

# Data Science

An Introduction

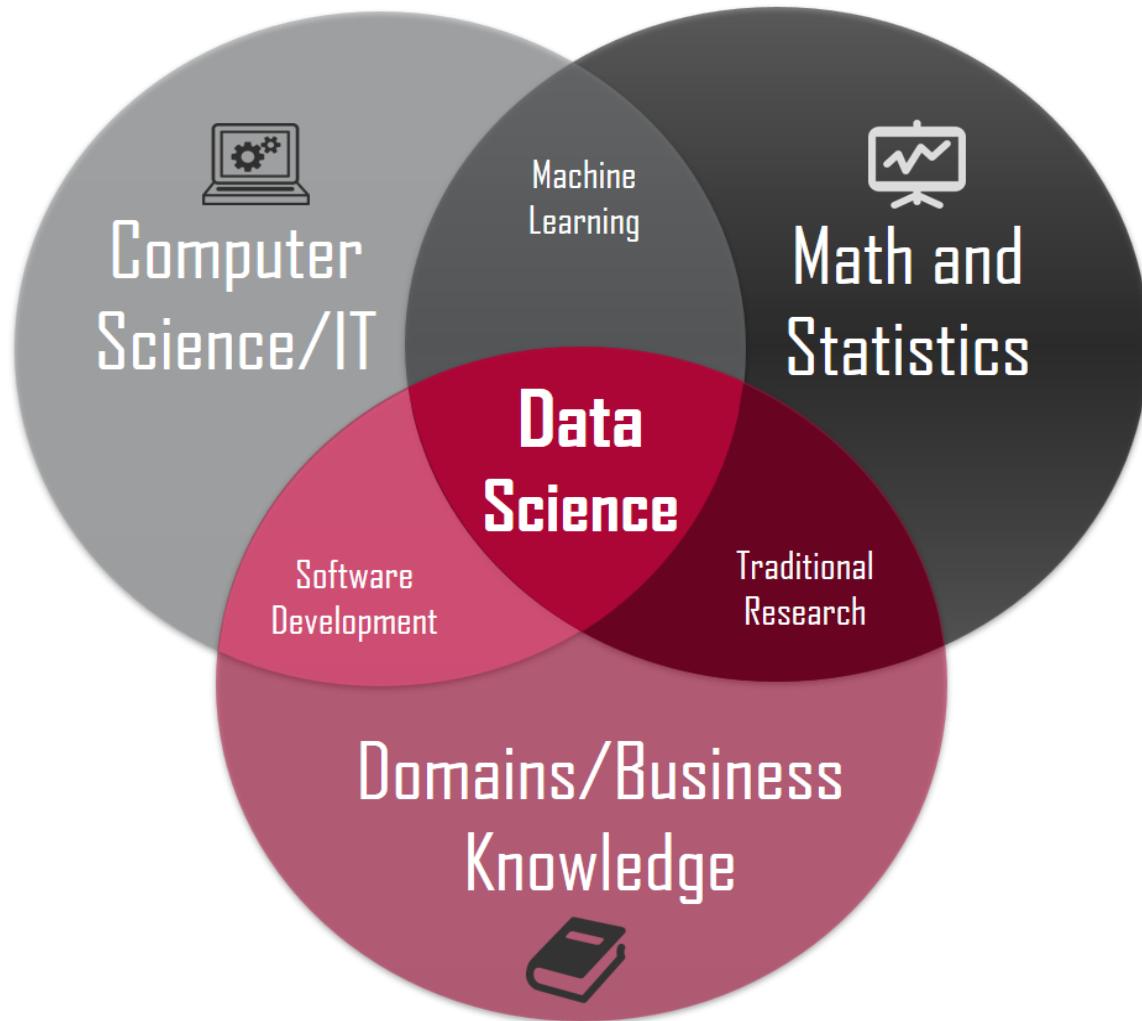
# Understand Terms...

- Data Science
- Analytics
- Artificial Intelligence
- Machine Learning
- Deep Learning

# What is Data Science?

- Application of Scientific Methods like Statistical and Machine Learning in order to understand the phenomena to gain control on it
- It employs techniques from both the fields computer science and statistics
- Data science involves Machine Learning, Clustering, Visualization and many other things related to data

# Data Science Composition

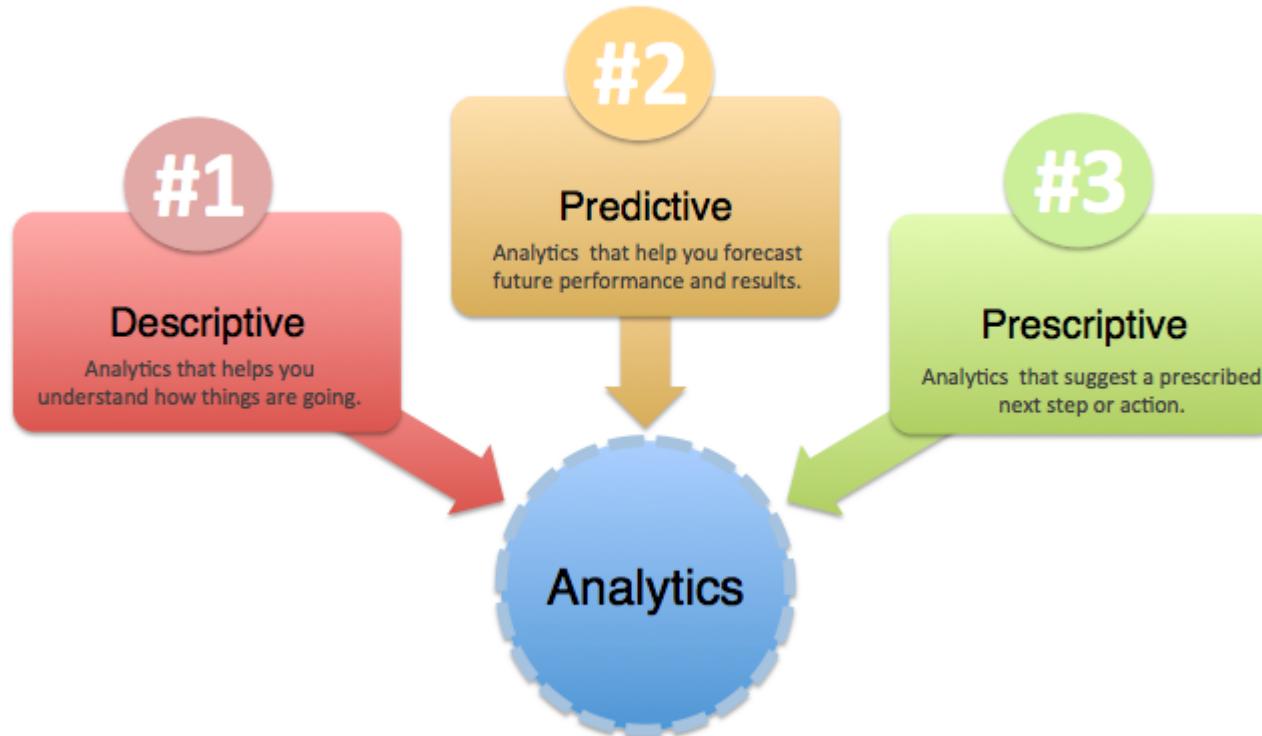


Courtesy: <https://www.fox.temple.edu/institutes-and-centers/data-science/>

# What is Analytics?

- Analytics is the discovery, interpretation, and communication of meaningful patterns in data.
- Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance

# Types of Analytics



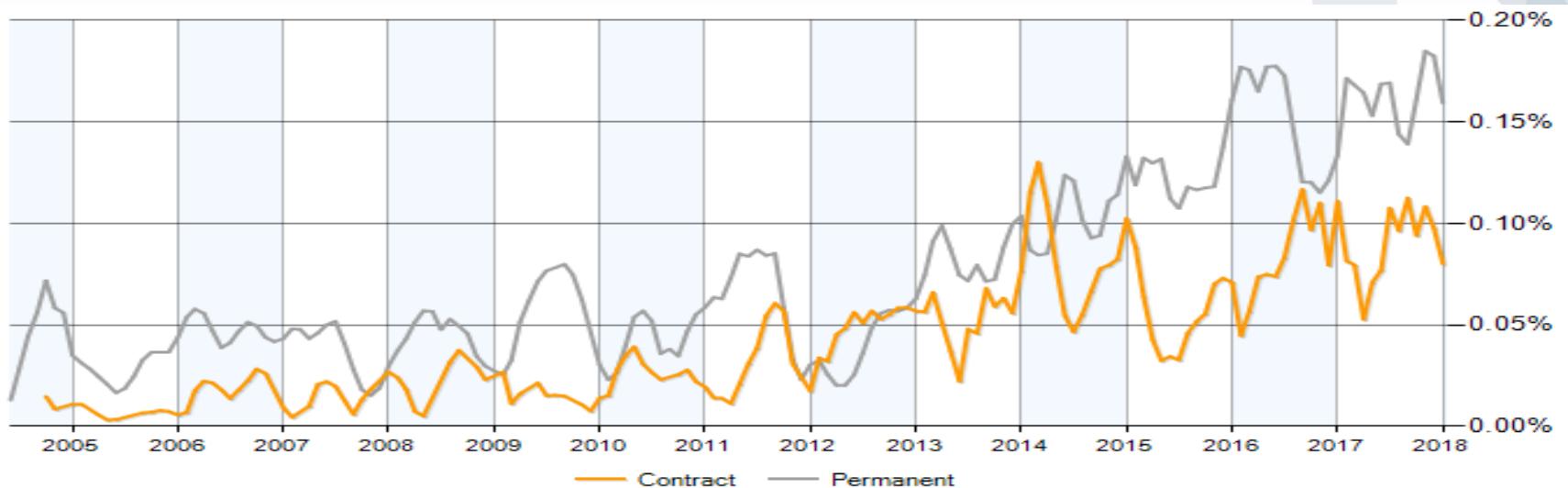
Courtesy: <https://moz.com/blog/when-it-comes-to-analytics-are-you-doing-enough>

