the distance or magnitude between two values is clearly known (meaningful).
- However, it lacks a true zero point (i.e., zero point is not meaningful). For example, temperature in degree centigrade or Fahrenheit of an object. If the temperature of an object is zero degree centigrade, it doesn't mean that the object lacks heat. Hence zero is arbitrary point in the scale. It doesn't make sense to say that 80° F is twice as hot as 40° F.
- We can do subtraction and addition on interval level data but division and multiplication are impossible.

Scales of measurements cont…

Ratio scale:

- It is the highest level of measurement scale.
- It shares the ordering, labeling and meaningful distance properties of interval scale.
- In addition, it has a true or meaningful zero point. The existence of a true zero makes the ratio of two measures meaningful. example includes, weight, height e t c.
- We can do subtraction, addition, multiplication and division on ratio level data

Scales of measurements cont…

- The more precise variable is ratio variable and the least precise is the nominal variable. Ratio and interval level data are classified under quantitative variable and, nominal and ordinal level data are classified under qualitative variable.

Addis Ababa Science and Technology University

School of Interdisciplinary program directorate

Department of Statistics

Probability and statistics for engineers

By Mulugeta G. (BSc, MPH)

Email: mullergaro@gmail.com

2015/2016

UNIT TWO

METHODS OF DATA COLLECTION AND PRESENTATION

Ashebir Feyisa

12/07/2025

After completing this unit you should be able to:

- organize data using frequency distribution.
- present data using suitable graphs or diagrams.

Methods of data collection

- Depending on the source, data can be classified in to two:

1. Primary data &
2. Secondary data

- **Primary data** refers to the statistical data which the investigator originates for the purpose of inquiry.
- **Secondary data** refers to data which is not originated by the investigator himself, but which he/she obtains from someone else records. Secondary data can be obtained from **published or unpublished documents: reports, journals, magazines, articles** e t c.

Methods of data collection cont…

- **Primary methods of data collection:** It includes data collection using *observation, personal interview, self administered questionnaire, mailed questionnaire etc.*

Classification and tabulation of data

- The uses of classifying and tabulating data are:
  - to display the points of similarity and dissimilarity;
  - to save mental strain by systematic condensation and suppression of irrelevant detail;
  - to enable one to form a mental picture of objects of perception; and
  - to prepare the ground for comparison and inference.

Types of classification

- Geographical- in terms of cities, districts, countries etc.
- Chronological - on the basis of time
- Qualitative - according to some qualitative characteristics.
- Quantitative – in terms of magnitude.

Tabulation

- **Tabulation:** tables may be classified according to the number of characteristics used for tabulation.
- ***Simple or one way table***: it uses only one characteristic or variable for classification.
- **Example 2.1:** Students who took introduction to statistics in 2014 G.C.by gender.

Gender	Number
Male	2000
Female	700

Tabulation cont…

- ***Two-way tables***: it uses two variables for classification.
- **Example 2.2:** Students who took introduction to statistics in 2007 E.C.by age and gender.

Age	**Gender**	
	Number of male	Number of female
19 and below	200	180
20-25	1415	385
26 and above	385	135

Frequency distributions

Frequency distribution is the easiest method of organizing data, which converts raw data into a meaningful pattern for statistical analysis.

The main uses of a frequency distribution are:

- to organize data in a meaningful way.
- to enable one to determine the nature or shape of the distribution; how the observations cluster around a central value; and how the values spread around the center of the data.
- to facilitate computational procedures for measures of average and spread.
- to enable one to draw charts and graphs for the presentation of data.
- to enable one to make comparisons between data sets.

Terminologies

- **Frequency distribution**: a grouping of data into categories showing the number of observations in each mutually exclusive category.
- **Array:** data put in an ascending or descending order of magnitude.
- **Grouped data**: data presented in the form of a frequency distribution.
- **Frequency**: the number of observations corresponding to a fixed value or to a class of values.
- **Relative frequency**: the number obtained when the frequency of a class is divided by total number of observations.

Components of a frequency distribution

- **Class limits:** the values of a variable which typically serve to identify the classes of a frequency distribution.
- **Class boundaries**: the precise points which separate various classes rather than the values included in any one of the classes.
- **Class mark**: the point which divides the class into two equal parts. This is also known as c*lass mid-point*. This can be determined by dividing the sum of the two limits or the sum of the two boundaries by 2.
- **Class width**: the length of a class

- **Example 2.3:** The following data are the weights in kg of 40 individuals participated in a diet program for weight loss:
- 70 64 99 55 64 89 87 65 62 38 67 70 60 69 78 39 75 56 71 51 99 68 95 86 57 53 47 50 55 81 80 98 51 36 63 66 85 79 83 70
- By grouping data into classes we can make the data much easier to **read and understand**. Considering 10 as a class width. The smallest weight is 36 kg, thus the first class of weights is 31 kg.

Class	Class boundary	Count (Frequency)
31 – 40	30.5-40.5	3
41 – 50	40.5-50.5	2
51 – 60	50.5-60.5	8
61 – 70	60.5-70.5	12
71 – 80	70.5-80.5	5
81 – 90	80.5-90.5	6
91 – 100	90.5-100.5	4
Total		40

Steps of constructing frequency distribution

1) Find the highest and the smallest value,
2) Compute the range; R = H – L,
3) Determine the number of classes using sturgges formula

   K= 1 + 3.322Log n; n= Total frequency
4) Find the class width (W) by dividing the range by the number of classes and round to the nearest integer.

$$W = R/K$$

5) Identify the unit of measure usually as 1, 0.1, 0.01,.....
6) Pick a minimum value as starting point. Your starting point is lower limit of the first class, then continue to add the class width to get the rest lower class limits.
7) Find the upper class limits $UCL_i = LCL_i + w-U$. then continue to add width to get the rest upper class limit
8) Finally find the class frequencies.

- **Example 2.4:** The following data are on the number of minutes to travel from home to work for a group of automobile workers:
- 28 25 48 37 41 19 32 26 16 23 23 29 36 31 26 21 32 25 31 43 35 42 38 33 28. Construct a frequency distribution for this data.

Solution:

- Range = 48 – 16 =32
- K=1+3.322log 25 =5.64≈6
- W=32/6=5.33 rounding up to the nearest integer i.e W=6.

- Let the lower limit of the first class be 16 then the frequency distribution is as follows:

Class limit	Class boundaries	Tally	Frequency
16-21	15.5-21.5	\\\	3
22-27	21.5-27.5	\\\\\ \	6
28-33	27.5-33.5	\\\\\ \\\	8
34-39	33.5-39.5	\\\\	4
40-45	39.5-45.5	\\\	3
46-51	45.5-51.5	\	1
Total			25

Types of frequency distributions

- Based on the type of frequency assigned to the classes we have three types of frequency distributions:
  - Absolute frequency distribution
  - Relative frequency distribution
  - Cumulative frequency distribution

  - The frequency distributions that we have seen in the previous examples are absolute frequency distributions because the frequencies assigned are absolute frequencies.

Relative frequency distribution

- **Definition 2.1:** A relative frequency distribution is a distribution which specifies the frequency of a class relative to the total frequency.
  - By dividing the absolute frequency to total frequency in example 2.4 we can get relative frequency distribution.

Time (in minute)	Relative frequency
16-21	0.12
22-27	0.24
28-33	0.32
34-39	0.16
40-45	0.12
46-51	0.04
Total	1

Cumulative frequency distribution

- **Definition 2.2**: Cumulative frequency refers to the number of observations that are below/above a specified value.
- **Note:** Class boundaries are mostly used to obtain cumulative frequencies. Based on whether the observations are bounded from above or from below, we can have a cumulative less than or a cumulative more than frequency distributions, respectively.

- **Example 2.6:** Convert the absolute frequency distribution in example 2.4 into:
  - a cumulative less than frequency distribution.
  - a cumulative more than frequency distribution.

Table: Less than cumulative frequency distribution of times

Time (in minute)	Less than cumulative frequency
15.5- 21.5	3
21.5-27.5	9
27.5-33.5	17
33.5-39.5	21
39.5-45.5	24
45.5-51.5	25

More than cumulative frequency distribution

- Table: More than cumulative frequency distribution

Time (In minute)	More than cumulative frequency
15.5-21.5	25
21.5-27.5	22
27.5-33.5	16
33.5-39.5	8
39.5-45.5	4
45.5-51.5	1

Ungrouped frequency distributions (Single-value grouping)

- **Example 2.7**: A demographer is interested in the number of children a family may have. He took a random sample of 30 families. The following data is the number of children in a sample of 30 families.
- 4 2 4 3 2 8 3 4 4 2 2 8 5 3 4
  4 5 4 3 5 2 7 3 3 6 7 3 8 4 5
- To group these data, we will use classes based on the single numerical value.

Ungrouped frequency distributions

- Table: Distribution of the number of children

Number of Children	Frequency	Relative frequency
2	5	.17
3	7	.23
4	8	.27
5	4	.13
6	1	.03
7	2	.07
8	3	.1
Total	30	1

Categorical frequency distributions

- **Note**: Up to now we have seen frequency distributions for quantitative data; we can have also frequency distributions for qualitative (categorical) data.
  - The categorical frequency distribution is used for data which can be placed in specific categories such as nominal or ordinal level data.
  - For example, data on political affiliation, religious affiliation, blood type, marital status, or major field of study would use categorical frequency distributions

Categorical frequency distributions cont...

- **Example 2.8:** The following data are on the political party affiliations of sample of 40 engineering students. D, R, and O stand for Democratic, Republican and Other, respectively.
- D D D D O R O R O R O R O D D R D D D R R O R D R R O R R R R R O O R R D R D D
- The classes for grouping are ‘Democratic’, ‘Republican’ and ‘Other’

Categorical frequency distributions cont...

- Table: Number of students by political party affiliations.

Class	frequency	Relative frequency
Democratic	13	0.325
Republican	18	0.45
Other	9	0.225
Total	40	1

Diagrammatic and graphical presentation of data

Graphs for quantitative data

- **Histogram**: it consists of a set of adjacent rectangles whose bases are marked off by class boundaries (not class limits) along the horizontal axis and whose heights are proportional to the frequencies associated with the respective classes.

To construct a histogram from a data set:

- Construct a frequency table.
- Draw adjacent bars having heights determined by the frequencies in step1.

- Histogram can often indicate how symmetric the data are; how spread out the data are; whether there are intervals having high levels of data concentration; whether there are gaps in the data; and whether some data values are far apart from others.

- **Example 2.9:** The following is a histogram for the frequency distribution in example 2.4.

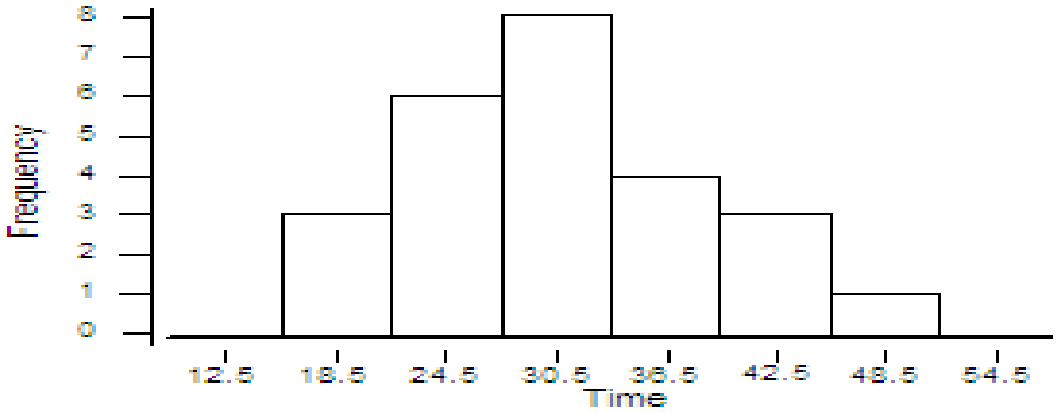
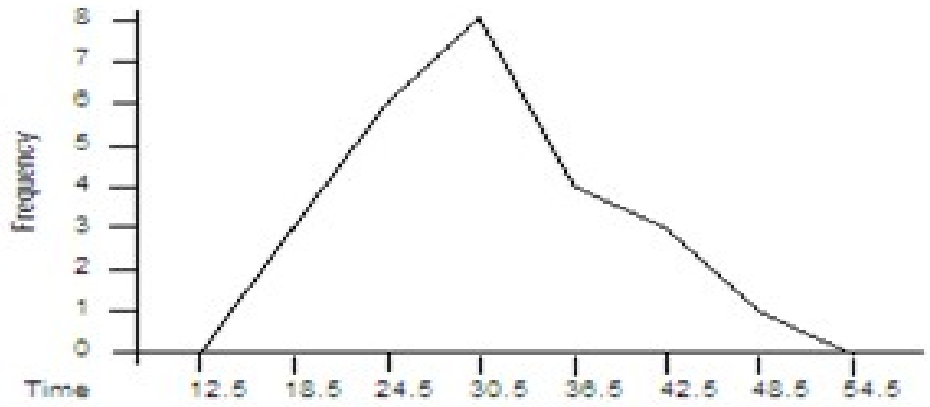


Figure: Distribution of number of minutes spent by the automobile workers

- **Frequency polygon**: is a graphic form of a frequency distribution. It can be constructed by plotting the class frequencies against class marks and joining them by a set of line segments.
- **Note:** we should add two classes with zero frequencies at the two ends of the frequency distribution to complete the polygon.

- **Example 2.10:** Construct a frequency polygon for the frequency distribution of the time spent by the automobile workers that we have seen in example 2.4



- Figure: Distribution of number of minutes spent by the automobile workers

Graphs useful for presenting qualitative data

- **Bar charts** are diagrammatic representation of data in which the data are represented by series of vertical or horizontal bars, the height (or length) of each bar indicating the size of the figure represented.
- **Example 2.11:** Draw a bar chart for the following coffee production data.

- Table: Coffee productions from 1990 to

Production year	1990	1991	1992	1993	1994	1995
Amounts of coffee (in 1000 tons)	50	75	92	64	100	120

- **Pie-chart: it** is a circle divided by radial lines into sections or sectors so that the area of each sector is proportional to the size of the figure represented.
- Pie-chart construction:
  - Calculate the percentage frequency of each component. It is given by $\frac{f_i}{\sum f_i} \times 100$
  - Calculate the degree measures of each sector. It is given by $\frac{f_i}{n} \times 360^0$
  - Then draw the circle.

- **Example 2.13:** Draw a pie-chart to represent the following data on a certain family expenditure.

Item	Food	Clothing	House rent	Fuel & light	Miscellaneous	Total
Expenditure (in birr)	50	30	20	15	35	150
Percentage frequencies	33.33	20	13.33	10	23.33	
Angles of the sector	120^0	72^0	48^0	36^0	84^0	360^0

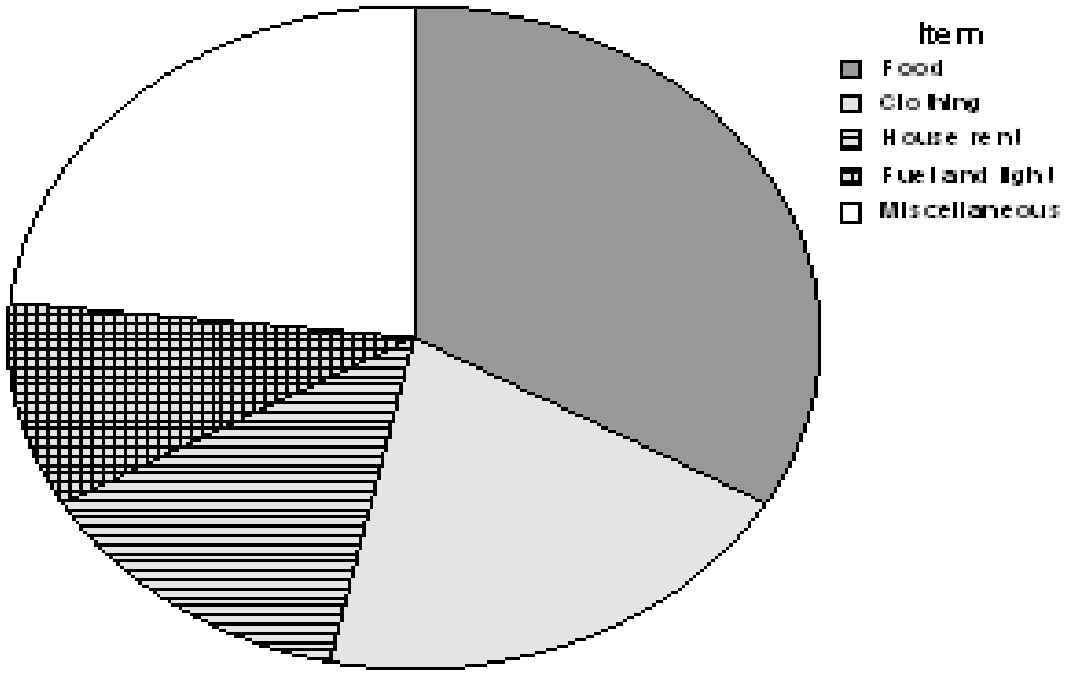


Figure: Family expenditure

Ashebir Feyisa

12/07/2025

MEASURES OF CENTRAL TENDENCY

Introduction and objectives of measuring central tendency

- In the pervious section, we have discussed how raw data can be organized in terms of tables, charts and frequency distributions in order to be easily understood and analyzed.
- Frequency distributions and their corresponding graphical displays roughly tell us some of the features of a data set.
- However, they don’t condense the mass of data in a way that we can easily understand and interpret.
- In this section, we will see how to summarize data using a descriptive measure called average. This will help us in condensing a mass of data into a single value which is in some sense representative of the whole data set.

- An average is a single value intended to represent a distribution as a whole.
- Note that the individual values of the distribution must have a tendency to cluster around an average. In view of this requirement an average is also referred to as a *measure of central tendency.*

- An average (a measure of central tendency) is considered satisfactory if it possesses all or most of the following properties. An average should be:
  - Rigidly defined (unique),
  - Based on all observation under investigation
  - Easily understood,
  - Simple to compute
  - Suitable for further mathematical treatment
  - Little affected by fluctuations of sampling
  - Not highly affected by extreme values.

The summation notation

- Suppose a variable is represented by X. The successive values of this variable may be represented by using subscripts or indexes as x_1, x_2, x_3,..., x_n. If the sum of these values or terms is required, we write $x_1+x_2+x_3+...+x_n$. The Greek letter $\sum$ (read as sigma) can be used to write the above sum in a compact form as $x_1 + x_2 + x_3 + \cdots + x_n = \sum_{i=1}^{n} x_i$

where 1= lower limit and n = upper limit.

When no confusion can result we shall often denote the above sum simply by $\sum x$ or $\sum x_i$ or $\sum_i x_i$.

Using this notation we may have also occasions to write terms like $\sum x_i^2, \sum_{i=1}^n x_i y_i, \sum x_i f_i, \sum x_i^2 f_i \ldots$

$$\sum_n x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$

$$\sum_{i=1}^n x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + \cdots + x_n y_n$$

$$\sum_{i=1}^n x_i f_i = x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_n f_n$$

$$\sum_{i=1}^n x_i^2 f_i = x_1^2 f_1 + x_2^2 f_2 + x_3^2 f_3 + \cdots + x_n^2 f_n$$

Rules of summation

$$\sum_{i=1}^{n}(x_i \pm y_i) = \sum_{i=1}^{n} x_i \pm \sum_{i=1}^{n} y_i \qquad \sum_{i=1}^{n} kx_i = k\sum_{i=1}^{n} x_i \text{ if } k \text{ is a constant.}$$

and $\sum_{i=1}^{n} k = nk$ if k is a constant.

Types of measures of central tendency

- **Arithmetic mean**

Definition 3.2:

i) Let $x_1, x_2, x_3, \ldots, x_n$ be the values of the variable X. The simple arithmetic mean denoted by $\bar{x}$ is the sum of these observations of X divided by the no values.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

ii) If the numbers $x_1, x_2, x_3, \ldots, x_k$ occur with frequencies $f_1, f_2, f_3, \ldots, f_k$, respectively. Then mean can be defined in a more compact form as

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_k f_k}{f_1 + f_2 + f_3 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

- Note that if the data refers to a population data the mean is denoted by the Greek letter μ (read as mu).

