

# **PROBABILITY AND STATISTICS**

## **[IT2110]**

*By SLIIT Mathematics Unit*  
*Faculty of Humanities and Sciences*

# COURSE

- Course Delivery
  - ✓ *2 Hour Lecture*
  - ✓ *1 Hour Tutorial*
  - ✓ *2 Hour Lab Session (R)*
- Course Web Enrollment Key
  - ✓ IT2110
- Module Outline
  - ✓ Refer course web

# COURSE EVALUATION

- *Continuous Assessments – 50 Marks*
  - *Mid Term Examination – 30%*
  - *Online R Assignment (Practical) – 20%*
- *Final Examination – 50 Marks*

# CONTACT INFORMATION

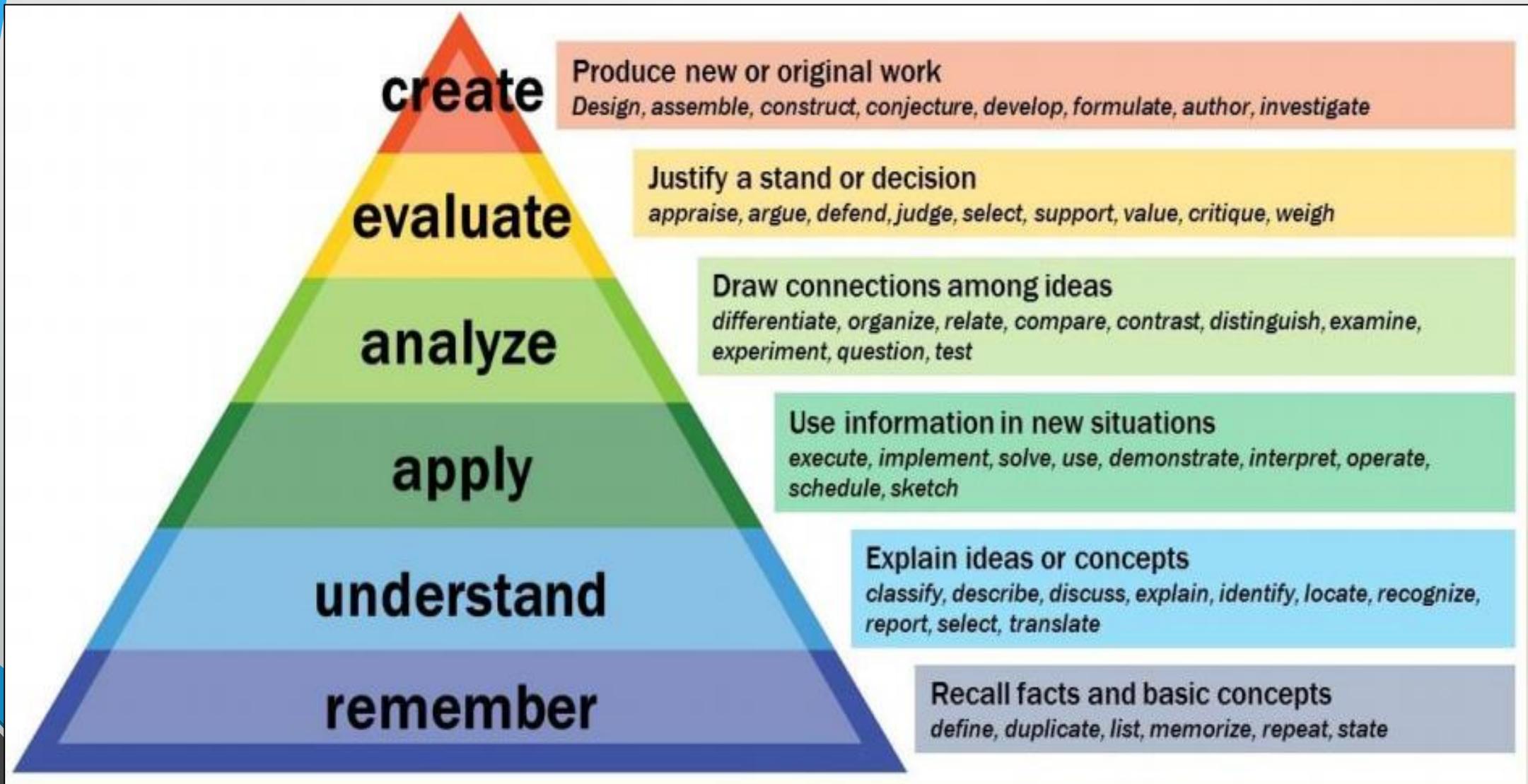
- Ms. Lakmali Guruge ([malika.l@sliit.lk](mailto:malika.l@sliit.lk))
- Ms. Nilushi Dias ([nilushi.d@sliit.lk](mailto:nilushi.d@sliit.lk))
- Ms. Shanika Ferdinandis ([shanikaferdinandis@yahoo.com](mailto:shanikaferdinandis@yahoo.com))

# COURSE CONTENT

- Introduction to Statistics
- Sampling Methods
- Descriptive Statistics
- Probability
- Random variables & Probability Distributions
  - Discrete Probability Distributions
    - ✓ *Binomial Distribution*
    - ✓ *Poisson Distribution*
  - Continuous Probability Distributions
    - ✓ *Exponential Distribution*
    - ✓ *Normal Distribution*

- Sampling Distributions
  - ✓ *Sampling Distribution of the Means*
  - ✓ *Central Limit Theorem*
- Statistical Inference
  - ✓ *Parameter Estimation*
  - ✓ *Hypothesis Testing*
- Chi squared Tests
  - ✓ *Goodness of Fit Test*
  - ✓ *Independence Test*
- Regression Analysis
  - ✓ *Scatter Plot*
  - ✓ *Correlation*
  - ✓ *Simple Linear Regression Model*

# BLOOM'S TAXONOMY

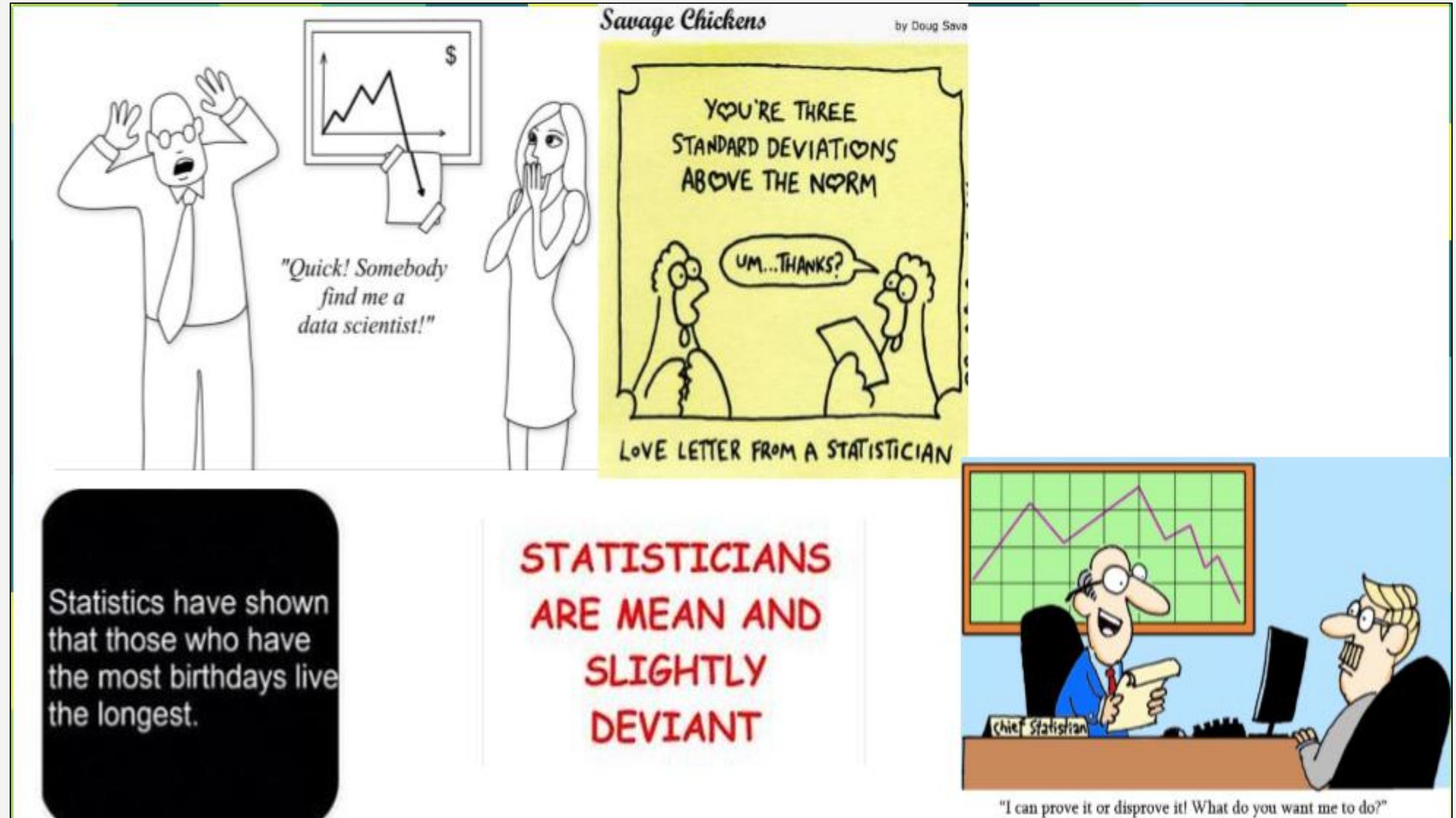


# REFERENCES

- Probability and Statistics for Engineers and Scientists (9<sup>th</sup> Edition)
  - Authors: Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers & K. Ye
  - Publisher: Prentice Hall
- Schaum's Outlines of Statistics (6<sup>th</sup> Edition)
  - Author: M. R. Spiegel & L.J. Stephens
  - Publisher: McGraw-Hill Education

# **1. INTRODUCTION TO STATISTICS [IT2110]**

*BY DEPARTMENT OF MATHEMATICS AND STATISTICS  
FACULTY OF HUMANITIES AND SCIENCES*



# APPLICATIONS OF STATISTICS

- Statistics can be applied in any field. Following are some examples for such applications.
  - ***Engineering and Sciences***
  - ***Medical Sciences***
  - ***Education***
  - ***Business Analytics***
  - ***Social Sciences***
  - ***Machine Learning***
  - ***Quality Control***
  - ***Actuarial Sciences etc.***

# DEFINITION - STATISTICS

- Statistics is the study of the ***collection, analysis, interpretation, presentation, and organization*** of data.  
-*Oxford: A Dictionary of Statistics-*
- Statistics are numbers that ***summarize*** raw facts and figures in some ***meaningful*** way.  
- *Head First Statistics* -

- Statistics is the ***study*** of ***uncertainty***.



- We need statistics to identify the ***variability*** in data.



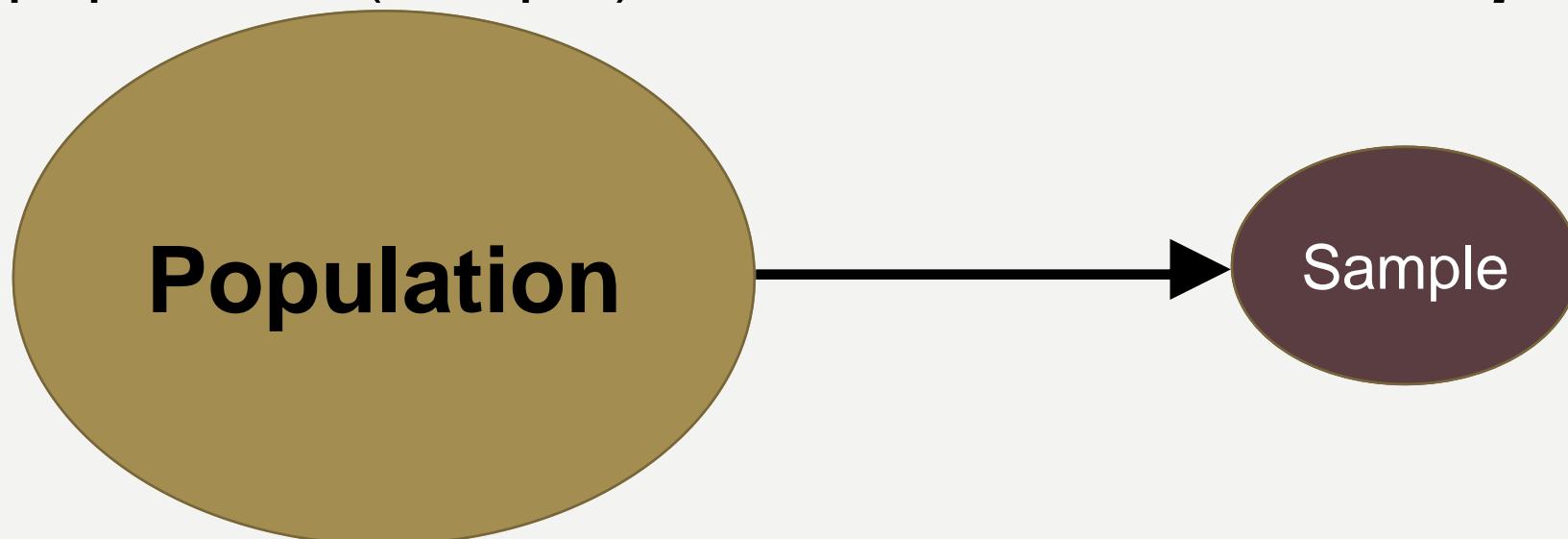
# TERMINOLOGY

# DEFINITION - POPULATION

- This is a collection of set of individuals or objects where researcher is interested about drawing inferences.
- Population can be finite or infinite.
- If you are going to collect data from all the individuals in the population, then it is known as a **census survey**.

# DEFINITION - SAMPLE

- A ***sub set*** of the population.
- If you are going to collect data from a part of the population (sample), then it is known as a ***sample survey***.



# DEFINITION - VARIABLE

- Variable is a ***characteristic/property*** of each individual in the population or a sample.
- ***Examples***      :-      Age, Gender, Temperature etc.
- We usually use capital letters to denote variables.

# DEFINITION – DATA (SINGULAR)

- The value of the variable associated with one element of a population or sample.
- This value may be a number, a word, or a symbol.

# DEFINITION - PARAMETER

- Parameter is a *summary characteristic* about the individuals in the *population*.
- Parameter is always related with the population.
- **Examples** :- Population mean ( $\mu$ ), Population variance ( $\sigma^2$ ), Population proportion (P) etc.

# DEFINITION - STATISTIC

- Statistic is a ***summary characteristic*** about the individuals in the ***sample***.
- Statistic is always related with the sample.
- ***Examples*** :- Sample mean ( $\bar{x}$ ), sample variance ( $s^2$ ), sample proportion ( $p$ ) etc.

# DEFINITION - EXPERIMENT

- An experiment is a planned activity whose results yield a set of data.

# EXAMPLES...

- A researcher is interested in finding the average weight of a first year student in SLIIT. He collected data from all first year students in computing faculty.

***Population*** : All the first year students in SLIIT

***Sample*** : All first year students in computing faculty

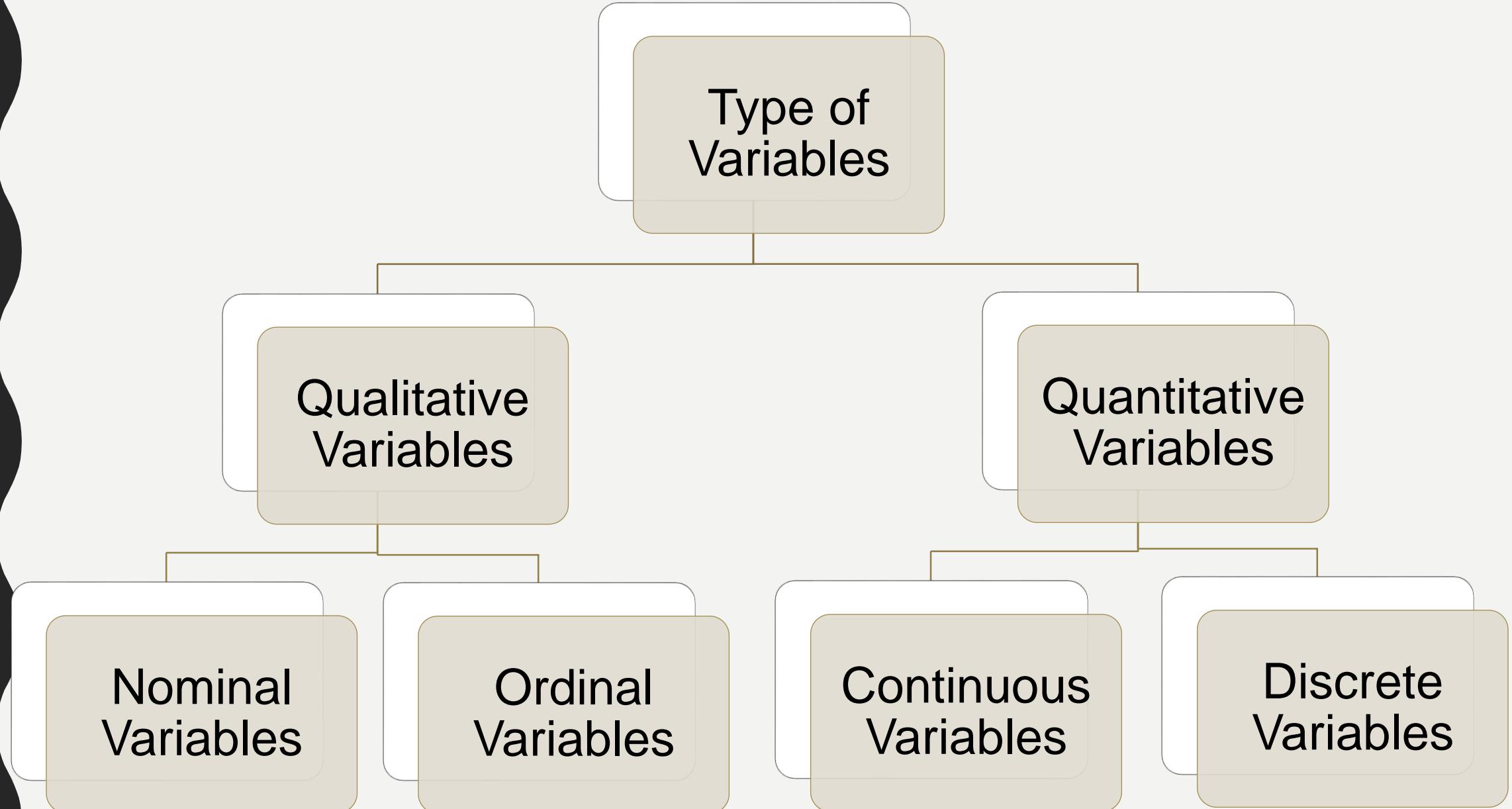
***Variable*** : Weight

***Summary Characteristic*** : Average Weight → **Statistic**

***Type of survey*** : Sample survey



# **TYPES OF VARIABLES**



- **Qualitative/Attribute/Categorical Variables :**

A variable that categorizes and describes an element. E.g. : Hair color, Gender, Marital status, Highest education qualification.

- **Quantitative/Numerical Variables :**

A variable that quantifies an element. E.g. : Marks for statistics, Age, Temperature, Time taken to travel to SLIIT from home.

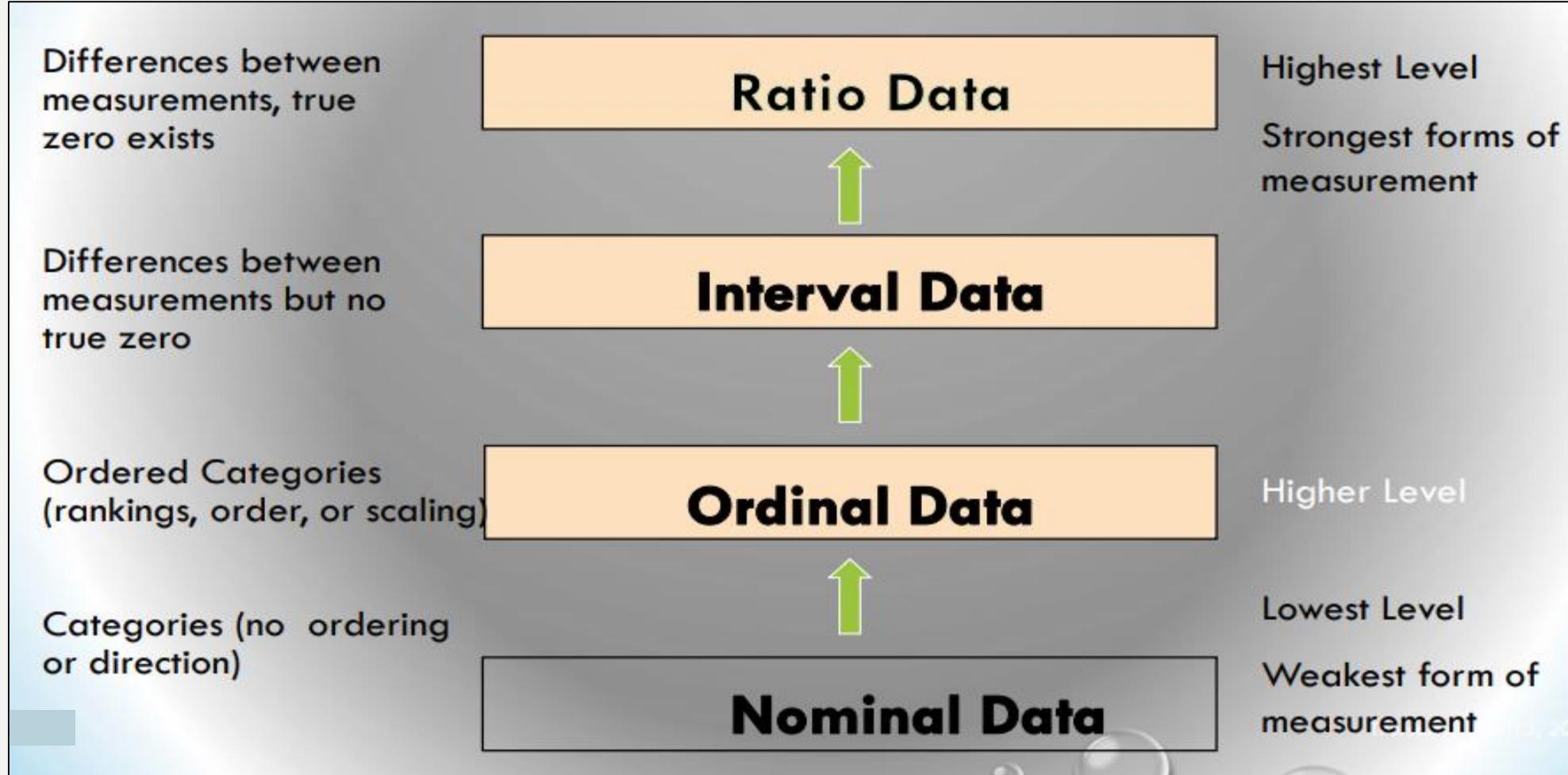
# Qualitative Variables

- **Nominal Variables:** Categories are not naturally ordered.  
E.g. : Gender, Hair Color, Marital Status
- **Ordinal Variables:** Categories are naturally ordered.  
E.g. : Satisfaction Rating, Pain Severity,  
Highest education qualification.

# Quantitative Variables

- **Discrete variables:** Distance between two values exists.  
E.g. : Age in years, No of children in a family, Number of accidents in a junction within an hour
- **Continuous variables:** This will contain any value within a given range.  
E.g.: Temperature, Heart beat of a patient etc.

# MEASUREMENT SCALES

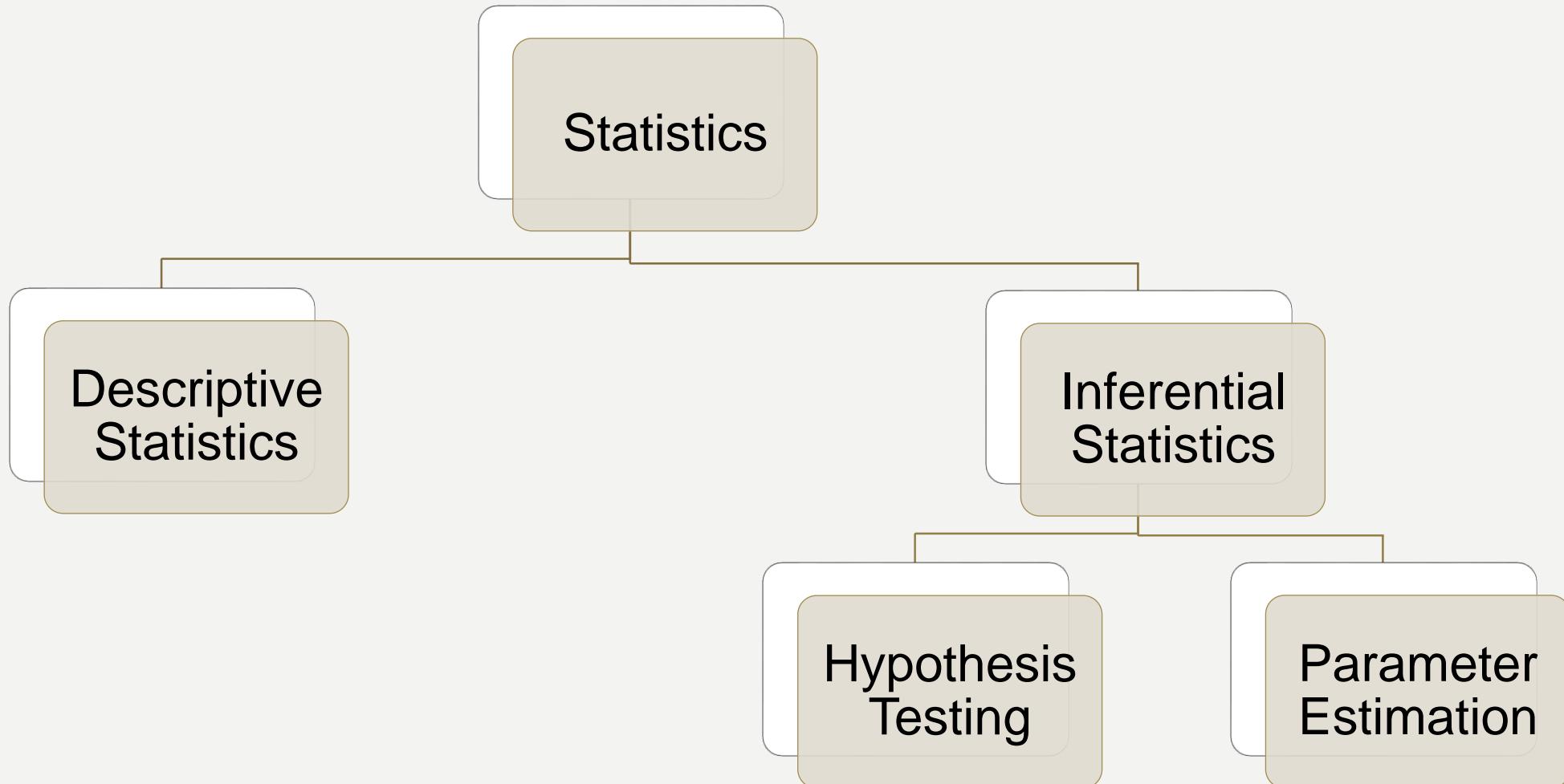


# INTERVAL SCALE VS. RATIO SCALE

Interval Scale	Ratio Scale
In this scale, variables can be added and subtracted. But ratio and multiplication is not possible.	Including ratio and multiplication of variables it has all characteristics of an interval scale.
Can calculate mean, median and mode.	Can calculate mean, median and mode.
Difference between variables can be evaluated.	Difference between variables can be evaluated.
Does not have a true zero point. (Eg:- Temperature can be below zero degree Celsius and negative)	True zero point exist. (Eg:- Weight can not be zero or below zero)
Examples:- Temperature in Celsius, Temperature in Farenhite, pH Value	Examples:- Height, Weight, Temperature in Kelvin, No of sales, Income of an individual, Heart Rate



# AREAS OF STATISTICS



- ***Descriptive Statistics*** :- This is also known as ***preliminary analysis / explanatory analysis***. This will give you a rough idea about the ***behavior of data***. It describes how each of the variables behave. There are ***two methods*** that you can use under descriptive statistics. They are,

- ***Graphical Methods***
- ***Numerical Methods***

- ***Inferential Statistics***:- This is ***drawing conclusions*** about population parameters by using sample statistics. Under this there are two main areas namely, ***parameter estimation*** and ***hypothesis testing***.

- You can analyze data by using some statistical package.
- It allows you to analyze data easily and precisely.
- Most commonly used statistical packages are **SPSS, SAS, Minitab, R, E-views and Matlab.**
- In this module we will discuss how to analyze data by using **R**.



# INTRODUCTION

## *R SOFTWARE*

- Independent and Open source.
- Initially developed at University of Auckland in the mid1990s.
- Distributed under the GNU open software license.
- Developed by the user community.
- Available On: Linux, Windows and Mac (OS X).
- Latest Version: 4.2.0 (Vigorous Calisthenics) - Released 22/04/2022.
- Terminal and GUI available.
- IDEs for R: R Studio, Rattle.

# THANK YOU!

Any questions?

# **2. Sampling Methods**

## **[IT2110]**

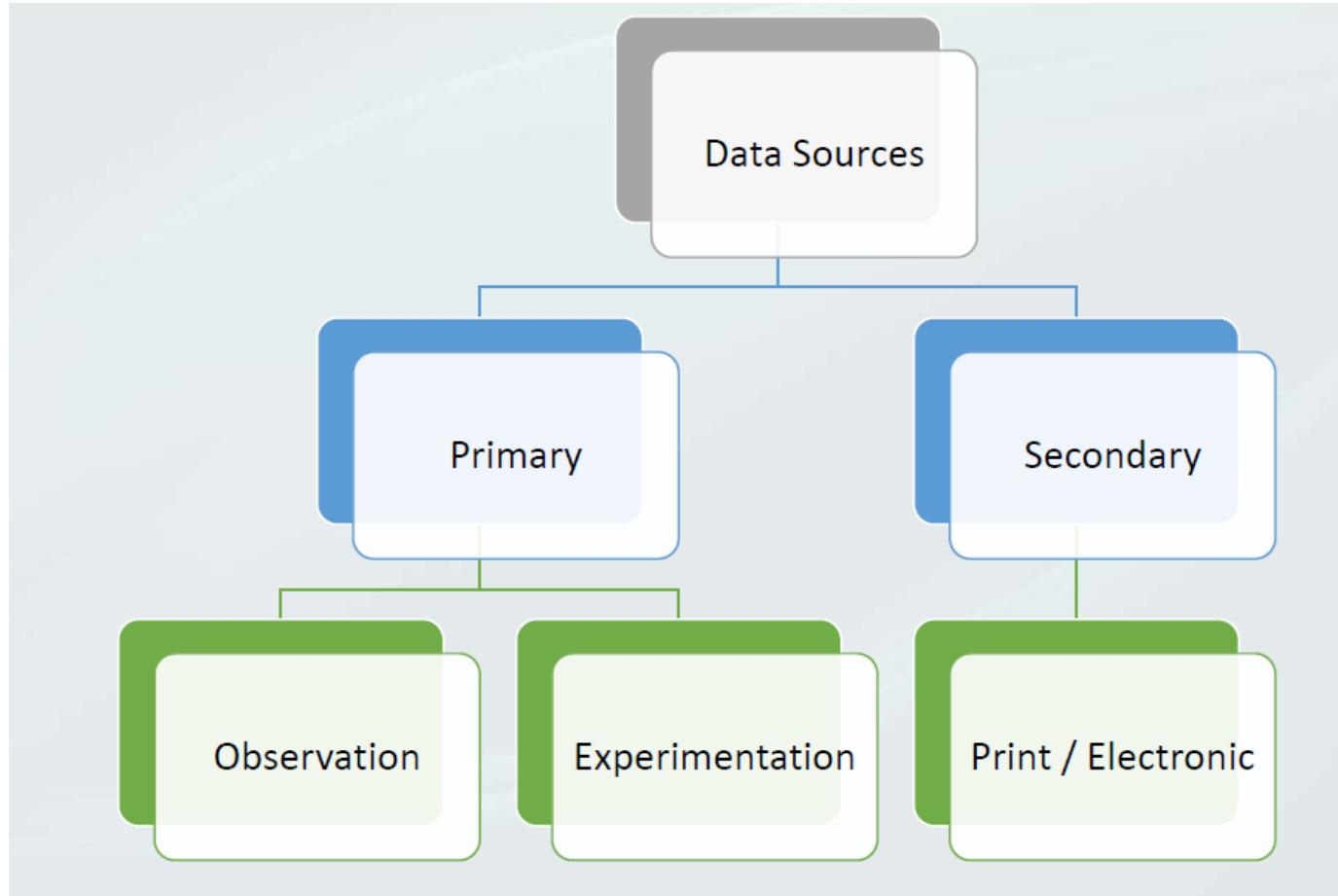
---

**By Department of Mathematics and Statistics**  
Faculty of Humanities and Sciences

# 1.

## Introduction to Sampling

# Data Sources



# Introduction to Sampling

---

- ❑ Data should be collected before describing.
- ❑ If a sampling survey is done, should plan how to select the sample.
- ❑ **Two types of sampling:**
  - ❑ Probability sampling.
  - ❑ Non-probability sampling.
- ❑ Why should a proper sample be selected?