# Descriptive Analytics



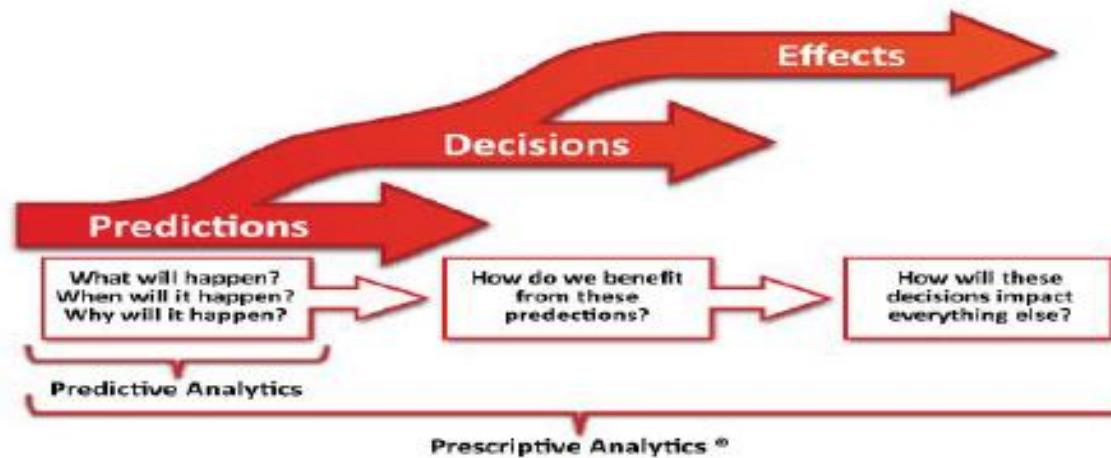
- Gain insight from historical data with reporting, scorecards, clustering etc.
- Can involve data visualization for knowing the basic characteristics of the data
- Descriptive analytics answers the questions what happened and why did it happen.
- Implementations : Business Intelligence, Visualizations
- Software: Informatica, Business Objects, TIBCO Spotfire, Tableau etc.

# Predictive Analytics



- Involves statistical and machine learning techniques
- Analyzing the historical patterns in the data and predicting the future patterns
- Predictive analytics answers the question what will happen
- Implementation: Machine Learning, Deep Learning
- Software: R, Python, Libraries like TensorFlow, h2o.ai etc.

# Prescriptive Analytics



- Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the implications of each decision option.
- Implementation: Optimization Techniques like Linear programming Problems, Non-linear programming Problems, Genetic Algorithm etc.

# What is AI?

- AI or artificial intelligence is the simulation of human intelligence processes by machines, especially computer systems.
- These processes include learning (the acquisition of information and rules for using the information), reasoning (using the rules to reach approximate or definite conclusions), and self-correction.
- AI is a discipline just like Physics.

# Role of Machine Learning (ML)

- Machine Learning is a tool set for implementing AI, today
- AI elements which don't include ML are expert systems
- ML Algorithms are driven by mathematical concepts
- ML Algorithms analyse the patterns in the captured data and can be used to build a predictive model on the existing phenomena in business
- Broadly, there are three types of ML Algorithms
  - Supervised Learning Algorithms
  - Unsupervised Learning Algorithms
  - Re-inforcement Learning Algorithms

# Supervised Learning

- Supervised learning algorithms are those used in classification and prediction.
- We must have data available in which the value of the outcome of interest (e.g., purchase or no purchase) is known.
- The objective is to predict the values of the outcome of interest

# Models for Supervised Learning

- We identify strong links between variables of a data table (columns).
- Such a link may translate into an expression between one variable  $y$  (the so-called "dependent" or "response" variable) and a group of other variables  $\{x_i\}$  (the so-called "independent variables" or "predictors" or "features") :

$$y = f(x_1, x_2, \dots, x_n) + \text{Small random noise}$$

# Types in Supervised Learning

- When the response variable is numerical, predictive modeling is called **Regression**.
- When the response variable is categorical (nominal / ordinal), predictive modeling is called **Classification**.

# Examples

- **Regression Case:** Sales are influenced by the variables like advertisement expenses, manpower deployed for sales, cost of products, number of dealers etc. Hence we see here  
$$\text{Sales} = \text{function}(\text{Adv. Exp , Manpower , Cost , Dealers , ...})$$
- **Classification Case:** The customer may purchase a particular product based on some conditions like his need, his age, his income, his place of residence etc. Hence we see here  
$$\text{Prob(Customer Purchases)} = \text{function}(\text{Age, Income, Residence,...})$$

# Algorithms of Supervised Learning

- Naïve Bayes
- K-NN
- Decision Trees
- Regression Models
- Neural Nets
- Support Vector Machines

# Unsupervised Learning

- Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify.
- Association rules, data reduction methods, and clustering techniques are all unsupervised learning methods.

# Examples

- Customer Segmentation like RFM (Recency, Frequency, Monetary)
- Market Basket Analysis
- Product Grouping

# Algorithms of Unsupervised Learning

- Clustering Techniques
  - Hierarchical
  - K-means
- Principal Component Analysis
- t-SNE
- Association Rules
- Auto-encoders

# Re-inforcement Learning

- In this type, there is an agent which/who receives information from the environment and learns to choose actions based on rewards or punishment received
- Examples include:
  - Self-driving cars
  - Robotics
- Algorithms:
  - Upper Confidence Bound
  - Thomson Sampling

# Technologies Used

For Machine Learning Algorithms Implementation

# Desktop Software

- Click and Drag (Menu Driven)

- KNIME
- RapidMiner
- SAS Enterprise Miner
- IBM SPSS Modeller

# Functional Programming Languages

- R
- Python
- Julia
- Scala

# R

- An open source project
- Fast on desktop with small sized data
- Add-ins (packages) available for every statistical/ML algorithm in the world
- Has been used since last 2 decades for statistical computing by statistical professionals community
- There are good IDEs available like RStudio, RTVS, R Commander, Tinn-R, STATET(Eclipse plug-in) etc.
- Among IDEs R Studio is most known
- Provides a scope for implementing or own algorithms being an open source language

# Python

- An open source project
- Fast on desktop with small sized data
- Add-ins (packages) available for every statistical/ML algorithm in the world
- The statistical aspects of Python have been developed recently
- There are good IDEs available like Spyder(Anaconda Installation), PyCharm etc.
- Provides a scope for implementing our own algorithms being an open source language

# Cloud-Based Platform

- Amazon Web Services
- Microsoft Azure
- Google Cloud AI

# Large Scale Data Processing Libraries

- Libraries are such kind of modules which are language independent.
- Using libraries, one can code in R / Python / Java
- Well known libraries for ML are
  - Apache Spark
  - h2o (by h2o.ai)
  - TensorFlow (by Google)
  - Theano (by University of Montreal)
  - CNTK (by Microsoft)
- All of the above provide support for GPU-based operations for algorithms in Deep Learning
- The superb feature which these libraries provide is the fast speed that too at low cost.



A large, abstract graphic in the upper right corner consists of a grid of light gray squares of varying sizes, creating a tessellated effect. Some squares are solid gray, while others have diagonal gray stripes, giving it a textured appearance.

Thank You



# Nature of Data

# Types of data

- Based on Data type
  - Numerical
    - Discrete
      - Likert Scale
      - Quantitative (counts)
    - Continuous
  - Categorical: Nominal and Ordinal
- Based on Time
  - Cross-sectional
  - Time-series
  - Panel Data

# Discrete: Likert Scale

- An ordered, one-dimensional scale from which respondents choose one option that best aligns with their view
- e.g.
  - Question: Eating pizza in the evening more preferable than eating it at the dinner.
    1. Strongly Disagree
    2. Disagree
    3. Neither agree nor disagree
    4. Agree
    5. Strongly Agree

# Discrete: Quantitative

- The data with particular values.
- e.g.
  - Count of Customers coming in a particular time slot
  - Count of patients of a particular disorder

# Continuous

- Data taking any real number value
- e.g.
  - Height
  - Weight
  - Volume in litres
  - Sales

# Short Quiz: Classify the data

Classify each set of data as discrete or continuous

1. The number of suitcases lost on a railway platform.
2. The height of sugarcane plants.
3. The number of sugarcane sticks grown in a plot.
4. The time it takes for a car battery to die.
5. The production of sugar by weight.
6. The opinion of passengers about the airport administration.

# Short Quiz : Classify the data

Classify each set of data as discrete or continuous

1. Discrete: The number of suitcases lost on a railway platform.
2. Continuous: The height of sugarcane plants.
3. Discrete: The number of sugarcane sticks grown in a plot.
4. Continuous: The time it takes for a car battery to die.
5. Continuous: The production of sugar by weight.
6. Likert Scale: The opinion of passengers about the airport administration.

# Categorical: Nominal

- **Nominal** level data is made up of values that are distinguished by name only.
- There is no standard ordering scheme to this data.
- e.g.
  - Colours in a fabric produced by a company
  - Flavours of corn flakes available in the market

# Categorical: Ordinal

- There is an ordering scheme in Ordinal Scale.
- e.g.
  - Movies on a certain TV show are classified as 2 thumbs up, 1 thumb up, or 0 thumbs up.
  - The dealers are classified as follows depending on the quality of service they provide



# Short Quiz : Identify the data

1. Letter grades on an English essay.
2. Flavors of yogurt.
3. Instructors classified as : Easy, Difficult or Impossible.
4. Employee evaluations classified as : Excellent, Average, Poor.
5. Political parties.
6. Ice cream flavour.
7. Students classified by their reading ability : Above average, Below average, Normal.

# Short Quiz : Identify the data

1. Ordinal: Letter grades on an English essay.
2. Nominal: Flavors of yogurt.
3. Ordinal: Instructors classified as : Easy, Difficult or Impossible.
4. Ordinal: Employee evaluations classified as : Excellent, Average, Poor.
5. Nominal: Political parties.
6. Nominal: Ice cream flavour.
7. Ordinal: Students classified by their reading ability : Above average, Below average, Normal.