Arithmetic mean for raw data (ungrouped data)

- **Example 3.1:** The following data is the weight (in Kg) of eight youths: 32,37,41,39,36,43,48 and 36. Calculate t

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{32 + 37 + 41 + \cdots + 36}{n} = \frac{312}{8} = 39$$

- **Example 3.2:** The ages of a random sample of patients in a given hospital in Ethiopia is given below:

Age	10	12	14	16	18	20	22
Number of patients	3	6	10	14	11	5	4

- Calculate the average age of these patients.
- **Solution:**

Age (x_i)	Number of patients (f_i)	fixi
10	3	30
12	6	72
14	10	140
16	14	224
18	11	198
20	5	100
22	4	88
Total	53	852

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \cdots + x_k f_k}{f_1 + f_2 + f_3 + \cdots + f_k}$$

$$= \frac{10 \times 3 + 12 \times 6 + 14 \times 10 + 16 \times 14 + 18 \times 11 + 20 \times 5 + 22 \times 4}{3 + 6 + 10 + 14 + 11 + 5 + 4}$$

$$= \frac{30 + 72 + 140 + 224 + 198 + 100 + 88}{53} = \frac{852}{53} = 16.075$$

Thus the mean age of these patients is 16.075.

The weighted arithmetic mean

- In some cases the data in the sample or population should not be weighted equally, and each value weighted according to its importance.
- There is a measure of average for such problems known as *weighted Arithmetic mean*.
- Weighted arithmetic mean is used to calculate the average when the relative importance of the observations differs.
- This relative importance is technically known as weight.
- Weight could be a frequency or numerical coefficient associated with observations.

Definition 3.3: If $x_1, x_2, x_3, \ldots, x_n$ have weights $w_1, w_2, w_3, \ldots, w_n$, respectively, then the weighted arithmetic mean denoted by $\bar{x}_w$, is defined as

$$\bar{x}_w = \frac{x_1 w_1 + x_2 w_2 + x_3 w_3 + \cdots + x_n w_n}{w_1 + w_2 + w_3 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Example 3.3:** The GPA or CGPA of a student is a good example of a weighted arithmetic mean. Suppose that Solomon obtained the following grades in the first semester of the freshman program at AASTU in 2006.

Course	Credit hour (w_i)	Grade
Math101	4	A=4
Stat2091	3	C=2
Chem101	3	B=3
Phys101	4	B=3
Flen101	3	C=2

- Find the GPA of Solomon.

$$\bar{x}_w = GPA = \frac{4(4) + 3(2) + 3(3) + 4(3) + 3(2)}{4 + 3 + 3 + 4 + 3} = \frac{49}{17} = 2.88$$

Properties of arithmetic mean

- It can be computed for any set of numerical data, it always exists, and unique.
- It depends on all observations.
- The sum of deviations of the observations about the mean is zero i.e. $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

- It is greatly affected by extreme values.
- It lends itself to further statistical treatment, for instance, combinations of means.
- It is relatively reliable, i.e. it is not greatly affected by fluctuations in sampling.
- Th
-

i.e. $\sum_{i=1}^{n}(x_i - \bar{X})^2 \leq \sum_{i=1}^{n}(x_i - A)^2$ for any constant A.

observations about the mean is the minimum

Geometric mean

Definition 3.4: The geometric mean of any n positive numbers is the n^{th} root of the products of the numbers. Symbolically if $x_1, x_2, x_3, \ldots, x_n$ are given their geometric (G.M) mean is given by

$$G.M = \sqrt[n]{x_1.x_2.x_3.\ldots.x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Example 3.5: Find the geometric mean of the following numbers 2, 4, 8.

Solution:

$$G.M = \sqrt[3]{2.4.8} = \sqrt[3]{64} = 4.$$

Note: The geometric mean is useful in finding the average of percentages, ratios, indexes, or growth rates.

- **Example 3.6**: During the beginning of an epidemic in a region 12 cases were reported in the first day, 18 on second day and 48 on the third day.
- Find the average growth rate of the epidemic disease.
- Assuming that the growth pattern continues, forecast the number of cases that would be reported on the 4^{th} and 8^{th} days.
- **Solution:**
- Find the 2 growth rates first.
- From first day to second day the rate is 18/12=1.5.
- From second day to third day the rate is 48/18=2.67.

- Therefore, the average rate $= \sqrt{\frac{18}{12}\cdot\frac{48}{18}} = \sqrt{4} = 2$
- b)

Day	Number of cases
1st	12
2nd	24=2×12
3rd	48=2× 2 × 12 = 2 × 24
4th	96=2× 2 × 2 × 12 = 2 × 48
5th	192=2× 2 × 2 × 2 × 12 = 2 × 96

The median

Definition 3.5: the median of a set of data is a value which divides the set in such a way that the number of observations below it is the same as the number of observations above it.

Median for raw data

i. If the number of observations, say n, is odd then the median is equal to the $\left(\frac{n+1}{2}\right)th$ observation of the array.

ii. If the number of observations n is even then the median is equal to the sum of $\left(\frac{n}{2}\right)th$ observation and $\left(\frac{n}{2}+1\right)th$ observation divided by two.

Notation: If X is the variable under consideration, then $\tilde{x}$ is used to denote the median.

Example 3.9: Find the median for the following sets of data:

i. 10 5 7 9 6 5 4

Solution: First arrange the data in the form of an array.

4 5 5 6 7 9 10

Here we have n=7 which is odd

Therefore, the median, $\tilde{x} = the\left(\frac{n+1}{2}\right)th$ observation = the 4^{th} observation = 6.

ii. 10 5 7 9 6 5 4 8

Solution: Arrange the data in ascending order.

4 5 5 6 7 8 9 10

Here n=8 which is even.

Therefore, $\tilde{x} = \frac{(\frac{8}{2})th\ observation + (\frac{8}{2}+1)th\ observation}{2}$

$$= \frac{4^{th}\ observation + 5^{th}\ observation}{2} = \frac{6+7}{2} = 6.5$$

iii. A shop keeper (sales person) recorded the number of video cassette recorders (VCRs) sold per month over a two year period. Find the median number VCRs sold.

Number of sets sold	Frequency (months)	Cumulative frequency
1	3	3
2	8	11
3	5	16
4	4	20
5	2	22
6	1	23
7	1	24

The number of observations n=24 , even.

$$\tilde{x} = \frac{12^{th}\,observation + 13^{th}\,observation}{2} = \frac{3+3}{2} = 3$$

Properties of median

- It is an average of position.
- It is affected by the number of observations than by extreme values.
- The sum of the deviations about the median, signs ignored, is less than the sum of deviations taken from any other value or specific average.

Definition 3.6: The mode (modal value) of an observed set of data is the value that occurs the largest number of times.

- **The mode for raw data**
- **Example 3.10:** Find the modal value for the following sets of data.
- 5 6 5 8 7 4 . In this data set, 5 is the most frequent value. Therefore, the mode is 5. Since the modal value is only one number, we call the distribution unimodal.
- 1 2 3 4 8 2 5 4 6. In this data the modal values are 2 and 4 since both 2 and 4 appear most frequently and they occur equal number of times. These kind distributions are called bimodal distribution.
- 1 2 4 3 5 6 8 7 In this data set, all values appear equal number of times so there is no modal value

- **Note**:
  - If a distribution has more than two modal values then we call the distribution multimodal.
  - If in a set of observed values, all values occur once or equal number of times, there is no mode.

Properties of modal value

- It is easy to calculate and understand.
- It is not affected by extreme values.
- It is not based on all observations.
- Is not used in further analysis of data.

- **The mean, median, and mode of grouped data**
  - The mean for grouped data can be found by considering the values in the interval are centered at the mid-point of the interval.
- **Example 3.12:** Consider the frequency distribution of the time spent by the automobile workers. Find the mean time spent by these workers from this frequency distribution.

Time (in minute)	Class mark (x_i)	Number of workers	$f \times x_i$
15.5- 21.5	18.5	3	55.5
21.5-27.5	24.5	6	147
27.5-33.5	30.5	8	244
33.5-39.5	36.5	4	146
39.5-45.5	42.5	3	127.5
45.5-51.5	48.5	1	48.5
Total		25	768.5

Solution:

$$\bar{x} = \frac{\sum_{i=1}^{6} f_i x_i}{\sum f_i} = \frac{768.5}{25} = 30.74$$

Note: In case of grouped data if any class interval is open, arithmetic mean cannot be calculated.

The median for grouped data can be approximated by the following formula.

$$\tilde{x} = L_m + \frac{\left(\frac{n}{2} - cf\right)}{f_m} w$$

where L_m= lower class boundary for the median class.

n= total number of observations in the distribution.

cf= less than cumulative frequency for the class preceding the median class.

w= class width for median class.

f_m=frequency for median class.

Note that the median class is the class containing the $(n/2)^{th}$ observation.

Example 3.13: Find the median for the following frequency distribution.

Class boundaries	Frequency (f)	Cumulative frequency
5.5-10.5	1	1
10.5-15.5	2	3
15.5-20.5	3	6
20.5-25.5	5	11
25.5-30.5	4	15
30.5-35.5	3	18
35.5-40.5	2	20

Solution: The class containing the $(n/2)^{th}$ observation or the 10^{th} observation is the median class. This class has class boundaries 20.5 & 25.5(4^{th} class).

$$\tilde{x} = L_m + \frac{\left(\frac{n}{2} - cf\right)}{f_m} w = 20.5 + \frac{(10-6)}{5} 5 = 24.5$$

Therefore, the median is 24.5.

Note:

- We approximate the median by assuming that the values in the median class are evenly distributed.
- We can compute the median for open-ended frequency distribution as long as the middle value does not occur in the open-ended class.

The mode for grouped data can be estimated by the following formula.

The modal value is denoted by $\hat{x}$. For grouped data we can compute the mode as follows:

$$\hat{x} = L_1 + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)w$$

where f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class next to the modal class

L_1 = lower class boundary of the modal class

W = class width of the modal class

Note: The modal class is the class with the highest frequency.

Example 3.14: Calculate the modal time spent by the automobile workers.

Time (in minute)	Number of workers
15.5- 21.5	3
21.5-27.5	6
27.5-33.5	8
33.5-39.5	4
39.5-45.5	3
45.5-51.5	1

The modal class is the class with largest frequency and it is the third class.

$$\hat{x} = 27.5 + \left(\frac{8-6}{16-6-4}\right)6 = 29.5$$

Therefore, the modal time spent is 29.5 minutes.

Note: The mode can be calculated for distributions with open ended classes.

Quartiles, Deciles and Percentiles

These are averages of position. They are collective known as fractile (quantile) points.

Definition 3.7

Quartiles are three points which divide an array into four parts in such a way that each portion contains an equal number of elements. The 1^{st}, the 2^{nd} and the 3^{rd} points are known as the 1^{st}, the 2^{nd} and the 3^{rd} quartiles and are usually denoted by Q_1, Q2 and Q_3, respectively.

Deciles are nine points which divide an array into 10 parts in such a way that each part contains equal number of elements. The 1^{st}, 2^{nd},…, and the 9^{th} points are known as the 1^{st}, 2^{nd},…, and the 9^{th} deciles and are usually denoted by $D_1, D_2, \ldots, D_9$, respectively.

Percentiles are 99 points which divide an array into 100 parts in such a way that each part consists of equal number of elements. The 1^{st}, 2^{nd}… and the 99^{th} points are known as the 1^{st}, 2^{nd}… and the 99^{th} percentiles and are usually denoted by P_1, P_2… P_{99}, respectively.

Note: The array should be in ascending order in order to get the quantiles.

i. Quantile points for raw data

First form an array in an ascending order and then apply the following procedure.

$$Q_k = \text{The value on the } \left\{k\left(\frac{n+1}{4}\right)\right\}^{th} \text{ position}$$

$$D_k = \text{The value on the } \left\{k\left(\frac{n+1}{10}\right)\right\}^{th} \text{ position}$$

$$P_k = \text{The value on the } \left\{k\left(\frac{n+1}{100}\right)\right\}^{th} \text{ position}$$

- **Example 3.15:** The following data relate to sizes of shoes sold at a stock during a week. Find the quartiles, the seventh decile and the 90th percentile.

Size of shoes	**5**	**5.5**	**6**	**6.5**	**7**	**7.5**	**8**	**8.5**	**9**	**9.5**
Number of pairs	2	5	15	30	60	40	23	11	4	1

- **Solution:** The total number of observations is 191.

$$Q_1 = The\ size\ on\ the\ \left\{1\left(\frac{191+1}{4}\right)\right\}^{th}\ position$$

$$= The\ size\ on\ the\ 48^{th}\ position = 6.5$$

$$Q_3 = The\ size\ on\ the\ \left\{3\left(\frac{191+1}{4}\right)\right\}^{th}\ position$$

$$= The\ size\ on\ the\ 144^{th}\ position = 7.5$$

$$D_7 = The\ size\ on\ the\ \left\{7\left(\frac{191+1}{10}\right)\right\}^{th}\ position$$

$$= The\ size\ on\ the\ 134.4^{th}\ position$$

$$\approx The\ size\ on\ the\ 134^{th}\ position = 7.5$$

$$P_{90} = The\ size\ on\ the\ \left\{90\left(\frac{191+1}{100}\right)\right\}^{th}\ position$$

$$= the\ size\ on\ 172.8^{th}\ position$$

$$\approx the\ size\ on\ 173^{th}\ position = 8$$

- **Note**: Relationships between fractile points
  - $Q_1=P_{25}$
  - $Q_2=P_{50}=D_5$
  - $Q_3=P_{75}$
  - $D_1=P_{10}$; $D2=P_{20}$ …$D_9=P_{90}$.

Ashebir Feyisa

Points of discussions

- State the types of data and discuss the difference between them
- write at least three sources of secondary data
- List methods of data collection for primary data
- What are the advantages of frequency distribution
- State types of frequency distribution based up on the frequency assigned for the class
- Differentiate grouped and ungrouped frequency distribution.
- What types of graphs do we use for quantitative and qualitative data.
-

MEASURES OF VARIATION

Ashebir Feyisa 12/07/2025

Objectives: Having studied this portion, you should be able to

- understand the importance of measuring the variability (dispersion) in a data set.
- measure the scatter or dispersion in a data set.
- understand ‘moments’ as a convenient and unifying method for summarizing several descriptive statistical measures.
- measure the extent to which the distribution of values in a data set deviate from symmetry.

Introduction and objectives of measuring variation

- We have seen that averages are representatives of a frequency distribution. But they fail to give a complete picture of the distribution. They do not tell anything about the spread or dispersion of observations within the distribution. Suppose that we have the distribution of yield (kg per plot) of two rice varieties from 5 plots each.

Variety 1:	45	42	42	41	40
Variety 2:	54	48	42	33	30

- The mean yield of both varieties is 42 kg. The mean yield of variety 1 is close to the values in this variety.
- On the other hand, the mean yield of variety 2 is not close to the values in variety 2.
- The mean doesn’t tell us how the observations are close to each other

Objectives of measuring variation

- To describe dispersion (variability) in a data.
- To compare the spread in two or more distributions.
- To determine the reliability of an average.

  - **Note**: The desirable properties of good measures of variation are almost identical with that of a good measure of central tendency.

Absolute and relative measures

- Measures of variation may be either **absolute or relative.**
  - Absolute measures of variation are expressed in the same unit of measurement in which the original data are given. These values may be used to compare the variation in two distributions provided that the variables are in the same units and of the same average size.

- In case the two sets of data are expressed in different units, however, such as quintals of sugar versus tones of sugarcane or if the average sizes are very different such as manager’s salary versus worker’s salary, the absolute measures of dispersion are not comparable.
- In such cases measures of relative dispersion should be used.
- A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate measure of central tendency.
- It is a unit less measure.

Types of measures of variation

- **The range and relative range**

Definition 4.1: Range is defined as the difference between the maximum and minimum observations in a set of data.

- Range is the crudest absolute measures of variation. It is widely used in the construction of quality control charts.

Definition 4.2: Relative range (RR) is defined as

Variance, standard deviation and coefficient of variation

- Definition 4.3: The variance is the average of the squares of the distance each value is from the mean.
- The symbol for the population variance is σ^2 (σ is the Greek lower case letter sigma). Let $x_1, x_2, \ldots, x_N$ be the measurements on N population units then, the population variance is given by the formula:
- where and N=Population size.

- Definition 4.4: The standard deviation is the square root of the variance. The symbol for the population standard deviation is The corresponding formula for the standard deviation is

.

- **Example 4.1:** The height of members of a certain committee was measured in inches and the data is presented below.
- Height(x): 69 66 67 69 64 63 65 68 72

	2	**-1**	**0**	**2**	**-3**	**-4**	**-2**	**1**	**5**
	4	**1**	**0**	**4**	**9**	**16**	**4**	**1**	**25**

$$\sigma = \sqrt{\sigma^2} = \sqrt{7.11} = 2.66$$

- Definition 4.5: The sample variance is denoted by S^2, and its formula is

.

- Definition 4.6: The sample standard deviation, denoted by S, is the square root of the sample variance

.

- **Example 4.2:** For a newly created position, a manager interviewed the following numbers of applicants each day over a five-day period: 16, 19, 15, 15, and 14. Find the variance and standard deviation.
- **Solution:**

- Note that the procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the mid-points of each class.

Properties of variance

- The unit of measurement of the variance is the square of the unit of measurement of the observed values. It is one of its limitations.
- The variance gives more weight to extreme values as compared to those which are near to mean value, because the difference is squared in variance.
- It is based on all observations in the data set.

Properties of standard deviation

- Standard deviation is considered to be the best measure of dispersion and is used widely.
- There is, however, one difficulty with it. If the unit of measurement of variables of two series is not the same, then their variability cannot be compared by comparing the values of standard deviation.

Uses of the variance and standard deviation

- The variance and standard deviations can be used to determine the spread of data, consistency of a variable and the proportion of data values that fall within a specified interval in a distribution.
- If the variance or standard deviation is large, the data is more dispersed.
- This information is useful in comparing two or more data sets to determine which is more (most) variable.
- Finally, the variance and standard deviation are used quite often in inferential statistics.

Coefficient of variation (CV)

- The standard deviation is an absolute measure of dispersion. The corresponding relative measure is known as the ***coefficient of variation (CV)*.**
- Coefficient of variation is used in such problems where we want to compare the variability of two or more different series. Coefficient of variation is the ratio of the standard deviation to the arithmetic mean, usually expressed in percent:

$$CV = \frac{S}{\bar{x}} \times 100\%$$

- A distribution having less coefficient of variation is said to be less variable or more consistent or more uniform or more homogeneous.

Example 4.3: Last semester, the students of Biology and Chemistry Departments took *Stat 273* course. At the end of the semester, the following information was recorded.

Department	Biology	Chemistry
Mean score	79	64
Standard deviation	23	11

Compare the relative dispersions of the two departments' scores using the appropriate way.

Solution:

Biology Department	Chemistry Department
$CV = \frac{23}{79} \times 100 = 29.11\%$	$CV = \frac{11}{64} \times 100 = 17.19\%$

Since the CV of Biology Department students is greater than that of Chemistry Department students, we can say that there is more dispersion in the distribution of Biology students' scores compared with that of Chemistry students.

- **Example 4.4:** The mean weight of 20 children was found to be 30 kg with variance of 16kg^2 and their mean height was 150 cm with variance of 25cm^2. Compare the variability of weight and height of these children.

- The weight of the children is more variable than their height.

Standard score

A standard score is a measure that describes the relative position of a single score in the entire distribution of scores in terms of the mean and standard deviation. It also gives us the number of standard deviations a particular observation lie above or below the mean.

Population standard score: $Z = \frac{x - \mu}{\sigma}$ where x is the value of the observation, μ and σ are the mean and standard deviation of the population respectively.

Sample standard score: $Z = \frac{x - \bar{x}}{S}$ where x is the value of the observation, $\bar{x}$ and S are the mean and standard deviation of the sample respectively.

- **Interpretation:**

Example 4.5: Two sections were given an exam in a course. The average score was 72 with standard deviation of 6 for *section 1* and 85 with standard deviation of 5 for *section 2*. Student A from *section 1* scored 84 and student B from *section 2* scored 90. Who performed better relative to his/her group?