# Reasons for Drawing a Sample

---

- ❑ Less time consuming than a census
- ❑ Less costly to administer than a census
- ❑ Less cumbersome and more practical to administer than a census of the targeted population

# 2.

## Non-Probability Sampling

# Non-Probability Sampling

---

- ❑ Uses a subjective (i.e., non-random) method.
- ❑ Does not require a sampling/survey frame.
- ❑ Fast, easy and inexpensive.
- ❑ Sample might not be representative of the population.
- ❑ Chance of each element being selected (i.e., probability), cannot be calculated.
- ❑ Can be applied to studies that are used as:
  - ❑ an idea generating tool.
  - ❑ a preliminary step.
  - ❑ a follow-up step.

3.

# Probability Sampling

# Probability Sampling

---

- Based on the principle of randomization or chance.
- More complex, time consuming and usually more costly.
- More reliable.
- Requires a sampling/survey frame.
- Can use computers or other methods to select elements randomly (e.g.: random number tables).

# Sampling Frame

---

- ❑ The list of elements from which a sample may be drawn.
- ❑ Also known as: ***working population***.
- ❑ Examples: Telephone directory, List of voters

# Probability Sampling (cont'd.)

---

- Commonly used probability sampling methods:
  - Simple Random Sampling (SRS).
  - Systematic Sampling (SYS).
  - Probability-Proportional-to-Size (PPS) Sampling.
  - Cluster Sampling.
  - Stratified Sampling (STR).
  - Multi-Stage Sampling.
  - Multi-Phase Sampling.
  - Replicated Sampling.

# Simple Random Sampling (SRS)

---

- Starting point for all probability sampling designs.
- Each unit in the sample has the same inclusion probability ( $n$  – Sample Size,  $N$  – Population Size).
- Sampling may be done with or without replacement (SRSWR or SRSWOR).
- Generally, SRSWOR yields more precise results and is operationally more convenient.

# SRS (cont'd.)

---

## □ Advantages of SRS

- Simplest sampling technique.
- Requires no additional (auxiliary) information on the frame in order to draw the sample.
- Needs no technical development.

## □ Disadvantages of SRS

- Makes no use of auxiliary information even if such information exists on the survey frame.
- Can be expensive.
- It is possible to draw a 'bad' SRS sample.

# Systematic Sampling (SYS)

---

- Units are selected from the population at regular intervals.
- A sampling interval ( $k = N/n$ ) and a random start are required.
- Every  $k^{\text{th}}$  individual thereafter.

## □ **Advantages**

- Can result in a sample that is better dispersed than SRS.
- Simpler than SRS.

## □ **Disadvantages**

- Can result in a ‘bad’ sample if the sampling interval matches some periodicity in the population.

# Stratified Sampling

---

- ❑ Divide population into two or more subgroups (called strata) according to some common characteristic.
  
- ❑ A simple random sample is selected from each subgroup, with sample sizes proportional to strata sizes.

# Cluster Sampling

---

- ❑ Population is divided into several “clusters,” each representative of the population.
- ❑ A simple random sample of clusters is selected.
- ❑ All items in the selected clusters can be used, or items can be chosen from a cluster using another probability sampling technique.

# Multistage Sampling

---

- With multistage sampling, we select a sample by using combinations of different sampling methods.
  
- **Example:-** In Stage 1, we might use cluster sampling to choose clusters from a population. Then, in Stage 2, we might use simple random sampling to select a subset of elements from each chosen cluster for the final sample.

# PROBLEM

---

An auto analyst is conducting a satisfaction survey, sampling from a list of 10,000 new car buyers. The list includes 2,500 Ford buyers, 2,500 GM buyers, 2,500 Honda buyers, and 2,500 Toyota buyers. The analyst selects a sample of 400 car buyers, by randomly sampling 100 buyers of each brand.

What type of sampling method have used in this scenario?

# Thank You

## Questins?

---

# **3. DESCRIPTIVE STATISTICS**

## **[IT2110]**

*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences*

**“** *This will give you an idea  
about the behavior of data.*

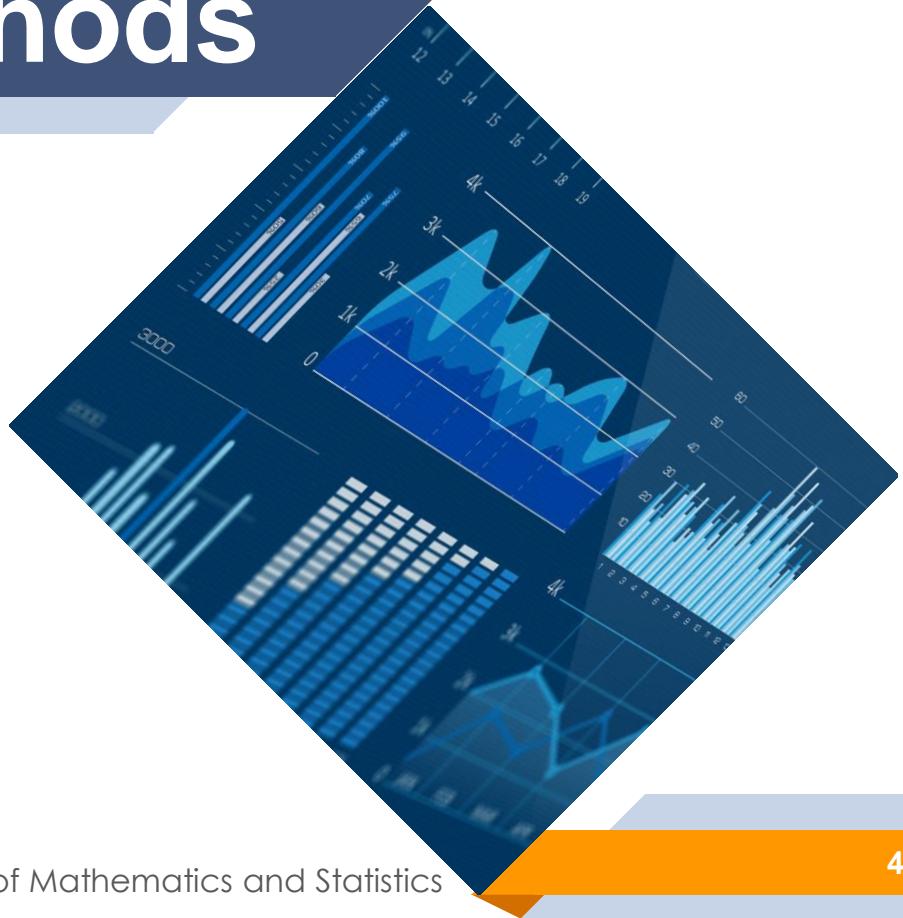


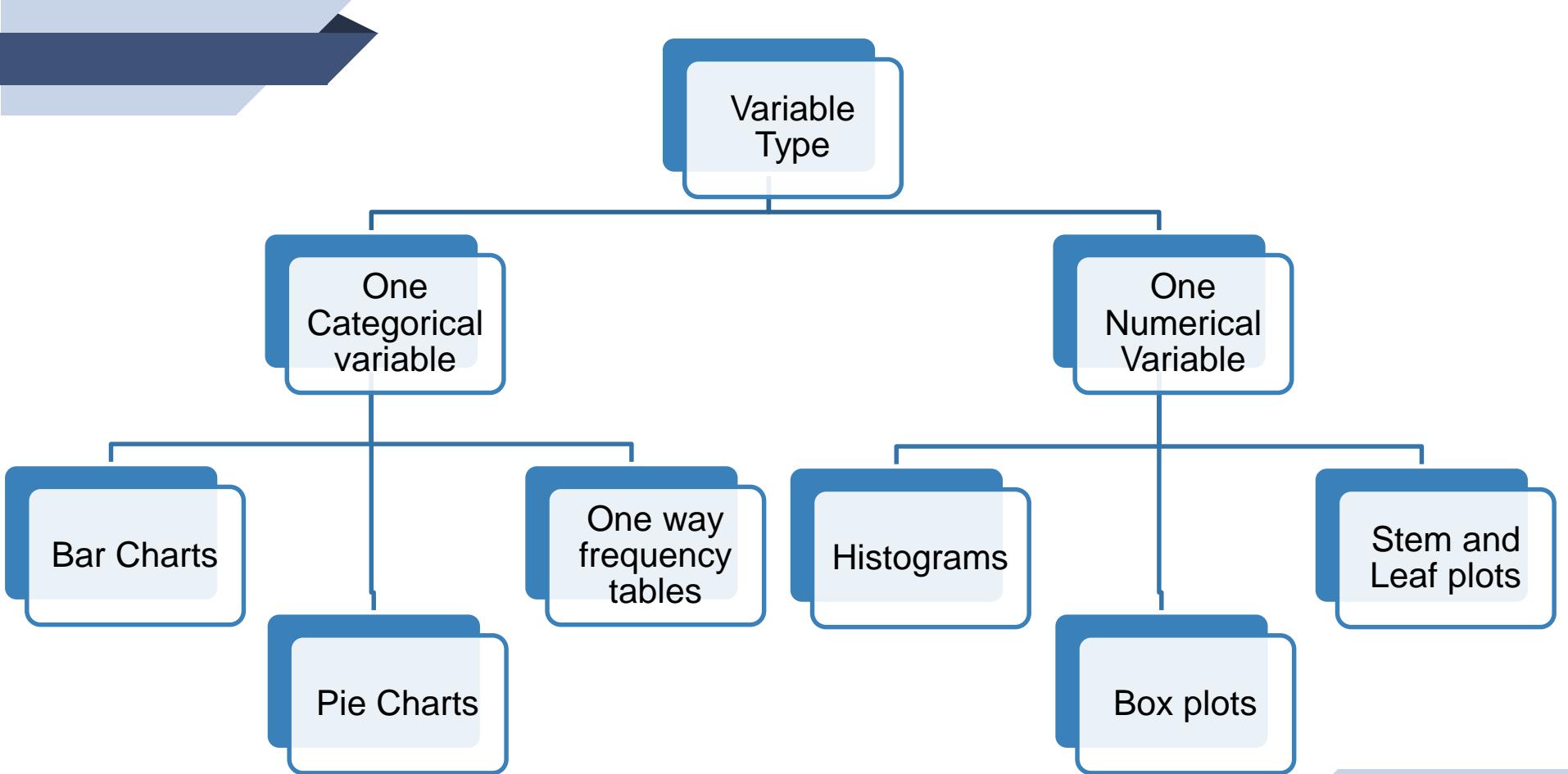
- This is also known as *preliminary analysis*.
- It describes how each of the variables in your analysis behave.
- There are *two methods* that you can use under exploratory analysis. They are,
  - ▶ *Graphical Methods &*
  - ▶ *Numerical Methods*
- Each method depends on the type of the data available



# Graphical Methods

- You can use graphical methods to analyze both categorical and numerical variables.
- Type of graph you use depends on the type of the data available







# One-way Frequency Tables

## Categorical Variable

Gender	Frequency
Male	48
Female	52

## Numerical Variable

Marks	Frequency
0-20	12
21-40	8
41-60	42
61-80	56
81-100	10

# Bar Charts

- In bar charts, each bar will represent each category level. These bars can be drawn in vertically or horizontally.
- Frequency, cumulative frequency or percentages can be used for the y axis while x axis will represent the categorical variable.
- Length of the bar will proportional to the value it represent.

# Bar Charts Cont...

- There are several type of bar charts. For example,

***Simple Bar Charts***

***Component bar charts / Stacked bar charts***

***Percentage component bar charts***

***Multiple bar charts / Clustered bar charts***

# Pie Charts

- Pie charts are used to analyze **one categorical** variable.
- In pie charts, area of each sector will proportional to the value of category it represent.
- This is **appropriate**, when there are **few number of categories** for the variable or when **value** of each **category** is **varying** widely.

# Histograms

- First, divide the given data set into suitable number of classes (intervals/categories) which have the same width.
- Classes with their frequencies (counts) is called a frequency distribution.
- Frequency, relative frequency or percentages can be used for the y axis while x axis will represent the classes of the variable.
- In histograms, each bar will represent each class and length of the bar will proportional to the frequency of respective class.
- In histograms, ***bars are drawn adjacent with each other*** (No gaps between two bars).

# *Example:-*

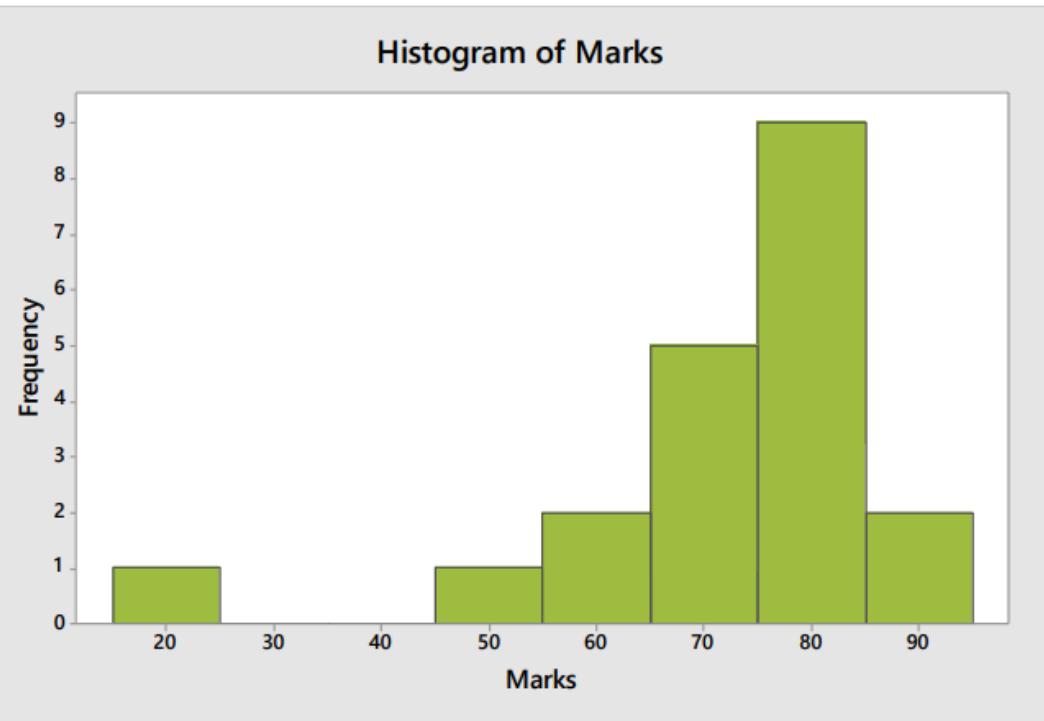
78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

- Range = Max – Min =  $91 - 16 = 75$
- Divide the range into required number of classes to find class width  
(Eg:- 8):-

$$75 / 8 = 9.375 \approx 10$$

- Classes can be selected by fixing the class width also

# Example:-



Class	Frequency
14.5 – 24.5	1
24.5 – 34.5	0
34.5 – 44.5	0
44.5 – 54.5	1
54.5 – 64.5	2
64.5 – 74.5	5
74.5 - 84.5	9
84.5 – 94.5	2

# Box Plots

- To draw a box plot, it is need to identify the ***five number summary & outliers*** for the variable.
- Five Number Summary:
  - ***Minimum***
  - ***Maximum***
  - ***Q1***
  - ***Q2 (Median)***
  - ***Q3***

# Outliers

- Before drawing the box-plot we should identify the potential outliers.
- A limit should be defined for the accepted range of values.

$$\text{Upper Bound} = Q_3 + 1.5 * IQR$$

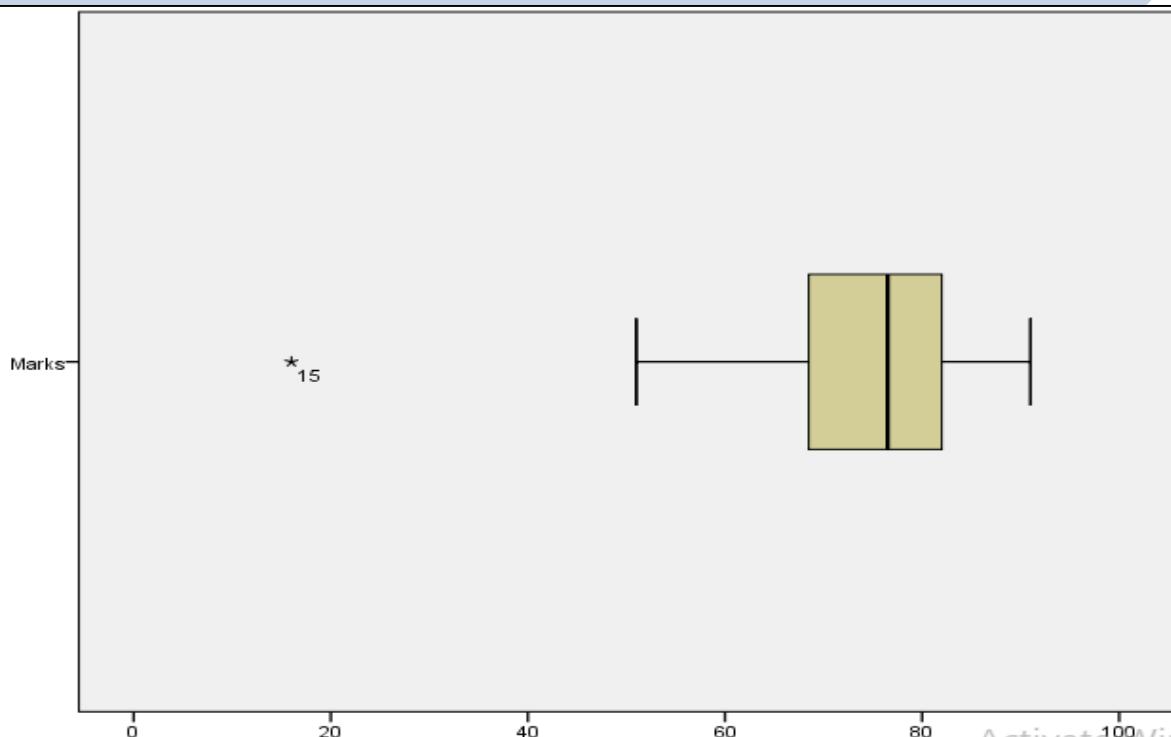
$$\text{Lower Bound} = Q_1 - 1.5 * IQR$$

- Values outside the range are considered as outliers and marked with asterisks (\*).

# Outliers

- $Q_1$ , Median,  $Q_3$  are marked as a box.
- Minimum & maximum values ***which are not outliers***, will be end point for whiskers of the box plot.

# Example:-



78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

# Stem-and-Leaf Plots

- These plots are useful when the data set is very small.
- Before drawing this plot, it is need to arrange the data in ascending order.
- Then, each data value will split into two parts known as ***stem*** and ***leaf***.
- The “leaf” is usually the last digit of the number.
- The other digits to the left of the “leaf” form the “stem”.
- These plots are not much use in preliminary analysis.

# *Example:-*

**Stem   Leaves**

1        6

2

3

4

5        1 5

6        4 6

7        1 1 4 4 5 8 8 9

8        0 2 2 3 4 8

9        1

**Key: 1 | 6 → 16**

78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

# Describing Two Variables

## Two categorical variables:

- ▶ *Two-way frequency table*
- ▶ *Clustered bar chart*
- ▶ *Stacked bar chart*

## One categorical & one numerical variable:

- ▶ *Parallel box plots*
- ▶ *Comparison of location measurements for each category.*

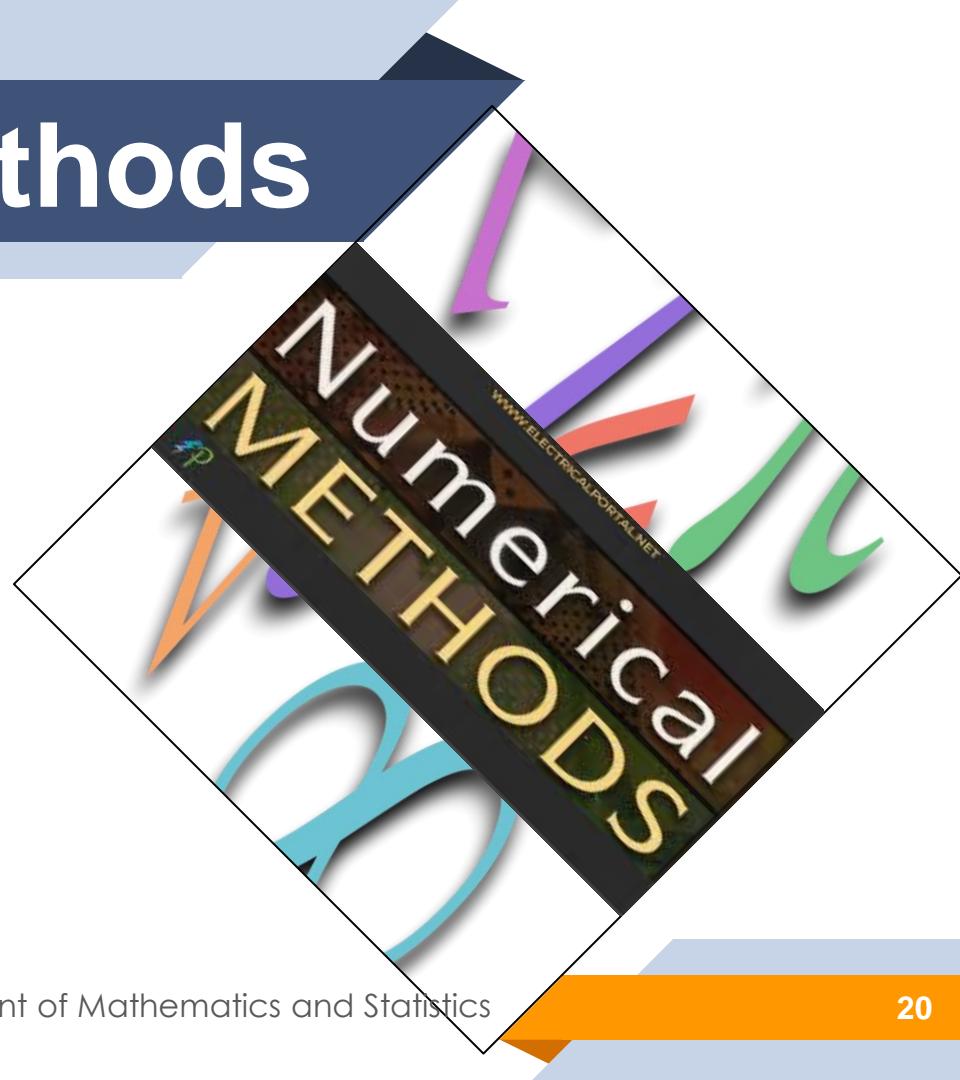
## Two numerical variables:

- ▶ *Scatter plot*



# Numerical Methods

- Numerical methods are applied only for numerical variables.
- These methods summarize the variable into a single value.



# Numerical Methods Cont...

- This has measurements under four main sections. They are,

***Measures of central tendency***

***Measures of dispersion***

***Measures of skewness***

***Measures of kurtosis***

# Measures of Central Tendency

- This gives an idea about the ***location*** of the data as a whole.
- Following three measurements can be used for this.
  - ***Mean***
  - ***Median***
  - ***Mode***
- Other location measurements :
  - ***Percentiles / Deciles / Quartiles***

# Mean

## ■ Different types of means

- ▷ Arithmetic mean
- ▷ Geometric mean
- ▷ Harmonic mean

## ■ Only the arithmetic mean is discussed (referred to as the mean).

- Mean of a population ( $\mu$ ), with  $N$  elements ( $x_1, x_2, \dots, x_N$ ),

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

- Mean of a sample ( $\bar{x}$ ), with  $n$  elements ( $x_1, x_2, \dots, x_n$ ),

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- If not specified, consider the data are coming from a sample.

# Examples:-

## Example 1.2 (revisited):

Find the median “marks for FCS” of each student at SLIIT Metro.

78	74	82	66	91	71	64	88	55	80
51	74	82	75	16	78	84	79	71	83

## Example 1.3 (revisited):

A load of aluminum sheets were purchased to construct a temporary shed. Twenty such sheets were examined for surface flaws. Find the median number of flaws in a sheet.

Number of flaws	Frequency
0	4
1	3
2	5
3	2
4	4
5	1
6	1

# Mode

- A value with the highest frequency in a data set.
- There can be multiple modes in a data set.
- If all the data values are different, the data set has no mode.

# Percentiles

- Divides the entire set of values into 100 equal sections.
- The values should be ordered in ascending order.
- Position of the  $k^{\text{th}}$  percentile. ( $n$  – number of observations)

$$\textbf{\textit{Position of }} P_k = \left( \frac{n+1}{100} \right) * k$$

- Find the value that corresponds to the found position from the ordered set of values.
- If the position is not an integer the following methods can be used.
  - ▷ *Nearest Rank method*
  - ▷ *Linear Interpolation*

# Deciles & Quartiles

- Deciles divides the entire set of values into 10 equal sections.
- Quartiles divides the entire set of values into 4 equal sections.
- Method is the same as what has used for percentiles.

# Measures of Dispersion

- This gives an idea about the ***dispersion / spread*** of the data as a whole.
- Following three measurements can be used for this.
  - ***Range (Max - Min)***
  - ***IQR (Q3 – Q1)***
  - ***Variance & Standard Deviation ( $\sqrt{\text{Variance}}$ )***
- Range is more suitable for small data sets.
- Range is highly sensitive for outliers while, IQR & variance are not sensitive for outliers.

# Variance & SD

- This is a measurement of ***dispersion/spread*** of the data. This describes how the data has dispersed around its mean.
- Not sensitive to outliers.(more robust for outliers).
- Variance of a population ( $\sigma^2$ ), with  $N$  elements ( $x_1, x_2, \dots, x_N$ ),

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

- Variance of a sample ( $s^2$ ), with  $n$  elements ( $x_1, x_2, \dots, x_n$ ),

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- If not specified, consider the data are coming from a sample.
- Standard deviation (SD) is the square-root of the variance.
  - ▷ Population SD –  $\sigma$
  - ▷ Sample SD -  $s$

# Measures of Skewness

- This gives an idea about the **asymmetry** of the data as a whole.
- If the data set is symmetric skewness will be zero.
- A negative skew means that the left tail is longer; the mass of the distribution is concentrated on the right.
- A positive skew means that the right tail is longer; the mass of the distribution is concentrated on the left.

# Measures of Kurtosis

- This gives an idea about whether the distribution is **peaked** or **flat**.

# Example

A sample of 25 plastic hinges was subjected to repeated stress cycles until failure. The number of cycles which each survived is given below.

72, 35, 63, 67, 87, 71, 64, 47, 60, 81, 39, 52, 57, 74, 43, 55, 37, 83, 48, 91, 53, 44, 94, 65, 75

- I. Find five number summary
- II. Find mode,  $p_{15}$ ,  $D_3$ , mean, variance & sd.
- III. Draw box plot & stem & leaf plot.
- IV. Comment on the distribution of data.



# THANKS!

Any questions?

# **4. PROBABILITY**

## **[IT2110]**

*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences*

# TERMINOLOGY

- ▶ **Experiment:** A process leading to a well-defined observations or outcomes that generates a set of data
- ▶ **Trial:** Each repetition, if the experiment can be repeated any number of times under identical conditions
- ▶ **Sample Space ( $S$ ):** The set containing all possible outcomes of an experiment
- ▶ **Finite sample space:** Sample space that contains a finite number of outcomes
- ▶ **Continuous Sample space:** Sample space that contains an interval of values

# EVENTS

- ▶ ***Event:*** A subset of the sample space. Usually denoted in upper case letters.
- ▶ ***Simple Event:*** An event that corresponds to a single possible outcome
- ▶ The null subset ( $\emptyset$ ) of  $S$  is called an impossible event.
- ▶ The event  $A \cup B$  consists of all outcomes that are in  $A$  or in  $B$  or in both.
- ▶ The event  $A \cap B$  consists of outcomes that are both in  $A$  and  $B$ .
- ▶ The event  $A^c$  (the complement of  $A$  in  $S$ ) consists of all outcomes not in  $A$ , but in  $S$ .

- ▶ **Mutually Exclusive Events:** Two events A and B are said to be mutually exclusive or disjoint if  $A \cap B = \emptyset$ . They cannot happen together.
- ▶ **Collectively exhaustive events:** One of the events must occur. The set of events covers the entire sample space.
- ▶ **Independent Events:** If the occurrence of one event not affect on the occurrence of other event then both events are said to be independent with each other.
- ▶ **Joint Events (Compound Events):** An event that corresponds to more than a single possible outcome is known as compound events.  
*Eg:- Getting an odd number by rolling a die*

# Example

1. A balanced/fair die (with all outcomes equally likely) is rolled. Let  $A$  be the event that an even number occurs.

***Experiment*** : Rolling a balanced die.

***Sample Space*** :  $S = \{1, 2, 3, 4, 5, 6\}$

***Event (A)*** :  $A = \{2, 4, 6\}$

***Type of the event***: Compound event

2. Consider a deck of cards. Let  $A$  - Aces,  $B$  - Black cards,  $C$  - Diamonds and  $D$  - Hearts. Find collectively exhaustive events and mutually exclusive events.

# PROBABILITY

- ▶ **Probability:** Measure of the chance that an uncertain event will occur.
- ▶ The notation for the statement “Probability of the event A” is denoted as  $Pr(A)$  or  $P(A)$ .
- ▶ The value for the probability is between 0 and 1.
- ▶ A probability of 1 means that we are 100% sure of the occurrence of an event.
- ▶ A probability of 0 means that we are 100% sure of the non-occurrence of an event.
- ▶ The probability of S is always 1 ( $Pr(S) = 1$ ).
- ▶ The probability of an impossible event is always 0 ( $Pr(\emptyset) = 0$ ).

## Classical Definition of Probability

If there are  $N$  equally likely outcomes, of which one must occur, and  $n$  of these are regarded as favourable to an event, then the probability of the event is given by  $\frac{n}{N}$ .

## Frequency (Empirical) Definition of Probability

The probability of an event is the proportion of times the event would occur in a long run of repeated experiments.

Probability of the Event = 
$$\frac{\text{Number of favourable outcomes observed}}{\text{Total number of outcomes observed}}$$

## Subjective Probability

An individual judgement or opinion about the probability of occurrence.