# Types of data

- Based on Data type
  - Numerical
    - Discrete
      - Likert Scale
      - Quantitative (counts)
    - Continuous
  - Categorical: Nominal and Ordinal
- Based on Time
  - Cross-sectional
  - Time-series
  - Panel Data

# Cross-sectional

- Refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time or without regard to differences in time
- e.g.
  - Data of Home and Automobile insurance policies sold at different branches and their total operating cost in a specific year

Branch	Home	Automobile	Operating_Cost
B01	400	1200	124000
B02	350	360	71000
B03	600	800	136000
B04	800	1800	219000
B05	900	1600	230000
B06	200	1000	75000
B07	120	900	56000
B08	340	1100	110000
B09	490	900	120000
B10	700	800	144000

# Time Series

- A sequence of data points, measured typically at successive times spaced at uniform time intervals
- e.g.
  - Month-wise Stock Price of a company



# Panel Data

- Panel Data is mixture of time series analysis and cross-sectional data

StoreID	Month	Sales
1	Jan	58000
2	Jan	59000
3	Jan	36000
1	Feb	15000
2	Feb	14700
3	Feb	15400
1	Mar	18000
2	Mar	12500

# Preliminary Data Analysis

## Measures of Central Tendency

# We will be covering...

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
  - Quartiles
  - Options of central tendency in pandas
- Measures of Dispersion
  - Range
  - Semi Inter-Quartile Range
  - Mean Deviation
  - Variance
  - Standard Deviation
  - Coefficient of Variation
  - Skewness
  - Kurtosis
  - Options of dispersion in pandas

# What are averages?

- These are statistical constants which enable us to comprehend in a single effort the significance of the whole thing.

# Measures of Central Tendency

- Mean
  - Arithmetic mean
- Median
- Mode
- Quartiles , Deciles and Percentiles

# Arithmetic mean

- Arithmetic mean of a given set of observations is their sum divided by the number of observations.
- e.g. A.M. of 5, 8, 10, 15, 24 and 28 is

$$\frac{5+8+10+15+24+28}{6} = \frac{90}{6} = 15$$

# Median

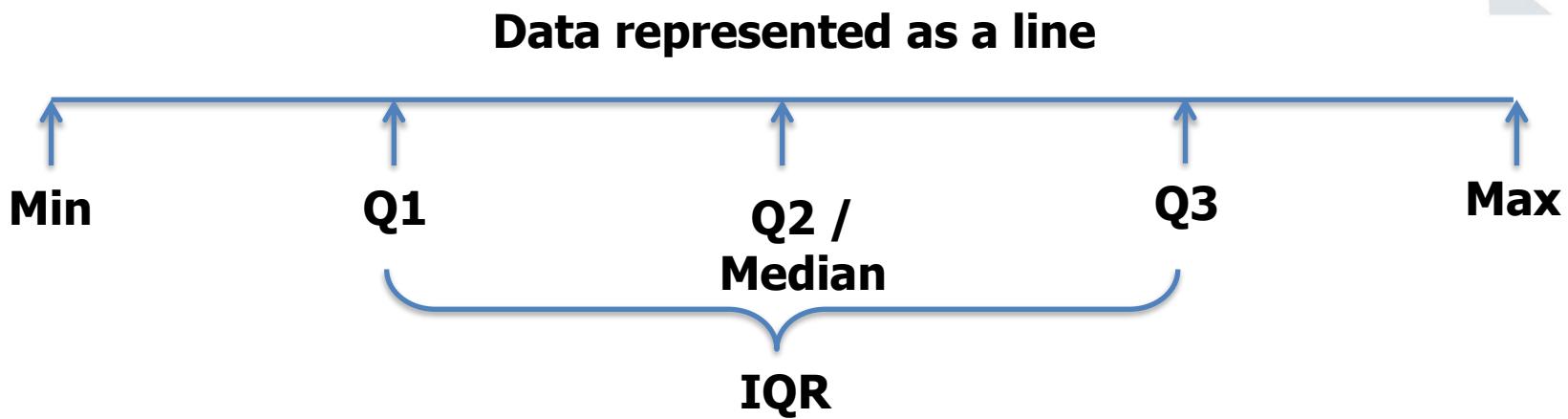
- Median is that value which divides the set of given numbers in two equal parts.
- E.g. Median of numbers 5, 10, 8, 15, 28 and 24 can be calculated as follows:
  - Arrange the given numbers in ascending/descending order as 5, 8, 10, 15, 24, 28
  - Count the numbers. They are 6 in number.
  - The middle two numbers are 10 and 15. Hence median is the arithmetic mean of 10 and 15. i.e. 12.5

# Mode

- The mode is the value which has greatest frequency.
- E.g. Mode of numbers [ 4, 5, 5, 6, 7, 8, 6 , 5 ] is 5

# Quartiles

- Quartiles divide the given data into four equal parts.



- Inter-quartile range (IQR) is given by the formula:

$$IQR = Q3 - Q1$$

# Preliminary Data Analysis

## Measures of Dispersion

# Measures of Dispersion

- Absolute Measures
  - Range
  - Quartile Deviation or Semi-Interquartile Range
  - Mean Deviation
  - Standard Deviation
- Relative Measures
  - Coefficient of Variation

# Range

- Range is defined as the difference between the two extreme observations in a distribution (i.e. greatest (maximum) and the smallest (minimum) observation.)
- E.g. Range of 5, 8, 10, 15, 24 and 28 is  $28 - 5 = 23$

# Quartile Deviation or Semi-Interquartile Range

- Quartile Deviation: It is calculated by a formula:

$$QD = \frac{Q_3 - Q_1}{2}$$

# Mean Deviation

- Mean Deviation (average deviation) is a measure of dispersion that is obtained on taking the average (arithmetic mean) of the absolute deviation of the given values from a measure of central tendency (mean).

# Standard Deviation

- Standard Deviation is defined as the positive square root of the arithmetic mean of the squares of the deviations of the given observations from their arithmetic mean.
- More is the magnitude of a standard deviation more is the dispersion.
- e.g. Data with  $SD=23.4$  can be said to be more dispersed than data with  $SD=12.7$ .

# Coefficient of Variation

- The ratio of SD and Mean
- CV is unit less quantity

$$CV = \frac{SD}{Mean} * 100$$

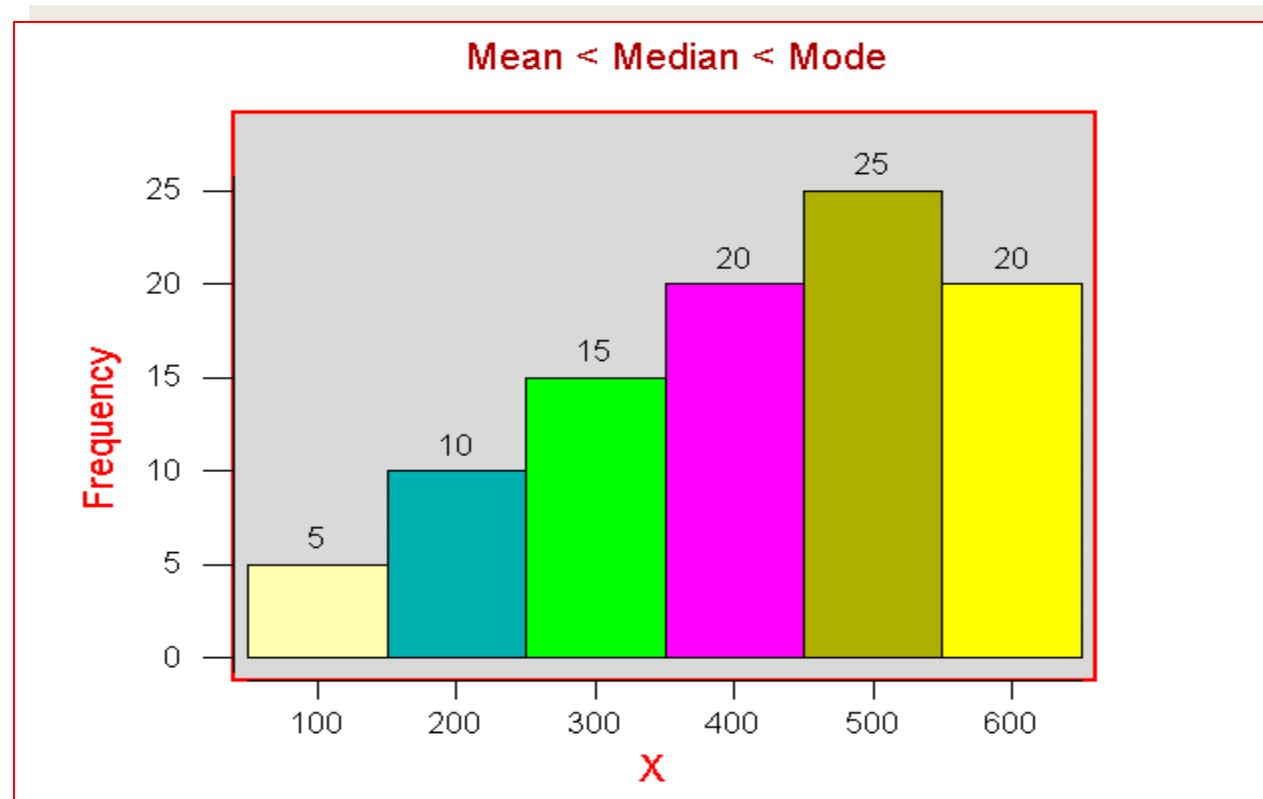
$$CV = \frac{\sigma}{\mu} * 100$$

# Skewness and Kurtosis

- **Skewness**
  - Measure of asymmetry of a frequency distribution
    - Skewed to left
    - Symmetric or unskewed
    - Skewed to right
- **Kurtosis**
  - Measure of flatness or peakedness of a frequency distribution
    - **Platykurtic** (relatively flat)
    - **Mesokurtic** (normal)
    - **Leptokurtic** (relatively peaked)