Solution: Section 1: $\bar{x} = 72, \quad S = 6$ and score of student A from Section 1; $x_A = 84$

Section 2: $\bar{x} = 85, \quad S = 5$ and score of student B from Section 2; $x_B = 90$

Z-score of student A: $Z = \frac{x_A - \bar{x}_1}{S_1} = \frac{84 - 72}{6} = 2.00$

Z-score of student B: $Z = \frac{x_B - \bar{x}_2}{S_2} = \frac{90 - 85}{5} = 1.00$

From these two standard scores, we can conclude that student A has performed better relative to his/her section students because his/her score is two standard deviations above the mean score of *selection 1* while the score of student B is only one standard deviation above the mean score of *section 2* students.

- **Example 4.6:** A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on each test.

  Solution: First, find the z-scores.

  For calculus the z-score is

  For history the z-score is

  Since the z-score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

Moments, skewness and kurtosis

Ashebir Feyisa 12/07/2025

Moments

- Definition 4.7: The average of deviations from an arbitrary origin raised to an integral power of the observations of a distribution is defined as a moment. Let $x_1, x_2, \ldots, x_n$ be observations, we define the r-th moment about A as:

- The most known moments are moments about the mean also known as the central moments and the moments about zero (also known as moments about the origin.)
- The r^{th} moment about the mean, μ_r, is given by:
- .
- Special ceases: $\mu_0=1$, $\mu_1=0$, $\mu_2=s^2$.
- The r^{th}moment about the origin,, is given by:
- .
- Special cases: ,,

- **Skewness:** it refers to lack of symmetry in a distribution.

  Note: for a symmetrical and unimodal distribution:
- Mean =median =mode
- The lower and upper quartiles are equidistant from the median, so also are corresponding pairs of deciles and percentiles.
- Sum of positive deviations from the median is equal to the sum of negative deviations (signs ignored).
- The two tails of the frequency curve are equal in length from the central value.
- If a distribution is not symmetrical we call it *skewed distribution*.

- **Measures of skewness**
- Pearsonian coefficient of skewness (P_{csk}) defined as:

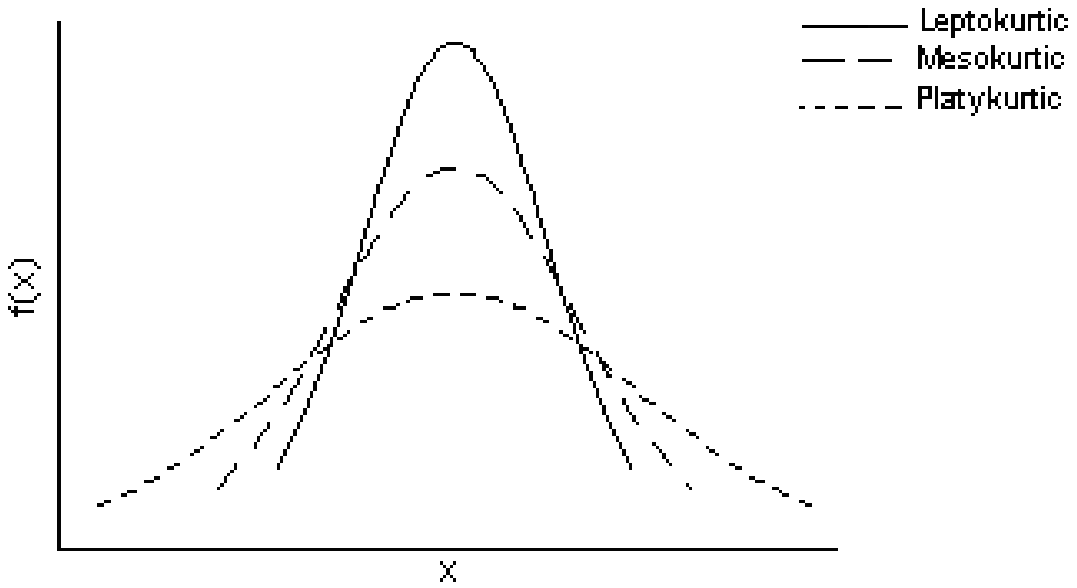
- In moderately skewed distributions: Mode = mean- 3(mean-median)

Interpretation:

- **Note:** in a negatively skewed distribution larger values are more frequent than smaller values. In a positively skewed distribution smaller values are more frequent than larger values.

- **Example 4.7:** If the mean, mode and s.d of a frequency distribution are 70.2, 73.6, and 6.4, respectively. What can one state about its skeweness
- .
- This figure suggests that there is some negative skewness

- **Kurtosis:** it refers to the degree of peakedness of a distribution.



- When the values of a distribution are closely bunched around the mode in such a way that the peak of the distribution becomes relatively high, the distribution is said to be *leptokurtic*. If it is flat topped we call it *platykurtic*. A distribution which is neither highly peaked nor flat topped is known as a *meso-kurtic* distribution (normal).

Measures of kurtosis

i. Moment coefficient of kurtosis (M_{ck}) is given by

$$M_{ck} = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{s^4}$$

where

$$\mu_4 = \frac{\sigma(x_i - \bar{x})^4}{n}, \quad \mu_2 = \frac{\sigma(x_i - \bar{x})^2}{n} = S^2.$$

Interpretation:

$$if\ M_{ck} \begin{cases} < 3, \text{the distribution is said to be platykurtic distribution.} \\ = 3, \text{the distribution is said to be mesokurtic distribution.} \\ > 3, \text{the distribution is said to be leptokurtic distribution.} \end{cases}$$

ELEMENTARY PROBABILITY

Ashebir Feyisa 12/07/2025

Introduction

- Why it is that science is not always certain?
- Due to complexity of Nature and full of unexplained variability.
- almost all methods of observation and experiment are imperfect.
- Observers are subject to human bias and error.
- Due to measurement fluctuation
- Science is a continuing story;
- subjects vary

Cont…

- Since Science is a continuing story; subjects vary; measurements fluctuate.
- Biomedical science, in particular, contains controversy and disagreement; with the best of intentions, biomedical data,
- medical histories,
- physical examinations,
- interpretations of clinical tests,
- descriptions of symptoms and diseases are somewhat inexact

Cont.

- But most important of all, we always have to deal with incomplete information:
- It is either impossible, or too costly, or too time consuming, to study the entire population;
- we often have to rely on information gained from a sample, that is, a subgroup of the population under investigation.

- So some uncertainty almost always prevails. Science and scientists cope with uncertainty by using the concept of probability

Cont...

- So By calculating probabilities, they are able to describe what has happened and predict what should happen in the future under similar conditions.
- In short, more often the quantities we are interested in will not be predictable in advance but, rather, will exhibit an inherent variation. Probability and statistics are concerned in the quantification of such quantities (or random phenomena).

Definition of some probability terms

- **Random experiment:** is an experiment in which the outcome cannot be determined or predicted exactly in advance, i.e. it is the process of observing or measuring the outcome of a chance event.
- Some of the characteristics of a random experiment are:
  - all the possible outcomes of the experiment can be specified in advance.
  - the experiment can be repeated indefinitely.
  - there is a sort of regularity in the outcomes observed in large repetitions of the experiment.

- **Sample point (outcome)**: The individual result of a random experiment.
- **Sample space**: The set containing all possible sample points (out comes) of the random experiment.
- The sample space is often called the *universe* and denoted by S.
- **Event**: The collection of outcomes or simply a subset of the sample space. We denote events with capital letters, A, B, C, etc.

- **Example 5.1:** If an experiment consists of flipping of a coin once, then
- $S = \{H, T\}$ where H means that the outcome of the toss is a head and T that it is a tail. A= {H} represents the event of head occurring.
- **Example 5.2:** If an experiment consists of rolling a die once and observing the number on top, then the sample space is $S = \{1, 2, 3, 4, 5, 6\}$ where the outcome i means that i appeared on the die, $i = 1, 2, 3, 4, 5, 6$. {1}, {2},{3},{4},{5} and {6}are elementary events i.e. events consisting of a single outcome. Let A represents the event of an odd number will occur, then A is simply the set containing 1, 3 and 5 i.e. A= {1, 3, 5}.

1. Union: The union of A and B, A u B, is the event containing all sample points in either A or B or both. Sometimes we use A or B for union.

2. Intersection: The intersection of A and B, A n B, is the event containing all sample points that are both in A and B. Sometimes we use AB or A and B for intersection.

3. Subset: If for any $\omega \in A$, then $\omega \in B$. Then $A \subseteq B$.

4. Empty set: If a set A contains no points, it will be called the null set, or empty set, and denoted by ϕ.

5. Complement: The complement of a set A denoted by A^c is the set where $\omega \in S$, $\omega \in A^c$ but, $\omega \notin A$.

6. Mutually Exclusive Events: Two events are said to be mutually exclusive (or disjoint) if their intersection is empty. (i.e. $A \cap B = \phi$). Subsets $A_1, A_2, \ldots$ are defined to be mutually exclusive if $A_i \cap A_j = \phi$ for every $i \neq j$.

4. Empty set: If a set A contains no points, it will be called the null set, or empty set, and denoted by ϕ.

5. Complement: The complement of a set A denoted by A^c is the set where $\omega \in S$, $\omega \in A^c$ but, $\omega \notin A$.

6. Mutually Exclusive Events: Two events are said to be mutually exclusive (or disjoint) if their intersection is empty. (i.e. $A \cap B = \phi$). Subsets $A_1, A_2, \ldots$ are defined to be mutually exclusive if $A_i \cap A_j = \phi$ for every $i \neq j$.

Theorem 5.1: Important elementary set theory results

i) $A \cup B = B \cup A$ and $A \cap B = B \cap A$

ii) $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$

iii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

iv) $(A^c)^c = A$

v) $A \cap S = A$; $A \cup S = S$; $A \cap \phi = \phi$; and $A \cup A = A$

vi) $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

Theorem 5.1: Important elementary set theory results

i) $A \cup B = B \cup A$ and $A \cap B = B \cap A$

ii) $A \cup (B \cup C) = (A \cup B) \cup C$ and $A \cap (B \cap C) = (A \cap B) \cap C$

iii) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ and $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

iv) $(A^c)^c = A$

v) $A \cap S = A$; $A \cup S = S$; $A \cap \phi = \phi$; and $A \cup A = A$

vi) $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$

Counting rules

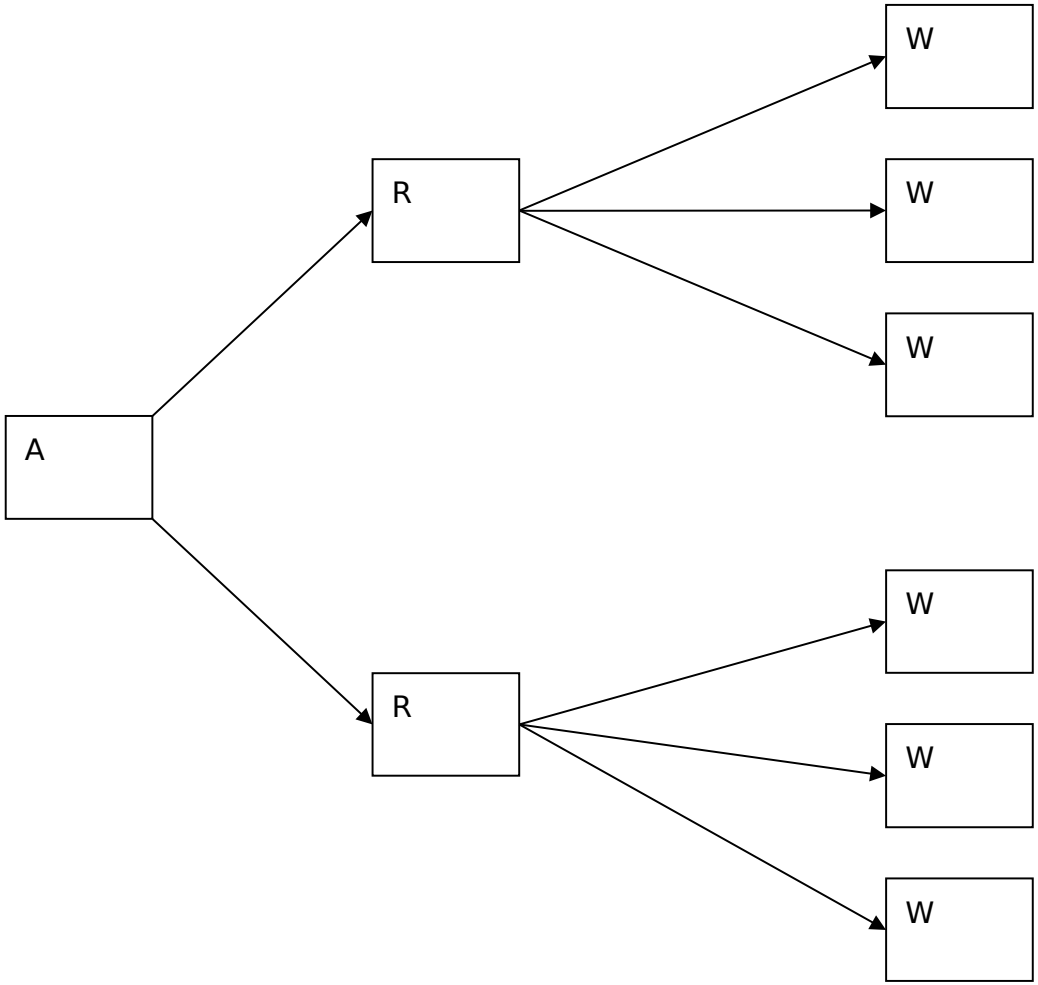
- to assign probabilities for an event, we might need to enumerate the possible outcomes of a random experiment and need to know the number of possible outcomes favoring the event. The following principles will help us in determining the number of possible outcomes favoring a given event.

Addition principle

- If a task can be accomplished by k distinct procedures where the i^{th} procedure has n_i alternatives, then the total number of ways of accomplishing the task equals $n_1 + n_2 + \ldots + n_k$.
- **Example 5.3:** Suppose one wants to purchase a certain commodity and that this commodity is on sale in 5 government owned shops, 6 public shops and 10 private shops. How many alternatives are there for the person to purchase this commodity?
- **Solution:** Total number of ways =5+6+10=21 ways

Multiplication principle

- If a choice consists of k steps of which the first can be made in n_1 ways, for each of these the second can be made in n_2 ways,..., and for each of these the k^{th} can be made in n_k ways, then the whole choice can be made in $n_1.n_2....n_k$ ways.
- **Example 5.4:** If we can go from Addis Ababa to Rome in 2 ways and from Rome to Washington D.C. in 3 ways then the number of ways in which we can go from Addis Ababa to Rome to Washington D.C. is 2x3 ways or 6 ways. We may illustrate the situation by using a tree diagram below:


A
R
R
W
W
W
W
W
W

- **Example 5.5:** If a test consists of 10 multiple choice questions, with each permitting 4 possible answers, how many ways are there in which a student gives his/her answers?

Solution: There are 10 steps required to complete the test.

First step: To give answer to question number one. He/she has 4 alternatives.

Second step: To give answer to question number two, he/she has 4 alternatives……

Last step: To give answer to last question, he/she has 4 alternatives.

Therefore, he/she has $4x4x4x…x4=4^{10}$ ways or1, 048, 576 ways of completing the exam. Note that there is only one way in which he/she can give correct answers to all questions and that there are 3^{10} ways in which all the answers will be incorrect.

- **Example 5.6:** A manufactured item must pass through three control stations. At each station the item is inspected for a particular characteristic and marked accordingly. At the first station, three ratings are possible while at the last two stations four ratings are possible. Hence there are 48 ways in which the item may be marked.
- **Example 5.7:** Suppose that car plate has three letters followed by three digits. How many possible car plates are there, if each plate begins with a H or an F?
- 2x 26x 26x 10x 10x 10 or 1, 352, 000 different plates.

- If n is a positive integer, we define n!= n(n-1)(n-2)…1 and call it *n-factorial* and 0!=1.
- **Permutations**
- Suppose that we have n different objects. In how many ways, say ${}_nP_n$, may these objects be arranged (permuted)? For example, if we have objects a, b and c we can consider the following arrangements: abc, acb, bac, bca, cab, and cba. Thus the answer is 6. The following theorem gives general result on the number of such arrangements.

- **Theorem 5.4: Permutation**
- The number of permutations of n different objects is given by ${}_nP_n = n!$
- A permutation of n objects, arranged in groups of size r, without repetition, and order being important is:

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

- **Example 5.8:** Suppose that we have five letters a, b, c, d.

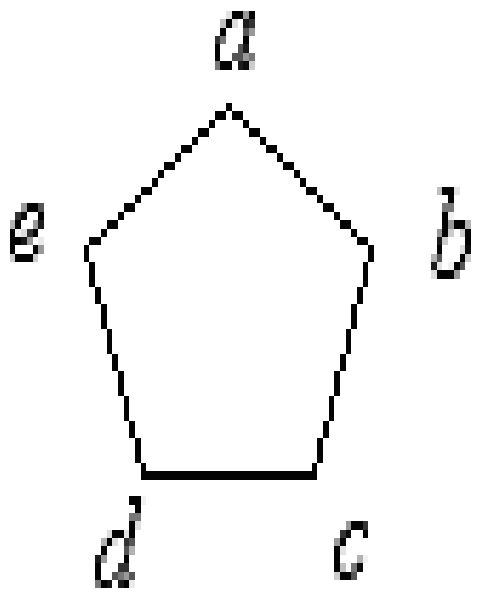
What is the number of possible arrangements of these letters taken all at a time?

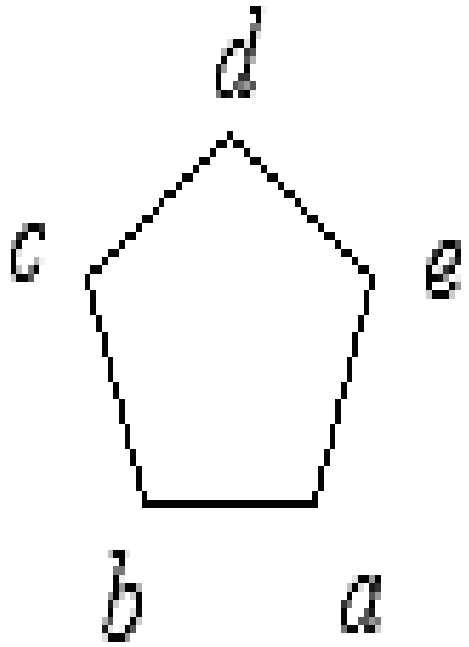
What is the number of possible arrangements of these letters if we use only three of the letters at a time?

- **Solution:**
- Using (i) of theorem 5.4, we have 4! ways of arranging the 4 letters, i.e. we have 24 possible arrangements.
- Using (ii) of theorem 5.4, we have ${}_4P_3$ ways of arranging 3 letters taken from the four letters, i.e. we have 24 possible arrangements.

- **Example 5.9:** In a class with 8 boys and 8 girls
- In how many ways can the children line up if they alternate girl-boy-girl-boy-… ?
- In how many ways can the children line up so that no two of the same sex are next to each other?
- **Solution:**
- The 8 girls can line-up in 8! ways, and likewise the 8 boys can line-up in 8! ways. For any single arrangement of the girls, all possible arrangements of the boys are possible, thus by multiplication principle we have 8!x 8! ways to arrange the children in girl-boy lines.
- Now we must include the case of boy-girl. So we have 2x8! x 8! ways of arranging.

- **Example 5.10:** If I have 5 different books on my shelf, in how many ways can I arrange these books? Solution: We can arrange the books in 5! different ways or 5x4x3x2x1 ways or 120 ways.
- **Remarks**
- i) The number of permutations of n distinct objects arranged in a circle is (n-1)!.
- This is because we consider two permutations the same if one is a rotation of the other. For n objects arranged around a circle, there a n rotations that give the same permutation. Dividing n! by n gives (n - 1)!. The two circular permutations below are considered the same; their order is a, b, c, d, e.


a
e
b
d
c


d
c
e
b
a

- ii) Permutations when not all objects are different

Given n objects of which n_1 are one kind, n_2 are another kind, …, n_k of another kind, then the total number of distinct permutations that can be made from these objects is $\frac{n!}{n_1!\,n_2!\cdots n_k!}$

- **Example 5.11**
- How many "words" (text strings or distinct arrangements) can be made from the letters b,k,o,o?
- How many permutations are there for the letters in the word banana?

- If we label the two o's as o_1 and o_2, and think of them as distinct, then the number of permutations is 4!. For each permutation there will be a matching permutation that switches the o's, that is for $o_1o_2$bk there is the matching $o_2o_1$bk permutation. We can see then that if we divide the number of distinct permutations by two, we have a count of the number of permutations of the 4 letters where we do not distinguish between the two o's. Therefore, there are distinct4!/2 text strings or 12 text strings.