# Examples

1. A balanced/fair die (with all outcomes equally likely) is rolled. Let  $A$  be the event that an even number occurs. What is the probability of  $A$ ?

$$\Pr(A) = \frac{\text{Number of outcomes in } A}{\text{Total number of outcomes in } S} = \frac{3}{6} = 0.5$$

# Examples

2. Suppose we toss two coins. Assume that all the outcomes are equally likely (fair coins). Let A be the event that at least one of the coins shows up heads. Find  $P(A)$ .

$$Pr(A) = \frac{\text{Number of outcomes in } A}{\text{Total number of outcomes in } S} = \frac{3}{4} = 0.75$$

# Basic Properties

Consider two events A and B in S. Then,

- ▶  $\Pr(A^c) = 1 - \Pr(A)$
- ▶  $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B).$
- ▶ If  $A \cap B = \emptyset$  (A and B are mutually exclusive) then,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

- ▶ If  $A_1, A_2, \dots, A_k$  are mutually exclusive then,

$$\Pr(A_1 \cup A_2 \cup \dots \cup A_k) = \Pr(A_1) + \Pr(A_2) + \dots + \Pr(A_k)$$

- ▶ If A and B are independent then,

$$\Pr(A \cap B) = \Pr(A) * \Pr(B)$$

# Example

1. In a large university, the freshman profile for one year's fall admission says that 40% of the students were in the top 10% of their high school class, and that 65% are white, 25% of the students were white as well as were in the top 10% of their high school class. What is the probability that a freshman student selected randomly from this class either was in the top 10% of his or her high school class or is white?

# Joint Probability

- ▶ The probability of events A and B occurring together is defined as Joint probability of A and B.

The probability of a joint event, A and B [ $\Pr(A \cap B)$ ]:

$$\Pr(A \text{ and } B) = \frac{\text{Number of outcomes satisfying } A \text{ and } B}{\text{Total number of outcomes in } S}$$

# Examples

1. Find the probability that you will get a Black-Ace from a playing deck of cards, if a card is drawn at random.
  
2. Find the probability that you will get a Red-Jack from a playing deck of cards, if a card is drawn at random.

# Marginal Probability

- ▶ The probability of a single event occurring ( $\text{Pr}(A)$ ), without the interference of another event (not conditioned on another event) is known as marginal probability.
- ▶ This can be thought of as an unconditional probability

# Examples

1. Find the probability that you will get a King from a playing deck of cards, if a card is drawn at random.
  
2. Find the probability that you will get a Black card from a playing deck of cards, if a card is drawn at random.

**Note:** *Contingency Tables* and *Tree Diagrams* can be used to visualize events and make calculations easier.

# Example

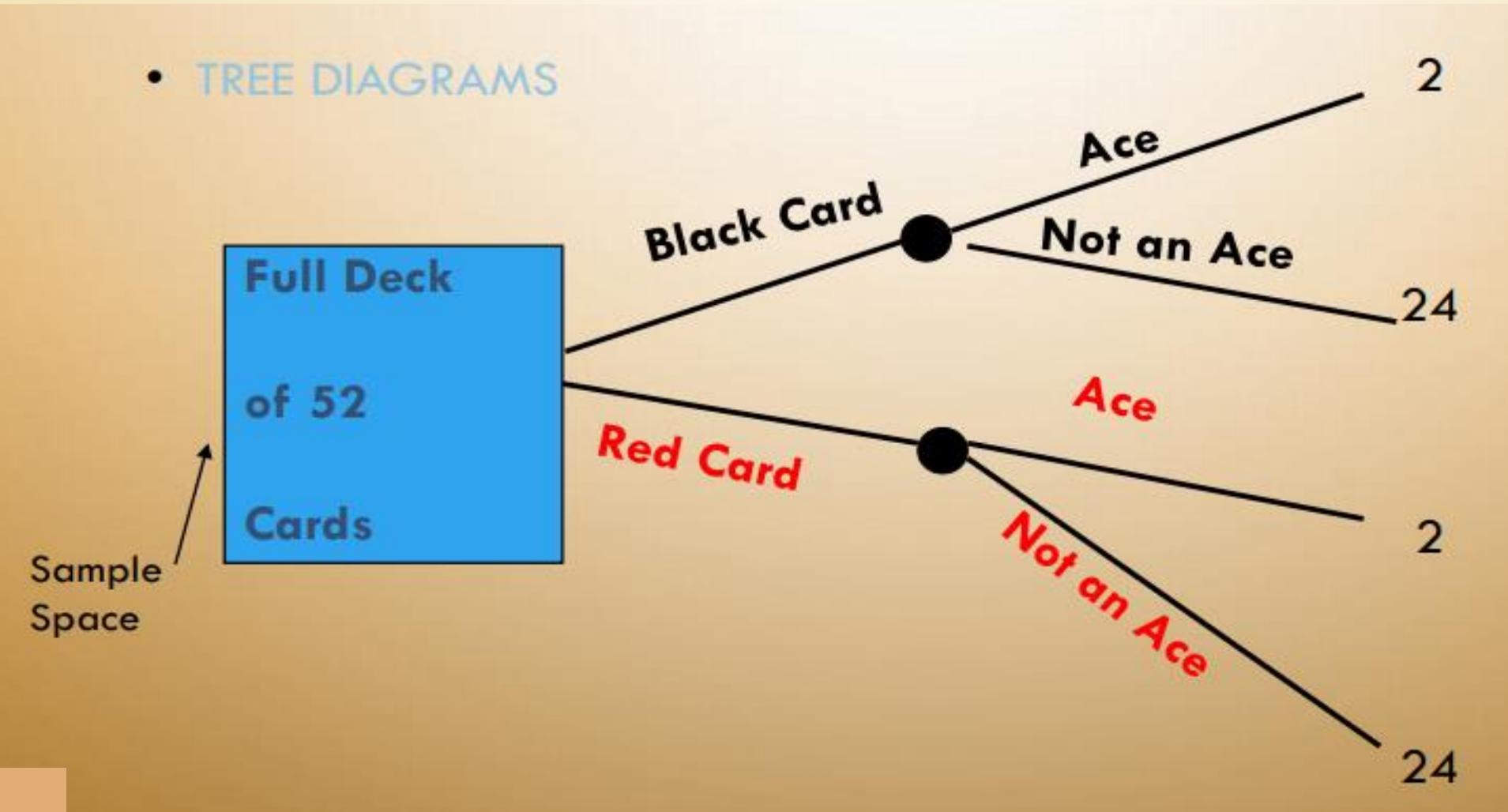
Type	Color		Total
	Red	Black	
Ace	2	2 [2/52]	4
Non-Ace	24 [24/52]	24	48 [48/52]
Total	26	26 [26/52]	52

*Joint Probabilities*

*Marginal Probabilities*

# Example

- TREE DIAGRAMS



# Conditional Probability

- ▶ This is the probability of one event, given that another event has already occurred.
- ▶ The conditional probability of an event A, given that an event B has already occurred is denoted by  $\Pr(A | B)$ .

$$\Pr(A | B) = \frac{\Pr(A \cap B)}{\Pr(B)} ; \Pr(B) > 0$$

$$\Pr(B | A) = \frac{\Pr(A \cap B)}{\Pr(A)} ; \Pr(A) > 0$$

;Where  $\Pr(A \cap B)$  = Joint probability of A and B

$\Pr(A)$  = Marginal probability of A

$\Pr(B)$  = Marginal probability of B

# Example

1. Of the cars on a used car lot, 70% have air conditioning and 40% have a CD player. 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC ?
2. If two balanced dice are tossed, find the probability that the sum of the face values is 8, if the face value of the first one is 3.

# Properties of Conditional Probability

- ▶  $\Pr(A|B) = 1 - \Pr(A^c|B)$
- ▶  $\Pr(B \cup C|A) = \Pr(B|A) + \Pr(C|A) - \Pr(B \cap C|A)$
- ▶ Multiplication law:

$$\Pr(A \cap B) = \Pr(B) * \Pr(A|B) = \Pr(A) * \Pr(B|A)$$

- ▶ If A and B are independent then,

$$\Pr(A|B) = \Pr(A) \quad \text{or} \quad \Pr(B|A) = \Pr(B)$$

$$\Pr(A \cap B) = \Pr(A) * \Pr(B)$$

- ▶ For independent events  $A_1, A_2, \dots, A_k$ ,

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_k) = \Pr(A_1) * \Pr(A_2) * \dots * \Pr(A_k)$$

# THANK YOU!

Any questions?

# **IT2110 – Probability and Statistics**

*Lecture 4- Probability*

# Unit Goals



**After completing this chapter, you should be able to:**

- ▶ Explain basic probability concepts and definitions
- ▶ Apply common rules of probability
- ▶ Compute conditional probabilities
- ▶ Determine whether events are statistically independent

# Important Terms



- ▶ **Experiment** - A Process leading to a well-defined observation or outcome that generates a set of data

Examples : 1. Tossing a Coin  
              2. Rolling a Die

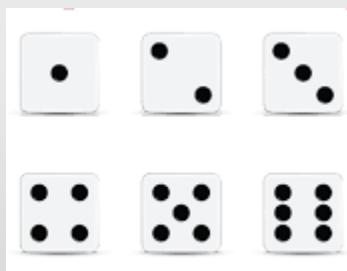
- ▶ **Sample Space** – the collection of all possible outcomes

Notations :  $S$ ,  $U$ ,  $\Omega$  ,  $\epsilon$

Ex.2 : Rolling a die



$$S = \{H, T\}$$



$$S = \{1, 2, 3, 4, 5, 6\}$$

# Important Terms (continued)



- ▶ **Finite sample space** – Sample space that contains a finite number of outcomes

Examples:

- Tossing a coin/Rolling a die
- Drawing from a bag of mixed-color balls
- Dealing with a regular 52-card deck
- ▶ **Continuous Sample space** – Sample space that contains an interval of values (Outcomes vary along continuous scale)

Examples:

- Height measurements
- Temperature readings

# Events



- ▶ Event – Each possible type of occurrence or outcome  
(Subsets of the sample space)
- ▶ Simple event
  - has a single outcome
- ▶ Empty set
  - { } or  $(\emptyset)$  is also called impossible event

## Example 1

Let's consider the example of tossing a fair coin. The 4 events are  $\{\}, \{H\}, \{T\}, \{H, T\}$  since all are subsets of the sample space =  $\{H, T\}$ .

$\{\}$  means empty set and denotes getting both H and T at once.

$\{H\}$  means getting only head, this possible.

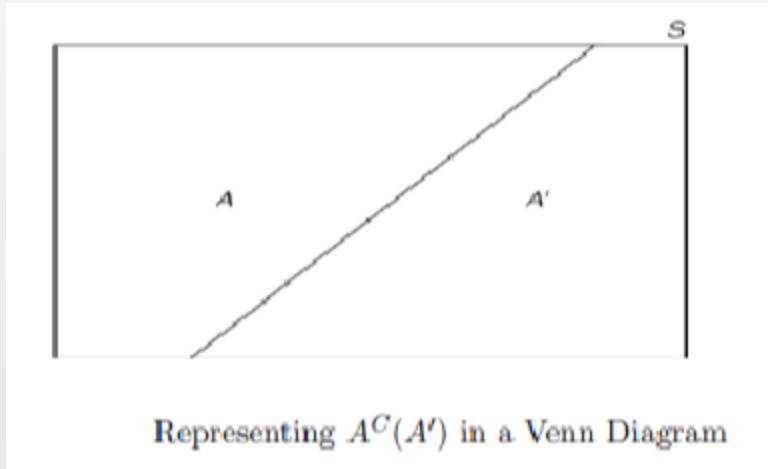
$\{T\}$  means getting only tail, this possible.

$\{H, T\}$  means getting head or tail, this possible and is called the certain event.

# Events (continued)



- ▶ Complement of an event A (denoted  $A'$ )
  - ▶ All outcomes that are not part of event A



- ▶ Joint events
  - ▶ Involves two or more characteristics simultaneously

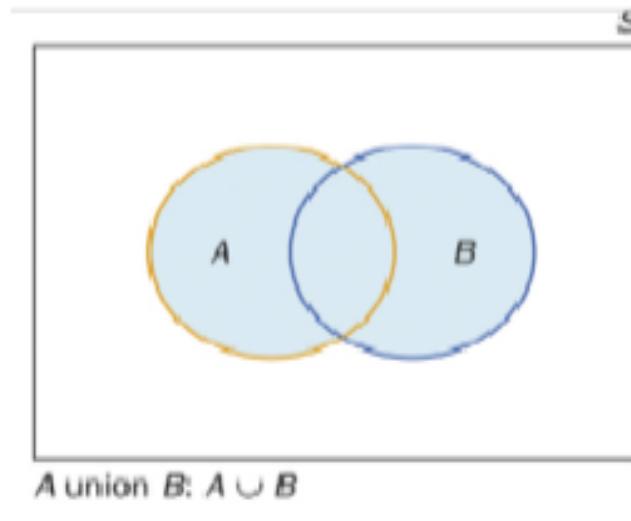
e.g. The long months that contain the letter r in their full name



# Events (continued)

## ► Union

The union of two events  $A$  and  $B$  denoted by  $A \cup B$  is the set of outcomes that belong either to  $A$  or  $B$  or to both. In words  $A \cup B$  means “ $A$  or  $B$ ”. Thus the event  $A \cup B$  occurs whenever either  $A$  or  $B$  (or both) occurs.



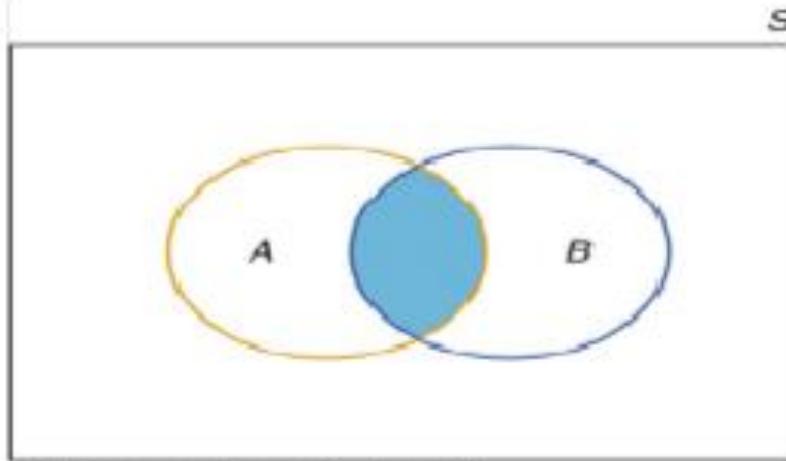
Representing  $A \cup B$  in a Venn Diagram

# Events (continued)



## ► Intersection

The intersection of two events  $A$  and  $B$ , denoted by  $A \cap B$  is the set of outcomes that belong both to  $A$  and  $B$ . In words  $A \cap B$  means “ $A$  and  $B$ ”. Thus the event  $A \cap B$  occurs whenever  $A$  and  $B$  (both) occur.



$A$  intersection  $B$ :  $A \cap B$

Representing  $A \cap B$  in a Venn Diagram

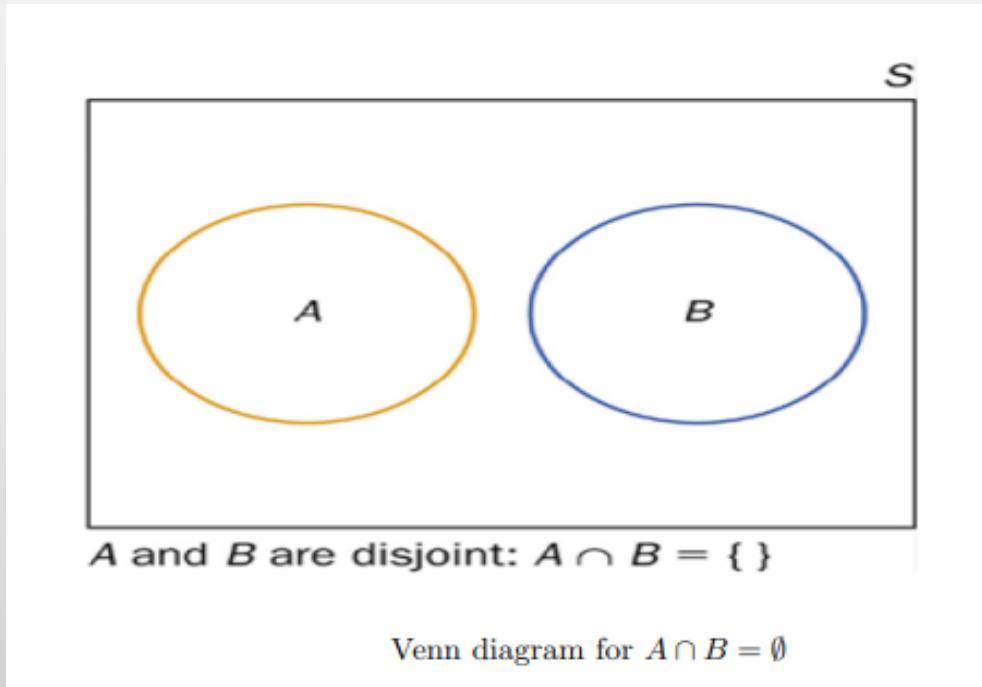


# Events (continued)

- ▶ Mutually exclusive/ Disjoint events

- ▶ Events that cannot occur together

e.g. The set of outcomes of a single coin toss, which can result in either head or tail, but not both.



# Events (continued)



## Independent Events

If occurring an event A does not affect to the occurring an another event B then A and B are independent.

e.g. Getting a head after tossing a coin and getting a 6 by rolling a die.

## Dependent events

If the outcome of the first event affects the outcome of the next event then these two events are dependent.

e.g. If a box containing different colored balls in each time a ball drawn to an out the chance of getting the next ball will change.

# Example 1



Let's consider rolling a fair die experiment. Then the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

Let A be the event that we observe an odd number. Then

$$A = \{1, 3, 5\}$$

Let B be the event that we observe a number greater than or equal to 3. Then

$$B = \{3, 4, 5, 6\}$$

Find  $A'$ ,  $B'$ ,  $A \cup B$  and  $A \cap B$ .

# Example 1 Answer



$A' = \{2, 4, 6\}$  and  $B' = \{1, 2\}$ .

$$A \cup B = \{1, 3, 4, 5, 6\}$$

$$A \cap B = \{3, 5\}$$

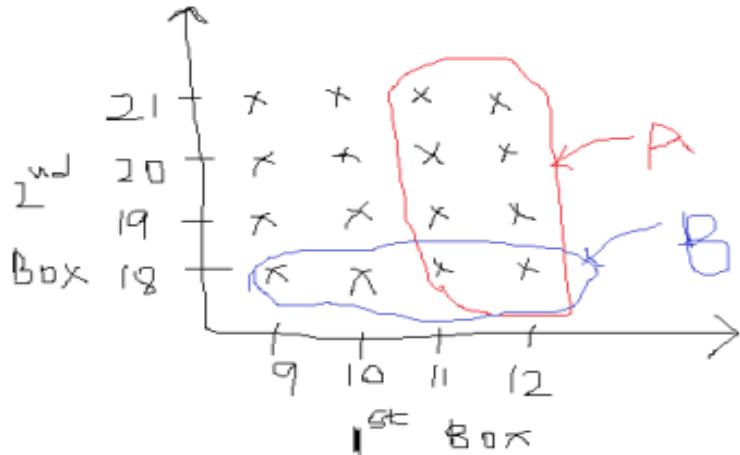
# Example 2



An electric engineer has on hand tow boxes of resistors, with four resistors in each box. The resistors in the first box are labeled  $10 \Omega$  (Ohmus), but in fact their resistances are 9, 10, 11, and  $12 \Omega$ . The resistances in the second box are labeled  $20 \Omega$ , but in fact their resistances are 18, 19, 20, and  $21 \Omega$ . The engineer chooses one resistor from each box and determine the resistance of each.

1. Write down the resulting sample space of this experiment.
2. Write down the following two events. Let A be the event that the first resistor has a resistance greater than  $10 \Omega$  and let B be the event that the second resistor has a resistance less than  $19 \Omega$ .
3. Find  $A \cup B$  and  $A \cap B$ .

# Example 2 Answers



$$S = \{(9, 18), (9, 19), (9, 20), (9, 21), (10, 18), (10, 19), (10, 20), (10, 21), (11, 18), (11, 19), (11, 20), (11, 21), (12, 18), (12, 19), (12, 20), (12, 21)\}$$

$$A = \{(11, 18), (11, 19), (11, 20), (11, 21), (12, 18), (12, 19), (12, 20), (12, 21)\}$$
$$B = \{(9, 18), (10, 18), (11, 18), (12, 18)\}$$

$$A \cup B = \{(9, 18), (10, 18), (11, 18), (11, 19), (11, 20), (11, 21), (12, 18), (12, 19), (12, 20), (12, 21)\}$$

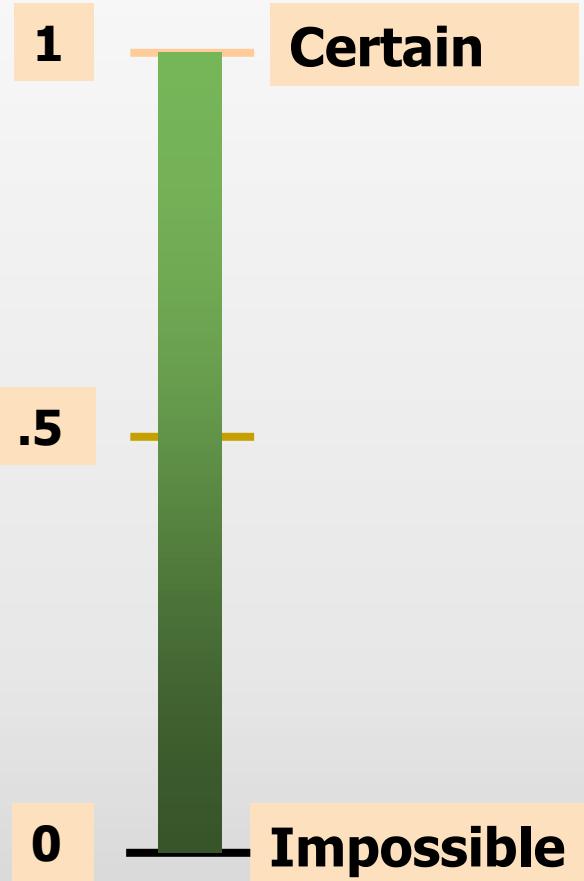
$$A \cap B = \{(11, 18), (12, 18)\}$$

# Probability



- ▶ Probability is the numerical measure of the likelihood(chance) that an uncertain event will occur
  - ▶ The probability of any event must be between 0 and 1, inclusively
- $$0 \leq P(A) \leq 1 \quad \text{For any event A}$$
- ▶ The sum of the probabilities of all mutually exclusive collectively exhaustive events is 1

$$P(S) = \sum_{i=1}^n P(i) = 1$$



# Equally Likely Outcomes



Whenever a sample space consists of  $N$  possible outcomes that are equally likely, the probability of each outcome is  $1/N$ .

## Example

In many experiments, such as tossing a fair coin or a balanced die, all the sample points have the same chance of occurring and are assigned equal probabilities.

Tossing a coin

$$S = \{H, T\}$$

$$P(H) = 1/2$$

$$P(T) = 1/2$$

Rolling a die

$$S = \{1, 2, 3, 4, 5, 6\}$$

Probability of getting each number =  $1/6$



# Probability of an Event

For a discrete sample space, the probability of an event A, denoted as  $P(A)$ , equals the sum of the probabilities of the outcomes in A.

## Example

A random experiment can result in one of the outcomes a, b, c, d with probabilities 0.1, 0.3, 0.5, and 0.1, respectively. Let A denote the event a, b, B the event b, c, d, and C the event d. Then find,

- |            |                   |                   |
|------------|-------------------|-------------------|
| 1. $P(A)$  | 6. $P(C')$        | 11. $P(A \cup C)$ |
| 2. $P(B)$  | 7. $P(A \cap B)$  | 12. $P(B \cup C)$ |
| 3. $P(C)$  | 8. $P(A \cap C)$  |                   |
| 4. $P(A')$ | 9. $P(B \cap C)$  |                   |
| 5. $P(B')$ | 10. $P(A \cup B)$ |                   |



# Probability of an Event (Continued)

## Example – Answer

$$A = \{a, b\} \quad B = \{b, c, d\} \quad C = \{d\}$$

$$A' = \{c, d\} \quad B' = \{a\} \quad C' = \{a, b, c\}$$

$$1. P(A) = 0.1 + 0.3 = 0.4$$

$$4. P(A') = 0.5 + 0.1 = 0.6$$

$$2. P(B) = 0.3 + 0.5 + 0.1 = 0.9$$

$$5. P(B') = 0.1$$

$$3. P(C) = 0.1$$

$$6. P(C') = 0.1 + 0.3 + 0.5 = 0.9$$

$$A \cap B = \{b\} \quad A \cap C = \{\} \quad B \cap C = \{d\}$$

$$7. P(A \cap B) = 0.3 \quad 8. P(A \cap C) = 0 \quad 9. P(B \cap C) = 0.1$$

$$A \cup B = \{a, b, c, d\} \quad A \cup C = \{a, b, d\} \quad B \cup C = \{b, c, d\}$$

$$10. P(A \cup B) = 0.1 + 0.3 + 0.5 + 0.1 = 1$$

$$11. P(A \cup C) = 0.1 + 0.3 + 0.1 = 0.5$$

$$12. P(B \cup C) = 0.3 + 0.5 + 0.1 = 0.9$$



# Classical Definition of Probability

probability of occurrence =  $\frac{\text{number of favorable outcomes observed}}{\text{total number of outcomes observed}}$

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

, where A is a subset of S

## Example 1

Probability of getting an even number when rolling a die

Let's consider event A= getting an even number

$$S = \{1, 2, 3, 4, 5, 6\} \quad A = \{2, 4, 6\}$$

$$\text{So, } P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2} = 0.5$$



## Classical Definition of Probability(Continued)

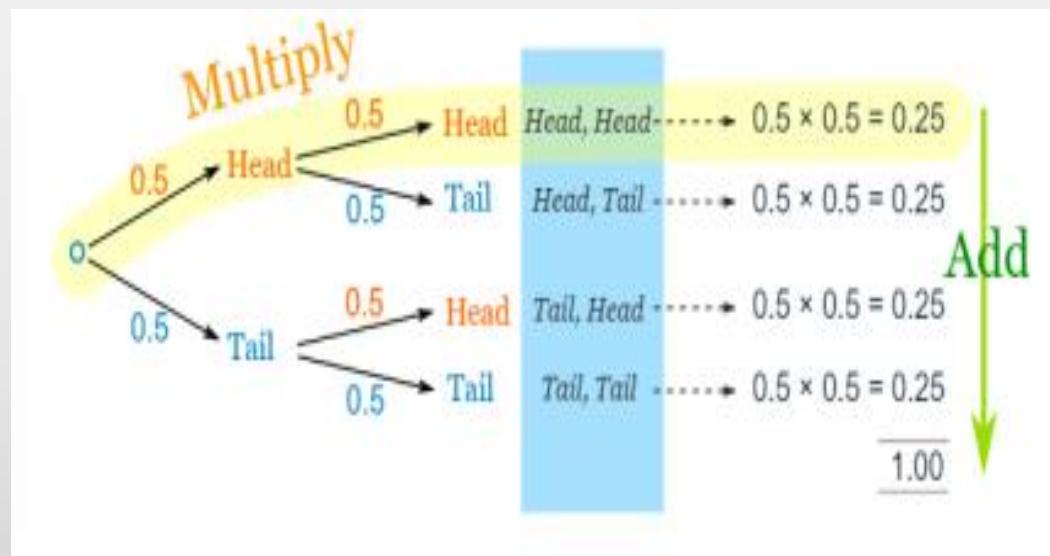
### Example 2

Suppose we toss two coins. Assume that all the outcomes are equally likely (fair coins). Let A be the event that at least one of the coins shows up heads. Find P(A).

$$S = \{HH, HT, TH, TT\}$$

$$A = \{HH, HT, TH\}$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{4} = 0.75$$



# Axioms of Probability



Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties. If  $S$  is the sample space and  $A$  is any event in a random experiment,

1.  $P(S) = 1$
2.  $0 \leq P(A) \leq 1$
3. For two events  $A$  and  $B$  with  $A \cap B = \emptyset$

$$P(A \cup B) = P(A) + P(B)$$

More generally, if  $A_1, A_2, \dots$ , are mutually exclusive events, then

$$P(A_1 \cup A_2 \cup A_3 \cup \dots) = P(A_1) + P(A_2) + P(A_3) + \dots$$

These axioms imply the following results.

$$P(\emptyset) = 0$$

$$P(A') = 1 - P(A)$$

# Additive Rule



If A and B are any two events, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \cup B)$  = probability that either events A or B occur

$P(A \cap B)$  = probability that both events A and B occur

## Example 1

John is going to graduate from an industrial engineering department in a university by the end of the semester. After being interviewed at two companies he likes, he assesses that his probability of getting an offer from company A is 0.8, and his probability of getting an offer from company B is 0.6. If he believes that the probability that he will get offers from both companies is 0.5, what is the probability that he will get at least one offer from these two companies?

# Additive Rule (Continued)



## Example 2

In a large university, the freshman profile for one year's fall admission says that 40% of the students were in the top 10% of their high school class, and that 65% are white, 25% of the students were white as well as were in the top 10% of their high school class. What is the probability that a freshman student selected randomly from this class either was in the top 10% of his or her high school class or is white?

# Example 1 and 2 Answers



## Example 1 Answer

Let  $A =$  getting an offer from company A

$B =$  getting an offer from company B

Using the additive rule, we have

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.8 + 0.6 - 0.5 = 0.9$$

## Example 2 Answer

Let  $H =$  students who were in the top 10% of their high school class and  $W =$  white Students in the class

$$P(H) = 0.4, P(W) = 0.65 \text{ and } P(H \cap W) = 0.25$$

$$P(H \cup W) = P(H) + P(W) - P(H \cap W) = 0.4 + 0.65 - 0.25 = 0.8$$



## Example 3

What is the probability of getting a total of 7 or 11 when a pair of dice are tossed?

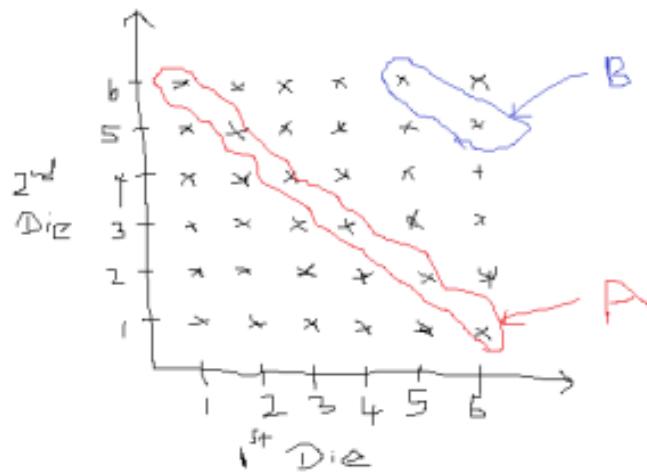


Figure 2: Rolling two fair Dies Experiment

Since all sample points are equally likely,  $P(A) = 1/6$  and  $P(B) = 1/18$ . The events A and B are mutually exclusive, since a total of 7 and 11 cannot both occur on the same toss. Therefore,

$$P(A \cup B) = P(A) + P(B) = \frac{1}{6} + \frac{1}{18} = \frac{2}{9}$$

This result could also have been obtained by counting the total number of points for the event , namely 8, and writing

$$P(A \cup B) = \frac{n}{N} = \frac{8}{36} = \frac{2}{9}$$

# Joint Probability



- ▶ The probability of events A and B occurring together is defined as Joint probability of A and B.

The probability of a joint event, A and B [ $\Pr(A \cap B)$ ]:

$$\Pr(A \text{ and } B) = \frac{\text{Number of outcomes satisfying } A \text{ and } B}{\text{Total number of outcomes in } S}$$

## Examples

1. Find the probability that you will get a Black - Ace from a playing deck of cards if a card is drawn at random.
2. Find the probability that you will get a Red-Jack from a playing deck of cards if a card is drawn at random.