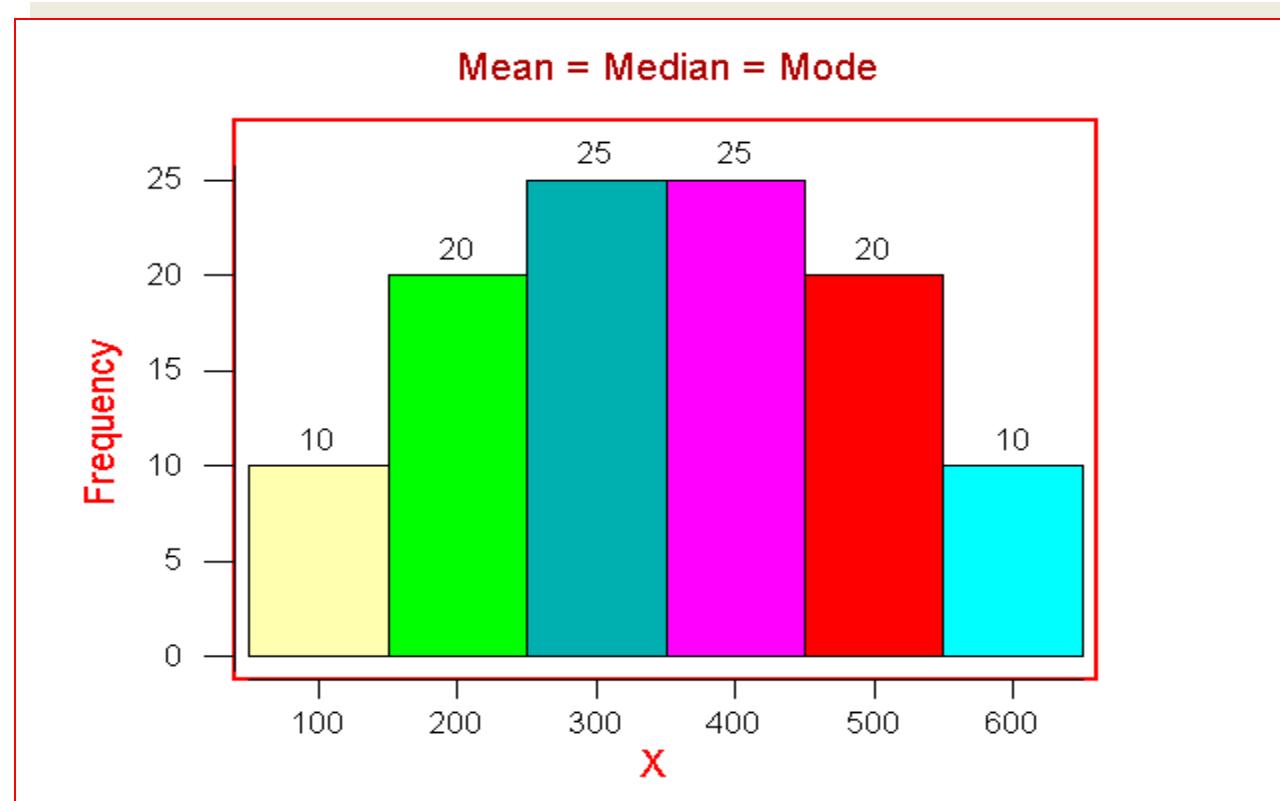
# Skewness

Skewed to left



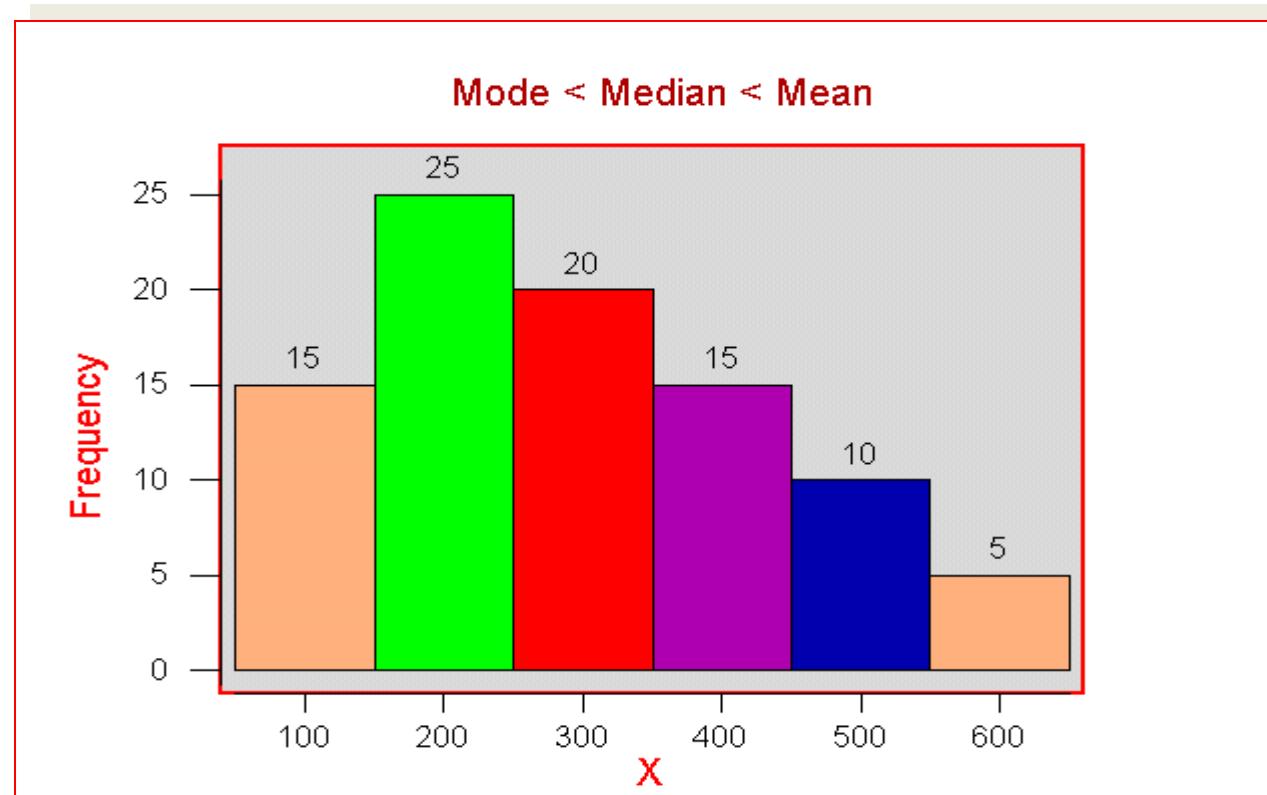
# Skewness

Symmetric



# Skewness

Skewed to right



## Coefficient of Skewness

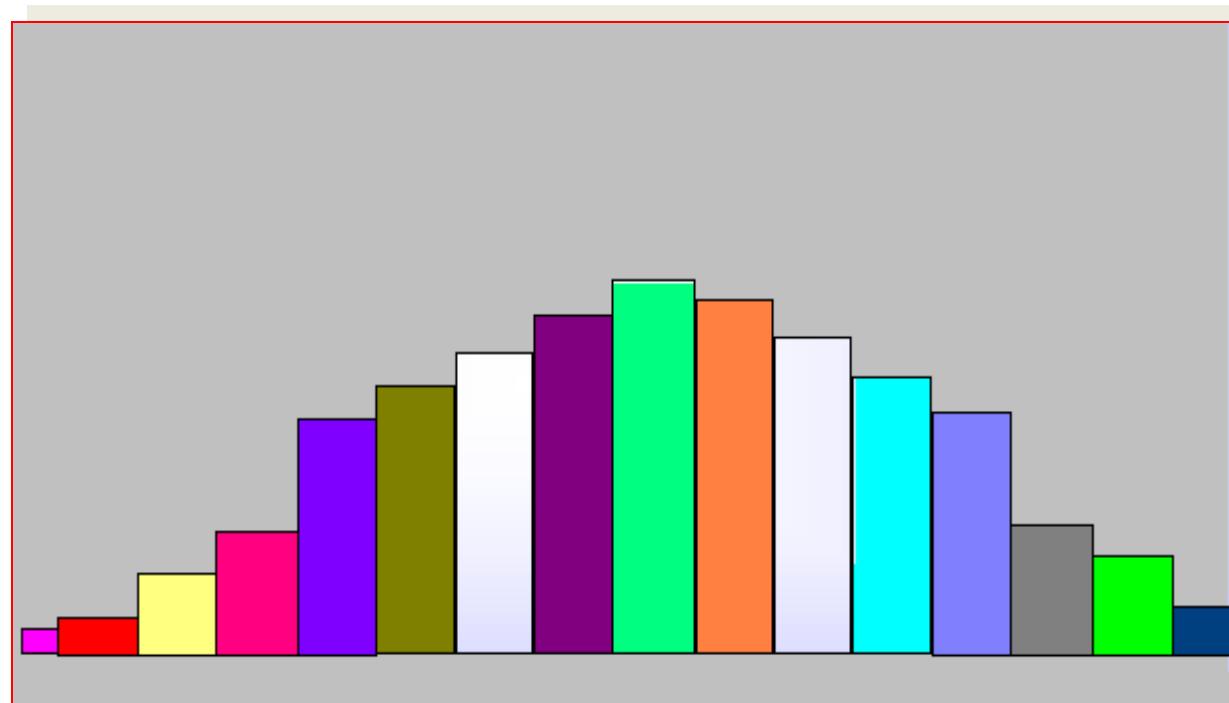
- Coefficient of Skewness (CS):

$$CS = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$$

- ▶ CS is negative for left-skewed data.
- ▶ CS is positive for right-skewed data.
- ▶  $|CS| > 1$  suggests high degree of skewness.
- ▶  $0.5 \leq |CS| \leq 1$  suggests moderate skewness.
- ▶  $|CS| < 0.5$  suggests relative symmetry.