- If we think of all 6 letters as distinct, then we would have 6! permutations. As in the preceding example for the two n's, we would need to divide 6! by 2. For the 3 a's, we would have 6 counts for a single permutation. For instance, each of the following would be a single word if the a's were not distinct. $a_1a_2a_3bnn$, $a_1a_3a_2bnn$, $a_2a_1a_3bnn$, $a_2a_3a_1bnn$, $a_3a_1a_2bnn$, and $a_3a_2a_1bnn$. Hence the number of distinct permutations of the word banana is

$$\frac{6!}{2!3!} = 60$$

.

Combinations

- Consider n different objects. This time we are concerned with counting the number of ways we may choose r out of these n objects without regard to order. For example, we have the objects a, b, c and d, and r=2; we wish to count ab, ac, ad, bc, bd, and cd. In other words, we do not count ab and ba since the same objects are involved and only the order differs.

- There are many problems in which we are interested in determining the number of ways in which r objects can be selected from n distinct objects without regard to the order in which they are selected. Such selections are called *combinations or r-sets*. It may help to think of combinations as committees. The key here is without regard for order.

Combination

The number of ways of choosing r out of n different objects, disregarding order, is given by $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

- **Example 5.12:** How many different committees of 3 can be formed from Hawa, Segenet, Nigisty and Lensa?
- **Solution:** The question can restated in terms of subsets from a set of 4 objects, how many subsets of 3 elements are there? In terms of combinations the question becomes, what is the number of combinations of 4 distinct objects taken 3 at a time? The list of committees: {H,S,N}, {H,S,L}, {H,N,L}, {S,N,L}.Therefore, we have ${}_4C_3$ or 4 possible number of committees.

- **Example 5.13:**
- (i) A committee of 3 is to be formed from a group of 20 people. How many different committees are possible?
- (ii) From a group of 5 men and 7 women, how many different committees consisting of 2 men and 3 women can be formed?

Solution: (i) There are $\binom{20}{3} = \frac{20!}{3!17!} = 1140$ possible committees.

(i) $\binom{5}{2}\binom{7}{3} = \frac{5!}{2!3!}\frac{7!}{3!4!} = 350$ possible committees.

Remarks:

i) $\binom{n}{r} = \binom{n}{n-r}$

ii) A set with n elements has 2^n subsets.

- **The Axioms of Probability**
- Probabilities are real numbers assigned to events (or subsets) of a sample space. We can think of the assignment of probabilities to events, or probability measure, as a function between the collection of subsets of the sample space and the real numbers.

- Mathematically, a probability measure P for a random experiment is a real-valued function defined on the collection of events that satisfies the following axioms:
- Axiom 1: The probability of an event is a nonnegative real number; that is, P(A) ≥ 0 for any subset A of S.
- Axoim 2: P(S) = 1
- Axiom 3: If A_1, A_2, A_3 ... is a finite or infinite sequence of mutually exclusive
- events of S, then P(A_1 u A_2 u A_3 u ...) = P(A_1) + P(A_2) + P(A_3) + ...= $\sum P(A_i)$

- Suppose that we have a random experiment with sample space S and probability function P and A and B are events. Then we have the following results:
- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(B \cap A^c) = P(B) - P(A \cap B)$
- If A subset of B then $P(A) \leq P(B)$.

- **The classical definition of probability**

If an experiment can result in any one of N equally likely and mutually exclusive outcomes, and if n of these outcomes constitute the event A, then the probability of event A is

$$P(A) = \frac{n}{N}$$

- Consider the experiment of tossing a fair die. A fair die means that all six numbers are equally likely to appear. Calculate the probabilities of the following events:
- A=One will occur ={1}
- B=Even number will occur ={2, 4, 6}
- C=Odd number will occur ={1, 3, 5}
- D=A number less than 3 will occur ={1,2}

Solution:

a) Since the die is fair

$$P(A) = P(\{1\}) = \frac{1}{6}$$

$$P(\{2\}) = P(\{3\}) = P(\{4\}) = P(\{5\}) = P(\{6\}) = \frac{1}{6}$$

b) $P(B) = \frac{3}{6} = 0.5;$ c) $P(C) = \frac{3}{6} = 0.5;$ d) $P(D) = \frac{2}{6}$

- **Example 5.15:** Suppose that we toss two coins, and assume that each of the four outcomes in the sample space $S = \{(H,H),(H, T), (T ,H), (T , T)\}$ are equally likely and hence has probability ¼. Let $A = \{(H, H),(H, T)\}$ and $B = \{(H,H), (T ,H)\}$ that is, A is the event that the first coin falls heads, and B is the event that the second coin falls heads. Then, calculate the probabilities of A, B, A^c, B^c, and S^c. The event that none of the outcomes will occur is the same as S^c.

Solution:

$$P(A) = \frac{2}{4} = 0.5$$

$$P(B) = \frac{2}{4} = 0.5$$

$$P(A^c) = 1 - P(A) = 1 - 0.5 = 0.5$$

$$P(B^c) = 1 - P(B) = 1 - 0.5 = 0.5$$

$$P(S^c) = 1 - P(S) = 1 - 1 = 0 = P(\phi)$$

- **Example 5.16:** From a group of 5 men and 7 women, it is required to form a committee of 5 persons. If the selection is made randomly, then

I. What is the probability that 2 men and 3 women will be in the committee?

II. What is the probability that all members of the committee will be men?

III. What is the probability that at least three members will be women?

Solution: The total number of possible committees is $\binom{12}{5} = \frac{12!}{5!7!} = 792$, i.e. the number of possible out comes in the sample space is 792.

i) Let A be the event that the committee will consist of two 2 men and 3 women. We need to know the number of possible outcomes favoring this event. The number of ways we can select 2 men from 5 men is $\binom{5}{2} = \frac{5!}{2!3!} = 10$ and the number of ways of selecting 3 women out of 7 women is $\binom{7}{3} = \frac{7!}{3!4!} = 35$. Using the multiplication principle, the number of elements favoring event A is 10x35 or 350.
Hence, using the classical definition of probability,

$$P(A) = \frac{\binom{5}{2}\binom{7}{3}}{\binom{12}{5}} = \frac{350}{792} = 0.44$$

ii) Let B be the event that all members of the committee will be men. Hence

$$P(A) = \frac{\binom{5}{5}\binom{7}{0}}{\binom{12}{5}} = \frac{1}{792}$$

iii) Let C be the event that at least three of the committee members will be women.

Basically, three different compositions of committee members can be formed in terms of sex: 3 women and 2 men, 4 women and 1 man, and all are women. Hence the number of possible outcomes favoring event C using the principle of combination together with the addition principle is

$$\binom{5}{2}\binom{7}{3} + \binom{5}{1}\binom{7}{4} + \binom{5}{0}\binom{7}{5} = 350 + 175 + 21 = 546 .$$

Therefore, $P(C) = \dfrac{\binom{5}{2}\binom{7}{3} + \binom{5}{1}\binom{7}{4} + \binom{5}{0}\binom{7}{5}}{\binom{12}{5}} = \dfrac{546}{792} = 0.69$

- **Relative Frequency Definition of probability**
- If an experiment is repeated a large number, n, of times and the event A is observed n_A times, the probability of A is $P(A) \approx n_A/n$.

- The above definition of probability is based on empirical data accumulated through time or based on observations made from repeated experiments for a

Some probability rules

- If A and B , then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- **Example 5.17:** Consider the experiment of tossing a fair die. Let
- A = Even number occurring = {2,4,6}
- B = A number greater than 2 occurring ={3, 4, 5, 6}
- C = Odd number occurring ={1, 3, 5}

i. What is the probability that A and B will occur?

ii. What is the probability that A or B will occur?

- **Solution:** We use the concept of set theory to help us solve probability questions very easily and vein diagrams are useful tools to depict the relations between events within the sample space. The shaded region on Fig 1. shows the event that both A and B will occur.
- A and B ≡ AnB ={4,6} Thus P(AnB)=2/6.
- A or B ≡ AUB ={2,3,4,5,6} AnB={4,6}
Hence,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{3}{6} + \frac{4}{6} - \frac{2}{6} = \frac{5}{6}.$$

- **Example 5.18:** Sixty percent of the families in a certain community own their own car, thirty percent own their own home, and twenty percent own both their own car and their own home. If a family is randomly chosen,

a) what is the probability that this family do not have a car?

b) what is the probability that this family owns a car or a house?

c) what is the probability that this family owns a car or a house but not both?

d) what is the probability that this family owns only a house?

e) what is the probability that this family neither owns a car nor a house?

- **Solution:** Let A represents that the family owns a car and B represents that the family owns a house. Given information: P(A)=0.6,P(B)=0.3, and P(AnB)=0.2.

a) Required: $P(A^c)$ = ? $P(A^c)=1-P(A) = 1-0.6 = 0.4$

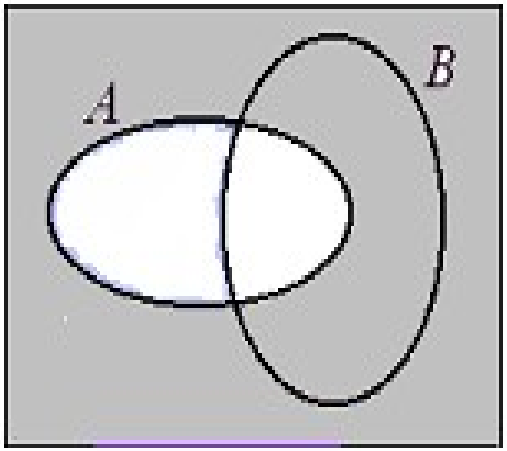
b) Required: P(AUB) = ? $P(AUB) = P(A)+P(B)-P(AnB) = 0.6+0.3-0.2 = 0.7$

c) Required: $P((AnB^c)U(A^cnB))$ = ? $P((AnB^c)U(A^cnB)) = P(AnB^c)+P(A^cnB) = [P(A)-P(AnB)]+[P(B)-P(AnB)] = [0.6-0.2]+[0.3-0.2]=0.5$

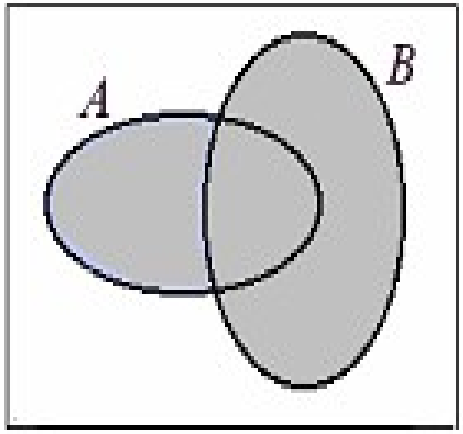
d) Required: $P(A^c \cap B) = ?$ $P(A^c \cap B) = P(B) - P(A \cap B) = 0.3 - 0.2 = 0.1$

e) Required: $P(A^c \cap B^c) = ?$ $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 0.7 = 0.3$

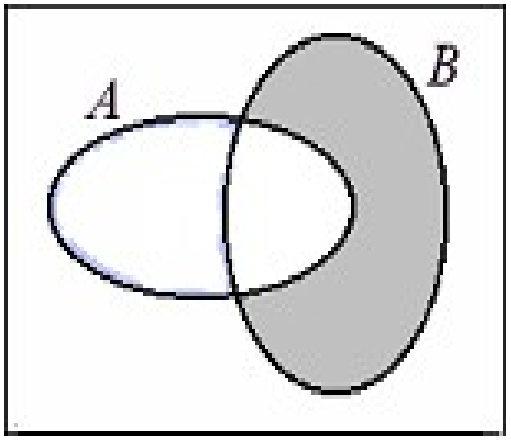
- We can represent various events by an informative diagram called vein *diagram.* If properly and correctly drawn, a vein diagram helps to calculate probabilities of events easily. The figure below shows various events represented by shaded regions. Note that the rectangle in each figure represents the sample space.



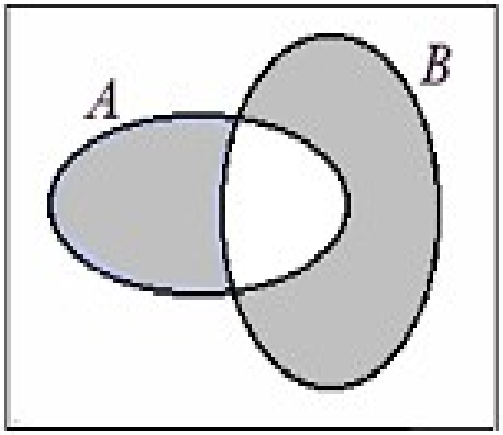
A^c



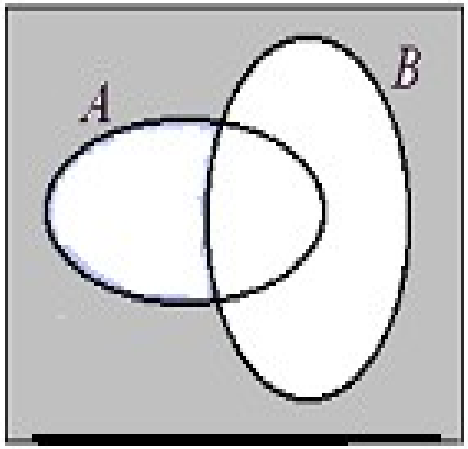
AUB



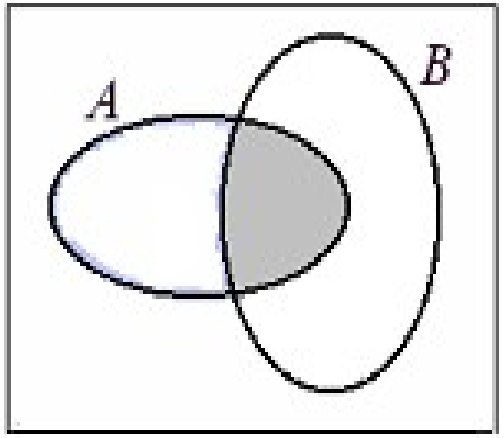
$A^c \cap B$



$(A \cap B^c) \cup (A^c \cap B)$



$(AUB)^c$



$A \cap B$

Conditional probability and independence

- **Conditional Probability**
- Conditional probability provides us with a way to reason about the outcome of an experiment, based on partial information. Here are some examples of situations we may have in our mind:

(a) What is the probability that a person will be HIV-Positive given he has tuberculosis?

(d) A spot shows up on a radar screen. How likely is it that it corresponds to an

- If P(B) > 0, the conditional probability of A given B, denoted by P(A|B), is

$$P(A/B) = \frac{P(AnB)}{P(B)}.$$

- **Example 5.19:** Suppose cards numbered one through ten are placed in a hat, mixed up, and then one of the cards is drawn at random. If we are told that the number on the drawn card is at least five, then what is the conditional probability that it is ten?
- **Solution:** Let A denote the event that the number on the drawn card is ten, and B be the event that it is at least five. The desired probability is $P(A|B)$.

$$P(A/B) = \frac{P(AnB)}{P(B)} = \frac{P(\{10\}n\{5,6,7,8,9,10\})}{P(\{5,6,7,8,9,10\})} = \frac{P(\{10\})}{P(\{5,6,7,8,9,10\})} = \frac{1/10}{6/10} = \frac{1}{6}$$

- **Example 5.20:** A family has two children. What is the conditional probability that both are boys given that at least one of them is a boy? Assume that the sample space S is given by $S = \{(b, b), (b, g), (g, b), (g, g)\}$, and all outcomes are equally likely. (b, g) means, for instance, that the older child is a boy and the younger child is a girl.

- **Solution:** Letting *A* denote the event that both children are boys, and *B* the event that at least one of them is a boy, then the desired probability is given by

$$P(A/B) = \frac{P(AnB)}{P(B)} = \frac{1/4}{3/4} = \frac{1}{3}$$

- **Law of Multiplication**
- The defining equation for conditional probability may also be written as:
- P(AnB) = P(B) P(A|B)
- This formula is useful when the information given to us in a problem is P(B) and P(A|B) and we are asked to find P(AnB). An example illustrates the use of this formula. Suppose that 5 good fuses and two defective ones have been mixed up. To find the defective fuses, we test them one-by-one, at random and without replacement. What is the probability that we are lucky and find both of the defective fuses in the first two tests?

- **Example 5.21**: Suppose an urn contains seven black balls and five white balls. We draw two balls from the urn without replacement. Assuming that each ball in the urn is equally likely to be drawn, what is the probability that both drawn balls are black?
- **Solution:** Let A and B denote, respectively, the events that the first and second balls drawn are black. Now, given that the first ball selected is black, there are six remaining black balls and five white balls, and so $P(B|A) = 6/11$. As $P(A)$ is clearly 7/12 , our desired probability is

$$P(AB) = P(A)P(B|A) = \frac{7}{12}\frac{6}{11} = \frac{7}{22}$$

Independence

We have introduced the conditional probability $P(A|B)$ to capture the partial information that event B provides about event A. An interesting and important special case arises when the occurrence of B provides no information and does not alter the probability that A has occurred, i.e., $P(A|B) = P(A)$. When the above equality holds, we say that A is independent of B. Note that by the definition $P(A|B) = P(A \cap B)/P(B)$, this is equivalent to $P(A \cap B) = P(A)P(B)$.

Independence

Two events A and B are said to independent if $P(A \cap B) = P(A)P(B)$. If in addition, $P(B) > 0$, independence is equivalent to the condition $P(A|B) = P(A)$.

PROBABILITY DISTRIBUTIONS

Ashebir Feyisa 12/07/2025

Definition of random variables and probability distributions

- Given an experiment and the corresponding set of possible outcomes (the sample space), a random variable associates a particular number with each outcome. Mathematically, a random variable is a real-valued function of the experimental outcome. The following are some examples of random variables:

- (a) In an experiment involving a sequence of 5 tosses of a coin, the number of heads in the sequence is a random variable.
- (b) In an experiment involving two rolls of a die, the following are examples of random variables: (1) The sum of the two rolls, (2) The number of sixes in the two rolls.
- (c) In an experiment involving the transmission of a message, the time needed to transmit the message, the number of symbols received in error, and the delay with which the message is received are all random variables.

- **Notation:** We will use capital letters to denote random variables, and lower case characters to denote real numbers such as the numerical values of a random variable.
- **Types of random variables:** Generally, two types of random variables exist: discrete and continuous. A random variable is called *discrete* if its range (the set of values that it can take) is finite or at most countably infinite. For instance, the number of children in a family, number of car accidents within given period of time in a certain locality, the number of bacteria in a cubic mm of agar, etc.
- If random variable assumes any numerical value in an interval or collection of intervals, then it is called a *continuous random variable*.

- Examples include body weight of new born baby, life time of a human being, height of a person, etc.
- The most important way to characterize a random variable is through the probabilities of the values that it can take. For a discrete random variable X, these are captured by the probability mass function (p.m.f. for short) of X, denoted $P_X(x)$. For a continuous random variable X it is done by the probability

Probability mass function

If x is any possible value of X, the probability mass of x, denoted $P_X(x)$, is the probability of the event $\{X = x\}$ consisting of all outcomes that give rise to a value of X equal to x. A probability mass function must satisfy the following conditions:

i. $P_X(x) \geq 0$ for any value of x of X.

ii. $\sum P_X(x) = 1$ where the summation is over all values of x .

- **Example 6.1:** Consider an experiment of tossing two fair coins. Letting X denote the number of heads appearing on the top face, then X is a random variable taking on one of the values 0, 1, 2 . The random variable X assigns a 0 value for the outcome (T,T), 1 for outcomes (T ,H) and (H, T), and 2 for the outcome (H,H). Thus, we can calculate the probability that X can take specific value/s as follows:
- $P(X = 0) = P(\{(T , T)\}) = ¼$
- $P(X = 1) = P(\{(T ,H),(H, T)\}) = 2/4,$
- $P(X = 2) = P(\{(H,H)\}) = ¼$

The table below shows the probability mass function X.

X	0	1	2
$P_X(x)$	¼	2/4	¼

We can justify that $P_X(x)$ is probability mass function.

$P_X(x) \geq 0$ for x=0,1,2 and

$P(X = 0) + P(X = 1) + P(X = 2) = ¼ + 2/4 + ¼ = 1$

Suppose we are interested to calculate the probability that $X \geq 1$. The values of X which are greater than or equal to 1 are 1 and 2. Thus, the probability that X is greater than or equal to 1, denoted $P(X \geq 1)$, is found as $P(X \geq 1) = P(X = 1) + P(X = 2) = 3/4$.