# Joint Probability (Continued)



## Standard Deck of Cards

	Ace ↓	Two ↓	Three ↓	Four ↓	Five ↓	Six ↓	Seven ↓	Eight ↓	Nine ↓	Ten ↓	Jack ↓	Queen ↓	King ↓
Diamonds →	A diamond symbol at the top of the card.	Two diamond symbols.	Three diamond symbols.	Four diamond symbols.	Five diamond symbols.	Six diamond symbols.	Seven diamond symbols.	Eight diamond symbols.	Nine diamond symbols.	Ten diamond symbols.	Jack of diamonds.	Queen of diamonds.	King of diamonds.
Clubs →	A club symbol at the top of the card.	Two club symbols.	Three club symbols.	Four club symbols.	Five club symbols.	Six club symbols.	Seven club symbols.	Eight club symbols.	Nine club symbols.	Ten club symbols.	Jack of clubs.	Queen of clubs.	King of clubs.
Hearts →	A heart symbol at the top of the card.	Two heart symbols.	Three heart symbols.	Four heart symbols.	Five heart symbols.	Six heart symbols.	Seven heart symbols.	Eight heart symbols.	Nine heart symbols.	Ten heart symbols.	Jack of hearts.	Queen of hearts.	King of hearts.
Spades →	A spade symbol at the top of the card.	Two spade symbols.	Three spade symbols.	Four spade symbols.	Five spade symbols.	Six spade symbols.	Seven spade symbols.	Eight spade symbols.	Nine spade symbols.	Ten spade symbols.	Jack of spades.	Queen of spades.	King of spades.

There are 52 cards in a standard deck. There are 4 suits; Diamonds, Clubs, Hearts, Spades and each consist with 13 cards. Black cards include all Clubs and Spades. Red cards include all Hearts and Diamonds.

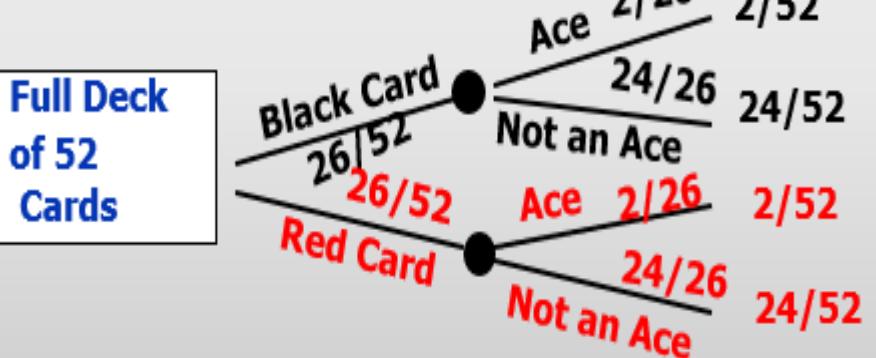
# Joint Probability (Continued)



## Examples – Answers

**Note:** Contingency Tables and Tree Diagrams can be used to visualize events and make calculations easier.

### Example 1 – Answer



Type	Color		Total
	Red	Black	
Ace	2	2 [2/52]	4
Non-Ace	24 [24/52]	24	48 [48/52]
Total	26	26 [26/52]	52

Joint Probabilities

Marginal Probabilities

This contingency table summarizes the joint and marginal probabilities for the event of drawing a card that is either an Ace or a Non-Ace, and either Red or Black. The joint probabilities are highlighted in red. The marginal probabilities for each row and column are also provided.

Probability of getting a Black Ace =  $2/52=1/26$

### Example 2 – Answer

Probability of getting a Red Jack =  $1/26$

# Marginal Probability



- ▶ The probability of a single event occurring ( $\text{Pr}(A)$ ), without the interference of another event (not conditioned on another event) is known as marginal probability.
- ▶ This can be thought of as an unconditional probability.

## Examples

1. Find the probability that you will get a King from a playing deck of cards, if a card is drawn at random.

$$4/52 = 1/13$$

2. Find the probability that you will get a Black card from a playing deck of cards, if a card is drawn at random.

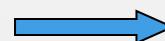
$$26/52 = 1/2$$

# Conditional Probability



- ▶ A **conditional probability** is the probability of one event, given that another event has occurred.

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



The conditional probability of A given that B has occurred

$$P(B | A) = \frac{P(A \cap B)}{P(A)}$$



The conditional probability of B given that A has occurred

Where  $P(A), P(B) > 0$

$P(A \cap B)$  = Joint probability of A and B.

$P(A)$  = Marginal probability of A.

$P(B)$  = Marginal probability of B.

# Conditional Probability (Continued)



## Example 1

Of the cars on a used car lot, 70% have air conditioning and 40% have a CD player. 20% of the cars have both. What is the probability that a car has a CD player, given that it has AC?

## Example 2

If two balanced dice are tossed, find the probability that the sum of the face values is 8, if the face value of the first one is 3.

## Example 3

The probability that a regularly scheduled flight departs on time is 0.83, the probability that it arrives on time is 0.82 and the probability that it departs and arrives on time is 0.78. Find the probability that a plane

- (i) arrives on time given that it departed on time and
- (ii) departed on time given that it has arrived on time



# Conditional Probability (Continued)

## Example – Answers

### Example 1 – Answer

Let  $A = \text{cars with air conditioning (AC)}$

$B = \text{cars with CD player}$

$$P(A) = 0.7 \quad P(B) = 0.4 \quad P(A \cap B) = 0.2$$

$$P(B | A) = P(A \cap B) / P(A) = 0.2 / 0.7 = 2/7$$

### Example 2 – Answer

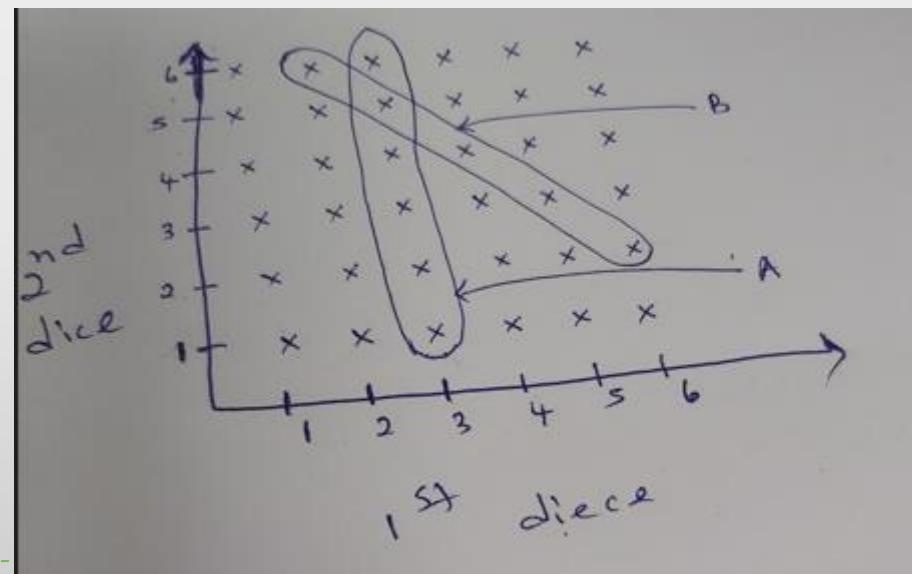
$A = \text{Face value of 1st die is 3}$

$B = \text{Sum of the face values is 8}$

$$P(A) = 6/36$$

$$P(B) = 5/36 \quad P(A \cap B) = 1/36$$

$$P(B|A) = P(A \cap B) / P(A) = 1/6$$





# Conditional Probability (Continued)

## Example 3 – Answer

Let , D = a regularly scheduled flight departs on time

A = a regularly scheduled flight arrives on time , then

$$P(D) = 0.83, P(A) = 0.82 \text{ and } P(A \cap D) = 0.78$$

(i)  $P(A | D) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.94$

(ii)  $P(D | A) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.95$



# Properties of Conditional Probability

$$\Pr(A|B) = 1 - \Pr(A^c|B)$$

The Product Rule, or the Multiplication Rule

$$\Pr(A \cap B) = \Pr(B) * \Pr(A|B) = \Pr(A) * \Pr(B|A)$$

## Example

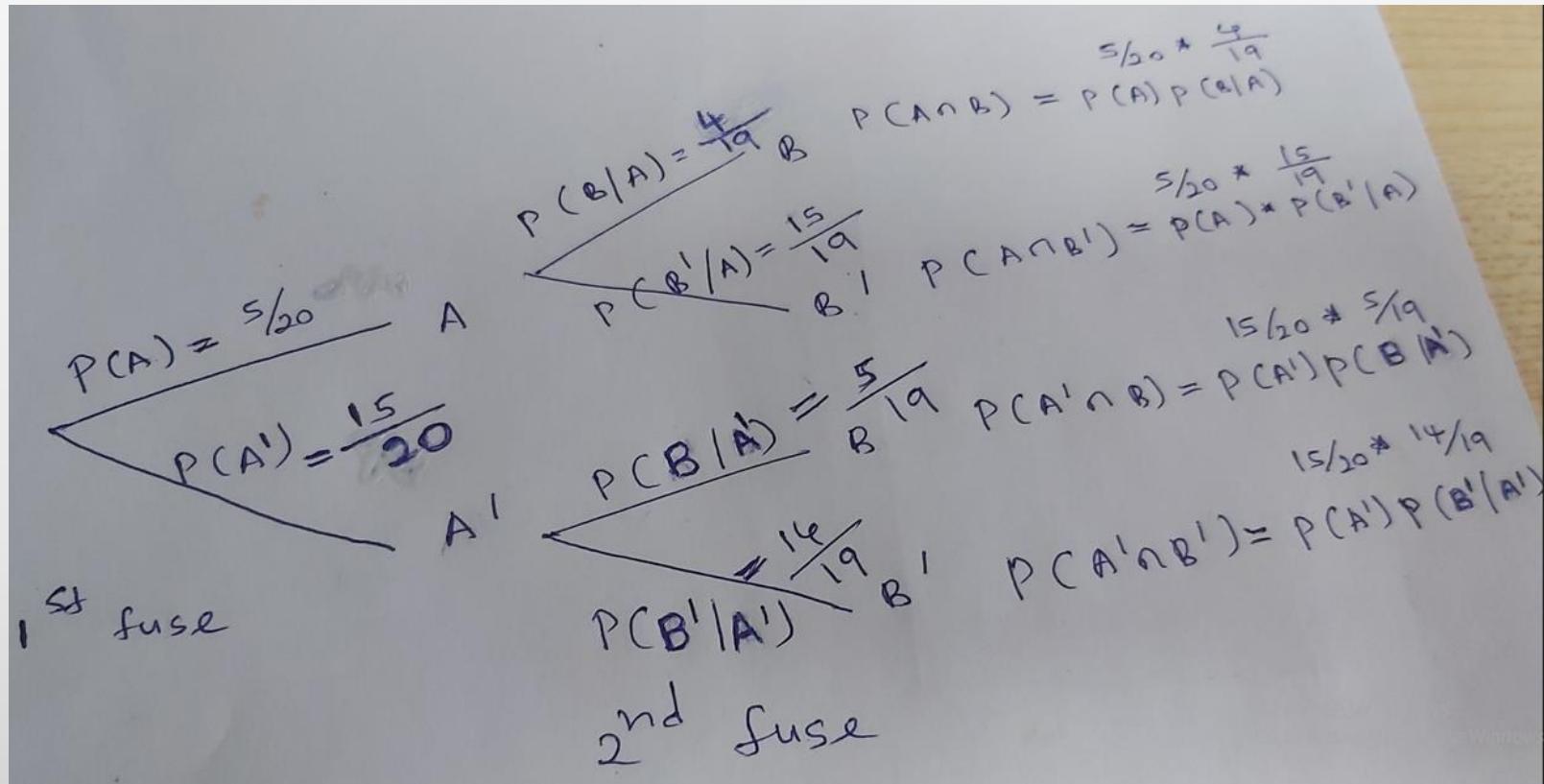
Suppose that we have a fuse box containing 20 fuses, of which 5 are defective. If 2 fuses are selected at random and removed from the box in succession without replacing the first, what is the probability that both fuses are defective?

# Properties of Conditional Probability (Continued)



Example – Answer

Let A = 1<sup>st</sup> fuse is a defective and B = 2<sup>nd</sup> fuse is a defective



$$P(A \cap B) = P(A) P(B|A) = \frac{5}{20} * \frac{4}{19} = 1/19$$

# Statistical Independence



- ▶ Two events A and B are **independent** if and only if:

$$P(A|B) = P(A) \text{ or } P(B|A) = P(B)$$

- ▶ If two events A and B are independent then,

$$P(A \cap B) = P(A)P(B)$$

- ▶ If the events A<sub>1</sub>, A<sub>2</sub>, ..., A<sub>k</sub> are independent, then

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1)P(A_2)\dots P(A_k).$$

# Statistical Independence (Continued)



## Example 1

Now consider an experiment in which 2 cards are drawn in succession from an ordinary deck, with replacement. The events are defined as A: the first card is an ace, B: the second card is a spade. Since the first card is replaced, our sample space for both the first and the second draw consists of 52 cards, containing 4 aces and 13 spades. Hence,

$$P(B|A) = 13/52 = \frac{1}{4}$$

$$P(B) = 13/52 = \frac{1}{4}$$

That is,  $P(B|A) = P(B)$ . Therefore, the events A and B are said to be independent

# Statistical Independence (Continued)



## Example 2

A small town has one fire engine and one ambulance available for emergencies. The probability that the fire engine is available when needed is 0.98, and the probability that the ambulance is available when called is 0.92. In the event of an injury resulting from a burning building, find the probability that both the ambulance and the fire engine will be available, assuming they operate independently.

Let A and B represent the respective events that the fire engine and the ambulance are available.

Then  $P(A) = 0.98$ ,  $P(B) = 0.92$ .

$$P(A \cap B) = P(A)P(B) = 0.98 \times 0.92 = 0.9016$$

# **5. RANDOM VARIABLES & PROBABILITY DISTRIBUTIONS**

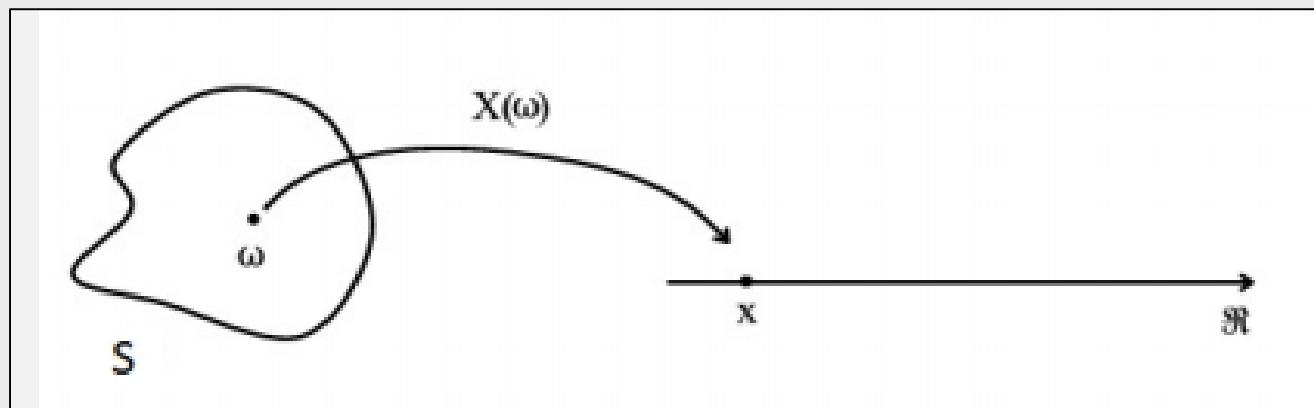
## **[IT2110]**

*By Department of Mathematics and Statistics*

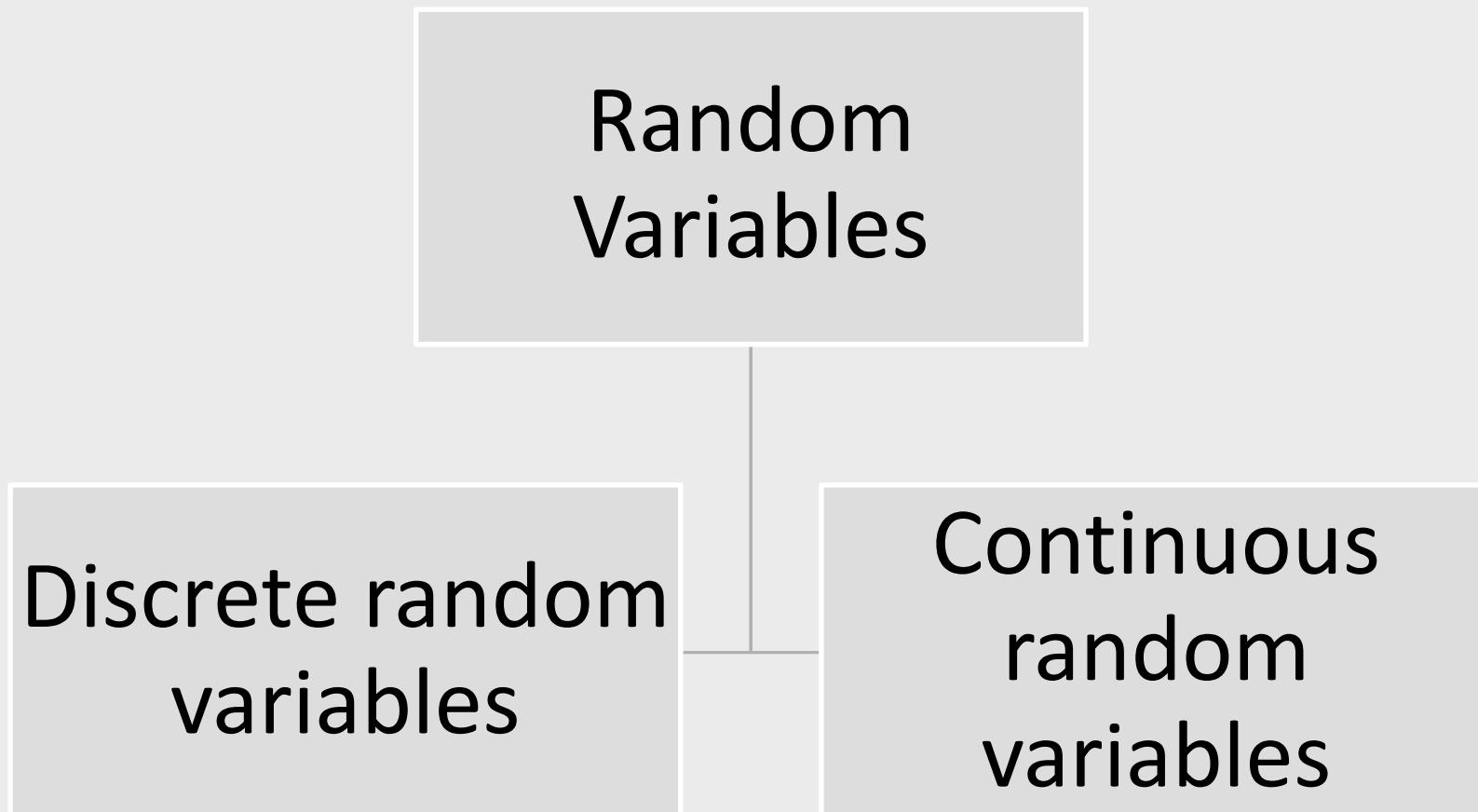
*Faculty of Humanities and Sciences*

# **RANDOM VARIABLES**

- A random variable (r.v.)  $X$  is a function defined on a sample space ( $S$ ), that associates a real number,  $X(\omega) = x$ , with each outcome  $\omega$  in  $S$ .
- ***Simple Definition*** : A random variable is a mapping between all the outcomes in the sample space with a set of real numbers.



- Random variables are denoted by using capital letters



# Discrete Random Variables

- A random variable is said to be discrete, if it can assume only *distinct* values.

*OR*

- In other words, it can assume only countable number of values.

# Examples

- Toss a coin 5 times. Let  $X$  be the number of heads appeared. Then,

$$X - 0,1,2,3,4,5$$

- Roll a die twice. Let  $X$  be the number of times 4 comes up.

$$X - 0,1,2$$

- Suppose we toss two coins. Assume that all the outcomes are equally likely (fair coins). Let  $Y$  be the number of heads appeared. Then,

$$Y - 0,1,2$$

# **PROBABILITY DISTRIBUTIONS**

- The set of all ordered pairs  $(x, \Pr(X = x))$  of a discrete r.v. ( $X$ ), is known as the probability distribution
- This is also known as the ***probability mass function*** (p.m.f.) and is denoted by  $P_X(x)$ .
- ***Simple Definition:*** All the possible values of a discrete random variable with their corresponding probability values is known as probability distribution.

# Properties

- $P_X(x)$  refers to  $P(X=x)$ .
- The probability distribution function is always non-negative.
- $\sum_{all \ x} P_X(x) = 1$
- The cumulative distribution function (c.d.f.)  $F$  of the random variable  $X$  is defined by

$$F_X(x) = \Pr(X \leq x)$$

# Example

- Suppose we toss two coins. Assume that all the outcomes are equally likely (fair coins). Let  $Y$  be the number of heads appeared. Then,

$Y$	0	1	2
$\Pr(Y=y)$	1/4	2/4	1/4
$F_Y(y) = \Pr(Y \leq y)$	1/4	3/4	1

# Expected Value & Variance

- This is same as mean of the random variable.
- Let  $X$  be a discrete random variable with p.m.f.  $P_X(x)$ . Then the expected value of  $X$ , denoted by  $E(X)$ , is defined by

$$E(X) = \sum_{\text{all } x} x * Pr(X = x)$$

- The variance of a random variable  $X$  is defined by

$$V(X) = E(X - E(X))^2 = E(X^2) - [E(X)]^2$$

# Properties [E(X)]

Let X & Y be two random variables. Then,

- $E(c) = c$
- $E[g(X)] = \sum_{all\ x} g(x) * Pr(X = x)$
- $E[g(X)+c] = E[g(X)] + c$
- $E[c*g(X)] = c*g(E[X])$
- $E[X+Y] = E[X] + E[Y]$

# Properties [V(X)]

Let X & Y be two random variables. Then,

- $V(c) = 0$
- $V[g(X)+c] = V[g(X)]$
- $V[c^*g(X)] = c^2 * V[g(X)]$
- $V[X+Y] = V[X] + V[Y] + 2Cov(X,Y)$
- $V[X-Y] = V[X] + V[Y] - 2Cov(X,Y)$
- *If X & Y are independent then,  $Cov(X,Y) = 0$*

# Covariance

- Covariance is a measure of how the changes in one variable are associated with the changes in second variable.

$$Cov(X, Y) = \sum_{i=1}^N [X_i - E(X)][Y_i - E(Y)]P(X_i Y_i)$$

# Examples

1. Suppose we toss two coins. Assume that all the outcomes are equally likely (fair coins). Let  $Y$  be the number of heads appeared. Find  $E(Y)$  and  $\text{Var}(Y)$ .

$Y$	0	1	2
$\Pr(Y=y)$	0.25	0.5	0.25

$$E(Y) = \sum_{\text{all } y} y * \Pr(Y = y) = (0*0.25) + (1*0.5) + (2*0.25) = 1$$

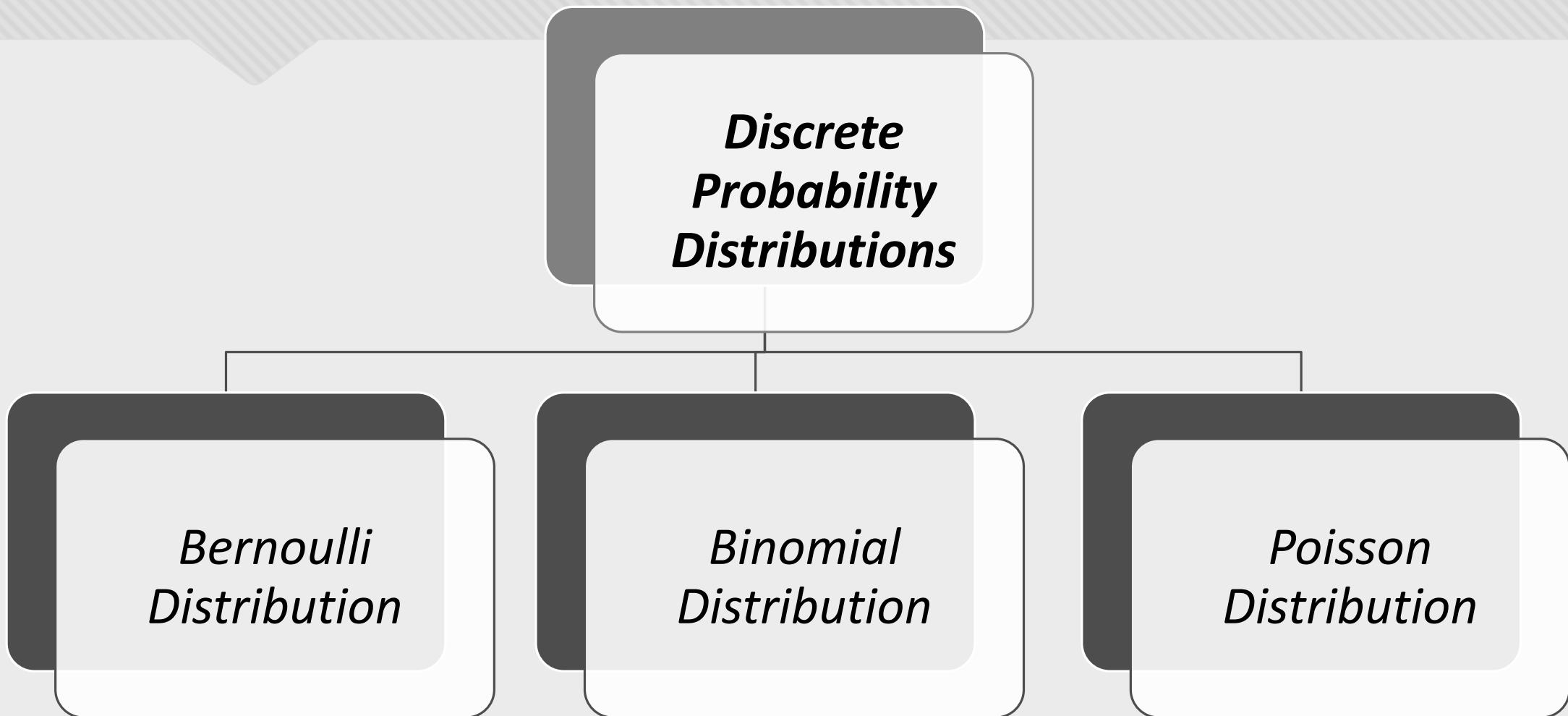
$$E(Y^2) = \sum_{\text{all } y} y^2 * \Pr(Y = y) = (0*0.25) + (1*0.5) + (4*0.25) = 1.5$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = 1.5 - 1^2 = 0.5$$

# Examples

2. To find out the prevalence of smallpox vaccine use, a researcher inquired into the number of times a randomly selected 200 people aged 16 and over in an African village had been vaccinated. He obtained the following figures: never, 16 people; once, 30; twice, 58; three times, 51; four times, 38; five times, 7. Assuming these proportions continue to hold exhaustively for the population of that village, what is the expected number of times those people in the village had been vaccinated, and what is the standard deviation?
3. Let  $X$  and  $Y$  be two independent random variables. Suppose that we know  $\text{Var}(2X-Y)=6$  and  $\text{Var}(X+2Y)=9$ . Find  $\text{Var}(X)$  and  $\text{Var}(Y)$ .

# Discrete Probability Distributions



# Conditions for Discrete Distributions

Bernoulli Distribution	Binomial Distribution	Poisson Distribution
Only two possible outcomes (Success & Failure)	For each trial, only two possible outcomes (Success & Failure)	For each trial, only two possible outcomes (Success & Failure)
Only one trial	No of trials (n) are fixed	Trials (n) is large
	Probability of success (p) is constant for each & every trial	The occurrences are independent of each other
	Trials are independent	<i>[Assume that the numbers of occurrences in disjoint Intervals]</i>
<i>Eg:- Tossing a coin once</i>	<i>Eg:- Tossing a coin 10 (n) times</i>	<i>Eg:- Number of defects in a lot</i>

# Discrete Distributions

Bernoulli Distribution	Binomial Distribution	Poisson Distribution
$X$ – Getting the success	$X$ - The number of successes in $n$ number of trials.	$X$ - The number of occurrences for a given rate of occurrence ( $\lambda$ )
$X \sim Bernoulli(p)$	$X \sim Bin(n,p)$	$X \sim Poisson(\lambda)$
$P_X(x) = p^x(1-p)^{1-x}$ [p.m.f.]	$P_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$ $;x = 0,1,2,\dots,n$ [p.m.f.]	$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ $;x = 0,1,2,\dots$ [p.m.f.]
$E(X) = p$	$E(X) = np$	$E(X) = \lambda$
$V(X) = p(1-p)$	$V(X) = np(1-p)$	$V(X) = \lambda$

# Binomial Distributions

- An expansion of the Bernoulli distribution.
- Each trial has a Bernoulli distribution.

# Examples

- 1) It is known that screws produced by a certain machine will be defective with probability 0.01 independently of each other. If we randomly pick 10 screws produced by this machine, what is the probability that
- a) exactly six screws will be defective?
  - b) at most 3 screws will be defective?
  - c) at least 2 screws will be defective?
  - d) What is the expected number of defectives?
  - e) What is the variance of defectives?

# Examples

- 2) Fifty seeds were planted and it is known that the probability of any seed germinating is 0.4. Assuming that the seeds have no other factors in germinating, find the following probabilities.
- a) More than 12 seeds germinate.
  - b) More than 15 but fewer than 30 seeds germinate.

# Poisson Distribution - Example

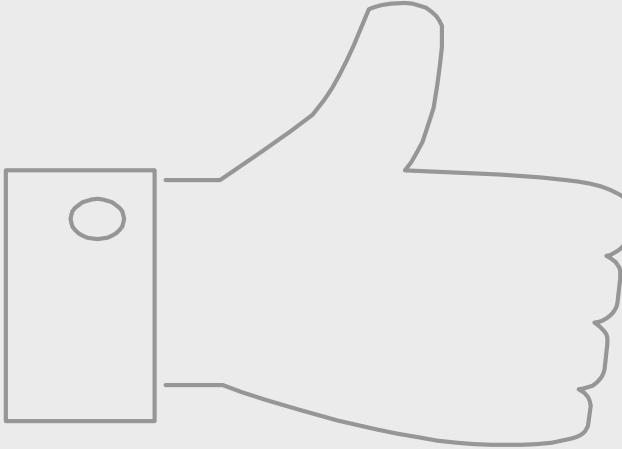
- 1) Suppose that, on average, in every two pages of a book there is one typographical error. What is the probability of at least one error on a certain page of the book?

# Poisson Approximation

- If  $X \sim \text{Bin}(n, p)$ , then  $X$  can be approximated with a Poisson distribution with the rate parameter ( $\lambda$ ) being equal to  $np$  if  $p$  is quite small and  $n$  is large.
- Usually this approximation can be used if  $p < 0.1$  and  $n > 50$ .

# Example

- 1) If the probability that an individual suffers an adverse reaction from a particular drug is known to be 0.001, determine the probability that out of 2000 individuals, (a) exactly three and (b) more than two individuals will suffer an adverse reaction.