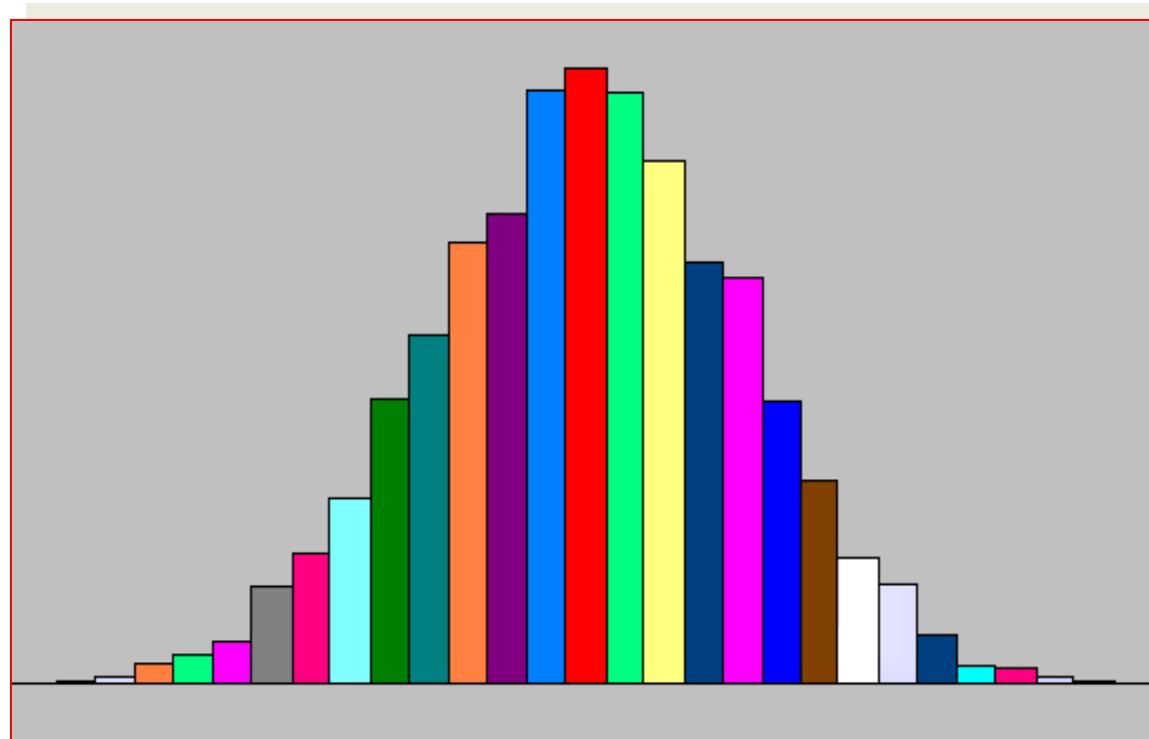
# Kurtosis

**Platykurtic** - flat distribution



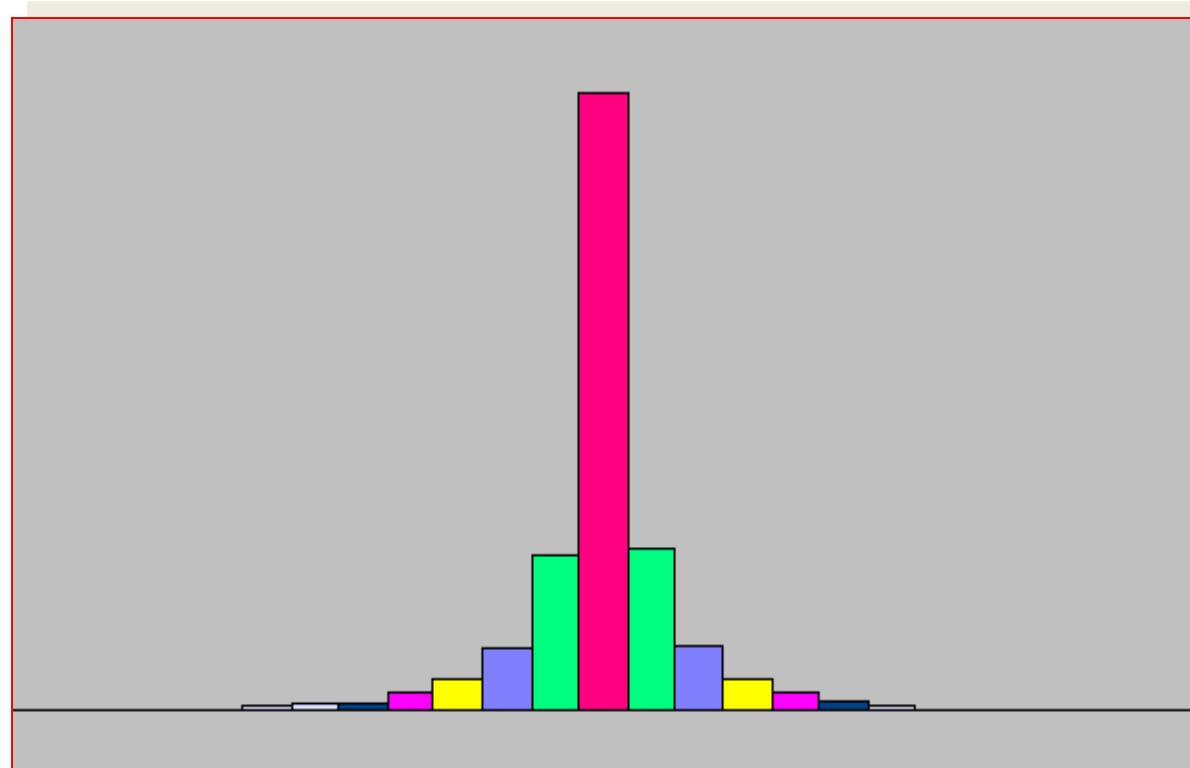
# Kurtosis

**Mesokurtic** - not too flat and not too peaked



# Kurtosis

**Leptokurtic** - peaked distribution



# Kurtosis

- **Kurtosis** refers to the peakedness (i.e., high, narrow) or flatness (i.e., short, flat-topped) of a histogram.
- The coefficient of kurtosis (CK) measures the degree of kurtosis of a population

$$CK = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4} - 3$$

- ▶ CK < 0 indicates the data is somewhat flat with a wide degree of dispersion.
- ▶ CK > 0 indicates the data is somewhat peaked with less dispersion.

# Correlation

# Need for correlation

- To find the association between the variables
- To find the degree of the association

# What is correlation?

- The relationship between two variables is called their correlation.
- Positive Correlation: As one variable becomes large, the other also becomes large, and vice versa.
- Negative Correlation: As one variable becomes small, the other becomes large, and vice versa.

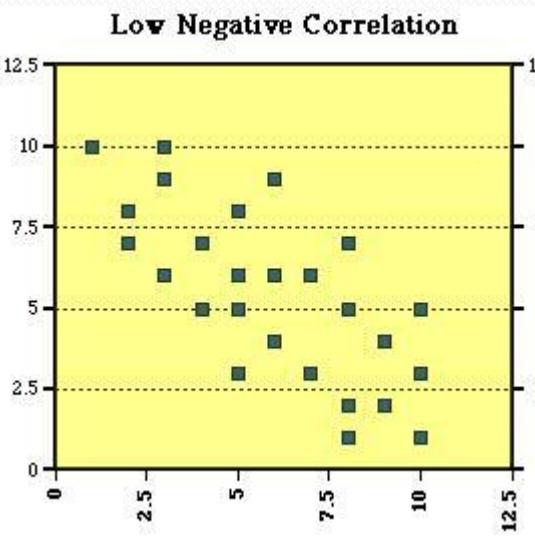
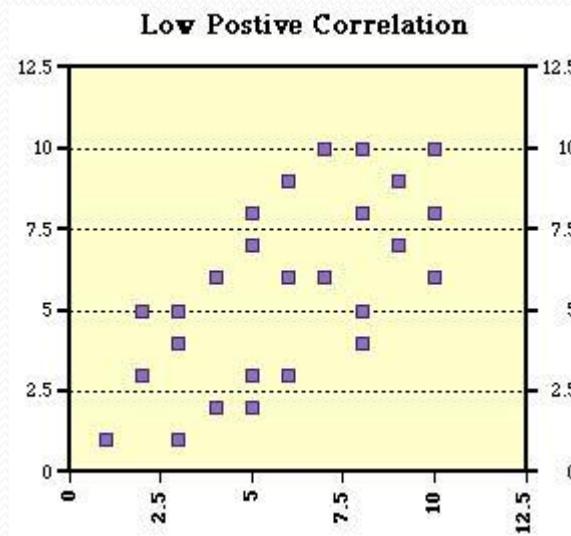
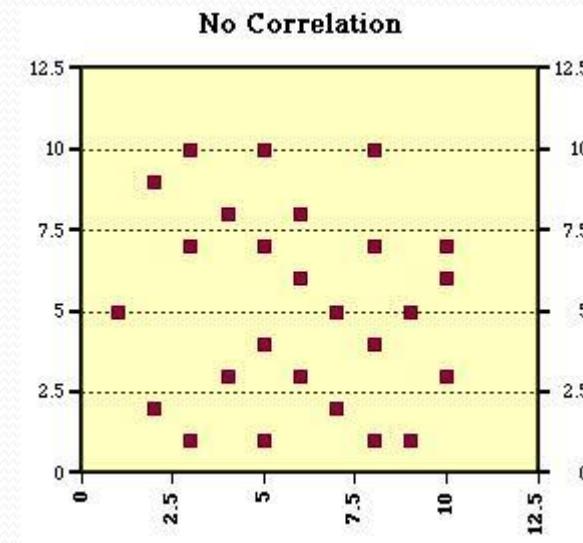
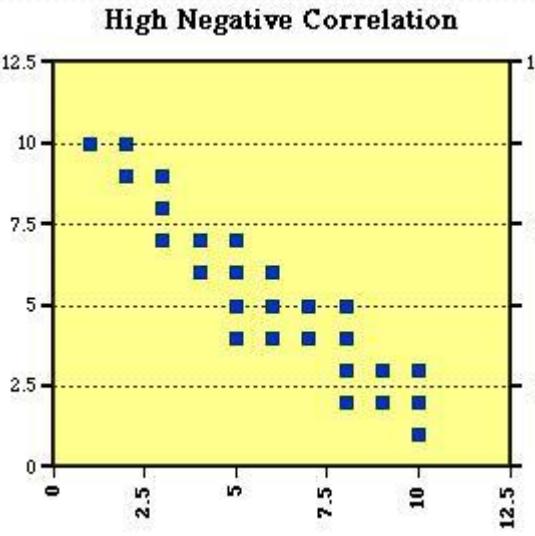
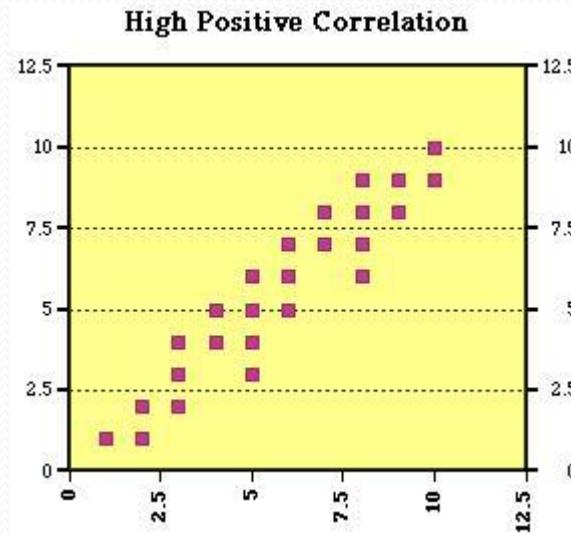
# Other types of correlation

- Linear: Corresponding to a unit change in one variable, there is a constant change in the other variable.
- Non-Linear: Corresponding to a unit change in one variable, the other variable doesn't change at a constant rate but it changes at a fluctuating rate.