Definition 6.2: Continuous random variable

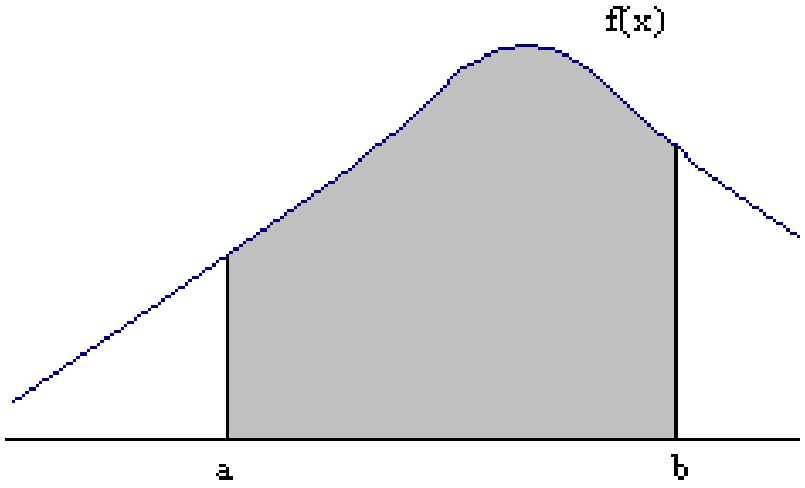
A random variable X is called *continuous* if there exists a function $f_X(x)$ called the probability density function of X which satisfies

a. $f_X(x) \geq 0$ for all x.

b. $\int_{-\infty}^{\infty} f_X(x)dx = 1$

We can use the probability density function to calculate probabilities of events expressed in terms of the random variable X. For instance, if we are interested in the probability that X lies between two points, say a and b, we can find it using integration of $f_X(x)$ on the interval [a,b],i.e.

$$P(a \le X \le b) = \int_a^b f_X(x)dx$$



- Figure: P (a≤ X ≤ b) is the shaded region

Remarks:

i) The area bounded under the graph of a probability density function and below by the horizontal axis is 1.

ii) The probability that a continuous random variable X will assume a specific value is zero,

i.e. $P(X = c) = \int_{c}^{c} f_X(x)dx = 0$ where c is a constant.

iii) The probability that a continuous random variable X will assume a value in a closed intervals is the same as the probability that it will assume in open interval or half open intervals, i.e., $P(a \leq X \leq b) = P(a < X < b) = P(a \leq X < b) = P(a < X \leq b)$, $P(X \leq c) = P(X < c)$, $P(X \geq c) = P(X > c)$ where a, b, and c are constants.

Introduction to expectation: mean and variance

- We can associate with each random variable certain “averages” of interest, such as mean and variance which give useful summary of a probability distribution.

The (mean) expected value of a random variable X denoted by E(X) or μ is given by

i) $E(X) = \sum xP_X(x)$ if X is discrete r.v.

ii) $\int_{-\infty}^{\infty} xf_X(x)dx$ if X is continuous r.v.

Variance

The variance of a random variable X denoted V(X) or σ^2 is defined as $V(X)=E[(X-\mu)^2] = E(X^2) - \mu^2$.

i) if X is discrete, $V(X) = [\sum x^2 P_X(x)] - \mu^2$

ii) if X is continuous, $V(X) = [\int_{-\infty}^{\infty} x^2 f_X(x)dx] - \mu^2$

- The variance provides a measure of dispersion of X around its mean. Another measure of dispersion is the standard deviation of X, which is defined as the square root of the variance and is denoted by σ.

Example 6.2: Calculate the mean and variance of the random variable X in example 6.1.

$$E(X) = \sum x P_X(x) = 0 \times \frac{1}{4} + 1 \times \frac{1}{2} + 2 \times \frac{1}{4} = 1$$

$$E(X^2) = \sum x^2 P_X(x) = 0^2 \times \frac{1}{4} + 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{4} = 1.5$$

$$V(X) = E(X^2) - \mu^2 = 1.5 - 1^2 = 0.5$$

Common discrete probability distributions – binomial and Poisson

- **The Binomial distribution**
- Many real problems (experiments) have two possible outcomes, for instance, a person may be HIV-Positive or HIV-Negative, a seed may germinate or not, the sex of a new born bay may be a girl or a boy, etc. Technically, the two outcomes are called *Success and Failure.*
- Experiments or trials whose outcomes can be classified as either a “success” or as a “failure” are called *Bernoulli trails*.

- Suppose that n independent trials, each of which results in a “success” with probability p and in a “failure” with probability 1 − p, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to have binomial distribution with parameters n and p. The probability mass function of a binomial distribution with parameters n and p is given by

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

- The mean and variance of the binomial distribution are np and np(1-p), respectively. Note that the binomial distributions are used to model situations where there are just two possible outcomes, success and failure. The following conditions also have to be satisfied.

I. There must be a fixed number of trials called *n*
II. The probability of success (called p) must be the same for each trial.
III. The trials must be independent

- **Example 6.3:** A fair coin is flipped 4 times. Let X be the number of heads appearing out of the four trials. Calculate the following probabilities:

I. 2 heads will appear
II. No head will appear
III. At least two heads will appear
IV. Less than two heads will appear
V. At most heads 2 will appear

- **Solution:** We can consider that the outcomes of each trial are independent to each other. In addition the probability that a head will appear in each trial is the same. Thus, X has a binomial distribution with number of trials 4 and probability of success (the occurrence of head in a trial) is ½. The probability mass function of X is given by

$$P_X(x) = \binom{n}{x} 0.5^x (1-0.5)^{n-x} = \binom{n}{x} 0.5^n,\ x = 0,1,2,3,4,$$ Note that n = 4 and p = 1/2

i) $P(X=2) = \binom{4}{2} 0.5^2 (1-0.5)^{4-2} = 0.3750$

ii) $P(X=0) = \binom{4}{0} 0.5^0 (1-0.5)^{4-0} = 0.0625$

iii) $P(X \geq 2) = P(X=2) + P(X=3) + P(X=4) = 0.3750 + 0.2500 + 0.0625 = 0.6875$

iv) $P(X < 2) = P(X=0) + P(X=1) = 0.0625 + 0.2500 = 0.3125$

v) $P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = 0.0625 + 0.2500 + 0.3750 = 0.6875$

- **Example 6.5:** Suppose it is known that the probability of recovery for a certain disease is 0.4. If random sample of 10 people who are stricken with the disease are selected, what is the probability that:

(a) exactly 5 of them will recover?

(b) at most 9 of them will recover?

- **Solution:** Let X be the number of persons will recover from the disease. We can assume that the selection process will not affect the probability of success (0.4) for each trial by assuming a large diseased population size. Hence, X will have a binomial distribution with number of trials equal to 10 and probability of success equal 0.4.

$$P(X=k)=\binom{10}{k}0.4^{k}0.6^{10-k}, k=0,1,2,...10$$

(a) $$P(X=5)=\binom{10}{5}0.4^{5}0.6^{10-5}=0.200658$$

(b) $$P(X\le 9)=1-P(X=10)=1-\binom{10}{10}0.4^{10}0.6^{10-10}=1-0.000105=0.9999$$

The Poisson Random Variable

- A random variable X, taking on one of the values 0, 1, 2, . . . , is said to have a Poisson distribution if its probability mass function is given by

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, 2, 3, \ldots \text{ and } \lambda > 0$$

- λ is the parameter of this distribution. The mean and variance of the poisson distribution are equal and their values are equal to λ. Note that poisson distributions is used to model situations where the random variable X is the number of occurrences of a particular event over a given period of time (or space).

- Together with this , the following conditions must also be fulfilled: events are independent of each other, events occur singly, and events occur at a constant rate (in other words for a given time interval the mean number of occurrences is proportional to the length of the interval).
- The poisson distribution is used as a distribution of rare events such as telephone calls made to a switch board in a given minute, number of misprints per page in a book, road accidents on a particular motor way in one day, etc.
- The process that give rise to such events are called poisson *processes*.

- **Example 6.6:** Suppose that the number of typographical errors on a single page of this lecture note has a Poisson distribution with parameter $\lambda = 1$. if we randomly select a page in this lecture note, calculate the probability that
    a) no error will occur.
    b) exactly three errors will occur.
    c) less than 2 errors will occur.
    d) there is at least one error.

Solution: Let X= Number of errors per page

$$P(X=k)=\frac{e^{-\lambda}\lambda^{k}}{k!},\ \lambda=1,\ k=0,1,2,\ldots$$

a) Required P(X≥1)=? $P(X=0)=\frac{e^{-1}1^{0}}{0!}=\frac{1}{e}=0.367879$

b) $P(X=3)=\frac{e^{-1}1^{3}}{3!}=0.061313$

c) $P(X<2)=P(X=0)+P(X=1)=0.73576$

D) $P(X\geq 1)=1-P(X=0)=1-0.367879=0.632121$

- **Example 6.7:** If the number of accidents occurring on a highway each day is a Poisson random variable with parameter $\lambda = 3$, what is the probability that no accidents will occur on a

r

Solution: Let X= number of accidents per day

$$P(X = k) = \frac{e^{-3}3^{k}}{k!}, \quad k = 0,1,2,\ldots$$

Required P(X= 0) = ? $P(X = 0) = \frac{e^{-3}3^{0}}{0!} = e^{-3} = 0.05$

- **Note:** The Poisson random variable has a wide range of applications in a diverse number of areas. An important property of the Poisson random variable is that it may be used to approximate a binomial random variable when the binomial parameter n is large and p is small. The probability that X will be k can be approximated by substituting λ by np in the poisson distribution, i.e.

$$P(X = k) = \frac{e^{-\lambda}\lambda^{k}}{k!}, \quad \lambda = np$$

Common examples of continous probability distribution

- Normal distribution
- Student’s T distribution
- F distribution

Normal distribution

- The normal distribution plays an important role in statistical inference because many real-life distributions are approximately normal;
- many other distributions can be almost normalized by appropriate data transformations (e.g., taking the log) and as a sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution.

- A continuous random variable X is said to follow normal distribution , if and only if , its probability density function (p.d.f.) is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

where $x \in (-\infty, \infty)$, $\mu \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$

- There are infinitely many normal distributions since different values of μ and σ define different normal distributions. For instance, when μ= 0 and σ =1 , the above density will have the following form

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

- This particular distribution is called *the standard normal distribution* and sometimes known as Z-distribution. The random variable corresponding to this distribution is usually denoted by Z. If X has a normal distribution with mean μ and variance σ^2, we denote it as

$$X \sim N(\mu, \sigma^2)$$

Properties of normal distribution

- The normal distribution curve is a bell shaped, symmetrical about μ and mesokurtic. The p.d.f. attains its maximum value at x= μ.
- Since for x= μ divides the area under the normal curve into two equal parts, μ is the mean, the median and the mode of the distribution.
- The mean and variance of the normal distribution are μ, and σ^2, respectively.

Properties of normal distribution cont….

- The total area under the curve and bounded from below by the horizontal axis is 1, i.e

$$\int f_X(x)dx = 1$$

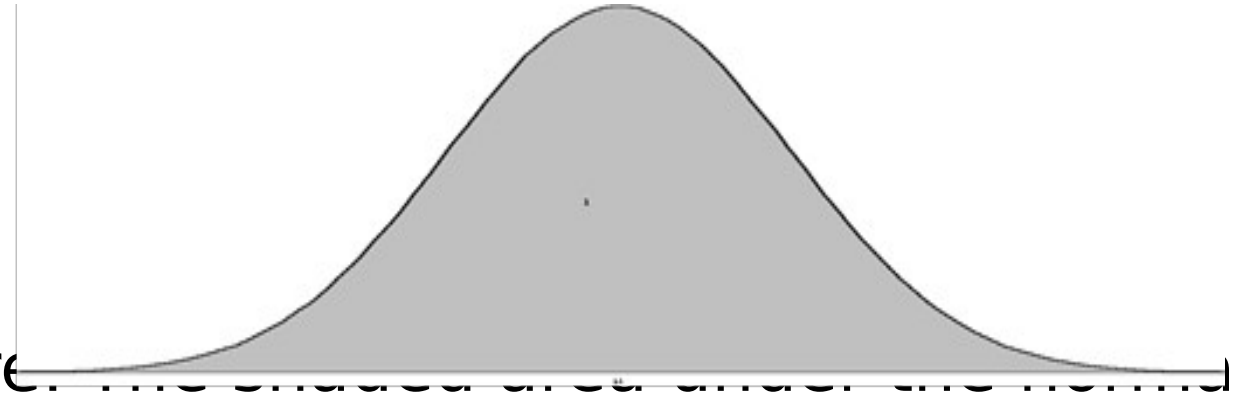


Figure: The shaded area under the normal curve is one

- Since a normal distribution is a continuous probability distribution, the probability that X lies between a and b is the area bounded under the curve, from left to right by the vertical lines x = a and x = b and below by the horizontal axis

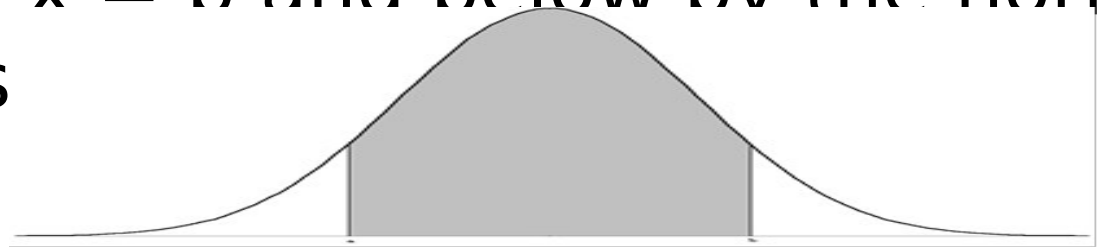


Figure: P(a<X<b) equals the shaded

- However, evaluating $P(a \leq X \leq b) = \int_a^b f_X(x)dx$ is very complicated.
- To facilitate this problem, we use the standard normal table which gives area values bounded by two points.
- Areas under the standard normal distribution curve are tabulated in various ways. The most common tables give areas bounded between Z=0 and a positive value of Z.

- In addition to the standard normal table, the properties of normal distribution and the following theorem are useful to make probability calculations very easy for any normal distribution.

Standardization of a normal random variable

- If X has a normal distribution with mean, μ and standard deviation ,σ , then

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

will have a standard normal distribution.

$$P(a < X < b) = P(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma})$$

$$= P(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma})$$

Let Z be the standard normal random variable. Calculate the following probabilities using the standard normal distribution table:

- a) $P(0<Z<1.2)$
- b) $P(0<Z<1.43)$
- c) $P(Z\leq 0)$
- d) $P(-1.2<Z<0)$
- e) $P(Z\leq -1.43)$
- f) $P(-1.43\leq Z<1.2)$
- g) $P(Z\geq 1.52)$
- h) $P(Z\geq -1.52)$

- **Solution:**

  The probability that Z lies between 0 and 1.2 can be directly found from the standard normal table as follows: look for the value 1.2 from z column (first column) and then move horizontally until you find the value of 0.00 in the first row. The point of intersection made by the horizontal and vertical movements will give the desired area (probability). Hence P(0<Z<1.2)= 0.3849. Refer the table below as a guide to find this probability.

  standard normal table.docx T test and F test.docx

z	0.00	0.01	0.02	0.03	0.04	.	.	.	0.09
	0.0000	0.0040	0.0080	0.0120	0.0160	.	.	.	
0.1	0.0398	0.0438							
0.2	0.0793	0.0832							
0.3									
.									
.									
.									
1.2	0.3849	0.3869							
1.3	0.4032	0.4049							
1.4	0.4192	0.4207	0.4222	0.4236					
1.5									
.									
.									
.									
3.4	0.4997	.	.	.	.	.	.	.	0.4998

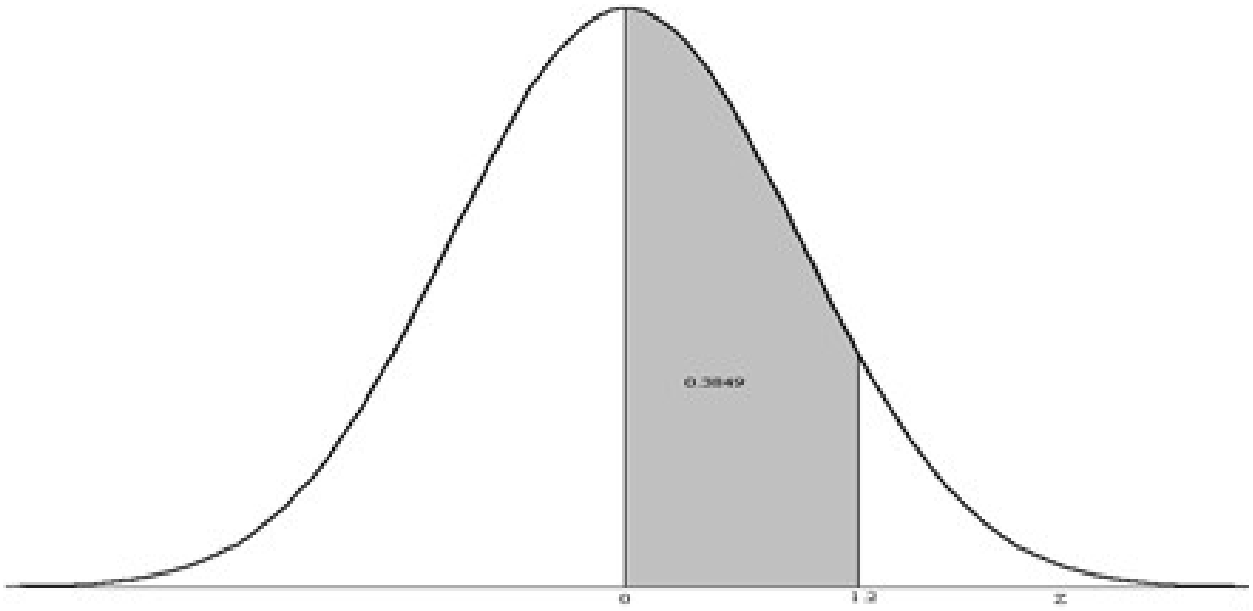
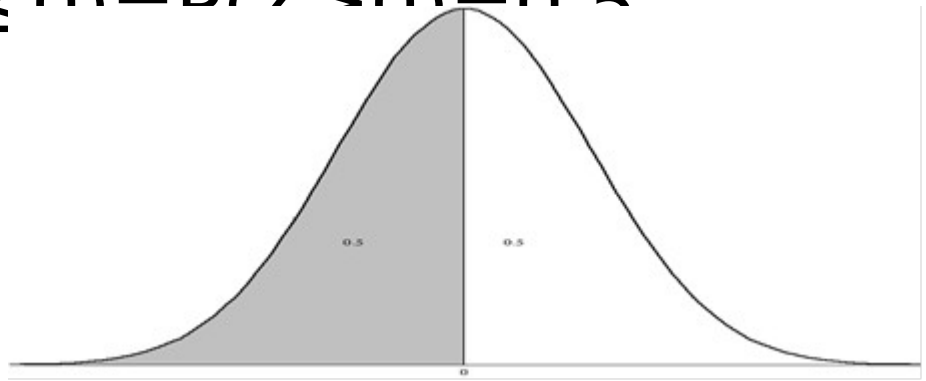


Figure: P(0<Z<1.2) is the shaded area

- In a similar way P(0<Z<1.43)= 0.4236.
- We know that the normal distribution is symmetric about its mean. Hence the area to the left of 0 and the to the right of zero are 0.5 each. Therefore P(Z≤0)=P(Z≥0)=0.5



d) $P(-1.2<Z<0)=P(0<Z<1.2)= 0.3849$ due to symmetry

e) $P(Z<-1.43)= 1- P(Z \geq -1.43)$ Using the probability of the complement event.

$= 1-[P(-1.43<Z<0)+P(Z\geq 0)]$ Since a region can be broken down

$=1-[P(0<Z<1.43)+P(Z \geq 0)]$ into non overlapping regions.