# THANKS!

**Any questions?**

# IT2110 - Probability and Statistics

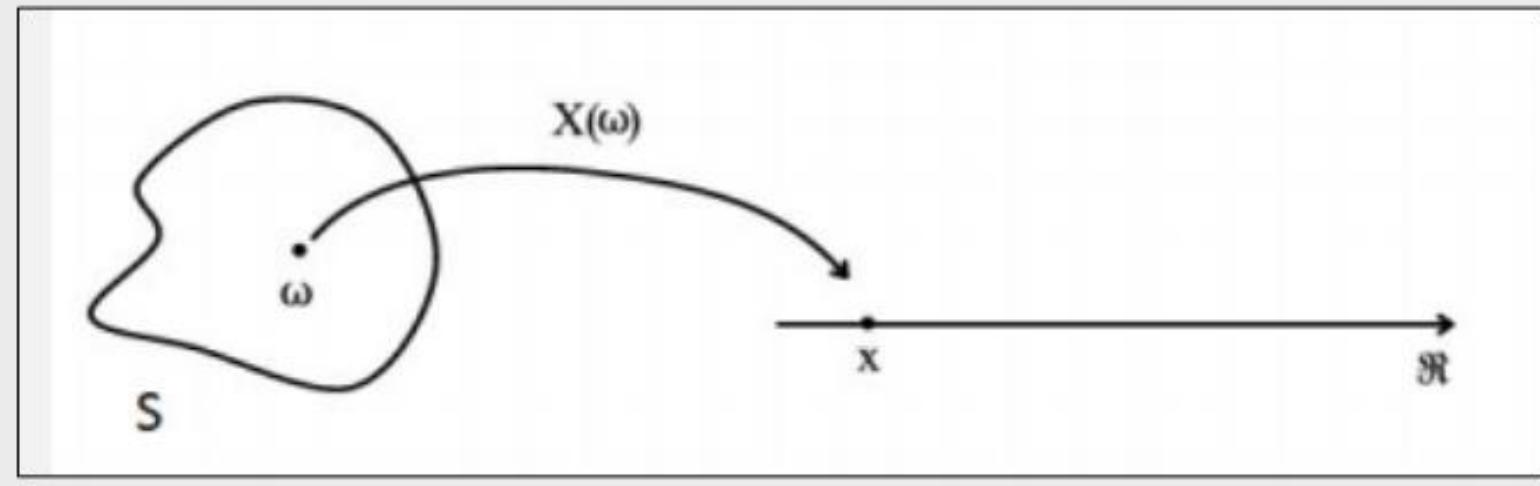
*Random variables and Discrete  
Probability Distributions*

*Chapter 5*

# Random Variables

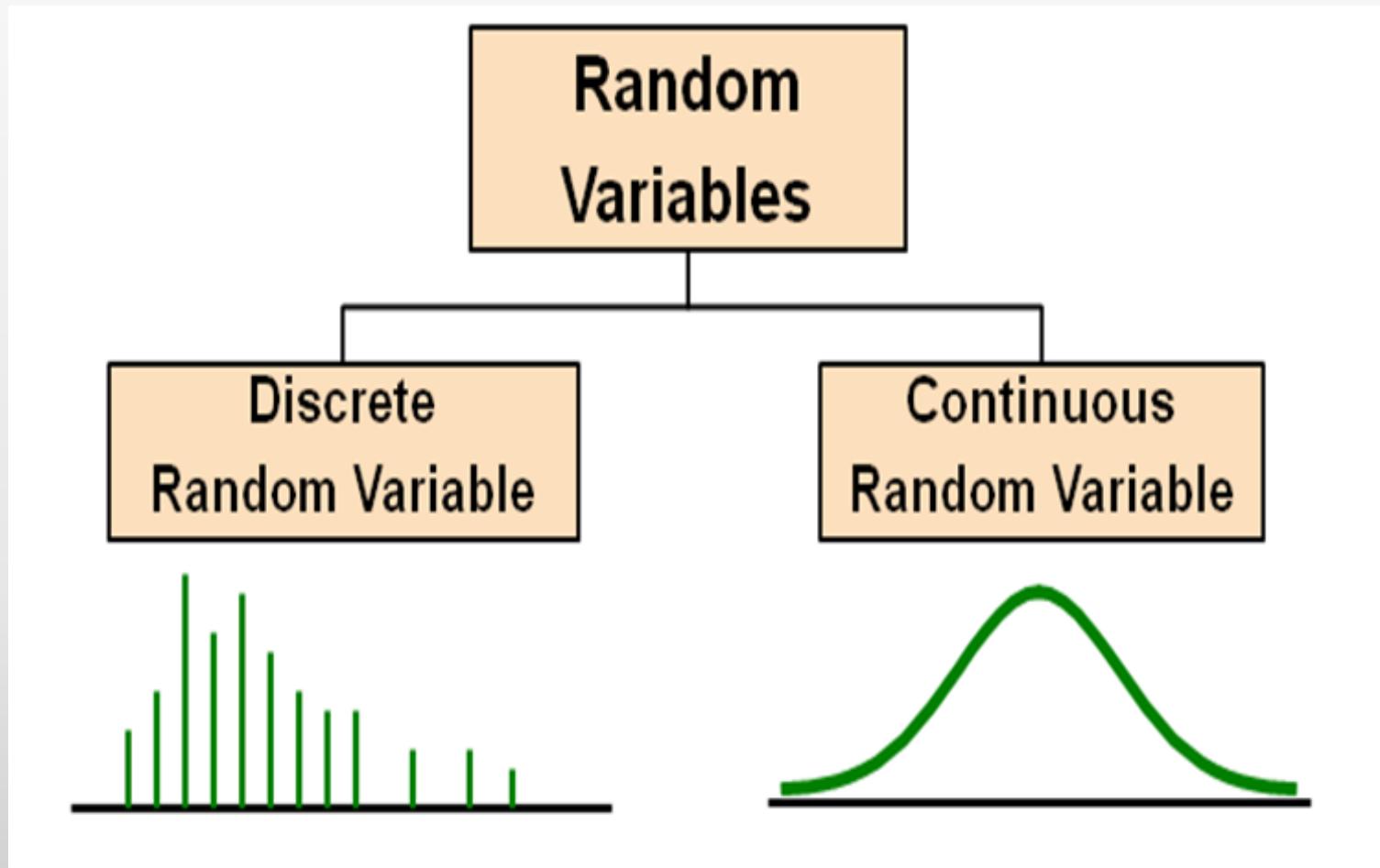
A random variable (r.v.)  $X$  is a function defined on a sample space ( $S$ ), that associates a real number,  $X(\omega) = x$ , with each outcome  $\omega$  in  $S$ .

***Simple Definition :*** A random variable is a mapping between all the outcomes in the sample space with a set of real numbers.



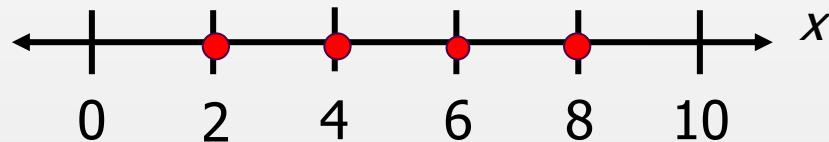
# Random Variables

- Random variables are denoted by using capital letters



# Discrete Random Variables

A **discrete random variable** is a random variable that has either a finite number of possible values or a countable number of possible values



## Notation

We use capital letter , like  $X$ , to denote the random variable and use small letter to list the possible values of the random variable.

## Example

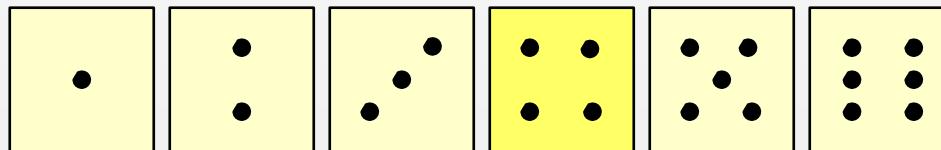
A single die is cast,  $X$  represent the number of pips showing on the die and the possible values of  $X$  are;

$$x=1,2,3,4,5,6.$$

# Discrete Random Variables (Continued)

## Examples

- Roll a die twice



Let  $X$  be the number of times 4 comes up  
(then  $x$  could be 0, 1, or 2 times)

- Toss a coin 5 times

Let  $X$  be the number of heads  
(then  $x = 0, 1, 2, 3, 4$ , or 5)

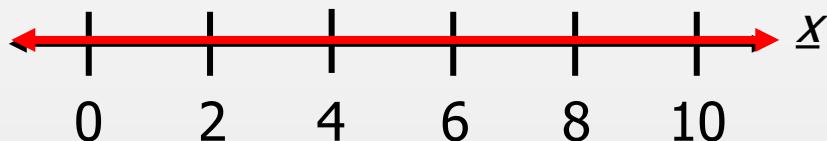
- Toss two fair coins

Let  $Y$  be the number of heads  
(then  $y = 0, 1, 2, , )$



# Continuous Random Variables

A **continuous random variable** is a random variable that has an infinite number of possible values that is not countable.



## Examples

Interest centers around the proportion of people who respond to a certain mail order solicitation. Let  $X$  be that proportion.  $X$  is a random variable that takes on all values  $x$  for which  $0 \leq x \leq 1$ .

Let  $X$  be the random variable defined by the waiting time, in hours, between successive speeders spotted by a radar unit. The random variable  $X$  takes on all values  $x$  for which  $x \geq 0$ .

# Exercise

Decide whether the given random variable is discrete or continuous

- a.) The distance your car travels on a tank of gas
  
- b.) The number of students in a statistics class

# Discrete Probability Distributions

A **discrete probability distribution** lists each possible value the random variable can assume, together with its probability.

- A discrete probability distribution is also known as the probability mass function (p.m.f.) and is denoted by  $P(X=x)$  or  $P_X(x)$ .
- A discrete probability distribution can be in the form of a **table**, **graph** or mathematical **formula**.

## Properties

- The probability distribution function is always non-negative  
 $0 \leq P(X = x) \leq 1$
- $\sum_x P(X = x) = 1$
- The cumulative distribution function (c.d.f.)  $F$  of the random variable  $X$  is defined by  $F_X(x) = P(X \leq x)$ .

# Discrete Probability Distributions (Continued)

## Guidelines to Construct a Discrete Probability Distribution

Let  $X$  be a discrete random variable with possible outcomes  $x_1, x_2, \dots, x_n$ .

1. Make a frequency distribution for the possible outcomes.
2. Find the sum of the frequencies.
3. Find the probability of each possible outcome by dividing its frequency by the sum of the frequencies.
4. Check that each probability is between 0 and 1 and that the sum is 1.

# Discrete Probability Distributions (Continued)

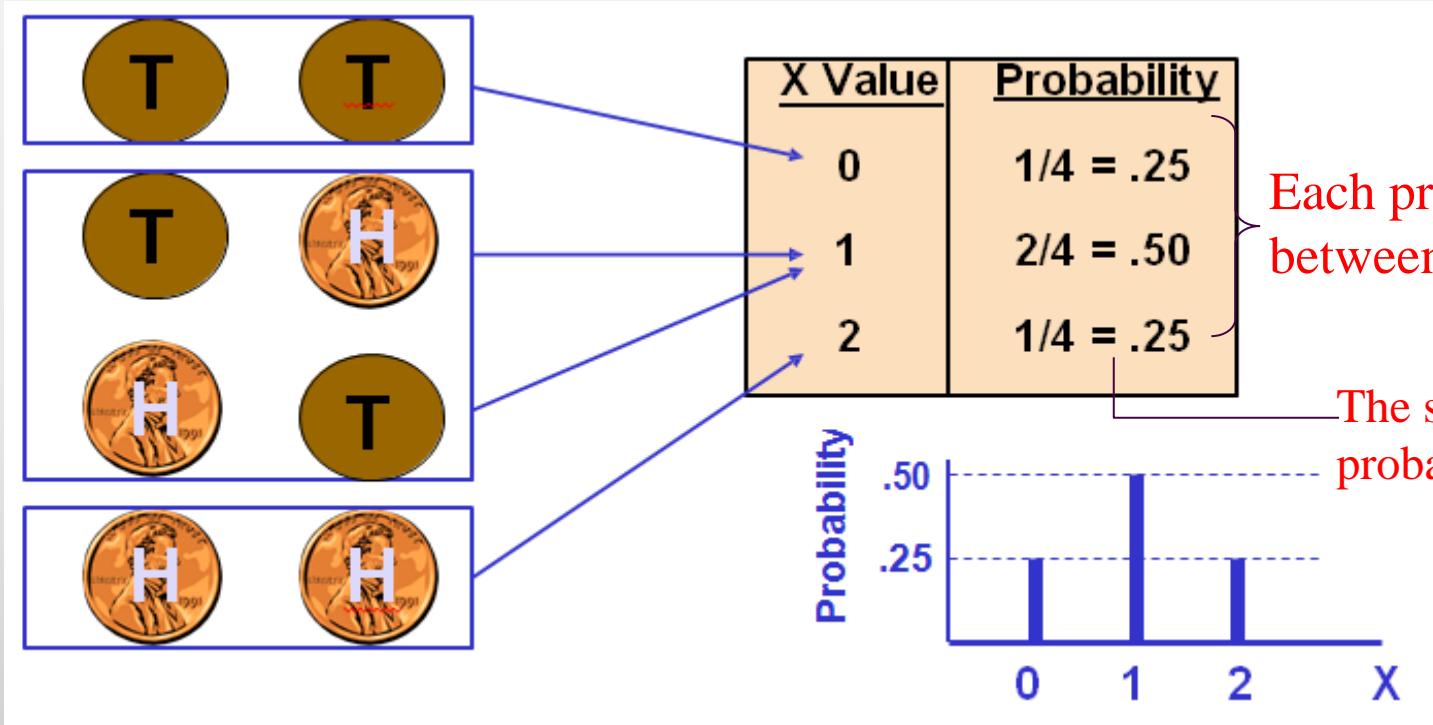
## Example 1

Experiment: Toss 2 Coins

4 possible outcomes

Let  $X$  = Number of heads

Probability Distribution



# Discrete Probability Distributions (Continued)

## Example 2

$X$  is the number of tails observed when a fair coin is tossed three times.

Outcomes (x)	Probability of Outcome
HHH (0)	$1/8 = 0.125$
HHT (1)	$1/8 = 0.125$
HTH (1)	0.125
THH (1)	0.125
HTT (2)	0.125
TTH (2)	0.125
THT (2)	0.125
TTT (3)	0.125

Probability distribution of  $X$

x	$P(X = x)$
0	0.125
1	$0.125+0.125+0.125 = 0.375$
2	$0.125+0.125+0.125 = 0.375$
3	0.125

$P(X = 3) = 0.125$ . This means that there is a 12.5% chance that we will observe three tails when we toss a fair coin three times.

# Discrete Probability Distributions (Continued)

## Calculating Cumulative Probabilities

What is the probability of observing at most 1 tail when a fair coin is tossed three times?

$$\begin{aligned}P(X \leq 1) &= P(X = 0) + P(X = 1) \\&= 0.125 + 0.375 \\&= 0.5\end{aligned}$$

## Calculating Upper Tail Probabilities

What is the probability of observing at least 2 tails when a fair coin is tossed three times?

$$\begin{aligned}P(X \geq 2) &= P(X = 2) + P(X = 3) \\&= 0.375 + 0.125 \\&= 0.5\end{aligned}$$

or

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - 0.5 = 0.5$$

# Discrete Probability Distributions (Continued)

## Calculating Interval Probabilities

What is the probability that the number of tails observed will be between 1 and 2 when a fair coin is tossed three times?

$$\begin{aligned}[P(1 \leq X \leq 2)] &= P(X = 1) + P(X = 2) \\ &= 0.375 + 0.375 \\ &= 0.75\end{aligned}$$

or

$$\begin{aligned}[P(1 \leq X \leq 2)] &= P(X \leq 2) - P(X \leq 0) \\ &= 0.875 - 0.125 \\ &= 0.75\end{aligned}$$

## Exercise

The table below shows the distribution for the random variable  $X$ , where  $X$  represents the number of DVDs a person rents from a video store during a single visit. Is this following a probability distribution?

$x$	$P(X=x)$
0	0.16
1	0.18
2	0.22
3	0.10
4	0.3
5	0.01

# Expected Value & Variance of a Random Variable

Let  $X$  be a discrete random variable with p.m.f.  $P_X(x)$ . Then the expected value (or mean) of  $X$ , denoted by  $E(X)$ , is defined by

$$E(X) = \sum_x x * P(X = x)$$

The variance of a random variable  $X$  is defined by

$$V(X) = E(X - E(X))^2 = E(X^2) - [E(X)]^2$$

# Properties of Expected Value

Let  $X$  &  $Y$  be two random variables. Then,

$$\textcircled{1} \quad E(c) = c$$

$$\textcircled{2} \quad E[g(X)] = \sum_{all \ x} g(x) * Pr(X = x)$$

$$\textcircled{3} \quad E[g(X)+c] = E[g(X)] + c$$

$$\textcircled{4} \quad E[c*g(X)] = c * E[g(X)]$$

$$\textcircled{5} \quad E[X+Y] = E[X] + E[Y]$$

# Properties of Variance

Let  $X$  &  $Y$  be two random variables. Then,

◎  $V(c) = 0$

◎  $V[g(X)+c] = V[g(X)]$

◎  $V[c^*g(X)] = c^2 * V[g(X)]$

◎  $V[X+Y] = V[X] + V[Y] + 2Cov(X,Y)$

◎  $V[X-Y] = V[X] + V[Y] - 2Cov(X,Y)$

◎ If  $X$  &  $Y$  are independent then,  $Cov(X,Y) = 0$

- Where  $Cov(X, Y)$  is a measure of how the changes in  $X$  are associated with the changes in  $Y$ .

# Expected Value & Variance of a Random Variable (Continued)

The mean, or **expected value**, of a **discrete random variable** is;

## Example 1

Toss 2 coins,  
 $X = \# \text{ of heads}$

X	P(X)
0	.25
1	.50
2	.25

## Example 2

x	f(x)	xf(x)
0	.40	.00
1	.25	.25
2	.20	.40
3	.05	.15
4	.10	.40

$$E(x) = 1.20$$

Compute expected value of X:

$$\begin{aligned} E(X) &= (0 \times .25) + (1 \times .50) + (2 \times .25) \\ &= 1.0 \end{aligned}$$

expected number of  
TVs sold in a day

## Expected Value & Variance of a Random Variable (Continued)

The variance of a **discrete random variable** is;

Let's consider the Example 1

$$E(X) = 1 \quad E(X^2) = \sum_x x^2 * P(X = x)$$

$$E(X^2) = 0^2 * 0.25 + 1^2 * 0.5 + 2^2 * 0.25 = 0.5 + 1 = 1.5$$

$$V(X) = E(X^2) - [E(X)]^2 = 1.5 - 1 = 0.5$$

### Exercise 1

Find the variance of Example 2.

### Exercise 2

A die is thrown; Find the mean (or expected) score and the variance of the score?

# Expected Value & Variance of a Random Variable (Continued)

## Exercise 3

To find out the prevalence of smallpox vaccine use, a researcher inquired into the number of times a randomly selected 200 people aged 16 and over in an African village had been vaccinated. He obtained the following figures: never, 16 people; once, 30; twice, 58; three times, 51; four times, 38; five times, 7. Assuming these proportions continue to hold exhaustively for the population of that village, what is the expected number of times those people in the village had been vaccinated, and what is the standard deviation?

## Exercise 4

Let  $X$  and  $Y$  be two independent random variables. Suppose that we know  $\text{Var}(2X-Y)=6$  and  $\text{Var}(X+2Y)=9$ . Find  $\text{Var}(X)$  and  $\text{Var}(Y)$ .

# Expected Value & Variance of a Random Variable (Continued)

## Exercise 2 - Answer

Let  $X$  = outcome of rolling one die

$X$	1	2	3	4	5	6
$P(X)$	1/6	1/6	1/6	1/6	1/6	1/6

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

$$E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = \frac{91}{6}$$

$$V[X] = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}$$

# Expected Value & Variance of a Random Variable (Continued)

## Exercise 3 - Answer

X	0	1	2	3	4	5
P(X=x)	0.08	0.15	0.29	0.255	0.19	0.035

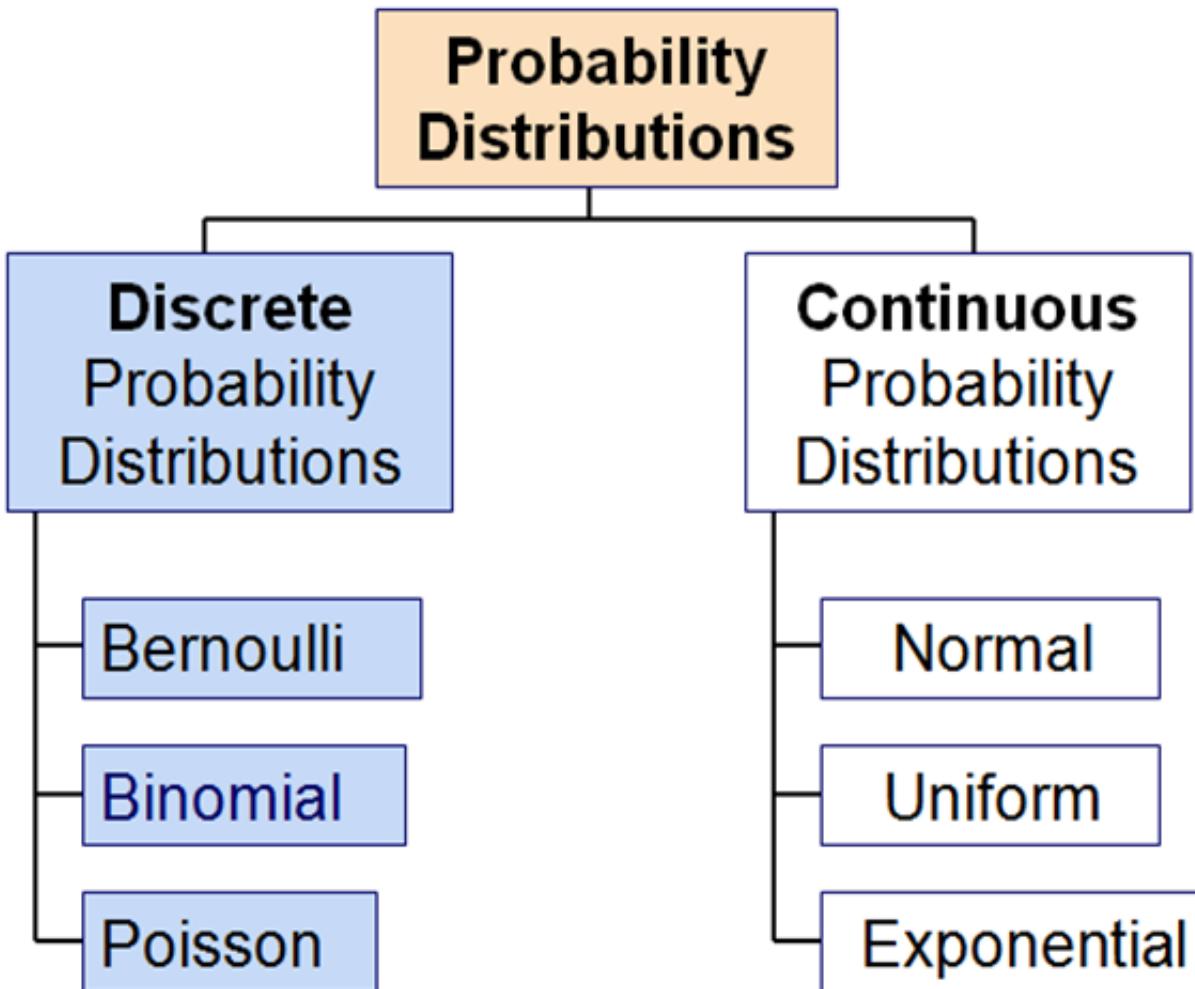
Do the rest!

## Exercise 4 - Answer

$$\begin{aligned} \sqrt{2x-y} &= 6 \\ 4\sqrt{x} + \sqrt{y} &= 6 \quad \text{--- (1)} \end{aligned}$$
$$\begin{aligned} \sqrt{x+2y} &= 9 \\ \sqrt{x} + 4\sqrt{y} &= 9 \quad \text{--- (2)} \end{aligned}$$
$$\begin{aligned} (1) \times 4 \Rightarrow \\ 16\sqrt{x} + 4\sqrt{y} &= 24 \quad \text{--- (3)} \end{aligned}$$

$$\begin{aligned} (3) - (2) \Rightarrow \\ 15\sqrt{x} &= 15 \\ \sqrt{x} &= 1 \quad \text{--- (1)} \\ \text{then assigning this in to} \\ 4\sqrt{x} + \sqrt{y} &= 6 \\ \sqrt{y} &= 2 \end{aligned}$$

# Probability Distributions



# Conditions for Discrete Probability Distributions

Bernoulli Distribution	Binomial Distribution	Poisson Distribution
Only two possible outcomes (Success & Failure)	For each trial, only two possible outcomes (Success & Failure)	For each trial, only two possible outcomes (Success & Failure)
Only one trial	No of trials ( $n$ ) are fixed	Trials ( $n$ ) is large
	Probability of success ( $p$ ) is constant for each & every trial	The occurrences are independent of each other
	Trials are independent	<i>[Assume that the numbers of occurrences in disjoint Intervals]</i>
Eg:- Tossing a coin once	Eg:- Tossing a coin 10 ( $n$ ) times	Eg:- Number of defects in a lot

# Discrete Probability Distributions

Bernoulli Distribution	Binomial Distribution	Poisson Distribution
$X$ – Getting the success	$X$ - The number of successes in $n$ number of trials.	$X$ - The number of occurrences for a given rate of occurrence ( $\lambda$ )
$X \sim Bernoulli(p)$	$X \sim Bin(n,p)$	$X \sim Poisson(\lambda)$
$P_X(x) = p^x(1-p)^{1-x}$ [p.m.f.]	$P_X(x) = \binom{n}{x} p^x(1-p)^{n-x}$ ; $x = 0, 1, 2, \dots, n$ [p.m.f.]	$P_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}$ ; $x = 0, 1, 2, \dots$ [p.m.f.]
$E(X) = p$	$E(X) = np$	$E(X) = \lambda$
$V(X) = p(1-p)$	$V(X) = np(1-p)$	$V(X) = \lambda$

# Examples

1. A fair coin is tossed. Let the variable  $x$  take values 1 and 0 according to the toss results in ‘Head’ or ‘Tail’.

Then,  $X$  is a Bernoulli variable with parameter  $p=1/2$ . Here,  $X$  denotes the number of heads obtained in the toss.

Probability of success =  $1/2$  and the probability of failure =  $1/2$ .

2. In a single throw of a dice, the outcome "5" is called a success and any other outcome is called a failure, then the successive throws of a dice will contain Bernoulli trials.

The probability of success =  $1/6$  and the probability of failure =  $5/6$

# Binomial Distribution

## Rule of Combinations

- The number of combinations of selecting X objects out of n objects is;

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}$$

Where;

$$n! = n(n - 1)(n - 2) \dots (2)(1)$$
$$X! = X(X - 1)(X - 2) \dots (2)(1)$$
$$0! = 1(\text{by definition})$$

# Binomial Distribution (Continued)

$$P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$P_X(x)$  = probability of  $X$  successes in  $n$  trials, with probability of success  $p$  on each trial

$X$  = number of ‘successes’ in sample, ( $X = 0, 1, 2, \dots, n$ )

$n$  = sample size (number of trials or observations)

$p$  = probability of “success”

**Example:** Flip a coin four times, let  $x$  = # heads:

$n = 4$

$p = 0.5$

$1 - p = (1 - .5) = .5$

$X = 0, 1, 2, 3, 4$

$$X \sim Bin(n, p)$$

# Binomial Distribution (Continued)

## Exercise

Decide whether the below experiments are binomial experiments. If it is, specify the values of  $n$ ,  $p$ , and  $q$ , and list the possible values of the random variable  $x$ . If it is not a binomial experiment, explain why.

1. You randomly select a card from a deck of cards, and note if the card is an Ace. You then put the card back and repeat this process 8 times.
2. You roll a die 10 times and note the number the die lands on.

# Binomial Distribution (Continued)

## Exercise - Answers

1. This is a binomial experiment. Each of the 8 selections represent an independent trial because the card is replaced before the next one is drawn. There are only two possible outcomes: either the card is an Ace or not.

$$n = 8 \quad p = \frac{4}{52} = \frac{1}{13} \quad q = 1 - \frac{1}{13} = \frac{12}{13} \quad x = 0, 1, 2, 3, 4, 5, 6, 7, 8$$

2. This is not a binomial experiment. While each trial (roll) is independent, there are more than two possible outcomes: 1, 2, 3, 4, 5, and 6.

# Binomial Distribution (Continued)

## Example 1

Suppose a quiz has 10 multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. Find the

- I. probability of no correct responses
- II. probability of one correct response
- III. expected number of correct responses
- IV. standard deviation of the correct responses

## Example 2

A biased coin is tossed 6 times. The probability of heads on any toss is 0.3. Let  $X$  denote the number of heads that come up. Calculate:

- I.  $P(X=2)$
- II.  $P(X>3)$
- III.  $P(1 < X \leq 5)$

# Binomial Distribution (Continued)

## Example 1 - Answer

Let  $X$  denote the number of correct responses. The probability of a correct response is 0.20 for a given question, so  $n = 10$  and  $p = 0.20$ .

$$P(X = 0) = \frac{10!}{0!10!} (0.20^0)(0.80)^{10} = (0.80)^{10} = 0.107$$

$$P(X = 1) = \frac{9!}{1!9!} (0.20^1)(0.80)^9 = 10(0.20)(0.80)^9 = 0.268$$

$$\mu = np = 10 \times 0.2 = 2$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{10 \times 0.2 \times 0.8} = 1.26$$

# Binomial Distribution (Continued)

## Example 2 - Answer

Using Binomial Formula,

$$X \sim Bin(6, 0.3)$$

(i)

$$\begin{aligned}P(X = 2) &= \binom{6}{2} 0.3^2 (1 - 0.3)^{6-2} \\&= \frac{6!}{2!4!} (0.3^2)(0.7^4) \\&= 15 \times (0.3^2)(0.7^4) \\&= 15 \times 0.09 \times 0.2401 \\&= 0.3241\end{aligned}$$

# Binomial Distribution (Continued)

## Example 2 – Answer (Continued)

(ii)

$$\begin{aligned}P(X > 3) &= 1 - P(X \leq 3) \\&= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)] \\&= 1 - [\binom{6}{0}(0.3^0)(0.7^6) + \binom{6}{1}(0.3^1)(0.7^{6-1}) + \binom{6}{2}(0.3^2)(0.7^{6-2}) + \binom{6}{3}(0.3^3)(0.7^{6-3})] \\&= 1 - [\frac{6!}{0!6!}(0.3^0)(0.7^6) + \frac{6!}{1!5!}(0.3^1)(0.7^5) + \frac{6!}{2!4!}(0.3^2)(0.7^4) + \frac{6!}{3!3!}(0.3^3)(0.7^3)] \\&= 1 - [(0.7^6) + 6 \times (0.3)(0.7^5) + 15 \times (0.3^2)(0.7^4) + 20 \times (0.3^3)(0.7^3)] \\&= 1 - [0.1176 + 0.3025 + 0.3241 + 0.1852] \\&= 1 - 0.9294 \\&= 0.0706\end{aligned}$$

or you can find

$$P(X > 3) = P(X \geq 4)$$

# Binomial Distribution (Continued)

## Example 2 – Answer (Continued)

(iii)

$$\begin{aligned}P(1 < X \leq 5) &= P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) \\&= 0.3241 + 0.1852 + \binom{6}{4}(0.3^4)(0.7^{6-4}) + \binom{6}{5}(0.3^5)(0.7^{6-5}) \\&= 0.5093 + \frac{6!}{4!2!}(0.3^4)(0.7^2) + \frac{6!}{5!1!}(0.3^5)(0.7^1) \\&= 0.5093 + 15 \times (0.3^4)(0.7^2) + 6 \times (0.3^5)(0.7^1) \\&= 0.5093 + 0.0595 + 0.0102 \\&= 0.579\end{aligned}$$

# Binomial Distribution (Continued)

## Example 2 – Answer using Binomial Table

$$X \sim \text{Bin}(6, 0.3)$$

I.  $P(X=2) = P(X \geq 2) - P(X \geq 3) = 0.57983 - 0.25569 = 0.32414 = 0.3241$

II.  $P(X > 3) = P(X \geq 4) = 0.07047$

III.  $P(1 < X \leq 5) = P(X \geq 2) - P(X \geq 6) = 0.57983 - 0.00073 = 0.57910 = 0.5791$

## Example 3

A company owns 300 laptops. Each laptop has a 10% probability of not working. If 20 laptops are selected at random. What is the probability that

- I. 3 laptops will be broken?
- II. at least 4 will be broken?