# How to find the correlation?

- ▶ Scatter plots show how much one variable is affected by another.
- ▶ Correlation Coefficients give the degree of correlation.

# Example – Numerical Data (Two variables) Scatter Plot



# Karl Pearson's Coefficient

- It is calculated a formula involving variance and covariance values.

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad -1 \leq \rho \leq 1$$



# Probability

# We will be covering...

- Concept of Probability
- Types of Events
  - Collectively Exhaustive
  - Mutually Exclusive
  - Independent
- Rules
  - Complement
  - Addition
  - Multiplicative

# Probability

- A probability in statistical theory is a number between 0 and 1 that measures the likelihood that some event will occur.
- A zero probability of an event signifies event not occurring and probability one signifies event being almost sure

# Collectively Exhaustive Events

- Events A<sub>1</sub>, A<sub>2</sub>, ... A<sub>n</sub> are said to be collectively exhaustive events if at least one of the events must occur
- Example:
  - A<sub>1</sub>: Rain
  - A<sub>2</sub>: Sunny Weather
  - A<sub>3</sub>: Cloudy
  - A<sub>4</sub>: Snowfall

# Rules

- If A and B are events then:
  - Rule of Complement :  $P(\text{Non-occurrence of } A) = 1 - P(A)$
  - Addition Rule:  $P(\text{Occurrence of at least one of the events}) = P(A) + P(B) - P(\text{Occurrence of both } A \text{ and } B)$
  - Multiplication Rule:  $P(\text{Occurrence of } A \text{ given that } B \text{ has occurred}) = P(\text{Occurrence of both } A \text{ and } B) / P(B)$

# Addition Rule

- $P(\text{Occurrence of at least one of the events}) = P(A) + P(B) - P(\text{Occurrence of both A and B})$

- By Notation:

$$P(A \text{ or } B \text{ or } A \text{ and } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Example:

- A: Ad Campaign will fail
  - B: Competitor launches a new product
  - $A \cap B$  : A and B both occur

Say  $P(A) = 0.23$ ,  $P(B) = 0.39$ ,  $P(A \cap B) = 0.18$  then

$$\begin{aligned}P(A \cup B) &= P(A) + P(B) - P(A \cap B) = 0.23 + 0.39 - 0.18 \\&= 0.44\end{aligned}$$

# Mutually Exclusive Events

- The two events A and B are said to be mutually exclusive (disjoint) if they both cannot occur simultaneously
- In case of mutually exclusive events A and B,
  - $P(A \cap B) = 0$
  - $P(A \cup B) = P(A) + P(B)$

Example: A: Snowfall, B: Temperature > 50 degrees Celsius

$P(A) = 0.25$ ,  $P(B) = 0.20$  then  $P(A \text{ or } B) = 0.45$

# Multiplication Rule

- $P(\text{Occurrence of both A and B}) = P(\text{Occurrence of A given that B has occurred}) * P(B)$
  - Or
  - $P(\text{Occurrence of A given that B has occurred}) = P(\text{Occurrence of both A and B}) / P(B)$
  - By Notations,  $P(A | B) = P(A \cap B) / P(B)$ ,  $P(A | B)$  is read as A given B
  - Similarly,  $P(B | A) = P(A \cap B) / P(A)$
  - $P(A|B)$  and  $P(B|A)$  are called conditional probabilities
  - Example:
    - A: Ad Campaign will fail
    - B: Competitor launches a new product
    - $A \cap B$  : A and B both occur
- $P(A) = 0.25$ ,  $P(B) = 0.32$ ,  $P(A \cap B) = 0.17$  then

$$P(A|B) = P(A \cap B) / P(B) = 0.17 / 0.32 = 0.5312,$$
$$P(B|A) = P(A \cap B) / P(A) = 0.17 / 0.25 = 0.68$$

# Independence of Events

- If A and B are independent events then:
    - $P(\text{Occurrence of both A and B}) = P(A) P(B)$
  - By Notations,  $P(A \cap B) = P(A) * P(B)$
  - Example:
    - A: Ad Campaign will fail
    - B: Sensex goes up
    - $A \cap B$  : A and B both occur
- Say,  $P(A) = 0.32$ ,  $P(B) = 0.64$ ,  $P(A \cap B) = P(A) * P(B) = 0.2048$



# Probability Distributions

# Covering...

- What is Random Variable?
- Types of Random Variables
- What is Probability Distribution?
- Expected Value of Random Variable

# Probability Distribution

- Random variable is a numerical quantity with uncertain values.
- The pattern of randomness of the random variable is the probability distribution of the random variable.

# Types of Random Variables

- Discrete: Random Variables that take particular values are discrete random variables
- Continuous: Random Variables that can take any real value are continuous random variables

# Properties of Probability Distributions

- For any probability distribution following things are always true:
  - Value of probability is between 0 and 1
  - Sum of all probabilities is 1
  - All the events are mutually exclusive
  - All the events are exhaustive

# Examples of Probability Distributions

- Discrete:

Xi	Pi
4	0.1
6	0.2
8	0.5
10	0.2

$$\sum_{i=1}^n p_i = 0.1 + 0.2 + 0.5 + 0.2 = 1$$

- Continuous:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$\int_a^b \frac{1}{b-a} dx = 1$$

# Probability Functions

- For Discrete Variables, functions are called Probability Mass Functions
- For Continuous Variables, functions are called Probability Density Function

# Mathematical Expectation

- Expected Value of a variable for discrete distribution is given by

$$E(X) = \sum x_i p_i$$

- Expected Value of a variable for continuous distribution is given by

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

# Expected Value Computation: Discrete

X <sub>i</sub>	P <sub>i</sub>
4	0.1
6	0.2
8	0.5
10	0.2

$$\begin{aligned}E(X) &= \sum x_i p_i \\&= 4 * 0.1 + 6 * 0.2 + 8 * 0.5 + 10 * 0.2 \\&= 7.6\end{aligned}$$

# Expected Value Computation: Continuous

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_a^b x \frac{1}{b-a} dx \\ &= \frac{1}{2}(a+b) \end{aligned}$$

# Binomial Distribution

# Binomial Distribution

- Considers experiment with two possible outcomes: success and failure.
- $p$  = Probability of Success of single trial
- $q = 1-p$  = Probability of Failure of single trial
- $n$  = No. of Trials of the Experiment
- $k$  = No. of successes out of  $n$  trials



Jakob Bernoulli

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

$$\text{Mean} = E(X) = np$$

$$\text{Variance}(X) = npq$$

# Python functions for Binomial Distribution

## Syntax:

`binom.pmf(k,n,p,...)`

Applies for  $P(X=k)$

`binom.cdf(k,n,p,...)`

Applies for  $P(X \leq k)$

`binom.sf(k,n,p,...)`

Applies for  $P(X > k)$

`binom.stats(n,p,...)`

Applies for extracting mean, variance and other moments

# Example

In a typical Month, an Insurance agent presents life insurance plans to 40 potential customers. Historically, one in four such customers chooses to buy Life Insurance from this agent. Based on the relevant binomial distribution , answer the following questions :

1. What is the probability that exactly 5 customers will buy life Insurance from this agent in the coming month ?
2. What is the probability that not more than 10 customers will buy life insurance from this agent in the coming month ?
3. What is the probability that at least 20 customers will buy life insurance from this agent in the coming month ?
4. Determine the mean and variance of the number of customers who will buy life insurance from this agent in the coming month.



# Poisson Distribution

# Poisson Distribution

- Considers experiment with two possible outcomes: success and failure.
- $n$  is infinitely large (or very large)
- $p$  is very small
- Characterized by a single parameter  $\lambda$
- Let  $X$  : No. of successes



Siméon Denis Poisson

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where

- $e$  is Euler's number ( $e = 2.71828\dots$ )
- $k!$  is the factorial of  $k$ .

$$\text{Mean} = E(X) = \lambda$$

$$\text{Variance}(X) = \lambda$$

## Python Functions for Poisson Distribution

`poisson.pmf(k,mu,...)`

Applies for  $P(X=k)$

`poisson.cdf(k,mu,...)`

Applies for  $P(X \leq k)$

`poisson.sf(k,mu,...)`

Applies for  $P(X > k)$

`poisson.stats(mu,...)`

Applies for extracting mean, variance and other moments

# Example

The annual number of industrial accidents occurring in a particular manufacturing plant is known to follow Poisson distribution with mean 12.

- a) What is the probability of observing exactly 5 accidents at this plant during the coming year ?
- b) What is the probability of observing not more than 12 accidents at this plant the coming year ?
- c) What is the probability of observing at least 15 accidents at this plant during the coming year ?
- d) What is the probability of observing between 10 and 15 accidents (inclusive) at this plant during the coming year ?