$=1-[0.4236 + 0.5]$

$=1-0.9236=0.0764$

F) $P(-1.43 \leq Z < 1.2) = P(-1.43 \leq Z < 0) + P(0 \leq Z < 1.2) = P(0 < Z \leq 1.43) + 0.3849 = 0.4236 + 0.3849 = 0.8085$

Figure: $P(-1.43 \leq Z < 1.2)$ is the shaded region

G) $P(Z \geq 1.52) = 0.5 - P(0 \leq Z < 1.52) = 0.5 - 0.4357 = 0.0643$

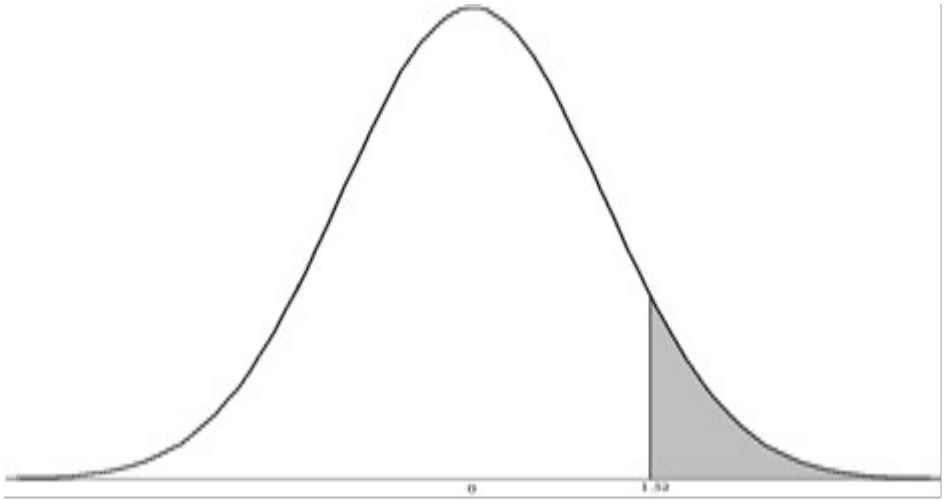
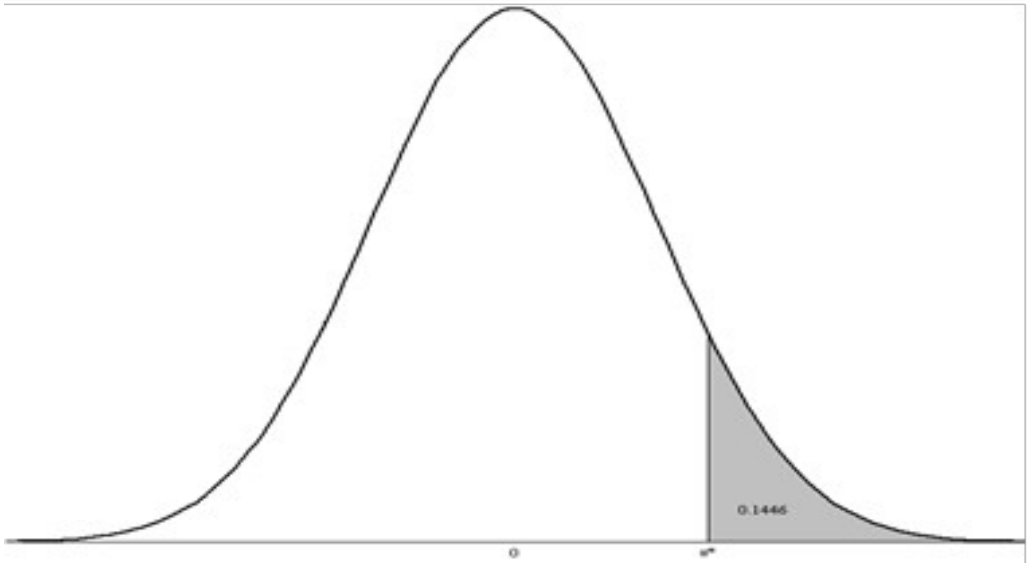


Figure: $P(Z \geq 1.52)$ is the shaded region

h) $P(Z \geq -1.52) = P(-1.52 \leq Z < 0) + P(Z \geq 0) = P(0 < Z \leq 1.52) + 0.5 = 0.4357 + 0.5 = 0.9357$

- **Example:** Find the following values of z^* of a standard normal random variable based on the given probability values:
- $P(Z > z^*) = 0.1446$
- $P(Z > z^*) = 0.8554$
- **Solution:** We need to find specific values of Z given some probability values.
- If the probability that $Z > z^*$ is 0.1446 implies that z^* is to the right of zero because $P(Z > 0) = 0.5$ is greater than $P(Z > z^*)$.



- P(Z > z*) = 0.1446 implies that P(0<Z≤z*) = 0.5 - 0.1446=0.3554.

Hence we can look for the value of z* satisfying the above condition form the standard normal table. Thus z* =1.06

- If the probability that Z>z* is 0.8554 implies that z* is to the left of zero because P(Z>0) = 0.5 is less than P(Z>z*). It implies that z* is a negative numbe

- $P(Z>z^*) = 0.8554 = P(z^* \leq Z < 0) + P(Z \geq 0) = P(0 \leq Z \leq -z^*) + 0.5$
- Implies $P(0 \leq Z \leq -z^*) = 0.8554 - 0.5 = 0.3554$. Hence the value $-z^*$ form the table satisfying the above condition is 1.06. Therefore $z^* = -1.06$.

- **Example:** If the total cholesterol values for a certain target population are approximately normally distributed with a mean of 200 (mg/100 ml) and a standard deviation of 20 (mg/100 ml), calculate the probability that a person picked at random from this population will have a cholesterol value
  a. greater than 240 (mg/100 ml)
  b. between 180 and 220(mg/100 ml)
  c. less 200 (mg/100 ml)

Solution: Let X be the cholesterol values in mg/100 ml, then $X \sim N(200, \quad 400)$

a) $$P(X > 240) = P(\frac{X-\mu}{\sigma} > \frac{b-\mu}{\sigma})$$

$$= P(Z > \frac{240-200}{20}) = P(Z > 2) = 0.5 - P(0 < Z < 2) = 0.5 - 0.4772 = 0.0228$$

$$P(180 < X < 220) = P(\frac{X-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma})$$

b) $$= P(\frac{180-200}{20} < Z < \frac{220-200}{20}) = P(-1 < Z < 1)$$

$$= 2P(0 < Z < 1) = 2 \times 0.3413 = 0.6826$$

c) $$P(X < 200) = P(Z < \frac{200-200}{20}) = P(Z < 0) = 0.5$$

- Assume that the test scores for a large class are normally distributed with a mean of 74 and a standard deviation of 10.

a. Suppose that you receive a score of 88. What percent of the class received scores higher than yours?

b. Suppose that the teacher wants to limit the number of A grades in the class to no more than 20%. What would be the lowest score for an A?

Solution: Let X be the score of a randomly picked student, then $X \sim N(74, \quad 100)$

a) $$P(X > 88) = P(\frac{X-74}{10} \geq \frac{88-74}{10}) = P(Z > 1.4)$$

$$= 0.5 - P(0 < Z \leq 1.4) = 0.5 - 0.4192 = 0.0808$$

Hence 8.08 percent of the students score more than you did?

b) Let X_A be the lowest mark to get letter grade A. We are given that

$$P(X \geq x_A) = 0.2 = P(\frac{X-74}{10} \geq \frac{x_A - 74}{10}) = P(Z > z_A)$$

$$\Rightarrow P(0 < Z \leq z_A) = 0.5 - 0.2 = 0.3 \Rightarrow z_A = 0.85 \Rightarrow z_A = 0.85 = \frac{x_A - 74}{10}$$

Hence, the lowest mark to get letter grade A is 82.5.

SAMPLING AND SAMPLING DISTRIBUTION OF SAMPLE MEAN

Ashebir Feyisa 12/07/2025

Objectives:

- After a successful completion of this unit, students will be able to:
- Differentiate the two major sampling techniques: probabilistic and non-probabilistic
- Apply simple random sampling technique to select sample
- Define sampling distribution of the sample mean

Methods of sampling

- **Definition of some basic terms**
- **Sampling:** is the technique of selecting representative sample from the whole.
- **Population:** is the totality of elements or units under study.
- **Sample:** is the part of the population.
- **Sampling Frame**: A complete list of all the units of the population is called the sampling frame.

- A unit of population is a relative term. If all the workers in a factory make a population, then a worker is a unit of the population. If all the factories in a country are being studied for some purpose, then a factory is a unit of the population of factories. The frame provides a base for the selection of a sample.

Major reasons to use sampling

- **Saves Time and Cost:** As the size of the sample is small as compared to the population, the time and cost involved on sample study are much less than the complete counts. Hence a sample study requires less time and cost.
- **To prevent destruction:** The destructive nature of some experiments (or inspection) do not allow to carryout complete enumeration, for instance, to check quality of beers, to study the efficacy of new drugs, testing the life length of a bulb, e t c.
- **Sample survey provides higher level of accuracy:** This accuracy can be achieved through more selective recruiting of interviewers and supervisors, more extensive training programs, a closer supervision of the personnel involved and a more efficient monitoring of the field work.

- **Types of sampling**

- Generally, two types of sampling methods exist:
  I. probability and
  II. non-probability sampling.

Probability Sampling

- The term probability sampling (or random sampling) is used when the selection of the sample is purely based on chance.
- There is no subjective bias in the selection of units. Every unit of the population has a known nonzero probability to be in the sample.

- The following are some of the random sampling methods: Simple random sampling, Stratified random sampling, Cluster sampling, Systematic random sampling.

Simple random sampling

- Simple random sampling is a method of selecting a sample from a population in such a way that every unit of the population is given an equal chance of being selected.
- In practice, you can draw a simple random sample of elements using either the 'lottery method' or 'tables of random numbers'.

Cont...

- For example, you may use the lottery method to draw a random sample by using a set of 'N' tickets, with numbers ' 1 to N' if there are 'N' units in the population. After shuffling the tickets thoroughly, the sample of a required size, say n, is selected by picking the required n number of tickets.

Cont...

- The best method of drawing a simple random sample is to use a table of random numbers. After assigning consecutive numbers to the units of population, the researcher starts at any point on the table of random numbers and reads the consecutive numbers in any direction horizontally, vertically or diagonally. If the read out numbers corresponds with the one written on a unit card, then that unit is chosen for the sample.

- Suppose that a sample of 6 study centers is to be selected at random from a serially numbered population of 60 study centers.

- The following table is portion of a random numbers table used to select a sample.

Row Column	1	2	3	4	5	.....	N
1	2315	7548	5901	8372	5993	.....	6744
2	0554	5550	4310	5374	3508	.....	1343
3	1487	1603	5032	4043	6223	.....	0834
4	3897	6749	5094	0517	5853	.....	1695
5	9731	2617	1899	7553	0870	.....	0510
6	1174	2693	8144	3393	0862	.....	6850
7	4336	1288	5911	0164	5623	.....	4036
8	9380	6204	7833	2680	4491	.....	2571
9	4954	0131	8108	4298	4187	.....	9527
10	3676	8726	3337	9482	1569	.....	3880
11	.....	.....	.....	.....	.....	.....	.....
12	.....	.....	.....	.....	.....	.....	.....
13	.....	.....	.....	.....	.....	.....	.....
14	.....	.....	.....	.....	.....	.....	.....
15	.....	.....	.....	.....	.....	.....	.....
N	3914	5218	3587	4855	4888	.....	8042

- If you start in the first row and first column, centers numbered 23, 05, 14,..., will be selected. However, centers numbered above the population size (60) will not be included in the sample. In addition, if any number is repeated in the table, it may be substituted by the next number from the same column.

- 1, the number to start with is 83. In this way you can select first 6 numbers from this column starting with 83.
- The sample, then, is as follows:

~~83~~	~~75~~
53	**33**
40	**01**
05	**26**

- Hence, the study centers numbered 53, 40, 05, 33, 01 and 26 will be in the sample.

- Simple random sampling ensures the best results. However, from a practical point of view, a list of all the units of a population is not possible to obtain.
- Even if it is possible, it may involve a very high cost which a researcher or an organization may not be able to afford. In addition, it may result an unrepresentative sample by chance.

Stratified sampling

- Stratified random sampling takes into account the stratification of the main population into a number of sub-populations, each of which is homogeneous with respect to one or more characteristic(s).
- Having ensured this stratification, it provides for selecting randomly the required number of units from each sub-population.
- The selection of a sample from each subpopulation may be done using simple random sampling. It is useful in providing more accurate results than simple random sampling.

Systematic sampling

- In this method, samples are selected at equal intervals from the listings of the elements.
- This method provides a sample as good as a simple random sample and is comparatively easier to draw a sample.
- For instance, to study the average monthly expenditure of households in a city, you may randomly select every fourth households from the household listings

Cluster sampling

- Cluster sampling is used when sampling frame is difficult to construct or using other sampling techniques (simple random sampling) is not feasible or costly.
- For instance, when the geographic distribution of units is scattered it is difficult to apply simple random sampling.
- It involves division of the population of elementary units into groups or clusters that serve as primary sampling units.
- A selection of the clusters is then made to form the sample.
- The precision of estimates made based on samples taken using this method is relatively low.

Non-probabilily sampling techniques

- In non-probability sampling, the sample is not based on chance.
- It is rather determined by personal judgment. This method is cost effective; however, we cannot make objective statistical inferences.
- Depending on the technique used, non-probability samples are classified into quota, judgment or purposive and convenience samples.

Sampling and non-sampling errors

- Sampling error is the difference between the value of a sample statistic and the value of the corresponding population parameter.
- On the other hand, non-sampling error is an error that occurs in the collection, recording and tabulation of data.
- Sampling error can be minimized by using appropriate sampling methods and/or increasing the sample size.
- The non-sampling error is likely to increase with increase in sample size.

Sampling distribution of the sample mean

- The value of the sample mean for any sample will depend on the elements included in that sample.
- Consequently, the sample mean is a random variable.
- Therefore, like other random variable, the sample means possess a probability distribution which is more commonly called the *sampling distribution of sample mean*.

- In general, the probability distribution of a sample statistic is called its sampling distribution.
- Sampling distribution is important in statistical inference.
- The important characteristics of the sampling distribution of the sample mean are its mean, variance and the form of the distribution.

- **Example:** Suppose we have a hypothetical population of size 3, consisting of three children namely: A is 3 years old, B is 6 years old and C is 9 years old. Construct sampling distribution of the sample mean of size 2 using sampling without replacement and with replacement.
- **Solution:** The mean and variance of the population are 6 and 6, respectively.
  - If sampling is without replacement we will have ${}_3C_2 = 3$ possible samples: (A, B), (A, C) and (B, C) and their corresponding sample means are (3+6)/2 = 4.5, 6 and 7.5, respectively. Hence the probability distribution (sampling distribution) of the sample mean is:

$\bar{x}$	4.5	6	7.5
$P(\bar{X} = \bar{x})$	1/3	1/3	1/3

$$E(\bar{X}) = \sum \bar{x} P(\bar{x}) = 4.5(1/3) + 6(1/3) + 7.5(1/3) = 6$$

$$V(\bar{X}) = \left(\sum \bar{x}^2 P(\bar{x})\right) - \mu_{\bar{x}}^2 = (6.75 + 12 + 18.75) - 36 = 1.5$$

2. If sampling is with replacement we will have $N^n = 3^2 = 9$ possible samples: (A, A), (A, B), (A, C), (B, A), (B, B), (B, C), (C, A), (C, B) and (C, C). Hence the probability distribution (sampling distribution) of the sample mean is:

$\bar{x}$	3	4.5	6	7.5	9
$P(X = \bar{x})$	1/9	2/9	3/9	2/9	1/9

$$E(\bar{X}) = \sum \bar{x} P(\bar{x}) = 3(1/9) + 4.5(2/9) + 6(3/9) + 7.5(2/9) + 9(1/9) = 6$$

$$V(\bar{X}) = \left(\sum \bar{x}^2 P(\bar{x})\right) - \mu_{\bar{x}}^2 = (1 + 4.5 + 12 + 12.5 + 9) - 36 = 3$$

- **Note:**
- The mean of the sampling distribution of the sample mean is the same as the population mean irrespective of the sampling procedure.
- The variance of the sampling distribution

$$\begin{cases} \dfrac{\sigma^2}{n}, & \text{if sampling is with replacement} \\ \dfrac{\sigma^2}{n}\left(\dfrac{N-n}{N-1}\right), & \text{if sampling is without replacement} \end{cases}$$

- The problem with using sample mean to make inferences about the population mean is that the sample mean will probably differ from the population mean.
- This error is measured by the variance of the sampling distribution of the sample mean and is known as the *standard error*.
- The standard error is the average amount of sampling error found because of taking a sample rather than the whole population.
- As sample size increases, the standard error decreases.

Central Limit Theorem

If $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2, then as n goes to infinity the distribution of the sample mean, $\overline{X}$, approximates normal distribution with mean μ and variance σ^2/n. That is, as n gets large, $\overline{X} \sim N(\mu, \sigma^2/n)$ and its standardized form is

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1).$$

Note: The central limit theorem is useful for approximating the distribution of the sample mean based on a large sample size and when the population distribution is non normal; however, if the population is normal, then the sampling distribution of the sample mean will be normal regardless of the sample size.

Example: If the uric acid values in normal adult males are normally distributed with mean 5.7 mgs and standard deviation of 1mg. Find the probability that

a) a sample of size 4 will yield a mean less than 5
b) a sample of size 9 will yield a mean greater than 6

Solution: Let X be the amount of uric acids in normal adult males with mean 5.7 and variance 1.

a) If a sample of size 4 is taken, then $\bar{X} \sim \mathrm{N}(5.7, 0.25)$ since the population is normally distributed.

$$P(\bar{X} < 5) = P(Z < \frac{5 - 5.7}{0.5}) = P(Z < -1.4)$$

$$= 0.5 - P(0 < Z < 1.4) = 0.0808$$

b) If a sample of size 9 is taken, then $\bar{X} \sim \mathrm{N}(5.7, 1/9)$ since the population is normally distributed.

$$P(\bar{X} > 6) = P(Z > \frac{6 - 5.7}{1/3}) = P(Z > 0.9)$$

$$= 0.5 - P(0 < Z < 0.9) = 0.1841$$

REGRESSION METHODS AND CORRELATION

Ashebir Feyisa 12/07/2025

Introduction

- The statistical methods discussed so far are used to analyze the data involving only one variable.
- Often an analysis of data concerning two or more variables is needed to look for any statistical relationship or association between them.
- Thus, regression and correlation analysis are helpful in ascertaining the probable form of the relationship between variables and the strength of the relationship.

Simple linear regression analysis

- Regression analysis is the statistical method that helps to formulate a functional relationship between two or more variables.
- It can be used for assessment of association, estimation and prediction.
- For instance one might be interested to formulate a statistical model to relate the height of fathers and their sons, blood pressure and age, fertilizer amount and yield, etc.

- A simple model to relate dependent (response) variable Y and with only one predictor variable X is to consider a linear relationship.
- The first step in regression analysis involving two variables is to construct a scatter plot (diagram) of the observed data.
- Scatter diagram is a plot of all ordered pairs (Xi,Yi) on the coordinate plane which is helpful for determining an apparent relationship between two variables.

- The simple linear regression of Y on X can be expressed with respect to the population parameters α and β as

$$Y = \alpha + \beta X + \varepsilon$$

- where α = y-intercept that represents the mean value of the dependent variable Y when the independent variable X is zero; β = slope of the regression line that represents the change in the mean of for a unit change in the value of ; ε = error term

- The population parameters α and β can be estimated from sample data using the least square technique. The estimators of α and β are usually denoted by a and b, respectively.
- The resulting regression line is:

$$\hat{Y} = a + bX$$

and the equation is known as the fitted regression line.

The estimated values of y are denoted by $\hat{Y}$.
The observed values of are denoted by y

The difference between the observed and the estimated values, $Y - \hat{Y}$, is known as error or residual, and is denoted by $\{\hat{\varepsilon}\}$.

The residual can be positive, negative or zero.

A best fitting line is the one for which the sum of squares of the residuals, $\sum \hat{\varepsilon}^2$ has the minimum value.

This is called the method of least squares.

According to this method, one would select a and b such that

$\sum \hat{\varepsilon}^2 = \sum (Y - \hat{Y})^2$ is minimum.