# Binomial Distribution (Continued)

## Example 4

Fifty seeds were planted, and it is known that the probability of any seed germinating is 0.4. Assuming that the seeds have no other factors in germinating, find the following probabilities.

- I. More than 12 seeds germinate.
- II. More than 15 but fewer than 30 seeds germinate.

## Example 3 – Answer

Let  $X$ = number of Laptops which are not working and

$P$ = probability of a laptop is not working

Then  $p = 0.1$  and  $n=20$ . So,  $X \sim \text{Bin}(20, 0.1)$

Let's use Binomial Table,

$$\text{I. } P(X=3) = P(X \geq 3) - P(X \geq 4) = 0.32307 - 0.13295 = 0.19012 = 0.1901$$

$$\text{II. } P(X \geq 4) = 0.13295$$

**Homework!** Do this question using Binomial formula

# Binomial Distribution (Continued)

## Example 4 – Answer

Let  $X$ = number of germinated seeds

$P$  = probability of a seed is germinating

Then  $p = 0.4$  and  $n=50$ . So,  $X \sim \text{Bin}(50, 0.4)$

I.  $P(X > 12) = P(X \geq 13) = 0.98675$

II.  $P(15 < X < 30) = P(16 \leq X \leq 29) = P(X \geq 16) - P(X \geq 30) = 0.9045 - 0.00336 = 0.90114$

# Poisson Distribution

The **Poisson distribution** evaluates the probability of a (usually small) number of occurrences out of many opportunities in a Period of Time, Area, Volume, Weight, Distance and Other units of measurement.

## Examples

- The number of cars arriving at a service station in 1 hour. (The interval of time is 1 hour.)
- The number of flaws in a bolt of cloth. (The specific region is a bolt of cloth.)
- The number of accidents in 1 day on a particular stretch of highway. (The interval is defined by both time, 1 day, and space, the particular stretch of highway.)
- Number of misprint per page of text

# Poisson Distribution (Continued)

The probability of exactly  $x$  occurrences in an interval is,

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Where,

$\lambda$  = mean number of occurrences in the given unit of time,  
area, volume, etc.

$e = 2.71828\dots$

# Poisson Distribution (Continued)

## Example 1

Births in a hospital occur randomly at an average rate of 1.8 births per hour.

- I. What is the probability of observing 4 births in a given hour at the hospital?
- II. What about the probability of observing more than or equal to 2 births in a given hour at the hospital?
- III. What is the probability of observing 5 births in a given 2 hour interval?
- IV. What is the expected number of births per hour and calculate it's variance.

# Poisson Distribution (Continued)

## Example 1 – Answer

Let  $X$ =number of births per hour average births per hour ( $\lambda$ ) = 1.8.

(i)  $Pr(X = 4) = \frac{e^{-1.8} \times 1.8^4}{4!} = 0.0723$

(ii)

$$\begin{aligned} Pr(X \geq 2) &= 1 - Pr(X < 2) \\ &= 1 - [Pr(X = 0) + Pr(X = 1)] \\ &= 1 - [\frac{e^{-1.8} \times 1.8^0}{0!} + \frac{e^{-1.8} \times 1.8^1}{1!}] \\ &= 1 - (0.16529 + 0.29753) \\ &= 0.537 \end{aligned}$$

(iii) Average rate of births per 2 hours =  $1.8 \times 2 = 3.6$

$$P(X = 5) = \frac{e^{-3.6} \times 3.6^5}{5!} = 0.1377$$

(iv)  $\lambda = 1.8$

$E(X) = \lambda = 1.8$  and  $V(X) = \lambda = 1.8$

# Poisson Distribution (Continued)

## Example 1 – Answer using Poisson Table

I.  $P(X=4) = P(X \geq 4) - P(X \geq 5) = 0.10871 - 0.03641 = 0.0723$

II.  $P(X \geq 2) = 0.53716 = 0.5372$

III. Let average births per 2 hours ( $\lambda$ ) =  $2 * 1.8 = 3.6$  and  $Y =$  number of births per 2 hour

Then  $P(Y=5) = P(X \geq 5) - P(X \geq 6) = 0.29356 - 0.15588 = 0.13768 = 0.1377$

## Example 2

Suppose there is a disease, whose average incidence is 2 per million people. What is the probability that a city of 1 million people has at least twice the average incidence.

# Poisson Distribution (Continued)

## Example 2 - Answer

$$\lambda = \text{Average incidents per million} = 2$$

Let  $X$  = number of students per million. Then,

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + p(X = 3)] \\ &= 1 - \left[ \frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-2}}{1!} + \frac{2^2 e^{-2}}{2!} + \frac{2^3 e^{-2}}{3!} \right] \\ &= 1 - [e^{-2}(1 + 2 + 2 + 4/3)] \\ &= 1 - 0.857 \\ &= 0.143 \end{aligned}$$

## Example 2 – Answer using Poisson Table

$$P(X \geq 4) = 0.14288 = 0.143$$

# Poisson approximation to the binomial distribution

Note that if  $X$  has a binomial distribution with parameters  $n$  and  $p$  that  $E(X) = np$  and  $V(X) = np(1 - p)$ , now if  $p$  is small then  $(1 - p)$  is close to one and  $np(1 - p) \approx np$ . This suggest that if  $p$  is small we may be able to approximate  $X$  by a Poisson random variable with mean  $np$ . So long as  $p$  is small ( $p < 0.1$ ) and  $n$  is large ( $n > 50$ ) a binomially distributed random variable is well approximated by a Poisson random variable of mean  $np$ .

If  $n > 50$  and  $p < 0.1$   $X \sim Bin(n, p)$  then  $X \approx Poi(np)$

# Poisson approximation to the binomial distribution (continued)

$$X \sim \text{Bin}(n=100, p=0.01)$$

Then from the Binomial table

$$\begin{aligned}P(X \geq 1) &= 0.63397 \\P(X \geq 2) &= 0.26424 \\P(X \geq 3) &= 0.07937\end{aligned}$$

$$X \approx \text{Poi}(\lambda = 1)$$

Then from the Poisson table

$$\begin{aligned}P(X \geq 1) &= 0.63212 \\P(X \geq 2) &= 0.26424 \\P(X \geq 3) &= 0.08030\end{aligned}$$

## Example 1

A manufacturer claims that a newly-designed computer chip has a 1% chance of failure because of overheating. To test their claim, a sample of 120 chips are tested. What is the probability that at least two chips fail on testing?

# Poisson approximation to the binomial distribution (continued)

## Example 1 – Answer

Let  $X$ = number of defective chips then  $X \sim Bin(120, 0.01)$

Since,  $n = 120$  is large and  $p = 0.01$  then  $np = 120 * 0.01$  and  $X \approx Poi(1.2)$

$$P(X \geq 2) = 0.33737 = 0.3374$$

## Example 2

The probability that a car has defective gearbox is 0.02. If it was checked the gearboxes of 140 cars find the probability that it found,

- (a) 2 defectives
- (b) more than 5 defectives
- (c) fewer than 4 defectives

# Poisson approximation to the binomial distribution (continued)

## Example 1 – Answer

Let  $X$  be the number of defective gearboxes that it found. Then  $X$  has a binomial distribution with  $n=140$  and  $p=0.02$ . Since  $n$  is large and  $p$  is small a Poisson random variable with mean  $\lambda = np = 2.8$  will give a good approximation to  $X$ .

$$(a) P(X = 2) = P(X \geq 2) - P(X \geq 3) = 0.76892 - 0.53055 = 0.2384$$

$$(b) P(X > 5) = P(X \geq 6) = 0.0651$$

$$(c) P(X < 4) = 1 - P(X \geq 4) = 1 - 0.30806 = 0.6919$$

# **6. CONTINUOUS PROBABILITY DISTRIBUTIONS**

## **[IT2110]**

*By Department of Mathematics and Statistics*

*Faculty of Humanities and Sciences*

# Random Variables

Discrete  
Random  
Variables

Continuous  
Random  
Variables

# Continuous Random Variables

- A random variable is said to be continuous, if it can take any value within a range.
- Continuous data are frequently measured in some way rather than counted.
- If  $X$  is a continuous random variable,  $Pr(X=a) = 0$  for any value of  $a$ .

# Examples

- Temperature
- Heart beat of a patient
- Rainfall
- Waiting time for a bus

# **PROBABILITY DISTRIBUTIONS**

- For continuous random variables, the probability distribution cannot be presented in a tabular form.
- Probability distribution function of a continuous random variable is known as probability density function (*pdf*).
- The area under the p.d.f. gives probability values.

# PDF - DEFINITION

- The function  $f_X(x)$  is a probability density function for the continuous random variable X, defined over the set of real numbers ( $\mathbb{R}$ ), if
  - $f_X(x) \geq 0$ , for all  $x \in R$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $Pr(a < X < b) = \int_a^b f_X(x)dx$

# Properties

- Let  $X$  be a continuous random variable with a p.d.f. ( $f_X(x)$ ), defined over the set of real numbers ( $\mathbb{R}$ ).
  - The c.d.f.  $F_X(x) = \Pr(X \leq x) = \int_{-\infty}^x f_X(x)dx$
  - $E[g(x)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$
  - $V[g(x)] = E[g(x)^2] - \{E[g(x)]\}^2$

# PDF - Example

Suppose that the error in the reaction temperature, in  $^{\circ}\text{C}$ , for a controlled laboratory experiment is a continuous random variable  $X$  having the probability density function,

$$f_X(x) = \begin{cases} cx^2 & -1 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

- 1) Find the value of  $c$
- 2) Find  $\Pr(0 < X \leq 1)$ .
- 3) Find the expected value and the variance.
- 4) Find the c.d.f.

# Continuous Probability Distributions

*Continuous  
Probability  
Distributions*

*Exponential  
Distribution*

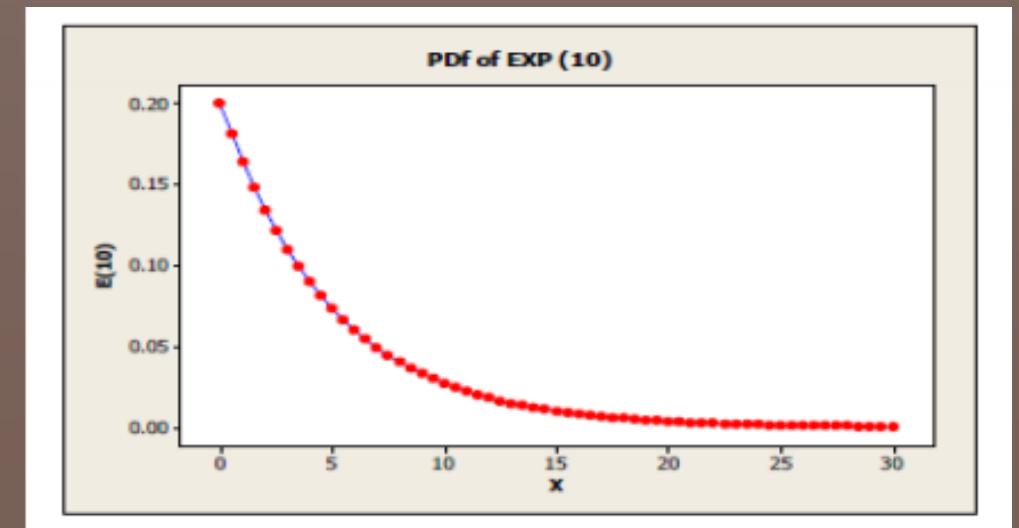
*Normal  
Distribution*

# Continuous Distributions

Exponential Distribution	Normal Distribution / Gaussian Distribution
The distribution is usually used to <b>model life times</b> . (There is a link to the Poisson distribution)	This is most commonly used distribution. This is bell shaped distribution and perfectly symmetric around $\mu$ .
$X \sim Exp(\lambda)$	$X \sim N(\mu, \sigma^2)$
$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty$
$E(X) = 1/\lambda$	$E(X) = \mu$
$V(X) = 1/\lambda^2$	$V(X) = \sigma^2$

# Exponential Distribution

- Widely used in waiting line (or queuing) theory to model the length of time between arrivals in process.
- Examples: duration between two customers at Bank ATMs, To model patients entering to an accident ward.



# Exponential Distribution - Example

- 1) The time, in hours, during which an electrical generator is operational is a random variable that follows an exponential distribution with a mean of 160. What is the probability that a generator of this type will be operational for,
  - a) Less than 40 hours?
  - b) Between 60 and 160 hours?
  - c) More than 200 hours?

# Standard Normal Distribution

- Normal distribution with  $\mu=0$  and  $\sigma^2=1$  is known as the Standard Normal Distribution.
- Evaluating probabilities with Normal requires complex integration.
- To simplify the procedure, statistical tables are defined.
- But, tables for each combination of  $\mu$  and  $\sigma^2$  cannot be created.
- So, tables are only for the standard normal distribution.

# Normal $\longrightarrow$ Standard Normal

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

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

# Normal Distribution - Examples

- 1) For  $Z \sim N(0, 1)$ , calculate  $Pr(Z \geq 1.13)$ .
- 2) For  $X \sim N(5, 4)$ , calculate  $Pr(-2.5 < X < 1.13)$ .
- 3) The actual marks for FCS of Metro students revealed that they were normally distributed with a mean mark of 45 and a standard deviation of 22. What is the probability that a randomly chosen student will pass? (Assume that pass mark is 45)

# Approximating Binomial Probabilities

## Normal Distribution

- For  $X \sim \text{Bin}(n, p)$  this approximation can be used if  $n$  is large and  $p$  is moderate.
- A general rule can be defined as,  $np$  and  $n(1 - p)$  is greater than 5.
- Can be approximated with a r.v. with a distribution  $N(np, np(1 - p))$ .
- A continuity correction is needed because a discrete distribution is approximated with a continuous distribution.

# Continuity Correction

- If  $X \sim Bin(n, p)$  is approximated with a r.v.  $Y \sim N(np, np(1 - p))$ ,
  - $Pr(X \leq a) = P(Y < a+0.5)$
  - $Pr(X \geq a) = P(Y > a-0.5)$
  - $Pr(X < a) = P(Y < a-0.5)$
  - $Pr(X > a) = P(Y > a+0.5)$
  - $Pr(X = a) = P(a-0.5 < Y < a+0.5)$

# Example

Suppose that a sample of  $n = 1,600$  tires of the same type are obtained at random from an ongoing production process in which 8% of all such tires produced are defective. What is the probability that in such a sample 150 or fewer tires will be defective?

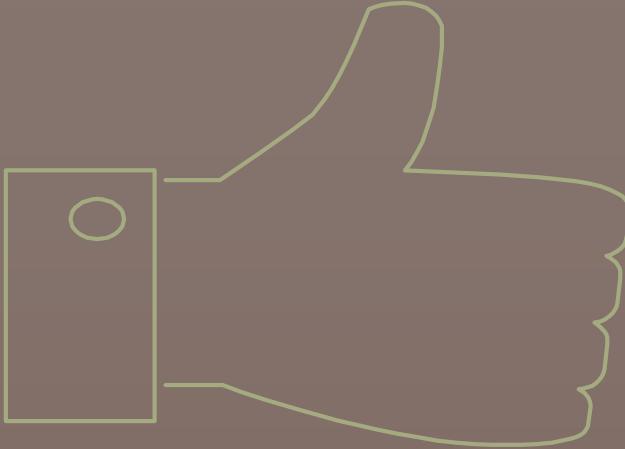
# Approximating Poisson Probabilities

Normal Distribution

- If  $X \sim \text{Poisson}(\lambda)$  then if  $\lambda$  is greater than 20, the approximation can be used.
- Can be approximated with a r.v. with a distribution  $N(\lambda, \lambda)$ .
- A continuity correction is needed because a discrete distribution is approximated with a continuous distribution (just as in the case of the Binomial to Normal approximation).

# Example

The annual number of earthquakes registering at least 2.5 on the Richter Scale and having an epicenter within 40 miles of down town Memphis follows a Poisson distribution with mean 22.5. What is the probability that at least 25 such earthquakes will strike next year?



# THANKS!

Any questions?

# 7. SAMPLING DISTRIBUTIONS

## [IT2110]

*By Department of Mathematics  
and Statistics  
Faculty of Humanities and Sciences*





# **Simple Random Sampling (SRS)**

“  
Each unit in the population ( $N$ ) has **same chance** of being selected to the sample ( $n$ ).  
“

**“ A SRS is a set of  
independent and  
identically  
distributed (iid)  
observable r.v.s.**

A close-up photograph of several yellow sticky notes scattered on a dark, textured surface. Some notes have faint, illegible handwriting on them.

# Statistic

2

**“** A function of observable r.v.s that does not depend on any unknown parameters is called a statistic.  
Eg: Sample Mean



# Sampling Distributions

3

“  
■ The probability distribution of a statistic is known as a **sampling distribution**.  
■



# **Sampling Distribution of the Mean**

# Sampling Distribution of the Mean

- Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a population with mean  $\mu$  and variance  $\sigma^2$ . Then,

$$E(\bar{X}) = \mu$$

$$V(\bar{X}) = \sigma^2/n$$

- The standard deviation of a sampling distribution of a statistic is called the **Standard Error (SE)**.
- Although the r.v.s were identically distributed, a specific distribution type was not needed.

# Sampling Distribution of the Mean

- Let  $X_1, X_2, \dots, X_n$  be a random sample of size  $n$  from a **Normal** population with mean  $\mu$  and variance  $\sigma^2$ . Then,

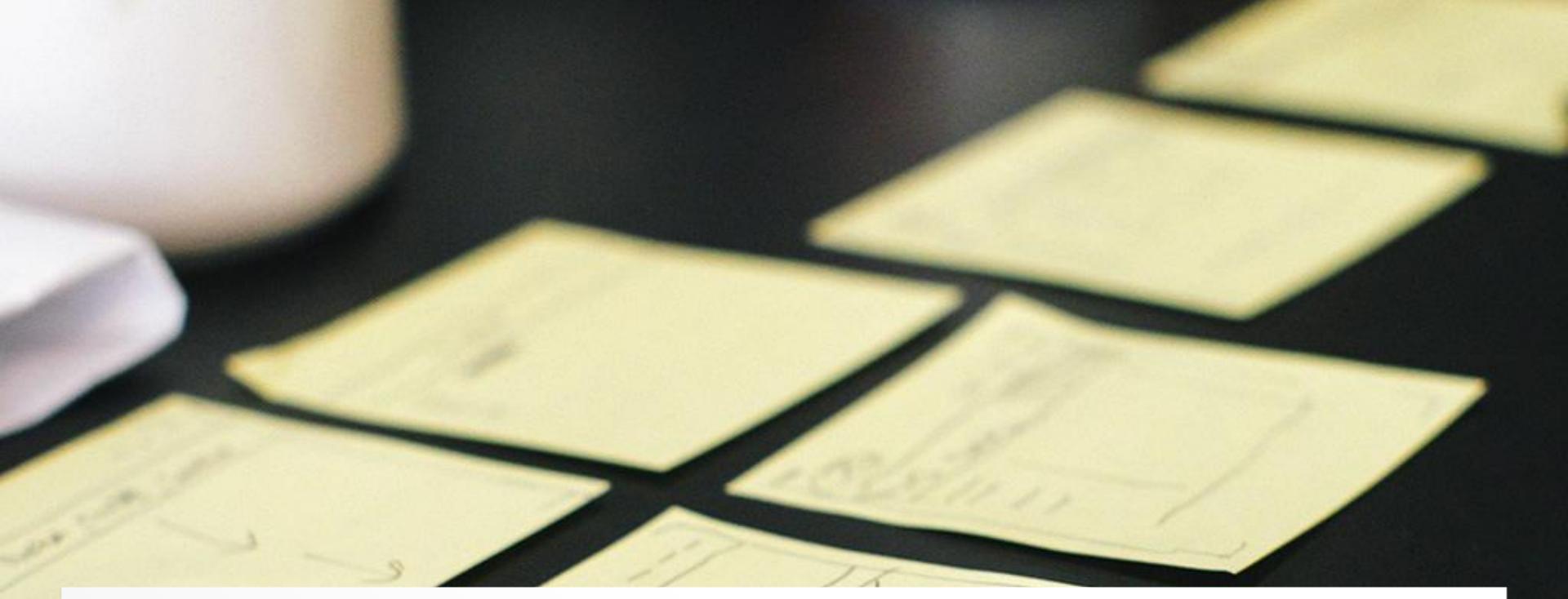
$$\bar{X} \sim N(\mu, \sigma^2/n)$$

# Examples

- 1) A particular brand of drink has an average of 12 ounces per can. As a result of randomness, there will be small variations in how much liquid each bottle really contains. It has been observed that the amount of liquid in these bottles is normally distributed with  $\sigma = 0.8$  ounce. A sample of 10 bottles of this brand of soda is randomly selected from a large lot of bottles, and the amount of liquid, in ounces, is measured in each. Find the probability that the sample mean will be within 0.5 ounce of 12 ounces.

# Examples

- 2) A company that manufactures cars claims that the gas mileage for its new line of hybrid cars, on the average, is 60 miles per gallon (mpg) with a standard deviation of 4 mpg. A random sample of 16 cars yielded a mean of 57 miles per gallon. If the company's claim is correct, what is the probability that the sample mean is less than or equal to 57 mpg? What assumptions did you make?



# **Central Limit Theorem (CLT)**

# Central Limit Theorem

- Let  $X_1, X_2, \dots, X_n$  be a large random sample of size  $n$  from a population with mean  $\mu$  and variance  $\sigma^2$ . Then,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

- A rule of thumb is that the sample size  $n$  must be at least **30**.
- Central Limit Theorem can be applied regardless of the distribution of the population.

# **Thanks!**

## **Any questions?**

# **8. STATISTICAL INFERENCE**

## **[IT2110]**

*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences, SLIIT*

- In most researches, we collect data through a sample survey over a census.
- Statistical inference is used when sample survey is conducted over a census.
- ***Inference:*** A conclusion reached on the basis of evidence and reasoning.

- Oxford University Press -

- ***Statistical Inference:*** Drawing conclusions about population parameters by using sample statistics.

## Statistical Inference

### Parameter Estimation

#### Point Estimation

#### Interval Estimation

### Hypothesis Testing

# PARAMETER ESTIMATION

research  
samples  
estimates  
two  
dependent  
groups  
level  
statistic  
techniques  
decision  
two-sided  
comparison  
dichotomous  
testing  
variable  
rule  
Simple  
Type  
interval  
Testing  
paired  
null  
interpret  
continuous  
outcome  
proportions  
Example  
errors  
General  
Estimate  
One-sided  
confidence  
values  
independent  
dependent  
test  
matched  
interpret

Hypothesis

# Parameter Estimation

- In distribution theory we assumed that distribution parameters are known.
- But practically they should be found or estimated.
- If estimated parameters are wrong, all calculated probabilities will be inaccurate.
- Estimation can be done in two methods.
  - *Point estimation*
  - *Interval estimation*

# Parameter Estimation

- Point estimation gives a single estimated value for the parameter.
- Interval estimation gives a range of values (interval) as the estimate.
- There are many point and interval estimation methods with their own criteria for use.
- Some interval estimates will be discussed later in this chapter.

# HYPOTHESIS TESTING

research means testing proportions Hypothesis Approach dichotomous Type statistic level groups decision dependent two-sided comparison variable testing significance rule Simple interval Testing hypothesis values independent null interpret matched test continuous outcome proportions Example errors General confidence One-sided Estimate continuous paired interpret

# Hypothesis Testing

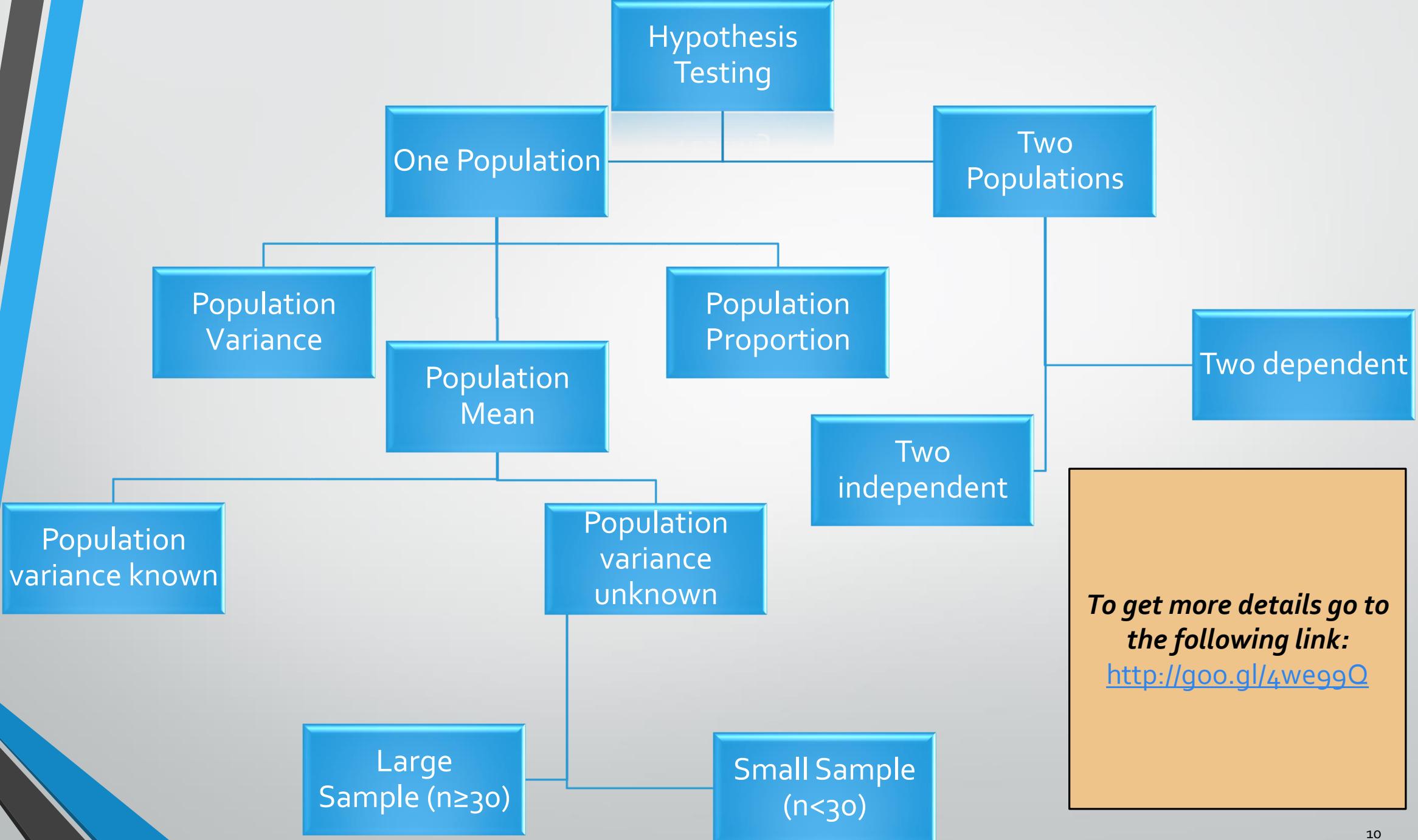
- ***Hypothesis:*** A supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation.

*-Oxford University Press-*

- Hypothesis testing is all about checking whether assumptions (research hypothesis) are correct.
- These assumption should be regarding population parameters.

## Major Steps under Hypothesis Testing

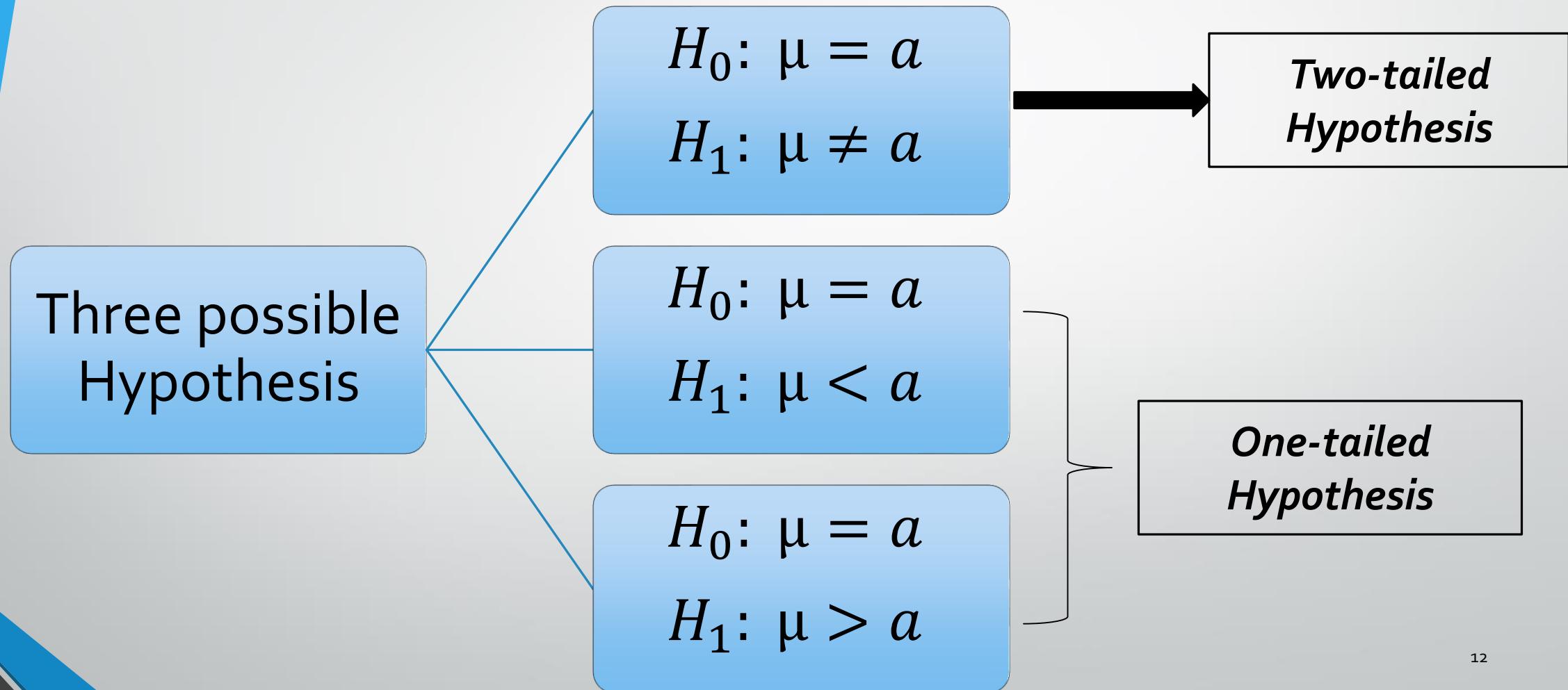
1. Define the hypothesis ( $H_0$  &  $H_1$ )
  2. Test statistic and its distribution
  3. Define the significance level ( $\alpha$ )
  4. Define the rejection region.
  5. Conduct the test (Calculate test statistic value)
  6. Conclusion
- 
- There are various cases under hypothesis testing. The test statistic that you should use depends on the case.
  - In this session, we will discuss the hypothesis testing for one population mean.



# Defining Hypothesis

- The assumption should be clearly stated in order to test.
- Two statements, null hypothesis ( $H_0$ ) and an alternative hypothesis ( $H_1$  or  $H_a$ ) are used for that.
- $H_0$  and  $H_1$  can be considered as opposites of each other.
- The statement with the equal (=) should always come to  $H_0$ . Usually if a claim is made, it is selected for  $H_1$ .