# Normal Distribution

# Probability Density Function

- PDF of Normal Distribution is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty < x < \infty,$$

$-\infty < \mu < \infty, \sigma > 0$

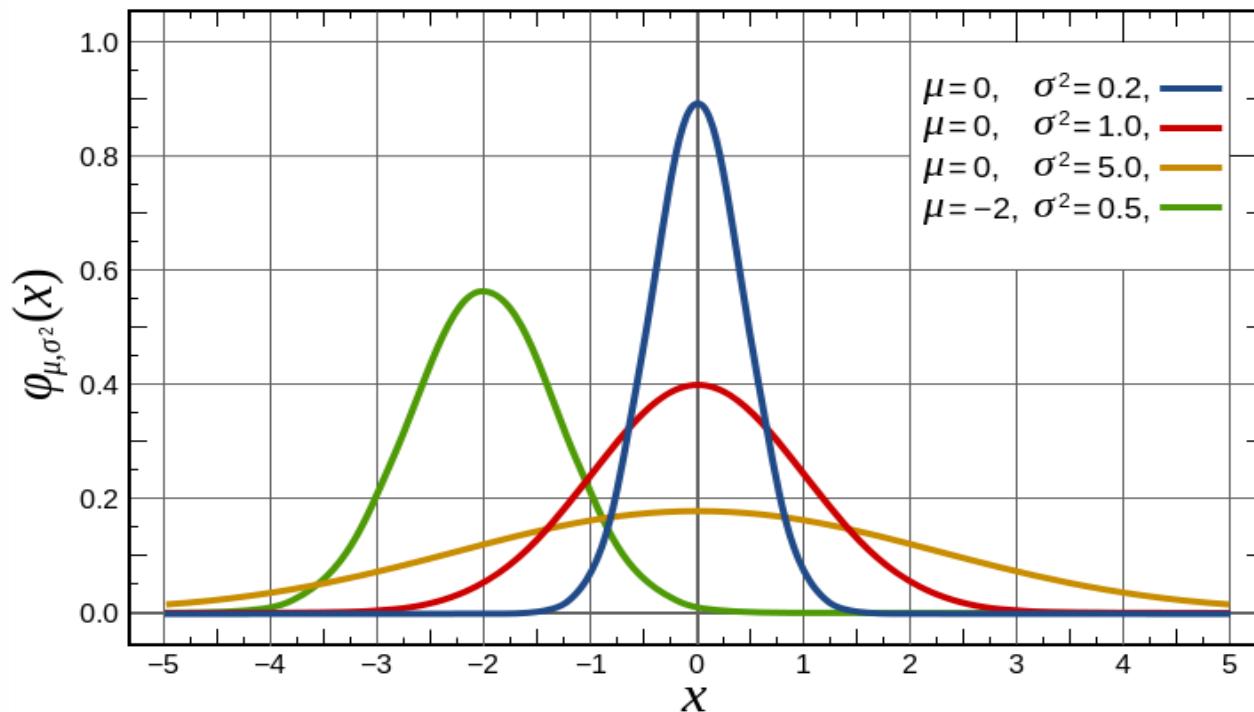


Image Courtesy: Wikipedia



Carl Friedrich Gauss

**Mean** =  $E(X) = \mu$

**Variance**(X) =  $\sigma^2$

# Standard Normal Distribution

- Standard Normal Distribution is distribution with mean 0 and standard deviation 1.
- Standard Normal variable Z is formed by transforming X as:

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

$$\therefore f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

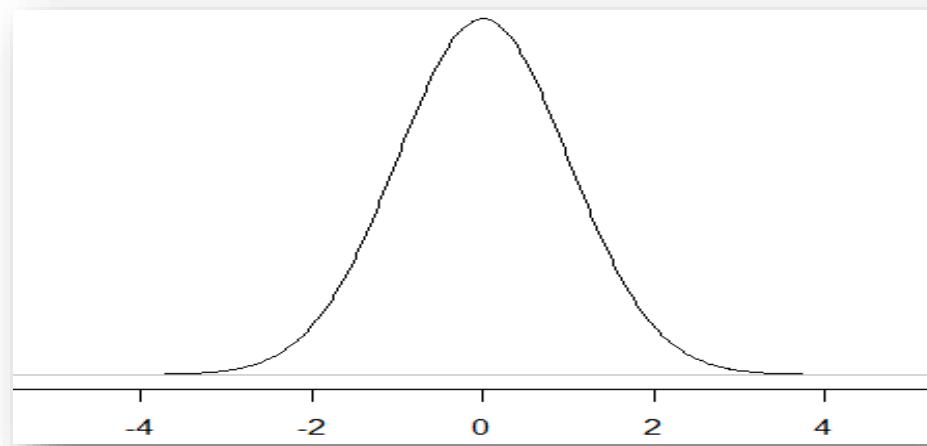


Image Courtesy: [mathsisfun.com](http://mathsisfun.com)

# Need for Standardization

- To measure variables with different means or standard deviations on a single scale.
- Easy to interpret Z-value
- Subtracting the mean from each value of the variable is called **centering**
- Dividing each value of the variable by standard deviation of the variable is called **scaling**
- We will often do **centering** and **scaling** for machine learning algorithms

# Python Functions for Normal Distribution

`norm.pdf( x,loc,scale )`

Applies for  $P(X=k)$

`norm.cdf( x,loc,scale )`

Applies for  $P(X \leq k)$

`norm.sf(x,loc,scale )`

Applies for  $P(X > k)$

`norm.stats(loc,scale)`

Applies for extracting mean, variance and other moments

`norm.ppf(q, loc, scale)`

Applies for inverse of cdf

By default if, loc and scale are not specified then standard normal distribution is assumed by all the functions

# Examples:

## Example 1:

Suppose that the height of a female in a geographical region is normally distributed with  $\mu = 64$  inches and  $\sigma = 4$  inches.

- What is the probability of finding a woman who will be less than 58 inches tall ?

## Example 2 :

Suppose the weight of a typical male in a geographical region follows a normal distribution with  $\mu = 180\text{lb}$  and  $\sigma = 30\text{lb}$ .

What fraction of males weigh more than 200 pounds?

## Example 3

A fast-food restaurant sells As and Bs. On a typical weekday the demand for As is normally distributed with mean 313 and standard deviation 57; the demand for Bs is normally distributed with mean 93 and standard deviation 22.

- A ) How many As must the restaurant stock to be 98% sure of not running out of stock on a given day ?
- B ) How many Bs must the restaurant stock to be 90% sure of not running out on a given day ?
- C ) If the restaurant stocks 450 As and 150 Bs for a given day, what is the probability that it will run out of As or Bs (or both) that day ? Assume that the demand for As and Bs are probabilistically independent.

# Simulation

# Covering...

- What is Simulation?
- Monte Carlo Technique
- Examples in Excel
- Examples in Python

# The Essence of Computer Simulation

- A **stochastic system** is a system that evolves over time according to one or more probability distributions.
- **Computer simulation** imitates the operation of such a system by using the corresponding probability distributions to *randomly generate* the various events that occur in the system.
- Rather than literally operating a physical system, the computer just records the occurrences of the *simulated* events and the resulting performance of the system.
- Computer simulation is typically used when the stochastic system involved is too complex to be analyzed satisfactorily by analytical models.

# Example 1: A Coin-Flipping Game

- ◎ Rules of the game:

1. Each play of the game involves repeatedly flipping an unbiased coin until the *difference* between the number of heads and tails tossed is three.
2. To play the game, you are required to pay \$1 for each flip of the coin. You are not allowed to quit during the play of a game.
3. You receive \$8 at the end of each play of the game.

- Examples:

---

HHH	3 flips	You win \$5
THTTT	5 flips	You win \$3
THHTHTHTTTT	11 flips	You lose \$3

---

# Computer Simulation of Coin-Flipping Game

- A computer cannot flip coins. Instead it generates a sequence of *random numbers*.
- A number is a **random number** between 0 and 1 if it has been generated in such a way that *every* possible number within the interval has an equal chance of occurring.
- An easy way to generate random numbers is to use the RAND() function in Excel.
- To simulate the flip of a coin, let half the possible random numbers correspond to heads and the other half to tails.
  - 0.0000 to 0.4999 correspond to heads.
  - 0.5000 to 0.9999 correspond to tails.

# Simulation Modeling

- One begins a simulation by developing a mathematical statement of the problem.
- The model should be realistic yet solvable within the speed and storage constraints of the computer system being used.
- Input values for the model as well as probability estimates for the random variables must then be determined.

# Random Variables

- Random variable values are utilized in the model through a technique known as Monte Carlo simulation.
- Each random variable is mapped to a set of numbers so that each time one number in that set is generated, the corresponding value of the random variable is given as an input to the model.
- The mapping is done in such a way that the likelihood that a particular number is chosen is the same as the probability that the corresponding value of the random variable occurs.

# Simulation Programs

- The computer program that performs the simulation is called a simulator.
- Flowcharts can be useful in writing such a program.
- While this program can be written in any general purpose language (e.g. BASIC, FORTRAN, C++, etc.) special languages which reduce the amount of code which must be written to perform the simulation have been developed.
- Special simulation languages include SIMSCRIPT, SPSS, DYNAMO, and SLAM.

# Experimental Design

- Experimental design is an important consideration in the simulation process.
- Issues such as the length of time of the simulation and the treatment of initial data outputs from the model must be addressed prior to collecting and analyzing output data.
- Normally one is interested in results for the steady state (long run) operation of the system being modeled.
- The initial data inputs to the simulation generally represent a start-up period for the process and it may be important that the data outputs for this start-up period be neglected for predicting this long run behavior.

# Example: Dynogen, Inc.

The price change of shares of Dynogen, Inc. has been observed over the past 50 trades. The frequency distribution is as follows:

<u>Price Change</u>	<u>Number of Trades</u>
-3/8	4
-1/4	2
-1/8	8
0	20
+1/8	10
+1/4	3
+3/8	2
+1/2	1
Total = 50	

# Example: Dynogen, Inc.

## ◎Relative Frequency Distribution and Random Number Mapping

<u>Price Change</u>	<u>Relative Frequency</u>	<u>Rnd Numbers</u>
-3/8	.08	00 - 07
-1/4	.04	08 - 11
-1/8	.16	12 - 27
0	.40	28 - 67
+1/8	.20	68 - 87
+1/4	.06	88 - 93
+3/8	.04	94 - 97
+1/2	<u>.02</u>	98 - 99
TOTAL	1.00	

# Example: Dynogen, Inc.

If the current price per share of Dynogen is 23, use random numbers to simulate the price per share over the next 10 trades.