The solution of this minimization problem using partial differentiation is as follows:

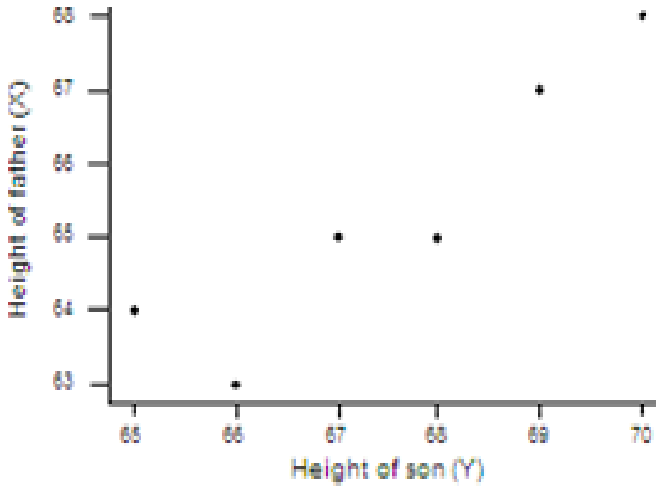
$$b = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{(\sum X)^2}{n}} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \quad \text{and} \quad a = \bar{Y} - b\bar{X}$$

Example: A researcher wants to find out if there is any relationship between height of the son and his father. He took random sample of 6 fathers and their sons. The height in inch is given in the table below:

Height of father (X)	63	65	64	65	67	68
Height of the son (Y)	66	68	65	67	69	70

Solution:

i)



From the scatter plot one can see that the points are roughly on straight line.

ii)

$$n=6 \quad \sum X=392,\ \sum Y=405,\ \sum X^2=25628,\ \sum XY=26476,\ \sum Y^2=27355$$

$$b=\frac{n\sum XY-\sum X\sum Y}{n\sum X^2-(\sum X)^2}=\frac{6(26476)-(392)(405)}{6(25628)-(392)^2}=0.923 \quad a=\bar{Y}-b\bar{X}=\frac{405}{6}-0.923\frac{392}{6}=7.2$$

Then the fitted (regression) line of Y on X is given by:

$$\hat{Y}=a+bX=7.2+0.923X$$

- ✓ The slope of the line, i.e. b=0.923, tells us that a unit (one inch) increase in the height of the father results in 0.923 inch increase in the height of the son.
- ✓ The y-intercept of the line, i.e. a=7.2, is the value of Y when the value of X is zero(do you think that the intercept is meaningful?)

iii) Y=7.2+0.923(66) =68.118, thus the height of the son is 68.118 inch.

The covariance and the correlation coefficient

- Correlation coefficient measures the degree of linear relationship between two variables. The population correlation coefficient is represented by ρ and its estimator is r.

- For a set of n pairs of sample values X and Y, Pearson's correlation coefficient is calculated as the ratio of the covariance of the variables X and Y to the product of the standard deviations of X and Y.

Symbolically,

$$r = \frac{Cor(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y).}} = \frac{\sum \frac{(X-\bar{X})(Y-\bar{Y})}{n-1}}{\sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}\sqrt{\frac{\sum (Y-\bar{Y})^2}{n-1}}}$$

$$= \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2}\sqrt{\sum (Y-\bar{Y})^2}}$$

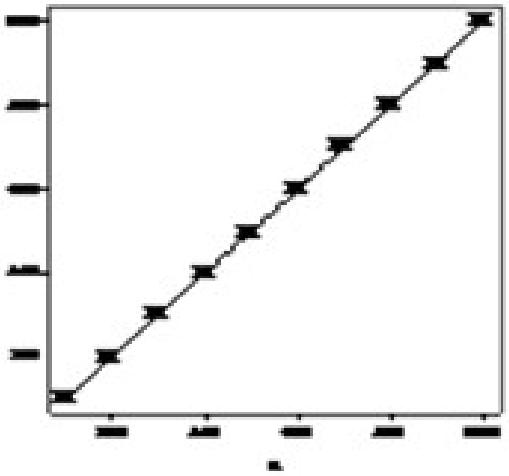
Alternatively, the Pearson's correlation coefficient r can be obtained as:

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

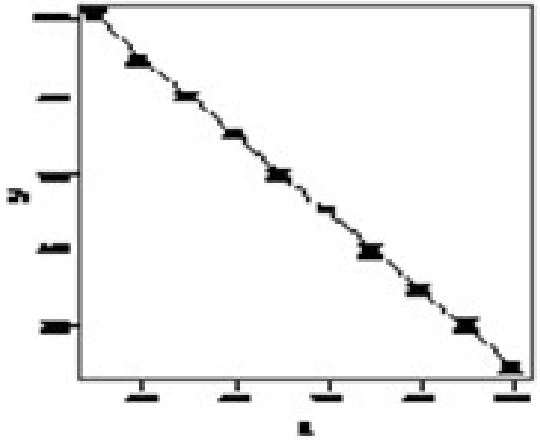
Properties of Pearson's correlation coefficient r,

- It is appropriate to calculate when both variables X and Y are measured on an interval or ratio scale.
- The value of r is independent of the unit in which X and Y are measured. i.e., it is a pure number.
- The value of r ranges from +1 to -1.
- r = +1 indicates a perfect linear relationship between X and Y with positive slope.
- r = -1 indicates a perfect linear relationship between X and Y with negative slope.
- r = 0 indicates no linear relationship between the two variables X and Y.
- as r approaches +1 indicates strong and positive linear relationship between the two variables
- as r approaches -1 indicates strong and negative linear relationship between the two variables
- as r approaches 0 indicates weak linear relationship between the two variables

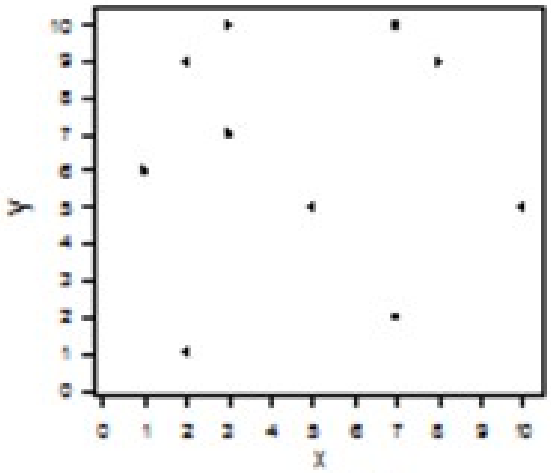
Examples of correlation coefficients:



r=+1



r=-1



r=0

Example: In some locations, there is strong association between concentrations of two different pollutants. An article reports the accompanying data on ozone concentration x (ppm) and secondary carbon concentration y ($\mu g / m^3$):

X	0.066	0.088	0.120	0.050	0.162	0.186	0.057	0.100
Y	4.6	11.6	9.5	6.3	13.8	15.4	2.5	11.8

0.112	0.055	0.154	0.074	0.111	0.140	0.071	0.110
8.0	7.0	20.6	16.6	9.2	17.9	2.8	13

a. Calculate the correlation coefficient and comment on the strength and direction of the relationship between the two variables.

Solution: The summary quantities are

$$n=16, \sum x_i = 1.656, \sum y_i = 170.6, \sum x_i y_i = 20.0397, \sum x_i^2 = 0.196912, \sum y_i^2 = 2253.56$$

The Person's correlation coefficient is

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{16(20.0397) - (1.656)(170.6)}{\sqrt{16(0.196912) - (1.656)^2}\sqrt{16(2253.56) - (170.6)^2}}$$

$$= \frac{320.6352 - 282.5136}{\sqrt{0.408256}\sqrt{6952.6}} = \frac{38.1216}{(.639)(83.38)}$$

$$= 0.716$$

The value of 0.716 indicates that there is somehow strong and positive relationship between ozone concentration and secondary carbon concentration.

chapter 8

- ESTIMATION AND HYPOTHESIS TESTING

Ashebir Feyisa 12/07/2025

Objectives:

Having studied this unit, you should be able to

- construct and interpret confidence interval estimates
- formulate hypothesis about a population mean
- determine an appropriate sample size for estimation

Introduction

- We now assume that we have collected, organized and summarized a random sample of data and are trying to use that sample to estimate a population parameter.
- Statistical inference is a procedure whereby inferences about a population are made on the basis of the results obtained from a sample.

Introduction

- Statistical inference can be divided in to two main areas: **estimation and hypothesis testing**.
- Estimation is concerned with estimating the values of specific population parameters;
- hypothesis testing is concerned with testing whether the value of a population parameter is equal to some specific value.

Point and interval estimation of the mean

Point estimate: In point estimation, a single sample statistic (such as $\bar{x}$, s *or* $\hat{p}$) is calculated from the sample to provide an estimate of the true value of the corresponding population parameters (such as μ, δ *or* p). Such a single statistic is termed as point estimator, and the specific value of the statistic is termed as point estimate. For example, the sample mean $\overline{X}$ is an estimator for population mean and $\overline{X} = 10$ is an estimate, which is one of the possible values of $\overline{X}$.

Interval estimate: In most practical problems, a point estimate does not provide information about 'how close is the estimate' to the population parameter unless accompanied by a statement of possible sampling errors involved based on the sampling distribution of the statistic. Hence, an interval estimate of a population parameter is a confidence interval with a statement of confidence that the interval contains the parameter value. An interval estimate of the population parameter θ consists of two bounds within which the parameter will be contained:

$$L \leq \theta \leq U$$

where L is the lower bound and U is the upper bound.

Case 1: When the population is normal.

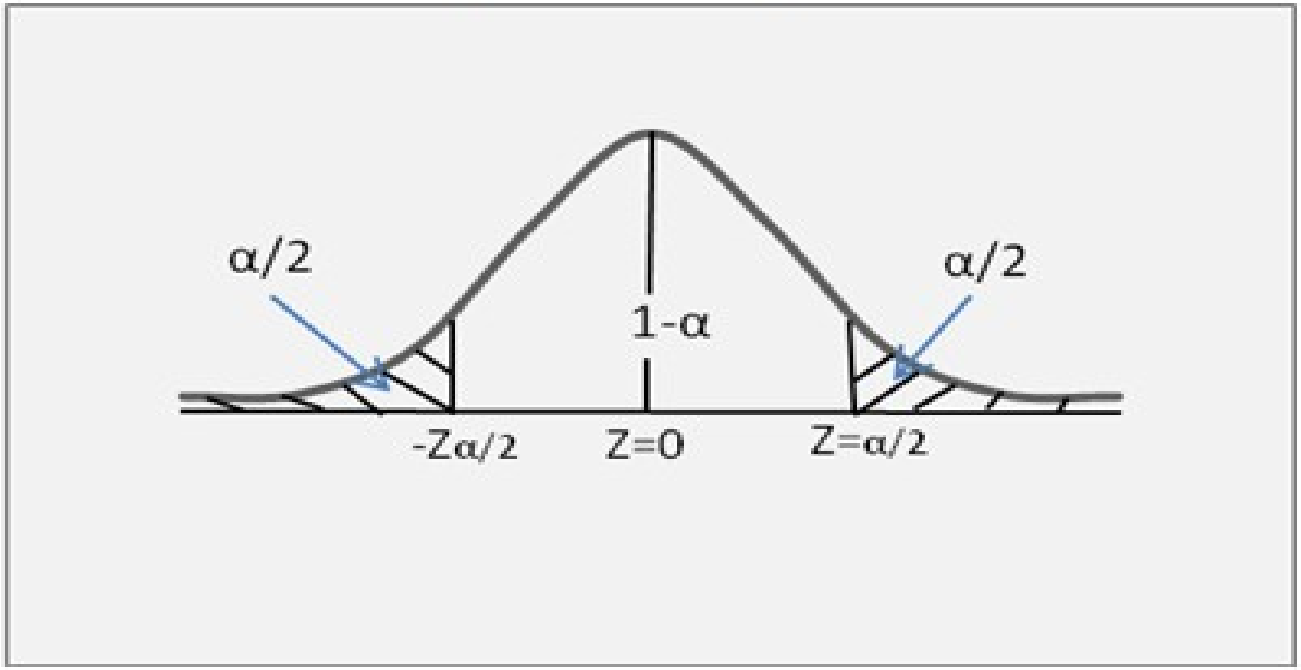
- ✓ If the variance σ^2 is known, the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and variance $\frac{\delta^2}{n}$. i.e., $\bar{X} \sim N\left(\mu, \frac{\delta^2}{n}\right)$ and $Z = \frac{\bar{X} - \mu}{\delta/\sqrt{n}} \sim \text{N}(0,1)$.
- ✓ If the variance σ^2 is unknown, $t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ will have t-distribution with n - 1 degrees of freedom. Moreover, as the sample size increases t is approximately the same as standard normal.

Consider the case σ^2 is known, we can derive a $(1-\alpha)100\%$ confidence interval for the population mean μ.

Let $Z_{\alpha/2}$ be a point on the standard normal curve that cuts an area of $\frac{\alpha}{2}$ to the right. i.e. $P(Z > Z_{\alpha/2}) = \frac{\alpha}{2}$. By the symmetric property of the normal distribution, $P(Z < -Z_{\alpha/2}) = \frac{\alpha}{2}$ (see the diagram below).

- From the standard normal distribution, w

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$$



To obtain the limit of the interval estimate, we use the standardized form of $\bar{X}$ in the above probability statement. i.e., letting $Z = \dfrac{\bar{X} - \mu}{\delta/\sqrt{n}}$

$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha$ Becomes

$$\Rightarrow P(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\delta/\sqrt{n}} < Z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(-Z_{\alpha/2}\frac{\delta}{\sqrt{n}} < \bar{X} - \mu < Z_{\alpha/2}\frac{\delta}{\sqrt{n}}) = 1 - \alpha$$

$$\Rightarrow P(-\bar{X} - Z_{\alpha/2}\frac{\delta}{\sqrt{n}} < -\mu < -\bar{X} + Z_{\alpha/2}\frac{\delta}{\sqrt{n}}) = 1 - \alpha$$

$$\Rightarrow P(\bar{X} - Z_{\alpha/2}\frac{\delta}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\delta}{\sqrt{n}}) = 1 - \alpha$$

We can assert with probability $1-\alpha$ that the interval $(\bar{X} - Z_{\alpha/2}\frac{\delta}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\delta}{\sqrt{n}})$ contains the population mean we are estimating.

Thus, $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

The end points of the interval, $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, are called *confidence limits* and the probability $1-\alpha$ is called the *degree of confidence*.

In a similar way a $(1-\alpha)100\%$ confidence interval for the population mean μ with unknown variance σ^2 is given by

$$\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\quad \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right)$$

Where $t_{\alpha/2}$ is the critical value of t-test statistic providing an area $\frac{\alpha}{2}$ in the right tail of the t-distribution with $n-1$ degrees of freedom, and $S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$.

Case 2: When the population is non normal.

We use the central limit theorem to approximate the distribution of the sample mean based on large sample ($n \geq 30$). Large sample size is a necessary condition to use the normal distribution. And hence, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim$ N(0, 1). If σ is unknown we can replace it by its sample estimate S. The resulting $(1-\alpha)100\%$ confidence interval of μ becomes

$$\begin{cases} \left(\bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right), & \text{when } \sigma \text{ is known} \\ \left(\bar{X} - Z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{S}{\sqrt{n}} \right), & \text{when } \sigma \text{ is unknown} \end{cases}$$

Example8.1: A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 millimeter of mercury (mmHg) with a standard deviation of 0.95 mmHg. What is the 95% confidence interval for the mean change in blood pressure? (Assume that the population is normal).

Solution: Given: $\bar{X} = 2.28$, $S = 0.95$, $n = 6$

$$(1-\alpha)100\% = 95\% \Rightarrow 1-\alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$$

- ✓ $\bar{X} = 2.28$ is a point estimate for the population mean drop in blood pressure μ.

A 95% confidence interval of population mean for unknown δ^2 and small sample size is:

$$\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\quad \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right).$$

And from the t distribution table, $t_{\alpha/2}(n-1) = t_{0.025}(5) = 2.571$

$$\left(2.28 - (2.571)\frac{0.95}{\sqrt{6}},\quad 2.28 + (2.571)\frac{0.95}{\sqrt{6}}\right)$$

$$\Rightarrow (2.28\text{-}0.997,\ 2.28\text{+}0.997)$$

$$\Rightarrow (1.28,\ 3.27)$$

We are 95% confident that the mean drop in blood pressure lies in between 1.28 mmHg and 3.27 mmHg for the sampled population.

Example 8.2: Punctuality of patients in keeping appointment is of interest to a research team. In a study of patients flow through the office of general practitioners, it was found that a sample of 35 patients were 17.2 minutes late for appointments, on the average. Previous research had shown the standard deviation to be about 8 minutes. The population distribution was felt to be not normal. What is the 90 percent confidence interval for the true mean amount of time late for appointment?

Solution: Given; $\bar{X} = 17.2,\ \delta = 8,\ \ n = 35$

$(1-\alpha)100\% = 90\% \Rightarrow 1-\alpha = 0.90 \Rightarrow \alpha = 0.1 \Rightarrow \alpha/2 = 0.05$

Since the sample size is fairly large (n > 30), and since the population standard deviation is known, according to the central limit theorem, the sampling distribution of sample mean is approximately normal. Thus, a confidence interval of the population mean is given by:

$$\left(\bar{X} - Z_{\alpha/2}\frac{\delta}{\sqrt{n}},\ \ \bar{X} + Z_{\alpha/2}\frac{\delta}{\sqrt{n}}\right)$$

And from the standard normal distribution table, $Z_{\alpha/2} = Z_{0.05} = 1.65$

$$\left(17.2 - (1.65)\frac{8}{\sqrt{35}},\ \ 17.2 + (1.65)\frac{8}{\sqrt{35}}\right)$$

$$\Rightarrow (17.2 - 2.2, 17.2 + 2.2)$$

$$\Rightarrow (15.0, 19.4)$$

Therefore, the 90% confidence interval for true mean amount of time late for appointment is between 15.0 and 19.4 minutes.

Hypothesis Testing about the Mean

- In many circumstances we merely wish to know whether a certain proposition is true or false.
- Different people can form different opinions by looking at data, but a hypothesis test provides a standardized decision-making process that will be consistent for all people.

- **Statistical hypothesis:** is a claim (belief or assumption) about an unknown population parameter values.
- Examples of hypothesis:
  - There is association between lung cancer and number of cigarettes an individual smokes.
  - The proportion of female students in AASTU is 0.35.
  - In sub-Saharan Africa 40% of individuals are leaving below poverty line.
- **Hypothesis testing:** is the procedure that enables decision-makers to draw inferences about population characteristics by analyzing the difference between the value of sample statistic and the corresponding hypothesized parameter value.

General procedure for hypothesis testing

- To test the validity of the claim or assumption about the population parameter, sample is drawn from the population and analyzed.
- The result of the analysis are used to decide whether the claim is valid or not.

Step 1: State the null hypothesis (H_0) and alternative hypothesis (H_1)

Null hypothesis (H_0): refers to a hypothesized numerical value of the population parameter which is initially assumed to be true. The null hypothesis is always expressed in the form of an equation making a claim regarding the specific value of the population parameter. That is, for example

$$H_0 : \mu = \mu_0$$

where μ_0 is hypothesized value of the population mean.

Alternative hypothesis (H_1): is the logical opposite of the null hypothesis. The alternative hypothesis states that specific population parameter value is not equal to the value stated in the null hypothesis. For example,

$H_1 : \mu \neq \mu_0$ (Two-sided test)

$H_1 : \mu < \mu_0$ *or* $H_1 : \mu > \mu_0$ (One-sided test)

Step 2: State the level of significance (alpha) for the test

- The level of significance is the probability to wrongly reject the null hypothesis when it is actually true. It is specified by the statistician or the researcher before the sample is drawn. The most commonly used values are 0.10, 0.50 or 0.01.

Step 3: Calculate the appropriate test statistic

Test statistic is a value computed from a sample that is used to determine whether the null hypothesis has to be rejected or not. The choice of suitable test statistic depends on the sampling distribution of the sample statistic. Accordingly, we have the following cases:

Case 1: When the population is normal.