# Defining Hypothesis



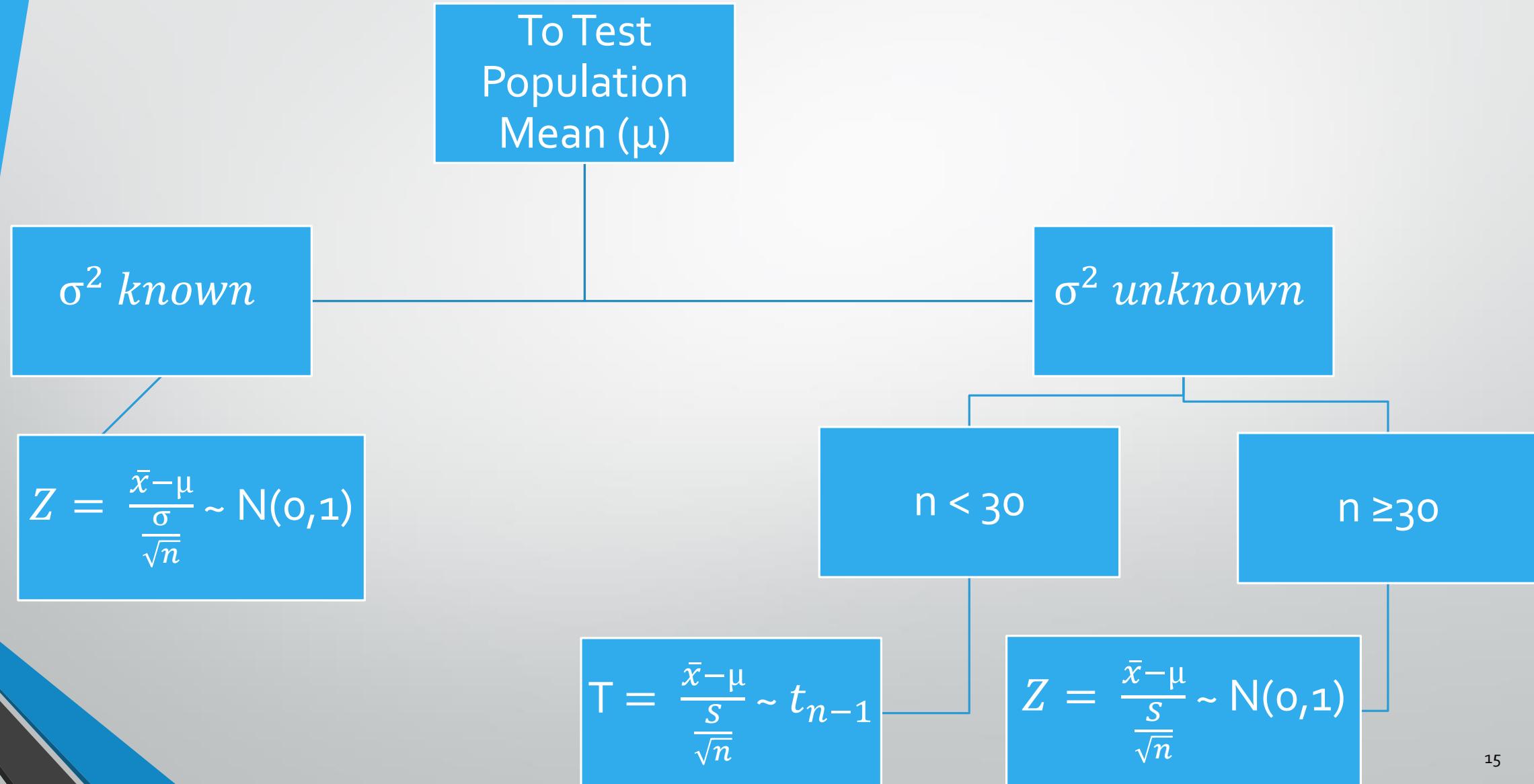
# Examples

- 1) In a coin tossing experiment, it should be found whether
  - a) it's fair coin or not.
  - b) it's biased in favor of heads.
  - c) it's biased in favor of tails.
- 2) A company that manufactures cars claims that the gas mileage for its new line of hybrid cars, on the average, is 60 miles per gallon (mpg) with a standard deviation of 4 mpg. It was also found out that the mpg was normally distributed. A random sample of 16 cars yielded a mean of 57 miles per gallon. Is the company's claim about the mean gas mileage per gallon of its cars, correct?

# Test Statistic

- *Recap:* A function of observable r.v.s that does not depend on any unknown parameters is called a statistic.
- A test statistic is a quantity associated with the sample.
- The test statistic will depend on the *parameter of interest* as well as the *characteristics of the population*.
- We assume that the assumption ( $H_0$ ) is correct and find a sampling distribution for the test statistic.

# Test Statistic & Distribution



# Test Statistic [For $\mu$ - When $\sigma^2$ known]

- **Recap:** Let  $X_1, \dots, X_n$  be a random sample of size  $n$  from a Normal population with mean  $\mu$  and variance  $\sigma^2$ . Then,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Then,

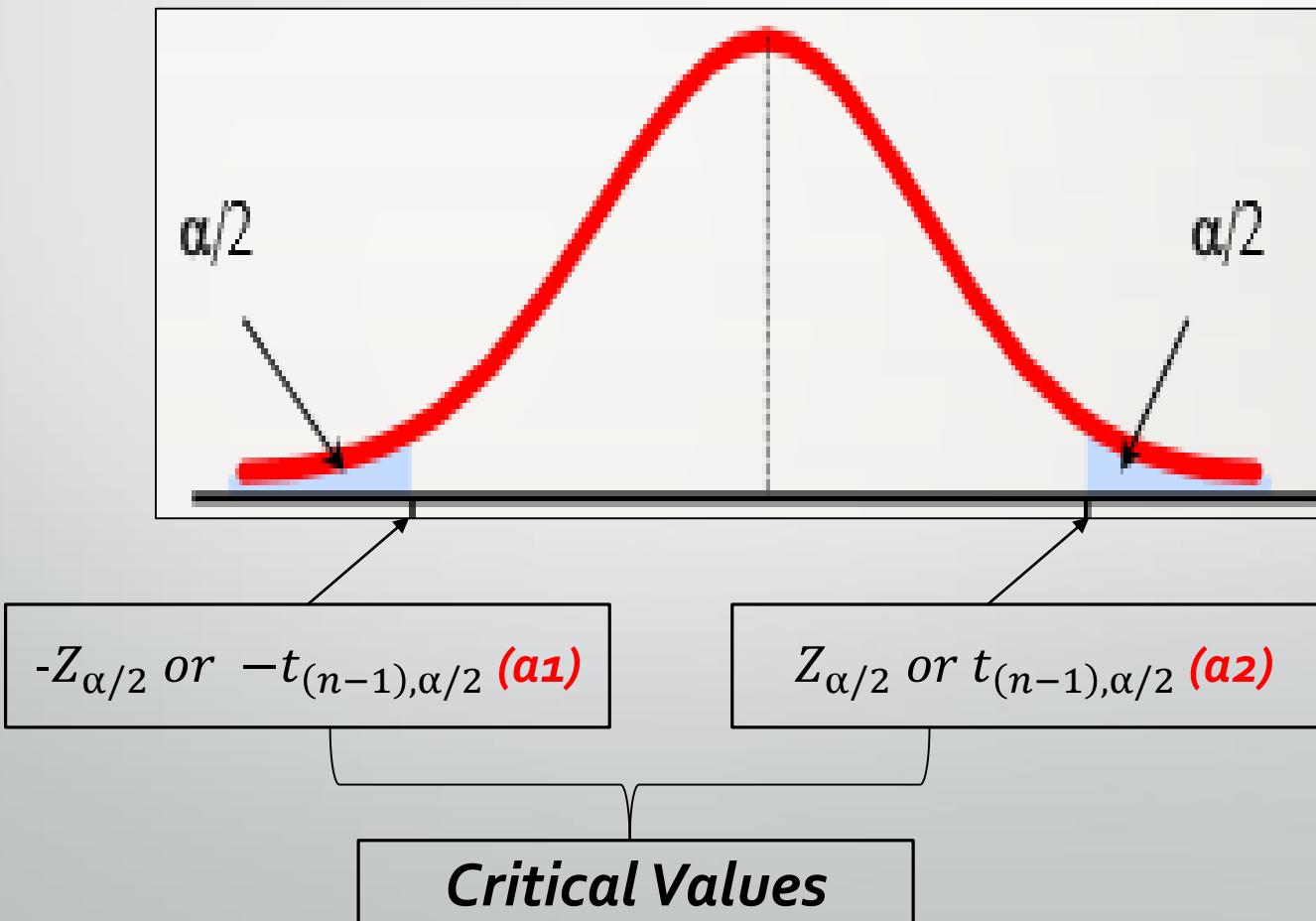
$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- If the hypothesis is,  $H_0: \mu = \mu_0$  vs.  $H_1: \mu \neq \mu_0$ , then under  $H_0$ ,

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

# Rejection Region [For $\mu$ ]

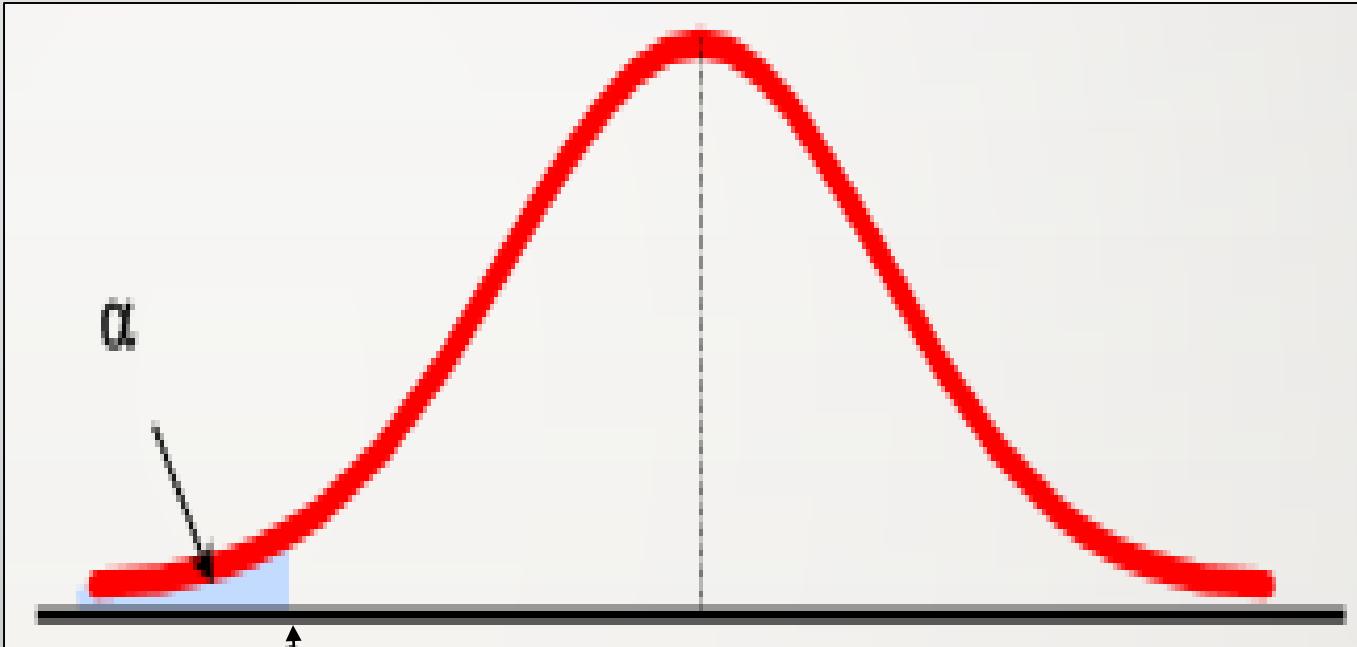
For a two-tailed hypothesis



Reject  $H_0$  if  $Z_{cal} \geq a2$   
OR if  $Z_{cal} \leq a1$

## For a one-tailed hypothesis

$$H_0 : \mu = a$$
$$H_1 : \mu < a$$



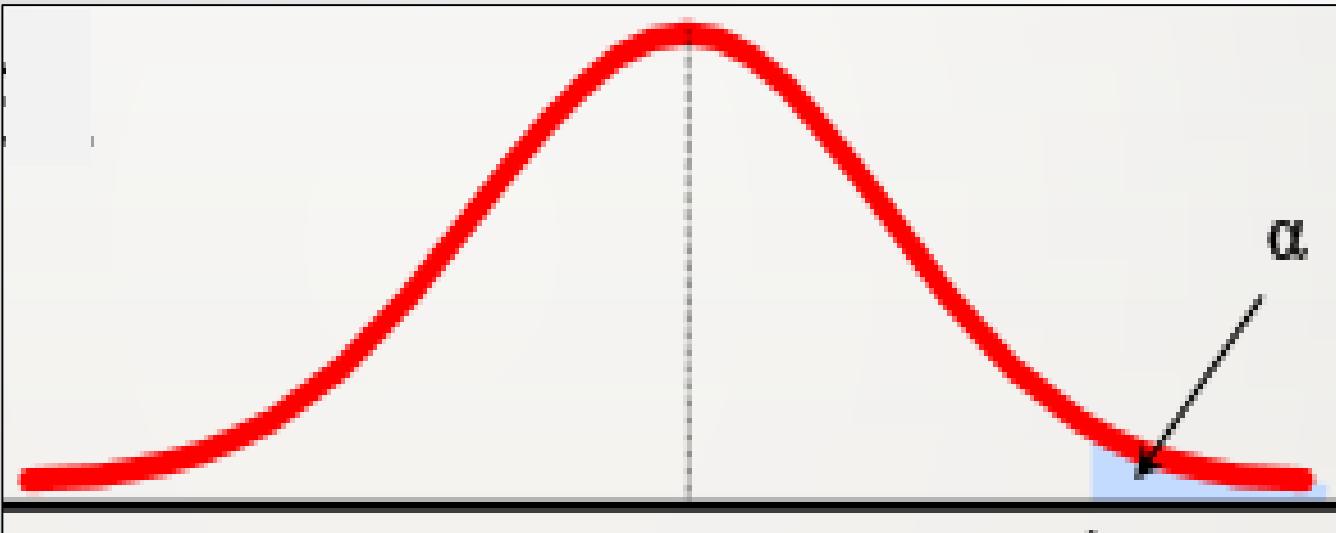
$-Z_\alpha$  or  $-t_{(n-1),\alpha}$  ( $a1$ )

**Critical Value**

Reject  $H_0$  if  $Z_{cal} < a1$

For a one-tailed hypothesis

$$H_0 : \mu = a$$
$$H_1 : \mu > a$$



**Critical Value**

$Z_\alpha$  or  $t_{(n-1),\alpha}$  ( $\alpha 1$ )

Reject  $H_0$  if  $Z_{cal} > \alpha 1$

## Example 02:

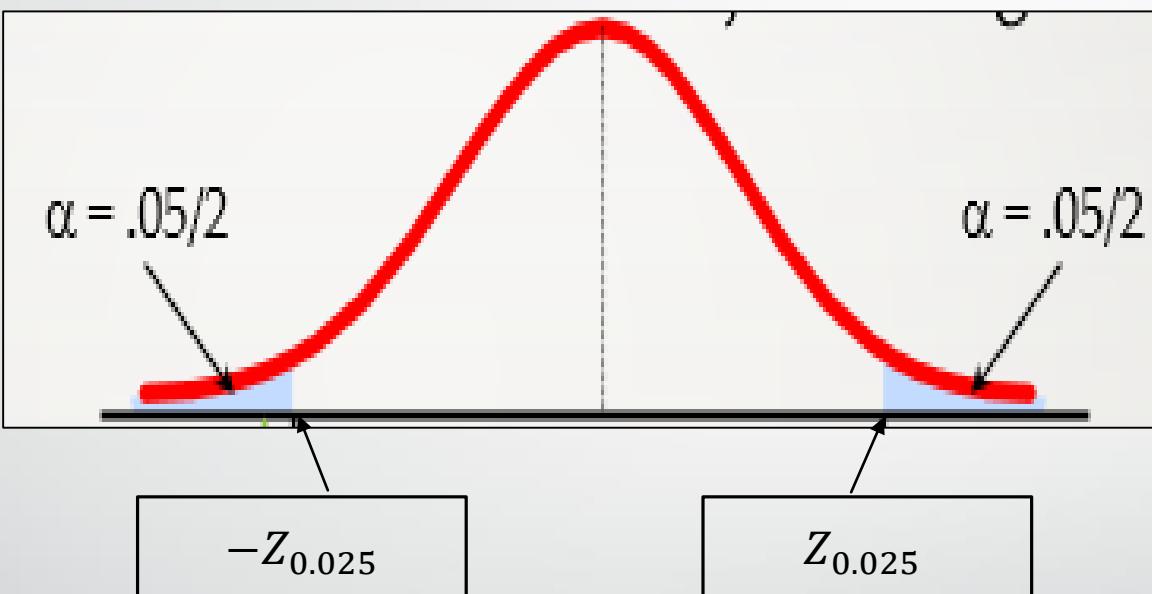
- $H_0: \mu = 60$
  - $H_1: \mu \neq 60$
- } **Two-tailed hypothesis**

- **Test Statistic:** Under  $H_0$ ,

$$Z = \frac{\bar{x} - 60}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Consider 5% level of significance.

- *Rejection Region:*



Reject  $H_0$  if  $Z_{cal} > Z_{0.025}$  **OR**  
if  $Z_{cal} < -Z_{0.025}$

$$Z_{0.025} = 1.96$$

- **Test:**

$$\bar{x} = 57, \sigma = 4 \text{ & } n = 16$$

Then,

$$Z_{Cal} = \frac{\bar{x}-60}{\frac{\sigma}{\sqrt{n}}}$$

$$Z_{Cal} = \frac{57-60}{\frac{4}{\sqrt{16}}}$$

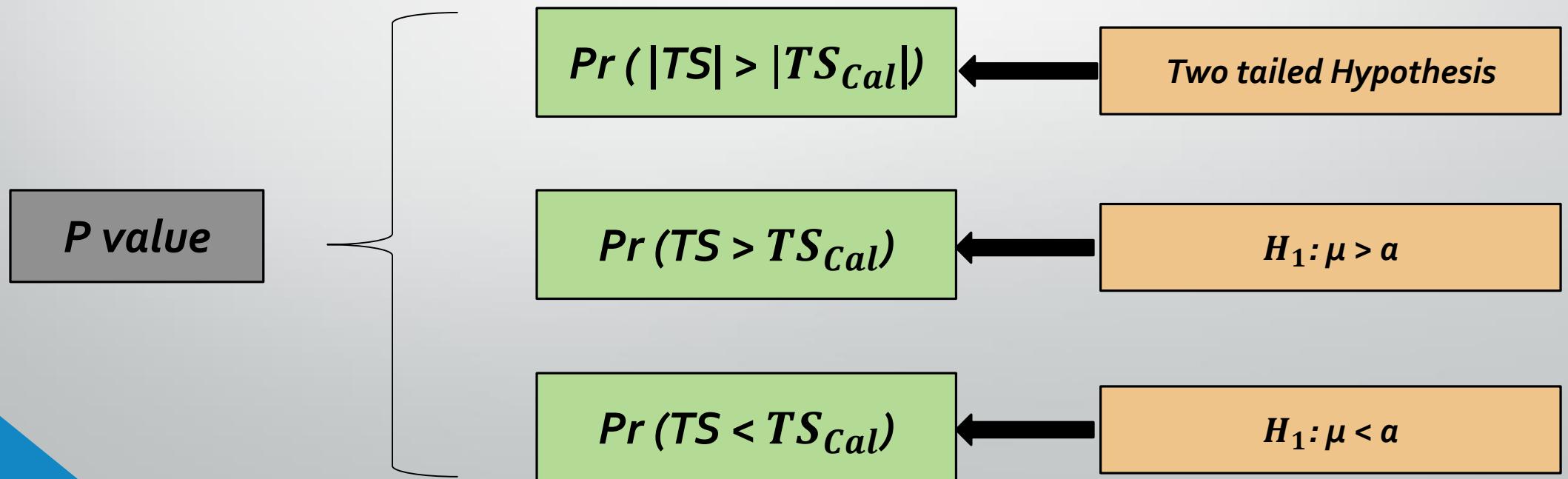
$$Z_{Cal} = -3$$

- **Conclusion:**

Since  $Z_{Cal} = -3 < -1.96 = Z_{0.025}$ , we reject  $H_0$  at 5% level of significance. Therefore, we can conclude that company's claim about the mean gas mileage per gallon of its cars is incorrect.

# P value Approach

- This is an alternative way of get the decision in hypothesis testing.
- **P value:** *The probability of obtaining a test statistic which is more extreme than observed test statistic value given when  $H_0$  is true.*



- For any test,

*If p value  $\leq$  significance level ( $\alpha$ )*  $\longrightarrow$  *Reject  $H_0$*

*If p value  $>$  significance level ( $\alpha$ )*  $\longrightarrow$  *Do not Reject  $H_0$*

- P value is a measure of the strength of evidence in the data against  $H_0$
- This is the smallest value of  $\alpha$  for which  $H_0$  can be rejected and *actual risk of committing type I error.*
- P value also known as *observed significance level.*

# Errors in Hypothesis Testing

<i>Statistical Decision</i>	<i>True State of the Null Hypothesis</i>	
	$H_0$ is True	$H_0$ is False
Reject $H_0$	Type I Error	Correct
Do not Reject $H_0$	Correct	Type II Error

$$\Pr(\text{Type I Error}) = \Pr(\text{Reject } H_0 | H_0 \text{ true}) = \alpha$$
$$\Pr(\text{Type II Error}) = \Pr(\text{Do not Reject } H_0 | H_0 \text{ false}) = \beta$$



# THANKS!

Any questions?

# **8. STATISTICAL INFERENCE (Part 2) [IT2110]**

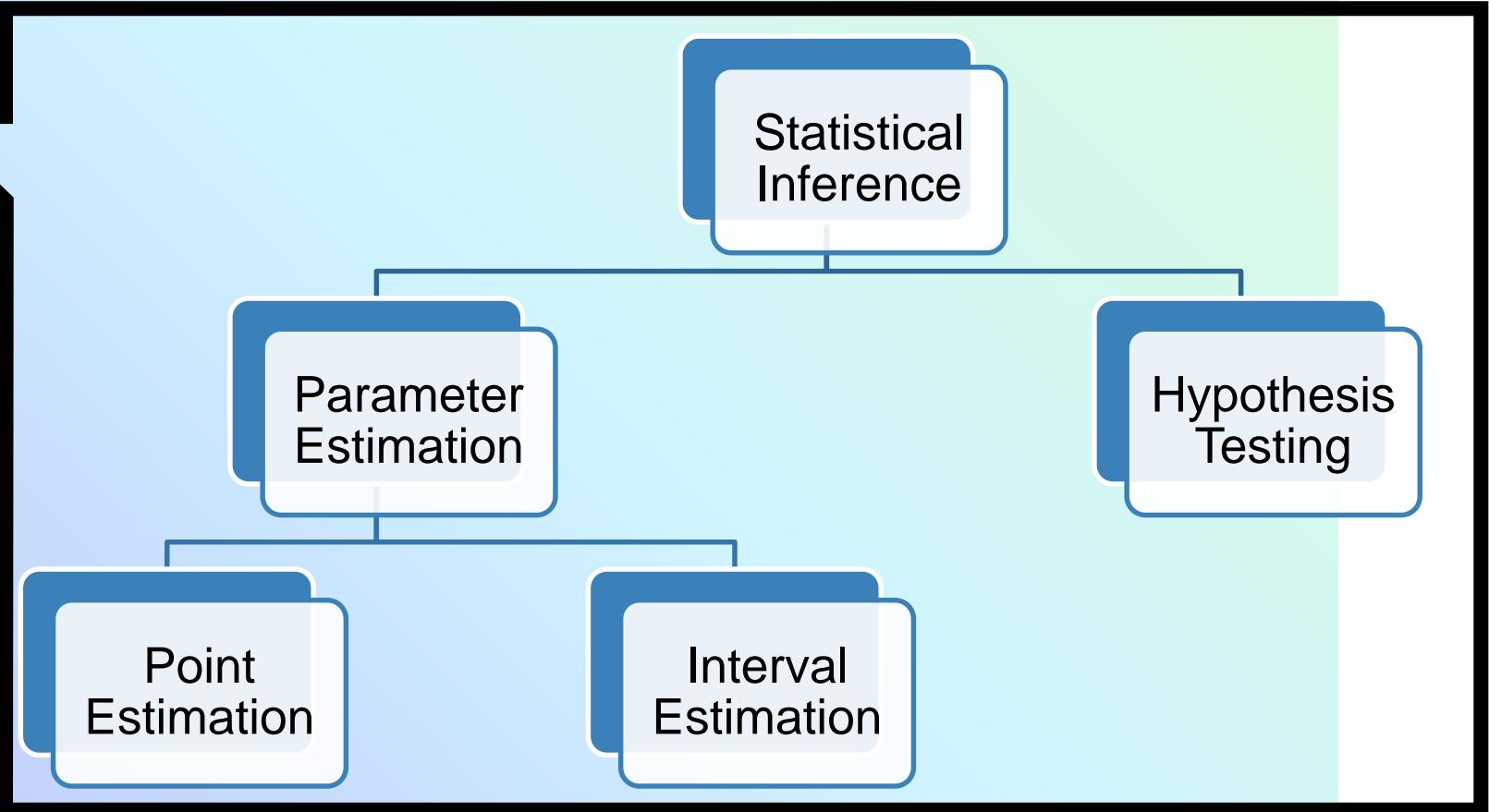
*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences, SLIIT*

## 8.2

# CONFIDENCE INTERVALS

A word cloud centered on the word "confidence". Other words include "samples", "error", "intervals", "proportions", "estimate", "difference", "parameter", "inference", "level", "interval", "margin", "standard", "point", "means", "formula", "mean", "paired", "independent", "matched", "samples", "population", and "estimate". The words are colored in shades of green, brown, and red.

confidence samples proportions estimate difference parameter intervals inference level interval margin standard point means formula mean paired independent matched samples population



# Introduction

- Estimates will differ from the true parameter values by varying amounts depending on the samples obtained.
- Point estimates do not convey any measure of reliability.

# Interval Estimation

- Interval estimation states that a ***population parameter*** is ***within two values*** ( an interval ) with a ***certain probability (Confidence Level)***.
- Interval Estimation is also known as ***Confidence Interval.***
- For a good interval estimate,
  - *The probability that the parameter is within the interval should be high.*
  - *The length of the interval should be small.*

- A confidence level for the interval should be defined first.

$$\text{Confidence Level} = 1 - \alpha$$

where  $\alpha$  is the significance level discussed in hypothesis testing.

- Let L and U be the lower and upper confidence limits for a parameter  $\theta$  based on a random sample  $X_1, \dots, X_n$ . Both L and U are functions of the sample. We can write the interval estimate of  $\theta$  as,

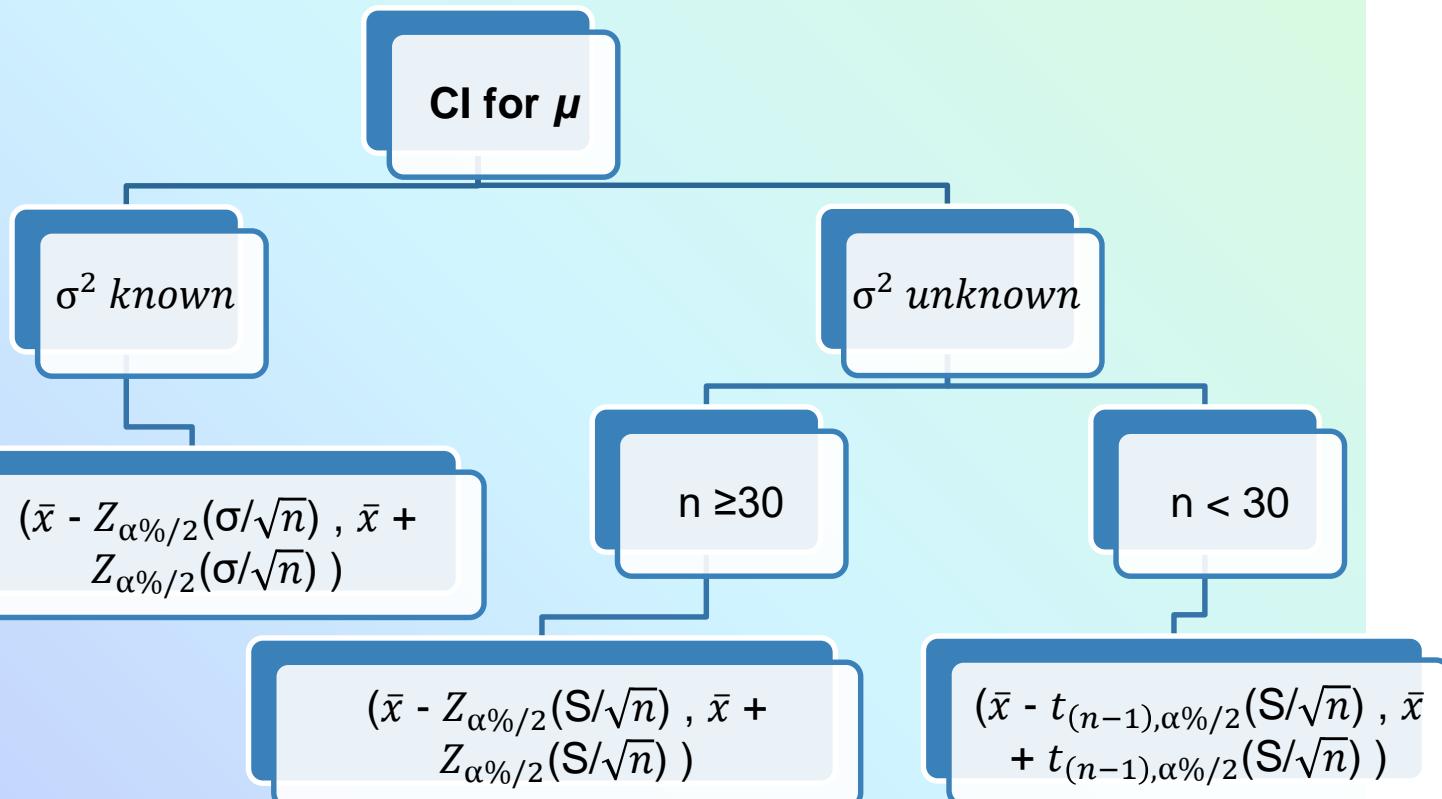
$$Pr(L \leq \theta \leq U) = 1-\alpha$$

- ***Interpretation:***

*We are  $(1 - \alpha)$  % confident that the true parameter  $\theta$  is located in the interval  $(L, U)$ .*

- In this session, we will discuss confidence intervals for population mean ( $\mu$ ) only.

# Confidence Intervals (CI)



### ***Example:***

A company that manufactures cars claims that the gas mileage for its new line of hybrid cars, has a standard deviation of 4 mpg. It was also found out that the mpg was normally distributed. A random sample of 16 cars yielded a mean of 57 miles per gallon. What is the interval estimation for the population mean at a 95% confidence level?

# Thanks!

Any questions?

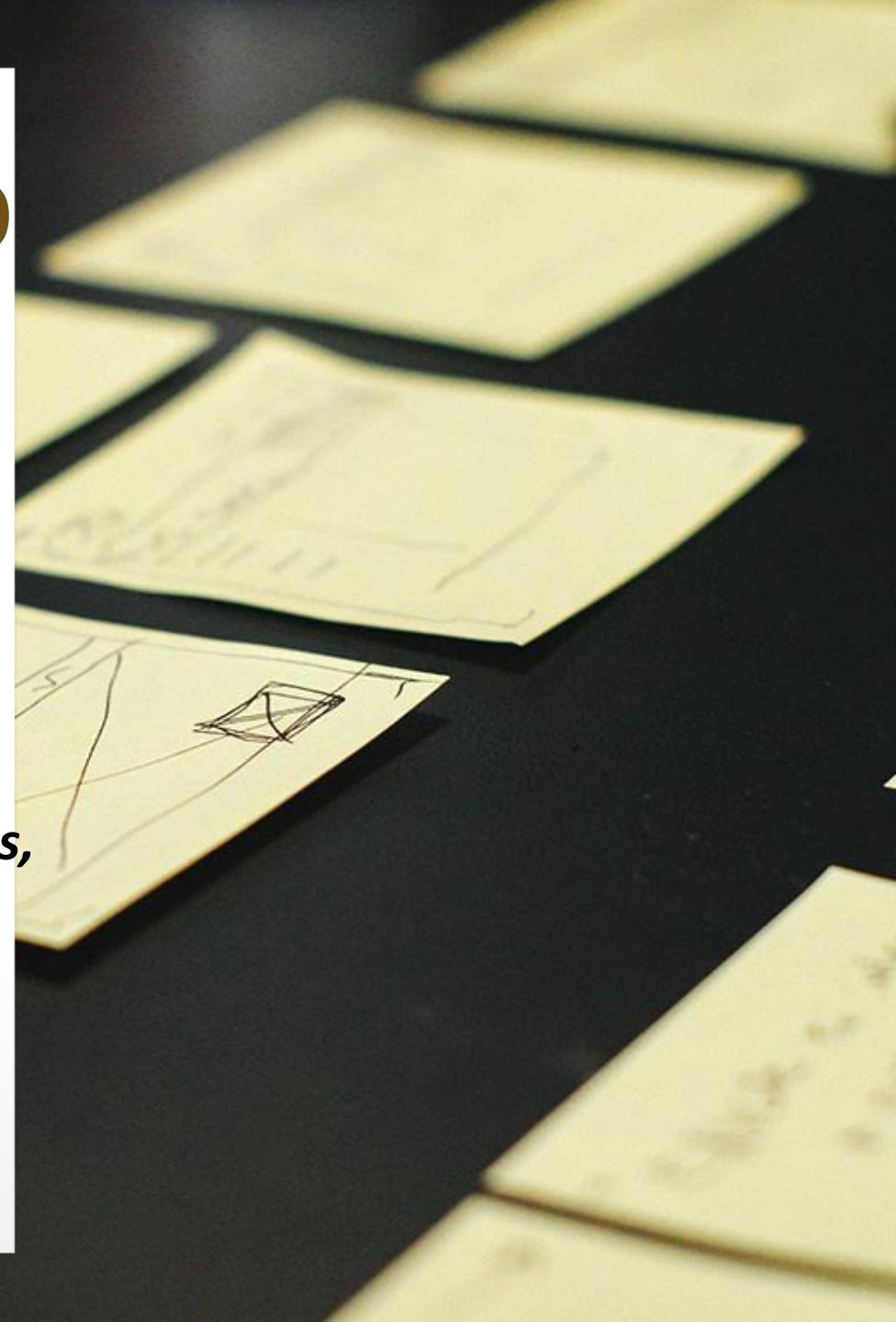


# 9. CHI SQUARED TESTS

## [IT2110]

*By Department of Mathematics  
and Statistics*

*Faculty of Humanities and Sciences,  
SLIIT*



- Chi-squared tests are used for,
  - *Discrete data*
  - *Categorical data*

# Chi Squared Tests

Goodness  
of Fit Test

Test for  
Association



# **Test for Association**

“

*Used to find  
whether two  
factors are  
independent.*

- The hypothesis for the test is,

$H_0$ : *The factors are independent.*

$H_1$ : *The factors are not independent.*

- Test Statistic,

***Under  $H_0$ ,***

$$X^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim X^2_{d.f.}$$

- $O_{ij}$  - Observed frequency for cell ij
- $E_{ij}$  - Expected frequency for cell ij
- $df = (\text{No of rows} - 1)(\text{No of columns} - 1)$

- **Reject  $H_0$ , if  $X_{cal}^2 > X_{df,\alpha\%}^2$**  (critical value)
- Test:
  - Find the **expected frequencies** for each cell.
  - Calculate **test statistic value**.
- Conclusion:

**Compare** calculated **test statistic value** with **critical value** and give the conclusion.

# Important

## Rule 01

- *All expected counts* should be *greater than 5*.

## Rule 02

- *All expected counts* should be *greater than 1 & at least 80% of the cells* should have *expected count* which is *greater than or equal to 5*.
- If not, categories can be joined.

# Example

The following table gives a classification according to religious affiliation and marital status for 500 randomly selected individuals. For  $\alpha = 1\%$ , test the null hypothesis that marital status and religious affiliation are independent.

		Religious Affiliation				
		A	B	C	D	None
Marital Status	Single	39	19	12	28	18
	Married	172	61	44	70	37



# Goodness of Fit Test

2

**“ Used to find  
whether a set of  
discrete or  
categorical data  
*follows a specified  
distribution.***

- The hypothesis for the test is,  
 $H_0$ : *The data are consistent with the specified distribution.*  
 $H_1$ : *At least one category deviates from the specified distribution.*

- Test Statistic,

***Under  $H_0$ ,***

$$X^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \sim X^2_{d.f.}$$

- $O_i$  - Observed frequency for cell i
- $E_i$  - Expected frequency for cell i

- $d.f. = \text{No of classes} - \text{No of parameters estimated} - 1$
- **Reject  $H_0$ , if  $X_{cal}^2 > X_{df,\alpha\%}^2$**  (critical value)
- Test:
  - Find the **expected frequencies** for each category.
  - Calculate **test statistic value**.
- Conclusion:

**Compare** calculated **test statistic value** with **critical value** and give the conclusion.

# Example

- 1) A die is rolled 60 times and the face values are recorded. The results are as follows.

Up Face	1	2	3	4	5	6
Frequency	8	11	5	12	15	9

Is the die balanced? Test using  $\alpha = 0.05$ .

# Example

- 2) The number of accidents in a month observed over a period of 10 years is given below.

No of accidents	0	1	2	3	4	5	6	$\geq 7$
Frequency	41	40	22	10	6	0	1	0

Is the data following a Poisson distribution?  
Test using  $\alpha = 0.05$ .

# Example

- 3) The grades of students in a class of 51 are given in the following table. Test the hypothesis that the grades are normally distributed with a mean of 75 and a standard deviation of 8. Use  $\alpha = 0.05$ .

Range	0-59.5	59.5-69.5	69.5-79.5	79.5-89.5	89.5-100
No of students	8	11	5	12	15

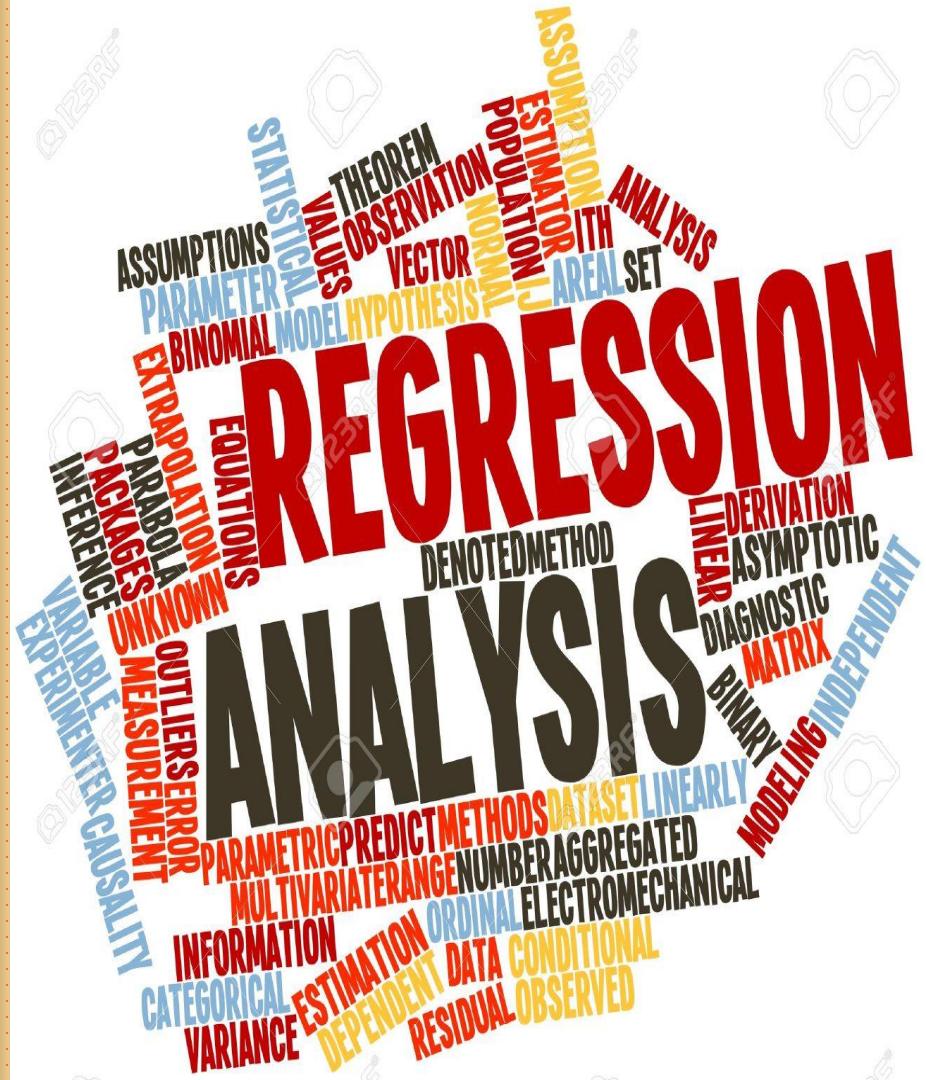
# Thanks!

**Any questions?**

# 10. REGRESSION ANALYSIS

## [IT2110]

*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences, SLIIT*



# ***Numerical Variables??***

- *Weight*
- *Height*
- *Temperature etc.*

# ***Paired Numerical Variables??***



<b>Paired Variables</b>			<b>Unpaired Variables</b>	
<b>ID_No (Females)</b>	<b>Age</b>	<b>Systolic BP</b>	<b>Age of Females</b>	<b>Systolic BP of Males</b>
001	45	151	45	149
002	25	138	25	150
003	48	143	48	138
004	37	140	37	142
005	24	136	24	139

## **How light Affects Plant Growth.**



# **Dependent Variable??**

*The variable we wish to explain*

# **Independent Variable??**

*The variable we use to explain the dependent variable*

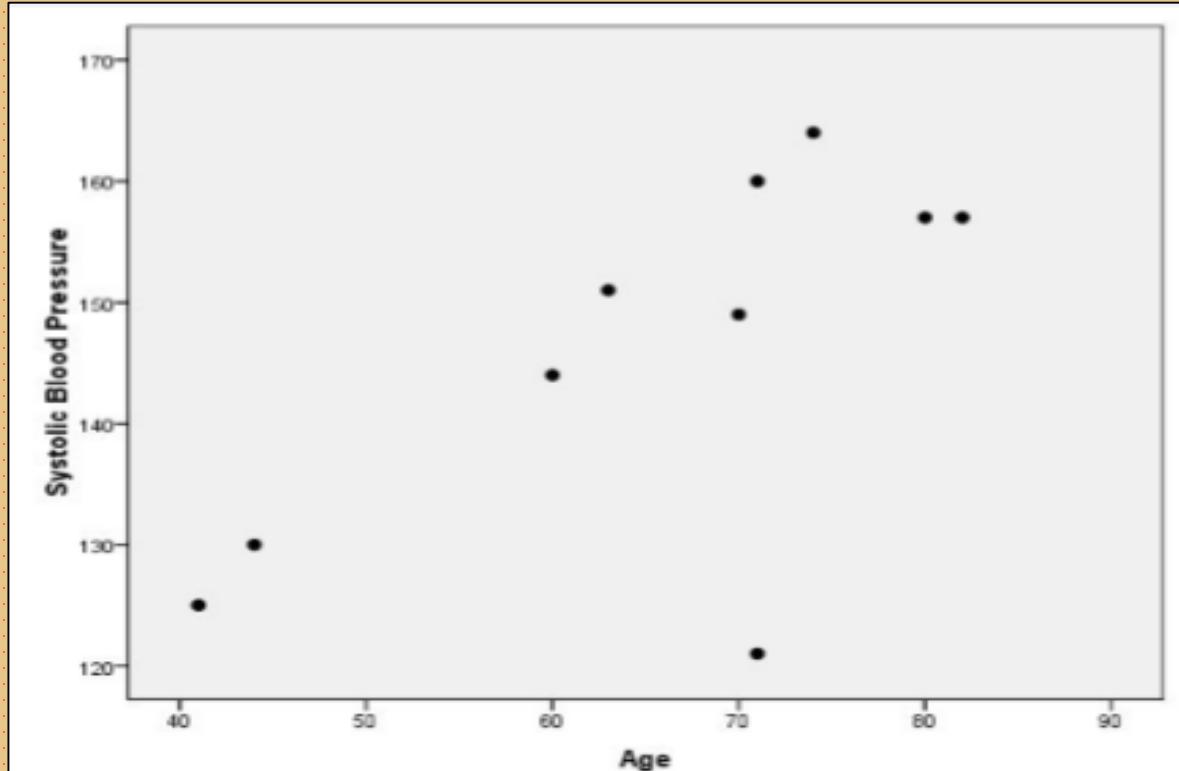
# How to identify Relationships??

- Basically, we will learn three main methods.  
They are,
  - Scatter plot (**Graphical Method**)
  - Correlation
  - Regression Analysis

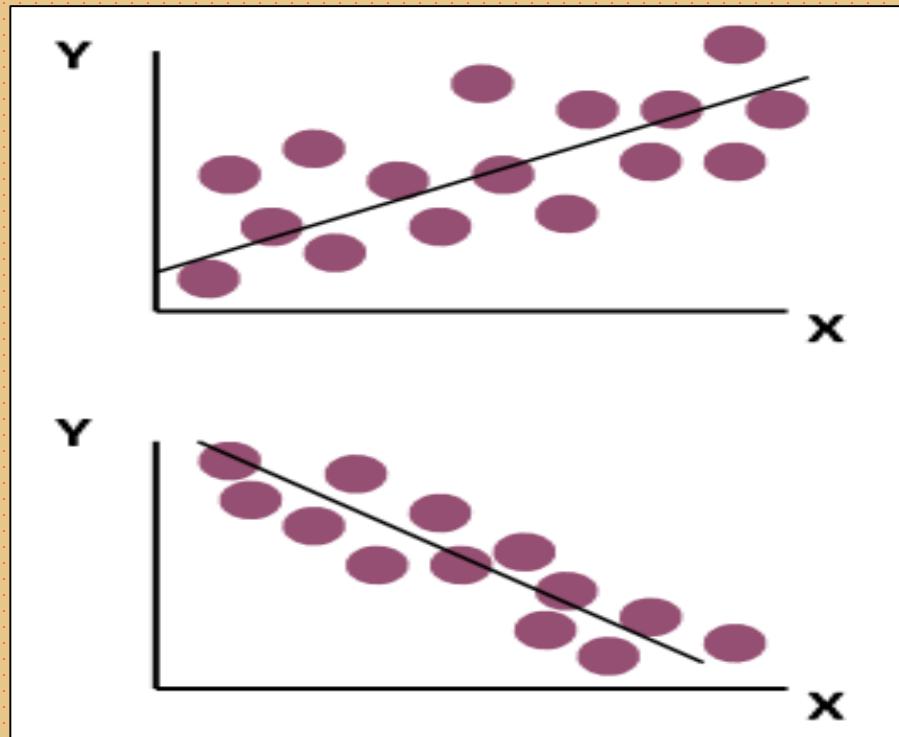


# Scatter Plot

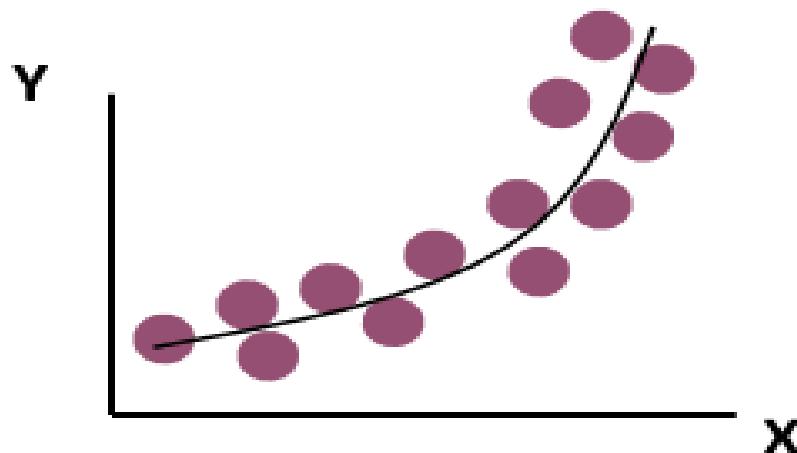
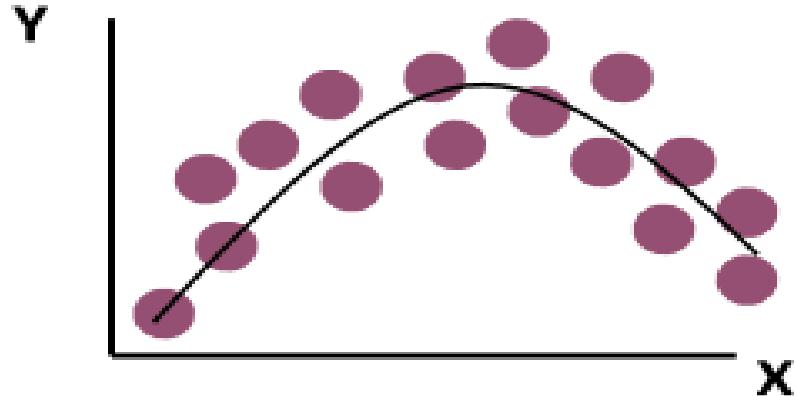
Age	Systolic BP
63	151
70	149
74	164
82	157
60	144
44	130
80	157
71	160
71	121
41	125



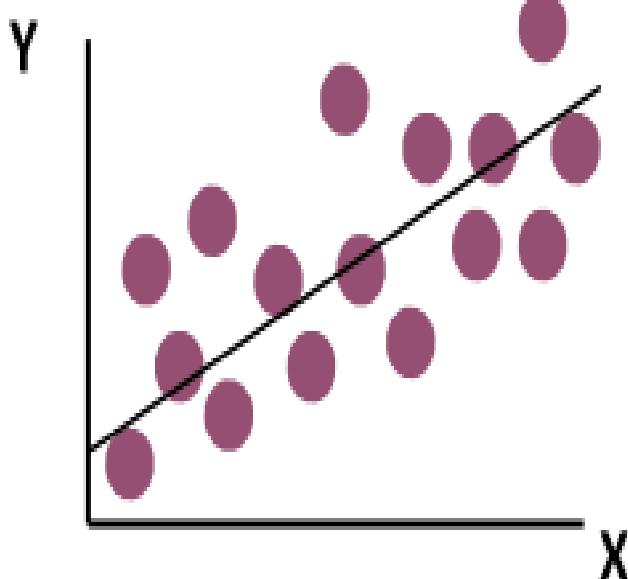
# Types of Relationships



***Linear  
Relationships***

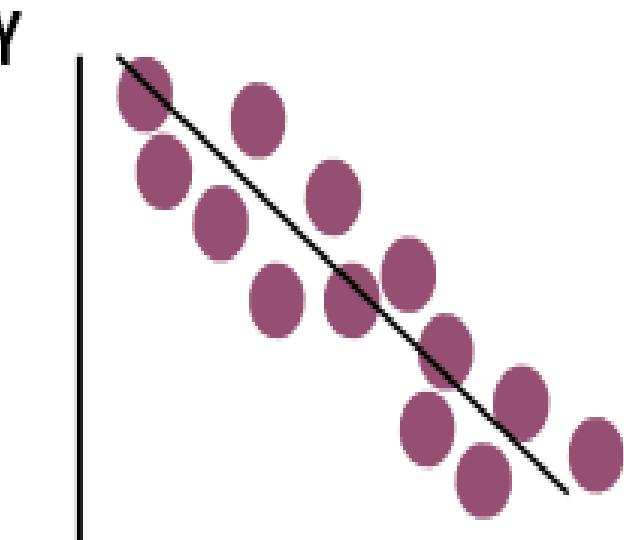


***Non-Linear  
Relationships***

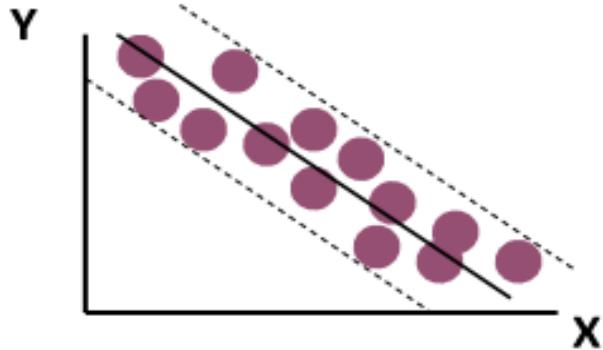
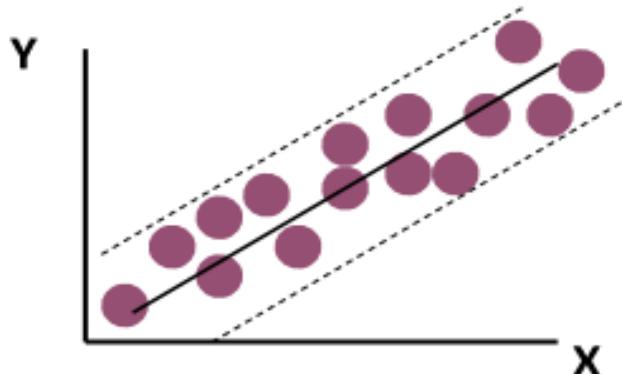


***Positive Linear  
Relationship***

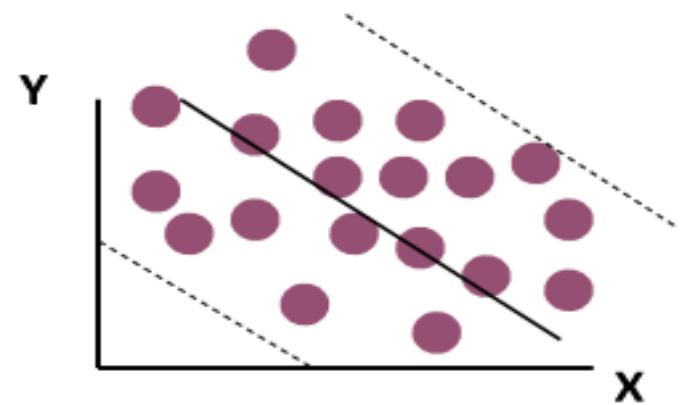
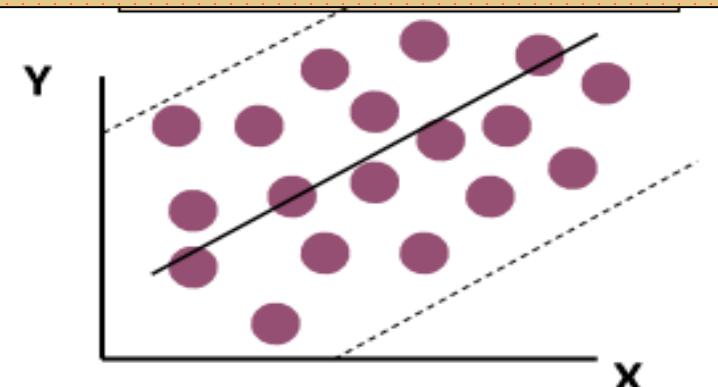
***Negative Linear  
Relationship***

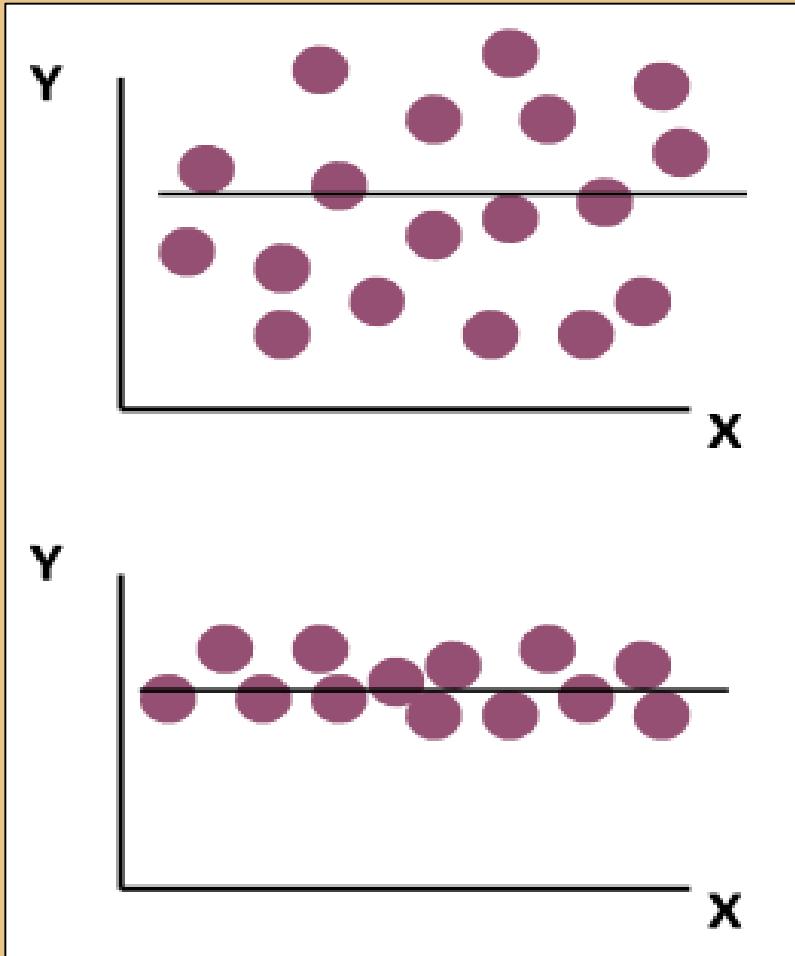


***Strong Relationships***



***Weak Relationships***





**No  
*Relationships***

# *Correlation??*



- This measures **strength** and the **direction** of the **linear relationship** between two numerical variables.
- Correlation is a **value** in between -1 & +1.



- This is also known as **Pearson product-moment correlation coefficient**.

## ***Sample correlation coefficient ( $r$ ),***

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2] [n(\sum y^2) - (\sum y)^2]}}$$

$r = -1$



*Perfect Negative  
Linear Relationship*

$r = 0$



*No Linear  
Relationship*

$r = +1$



*Perfect Positive  
Linear Relationship*

# **Exercise:**

In the pursuit of finding whether the age affects the systolic blood pressure of females, the following data were observed from 10 randomly selected females between ages 40 and 82.



<i>Age</i>	<i>Systolic BP</i>
63	151
70	149
74	164
82	157
60	144
44	130
80	157
71	160
71	121
41	125

# Correlation – Hypothesis Testing

- A hypothesis test can be carried out to find whether the population correlation is zero.

$$H_0: \rho = 0 \quad \text{Vs.} \quad H_1: \rho \neq 0$$

- Under  $H_0$ ,

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$$

# *Regression??*



- The process of ***finding a mathematical equation that best fits the noisy data*** is known as ***regression analysis***.
- In this session, only ***Simple Linear Regression models*** are discussed.
- The ***primary usage*** of a regression model is ***prediction***.

# Simple Linear Regression Model

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

- $\alpha$  - y Intercept
- $\beta$  - Regression Coefficient (Slope)
- $\epsilon$  - Random Error
- This model is defined for population data.
- Should be careful when making predictions outside the observed range.

- $\alpha$  and  $\beta$  in the regression model are population characteristics which cannot be measured straightaway.
- Therefore, they should be estimated by using sample data.
- Estimated regression model would be as follows.

$$\hat{y} = \hat{\alpha} + \hat{\beta}X$$

$$\hat{\beta} = b = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2}$$

$$\hat{\alpha} = a = \bar{y} - b \bar{x}$$

# Significance of Regression Coefficient

- A hypothesis test can be carried out to find whether the true slope ( $\beta$ ) is actually zero (This is same as testing whether the regression model is significant).
- An **ANOVA** table is used to evaluate the ***test statistic*** for this test.

# ANOVA Table

Model	Sum of Squares (SS)	Df (Degrees of Freedom)	Mean Sum of Square (MSS)	F Statistic	P Value
Regression	SSR	1	MSSR	F Statistic	
Error / Residual	SSE	n-2	MSSE		
Total	SST	n-1			

- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$
- $SST = SSR + SSE$
- $MSSR = SSR / 1$
- $MSSE = SSE / (n - 2)$
- $F Statistic = MSSR / MSSE$
- $P value = \Pr(F > F_{Cal})$

## Coefficient of Determination ( $R^2$ )

- One way to measure the strength of the relationship between the response variable ( $y$ ) and the predictor variable ( $x$ ) is to calculate coefficient of determination.
- This refers to the proportion of the total variation that is explained by the linear regression of  $y$  on  $x$ . In other words,  $R^2$  is percentage of variation of  $Y$  explained by the  $X$  variable in the fitted model.

$$R^2 = \frac{SSR * 100}{SST}$$

# Regression Assumptions



- The model is linear in parameters
- $E(\varepsilon_i) = 0$  (Mean of residuals is zero)
- $V(\varepsilon_i) = \sigma^2$  (Variance of residuals are constant)
- The residuals ( $\varepsilon_i$ ) are normally distributed.
- The residuals ( $\varepsilon_i$ ) are independent.

# Important



Remember that, ***neither correlation nor regression imply any causation*** between variables.



# Thanks!

**Any questions?**

# 11. TIME SERIES ANALYSIS [IT2110]

• • •

*By Department of Mathematics and Statistics  
Faculty of Humanities and Sciences, SLIIT*

# CONTENTS

- Introduction
- Components of Time Series
- Time Series Analysis
  - Additive Model
  - Multiplicative Model

# INTRODUCTION

- A time series is a collection of observations made sequentially in time.
- Examples,
  - Monthly inflation rates
  - Daily temperature
  - Annual sales of breads
  - Annual birth rates

# Discrete & Continuous Time Series

- A time series is said to be discrete when observations are taken only at specific time points. (Eg: Daily Temperature)
- A time series is said to be continuous when observations are made continuously in time. (Heart beat of a patient in every second)
- In both cases, the measured variable can be either discrete or continuous.

# Objectives of T.S. Analysis

- Description
  - Simple descriptive measures of time series. Eg: trend, seasonality
- Explanation
  - Use variation in one time series to explain another
- Forecasting (Most important)
- Control
  - Applicable in quality control

# COMPONENTS OF TIME SERIES

- A time series is made up of one or more components mentioned below.
  - Trend
    - Measures the average change in the variable per unit time
  - Seasonality
    - Periodic variations that recur with some degree of regularity within a year or shorter
  - Cyclical variations
    - Recurring up and down movements which are extended over long period (Usually 2 yrs or more).
  - Irregular variations
    - Random fluctuations

# Time Series Analysis

- There are two main classical methods of analyzing time series data.
  - *Additive Model*
  - *Multiplicative Model*
- Other classical Methods :
  - *Curve Fitting*
  - ✓ Polynomial Models
  - ✓ Exponential Models

# Model Selection

- *Additive Model* : Magnitude of the seasonal component is constant over the time
- *Multiplicative Model* : Magnitude of the seasonal component is increasing / deceasing with time
- If the model can not be clearly identify, fit the both models and use forecasts to choose the better model.

# Fitting an Additive Model

- *Additive Model :*

$$Y_t = T + S + C + I$$

- Can be fitted only when magnitude of the seasonal component is constant over the time

# Fitting a Multiplicative Model

- *Multiplicative Model :*

$$Y_t = T \times S \times C \times I$$

- Can be fitted only when magnitude of the seasonal component is increasing / decreasing with time

# THANK YOU!

• • •

## Good Luck for the Exam!