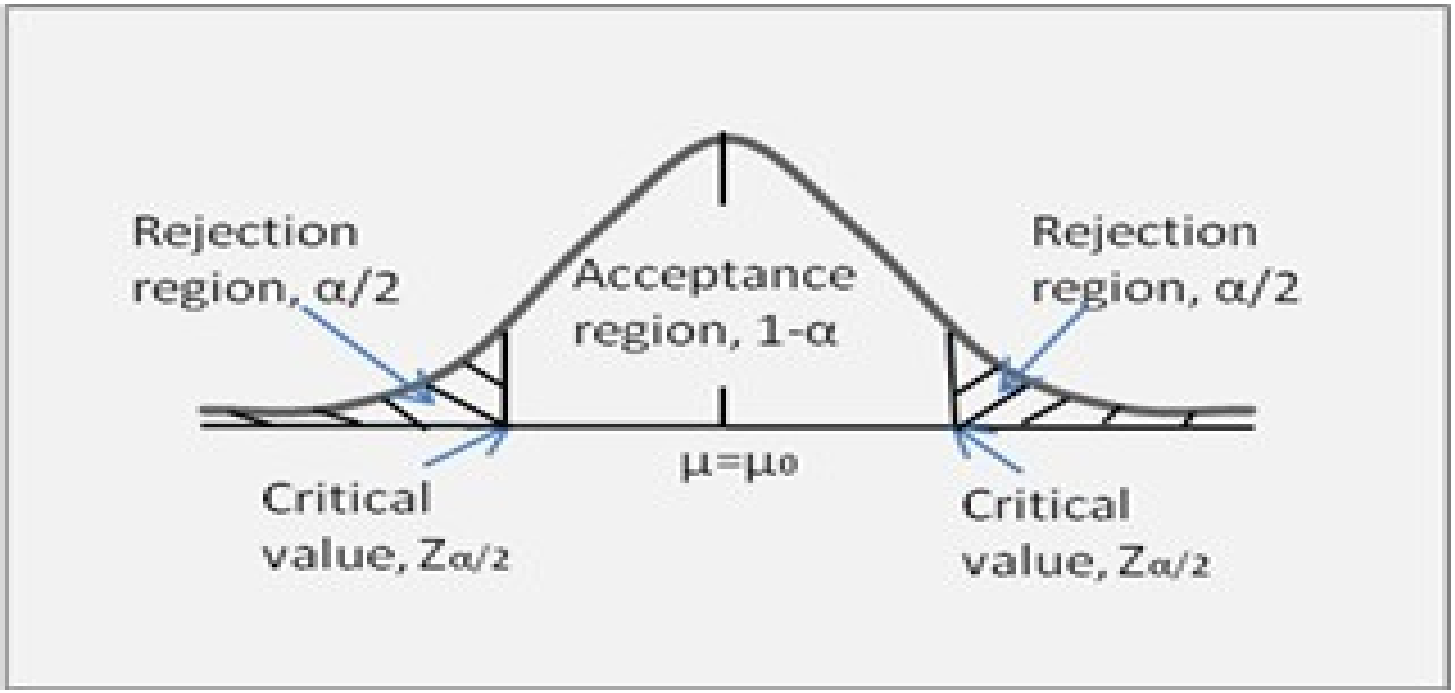
- ✓ If the variance σ^2 is known, the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and variance $\frac{\sigma^2}{n}$. i.e., $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and the test statistic is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
- ✓ If the variance σ^2 is unknown the test statistic is, $t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t\,(n-1)$.

Case 2: When the population is non normal.

We use the central limit theorem to approximate the distribution of the sample mean based on large sample ($n \geq 30$). Large sample size is a necessary condition to use the normal distribution. And hence the test statistic is

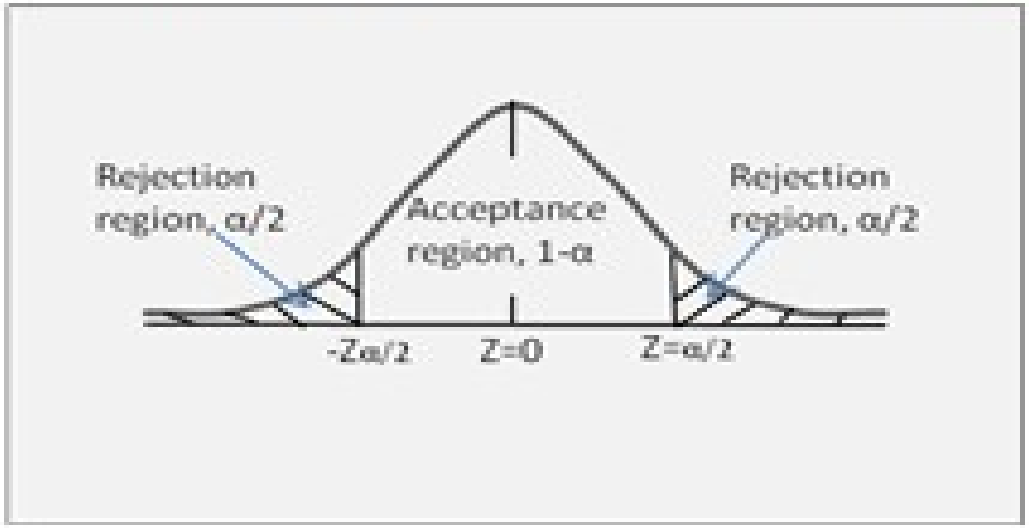
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$. If σ is unknown we can replace it by its sample estimate S.

- **Step 4: Establish a decision rule (critical or rejection region)**
- The cut-off point to reject or not reject depends on the level of significance **,** the type of test statistic chosen and the form of the alternative hypothesis. If the value of the test statistic falls in the rejection region, the null hypothesis is rejected, otherwise we do not reject (see the next fig).
- The value of the sample statistic that separates the regions of acceptance and rejection is called *critical value*. For a specified , we read the critical values from the Z or t tables, depending on the test statistic chosen.

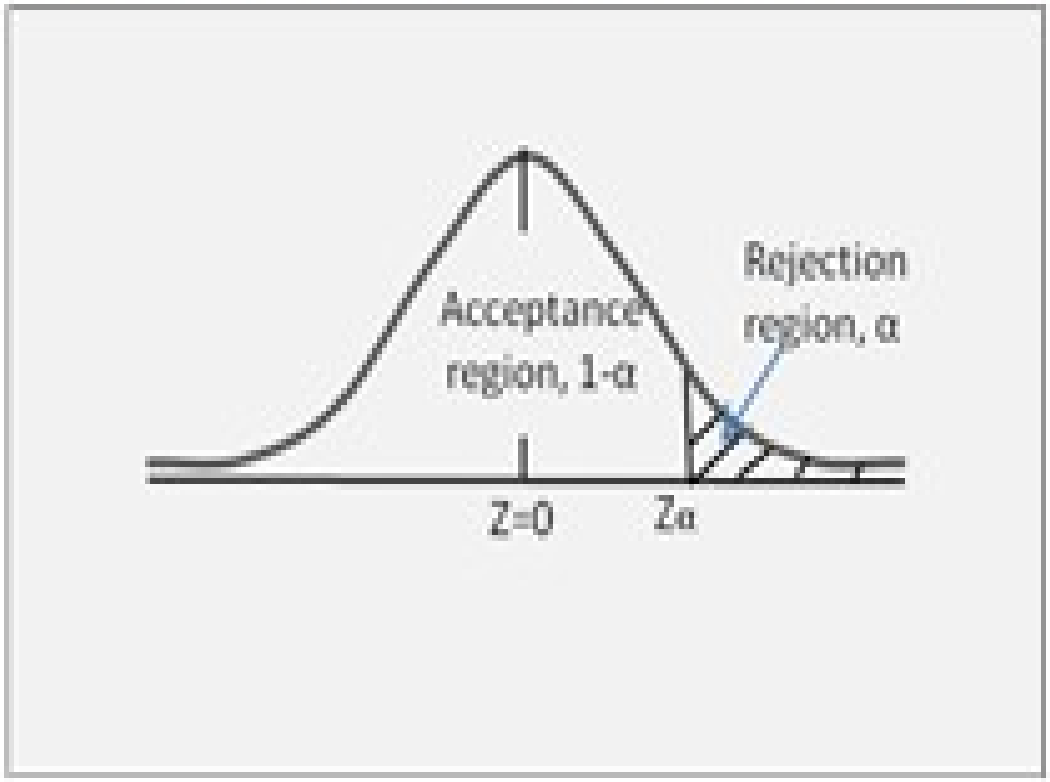

Rejection region, α/2
Acceptance region, 1-α
Rejection region, α/2
μ=μ0
Critical value, Zα/2
Critical value, Zα/2

- Based on the form of the alternative hypothesis and the test statistic we can make the following decisions:

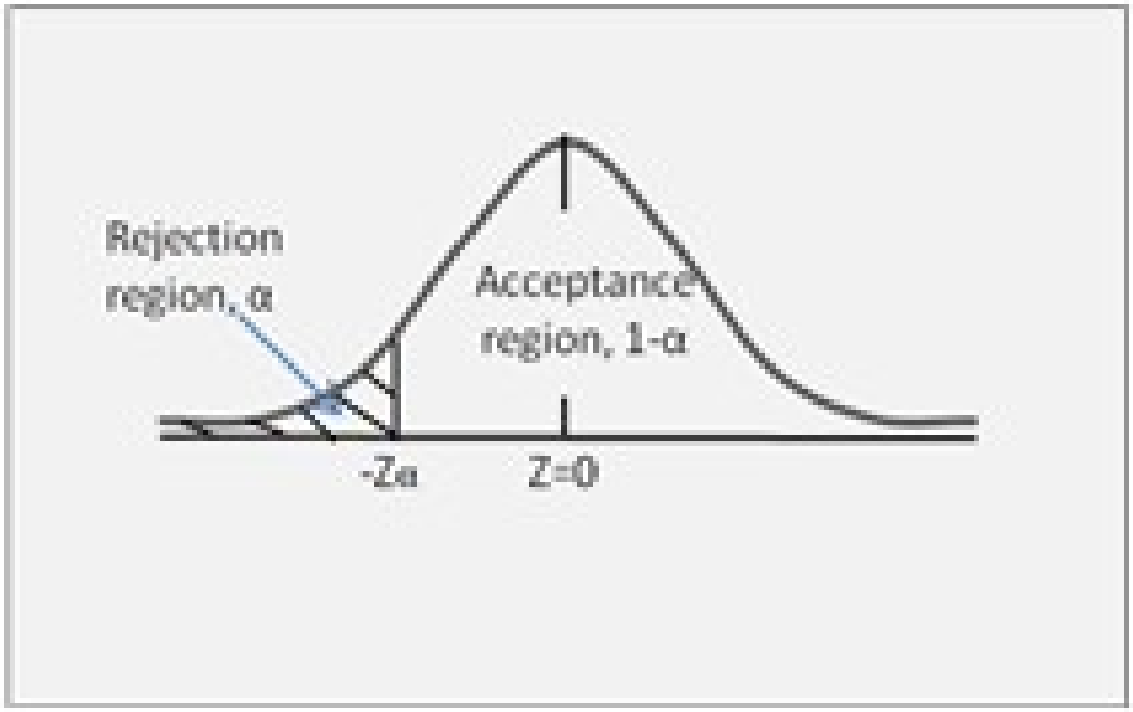
For $H_1 : \mu \neq \mu_0$ (two-tailed test) reject H_0 if $|Z| > Z_{\alpha/2}$.



ii. For $H_1 : \mu > \mu_0$ (right-tailed test) reject H_0 if $Z > Z_\alpha$.



iii. For $H_1 : \mu < \mu_0$ (left-tailed test) reject H_0 if $Z < -Z_\alpha$.



We can summarize the decsion rules as follows:

Decision	Alternative hypotheses		
	$H_1 : \mu \neq \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu < \mu_0$
Reject $H_0 : \mu = \mu_0$ if	$\lvert Z \rvert > Z_{\alpha/2}$	$Z > Z_{\alpha}$	$Z < -Z_{\alpha}$
Reject $H_0 : \mu = \mu_0$ if	$\lvert t \rvert > t_{\alpha/2}(n-1)$	$t > t_{\alpha}(n-1)$	$t < -t_{\alpha}(n-1)$

- **Step 5: Interpret the result.**
- **Errors in Hypothesis Testing**
- Ideally the hypothesis testing procedure should lead to the rejection of the null hypothesis when it is false and non rejection of when it is true.
- However, the correct decision is not always possible. Since the decision to reject or do not reject a hypothesis is based on sample data, there is a possibility of committing an incorrect decision or error.

- Hence, a decision-maker may commit one of the two types of errors while testing a null hypothesis. These errors

Decision	Null Hypothesis (H_0)	
	True	False
Reject H_0	Type I error (α)	Correct decision
Accept H_0	Correct decision	Type II error (β)

- Type I error is committed if we reject the null hypothesis when it is true. The probability of committing a type I error, denoted by is called the *level of significance*.
- The probability level of this error is decided by the decision-maker before the hypothesis test is performed. Type II error is committed if we do not reject the null hypothesis when it is false.
- The probability of committing a type II error is denoted by (Greek letter beta). As type one error increases type two errors will decrease (they are inversely proportional).
- Hence we cannot reduce both errors simultaneously.
- As the sample size increases both errors will decrease.

- **Example 8.3:** The life expectancy of people in the year 1999 in a country is expected to be 50 years. A survey was conducted in eleven regions of the country and the data obtained, in years, are given below:
- Life expectancy (years): 54.2, 50.4, 44.2, 49.7, 55.4, 47.0, 58.2, 56.6, 61.9, 57.5, and 53.4.
- Do the data confirm the expected view? (Assuming normal population) Use 5% level of significance

Solution: Let μ be the life expectancy of people in the year 1999 in a country.

1. $H_0 : \mu = 50$ (The life expectancy of people in the year 1999 in a country is 50 years)

   $H_1 : \mu \neq 50$ (The life expectancy of people in the year 1999 in a country is different from 50 years)
2. Level of significance, $\alpha = 0.05$.
3. Since σ is unknown and the population is normal, the t-test statistic is appropriate.

   Given: $n = 11$; $\mu_0 = 50$ and we need to compute $\bar{X}$ and s.

$$\bar{X} = \frac{\sum_{i=1}^{11} x_i}{n} = \frac{54.2 + 50.4 + \ldots.. + 57.5 + 53.4}{11} = \frac{598.5}{11} = 54.41$$

$$\sum_{i=1}^{11} x_i^2 = 54.2^2 + 50.4^2 + \ldots.. + 57.5^2 + 53.4^2 = 3279991$$

$$S^2 = \frac{1}{n-1}\left[\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right] = \frac{1}{10}\left[3279991 - \frac{(598.5)^2}{11}\right]$$

$$= \frac{1}{10}(236.07) = 23.607$$

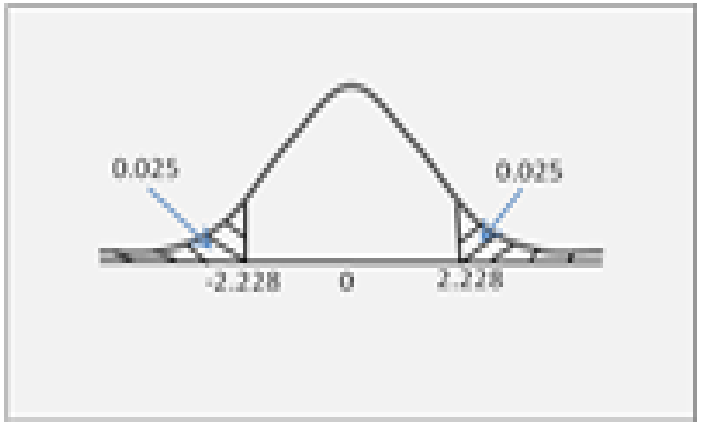
$$\Rightarrow S = \sqrt{23.607} = 4.859$$

Then, the t-test statistic is calculated as:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{54.41 - 50}{4.859/\sqrt{11}} = \frac{4.41}{1.465} = 3.01$$

4. For $\alpha = 0.05$ and two-tailed test, the critical (table) value is:

$$t_{\alpha/2}(n-1) = t_{0.05/2}(11-1) = t_{0.025}(10) = 2.228$$



Since $|t| = 3.01 > t_{\alpha/2}(n-1) = 2.228 \Rightarrow$ reject the null hypothesis H_0. That is, the calculated t value lies in the rejection region (the shaded region).

5. **Conclusion:** The data do not confirm the expected view. That is, the life expectancy is different from 50 years at 5% level of significance.

- **Example 8.4:** Suppose that we want to test the hypothesis with a significance level of .05 that the climate has changed since industrialization. Suppose that the mean temperature throughout history is 50 degrees.
- During the last 40 years, the mean temperature has been 51 degrees and the population standard deviation is 2 degrees. What can we conclude?

Solution:

Let μ be the mean temperature.

1. $H_0 : \mu = 50$ (There is no change in temperature since industrialization)

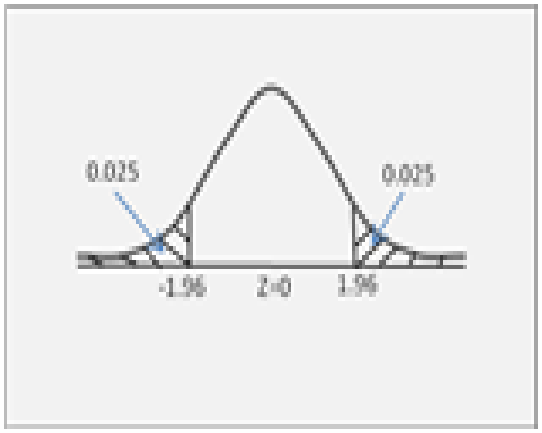
   $H_1 : \mu \neq 50$ (There is change in temperature since industrialization)
2. Level of significance, $\alpha = 0.05$.
3. Since $n = 40$ is large, the Z-test statistic is appropriate.

   Given: $n = 40$; $\delta = 2$; $\bar{X} = 51$; $\mu_0 = 50$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{51 - 50}{2/\sqrt{40}} = \frac{1}{0.316} = 3.16$$

4. For $\alpha = 0.05$ and two-tailed test, the critical (table) value is:

$$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$$



Since $|Z| = 3.16 > Z_{\alpha/2} = Z_{0.025} = 1.96 \Rightarrow$ reject the null hypothesis H_0. That is, the calculated Z value lies in the rejection region (the shaded region).

5. **Conclusion:** There has been a change in temperature since industrialization, at 5% level of significance.

- **Example 8.5:** A study was conducted to describe the menopausal status, menopausal symptoms, energy expenditure and aerobic fitness of healthy midwife women and to determine relationship among these factors. Among the variables measured was maximum oxygen uptake (Vo_{2max}). The mean Vo_{2max} score for a sample of 242 women was 33.3 with a standard deviation of 12.14. On the basis of these data, can we conclude that the mean score for a population of such women is greater than 30? Use 5% level of significance.

Solution:

Let μ be the mean Vo_{2max} score for a population of healthy midwife women.

1. $H_0 : \mu = 30$ (The mean score for a population of healthy midwife women is 30)

   $H_1 : \mu > 30$ (The mean score for a population of healthy midwife women is greater than 30).

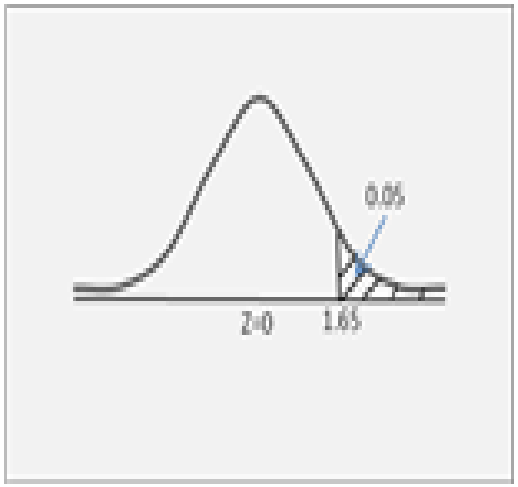
2. Level of significance, $\alpha = 0.05$.
3. Since $n = 242$ is large, the Z-test statistic is appropriate.

   Given: $n = 242$; $S = 12.14$; $\bar{X} = 33.3$; $\mu_0 = 30$

$$Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{33.3 - 30}{12.14/\sqrt{242}} = \frac{3.3}{0.7804} = 4.23$$

4. For $\alpha = 0.05$ and right-tailed test, the critical (table) value is:

$Z_\alpha = Z_{0.05} = 1.65$



Since $Z = 4.23 > Z_\alpha = 1.65 \Rightarrow$ reject the null hypothesis H_0. That is, the calculated Z value lies in the rejection region (the shaded region).

5. **Conclusion:** The mean Vo_{2max} score for the sampled population of healthy midwife women is greater than 30 at 5% level of significance.