Use the following stream of random numbers: 21, 84, 07, 30, 94, 57, 57, 19, 84, 84

# Example: Dynogen, Inc.

- **Simulation Worksheet**

Trade <u>Number</u>	Random <u>Number</u>	Price <u>Change</u>	Stock <u>Price</u>
1	21	-1/8	22 7/8
2	84	+1/8	23
3	07	-3/8	22 5/8
4	30	0	22 5/8
5	94	+3/8	23
6	57	0	23
7	57	0	23
8	19	-1/8	22 7/8
9	84	+1/8	23
10	84	+1/8	23 1/8

# Example: Dynogen, Inc.

- Spreadsheet for Stock Price Simulation

Lower	Upper	Price	Trade Number	Price Change	Stock Price
Random Number	Random Number	Change			
0.00	0.08	-0.375	1	0.125	23.125
0.08	0.12	-0.250	2	0.375	23.500
0.12	0.28	-0.125	3	0.000	23.500
0.28	0.68	0.000	4	0.000	23.500
0.68	0.88	0.125	5	0.000	23.500
0.88	0.94	0.250	6	0.000	23.500
0.94	0.98	0.375	7	0.125	23.625
0.98	1.00	0.500	8	0.125	23.750
			9	0.000	23.750
			10	0.125	<b>23.875</b>

# Example: Dynogen, Inc.

## ◎Theoretical Results and Observed Results

Based on the probability distribution, the expected price change per trade can be calculated by:

$$\begin{aligned} & (.08)(-3/8) + (.04)(-1/4) + (.16)(-1/8) + (.40)(0) \\ & + (.20)(1/8) + (.06)(1/4) + (.04)(3/8) + (.02)(1/2) = +.005 \end{aligned}$$

The expected price change for 10 trades is  $(10)(.005) = .05$ . Hence, the expected stock price after 10 trades is  $23 + .05 = 23.05$ .

Compare this ending price with the spreadsheet simulation and “manual” simulation results on the previous slides.

# Sick drivers problem

- At a bus terminal every bus should leave with the driver. At the terminus they keep 2 drivers as reserved if any one on scheduled duty is sick and could not come. Following is the probability distribution that driver becomes sick:

No. of Absent Drivers	0	1	2	3	4	5
Probability	0.30	0.20	0.15	0.10	0.13	0.12

Simulate the data for a week and find utilization of reserved drivers.  
Also find how many days and how many buses cannot run because of non-availability of drivers.

# Purchase - sale

- A trader deals in a perishable commodity, the daily demand and supply of which are random variables. Records of the past 500 trading days are shown below:

Supply		Demand	
Availability(Kg)	No. of days	Demand (kg)	No. of days
10	40	10	50
20	50	20	110
30	190	30	200
40	150	40	100
50	70	50	40

The trader buys the commodity at Rs. 20 per kg and sells it at Rs.30 per kg. Any commodity remaining at the end of a day results in a loss of Rs.8 per kg (after resale). Simulate the supply-demand data for 30 days. Calculate corresponding purchases, sales and profit/loss. Also calculate total profit/loss for those 30 days.

# References used

- Introduction to Management Science
  - By Hillier and Hillier
- Introduction to Management Science Quantitative approaches in decision making
  - By Anderson Sweeney Williams
- Statistical and Quantitative Methods
  - By Ranjeet Chitale

# Thank You

# Statistical Inference

Essentials

# Essential Terms

- Independence of Variables
- Independent Identically Distributed Variables
- Law of Large Numbers
- Types of Sampling
- Sampling Distribution
- Central Limit Theorem

# Independence of Variables

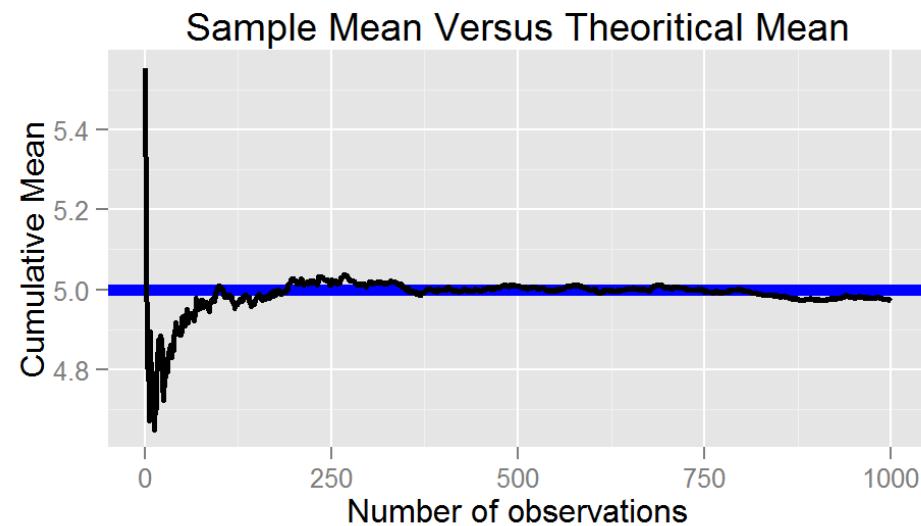
- The variables are said to be independent if any of the variable taking any value does not depend on the values of other variables
- e.g. In a table of employees, Variable ***Gender*** has values Male and Female, Variable ***City*** has values Pune, Bangalore and Hyderabad. Here ***Gender*** and ***City*** residing are independent of each other.

# Identically Distributed Variables

- Identically Distributed variables are the variables with same probability distribution
- Independent Identically Distributed variables (iid) are the variables with same probability distribution and are mutually independent
- e.g. Variables  $X_1, X_2, X_3$  all following same Normal Distribution with mean 90 and variance 140, where  $X_1, X_2, X_3$  represent the scores of three batsmen

# Law of large Numbers

- Law of large numbers states that, as the number of identically distributed, randomly generated observations increases, their sample mean (average) approaches their theoretical mean.
- The law of large numbers was first proved by the Swiss mathematician Jakob Bernoulli in 1713.



# Sampling Distribution

# A Sample

- When we draw a sample and study it, we find its characteristics by calculating its measures.
- We may calculate its mean and standard deviation.
- By calculating the measures, we intend to find the estimates of corresponding population parameters.

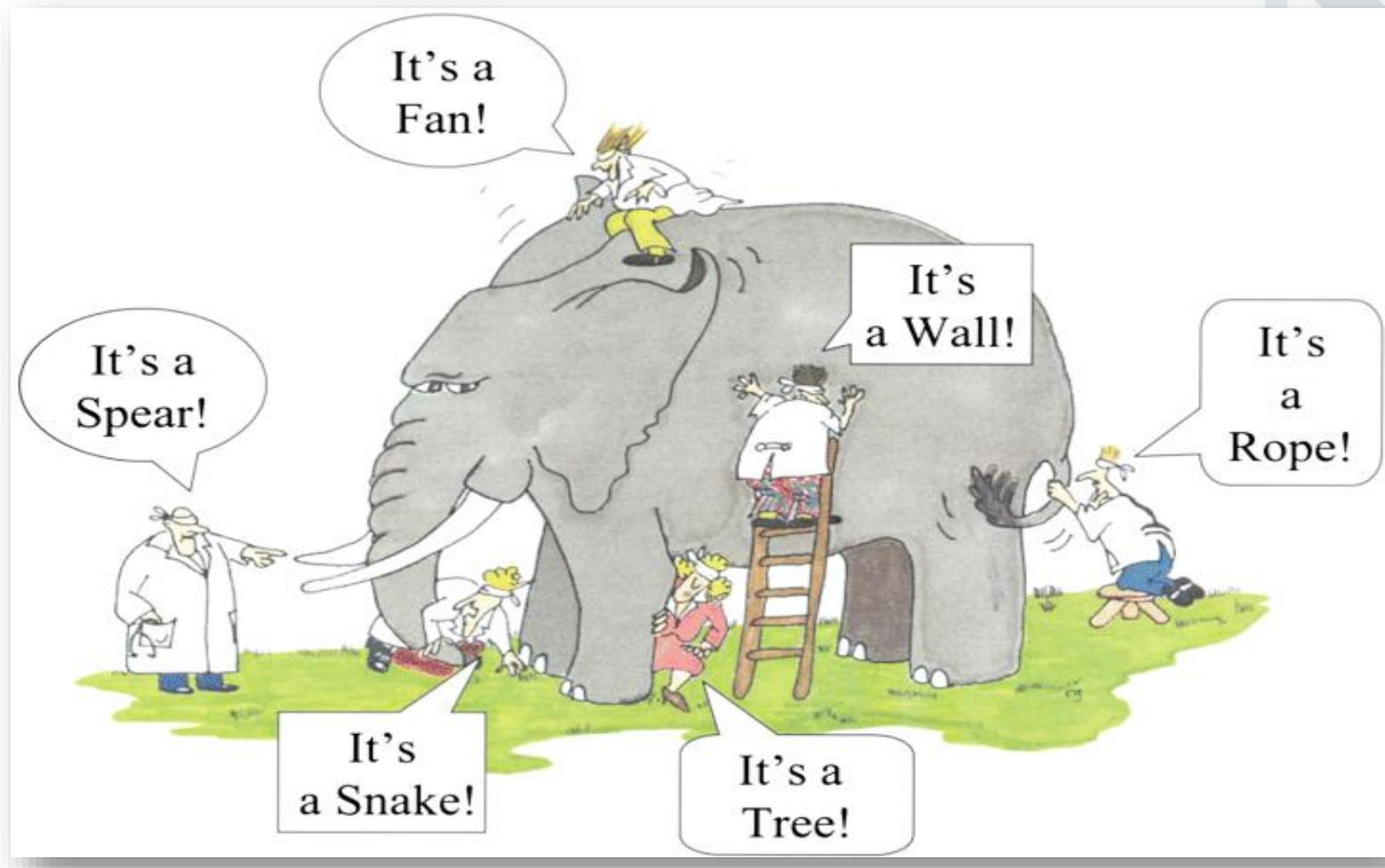
# Example

- Consider that we want to estimate the average salaries of Oracle Functional Consultants working in IT industry in India with 3-4 years experience.
- So we draw a random sample of size 10 and calculate the measures.

8.9	9.3	6.7	8.5	5.66	7.66	10.2	11.3	12.4	9.21
			Mean =	8.983					

- Hence we have the sample mean as 8.983 lakhs.

# To what extent can we believe in the sample?



Our sample  
Mean = 8.983

Sample from  
person C:  
Mean = 10.09

Sample from  
person A:  
Mean = 8.64

Sample from  
person B:  
Mean = 9.683

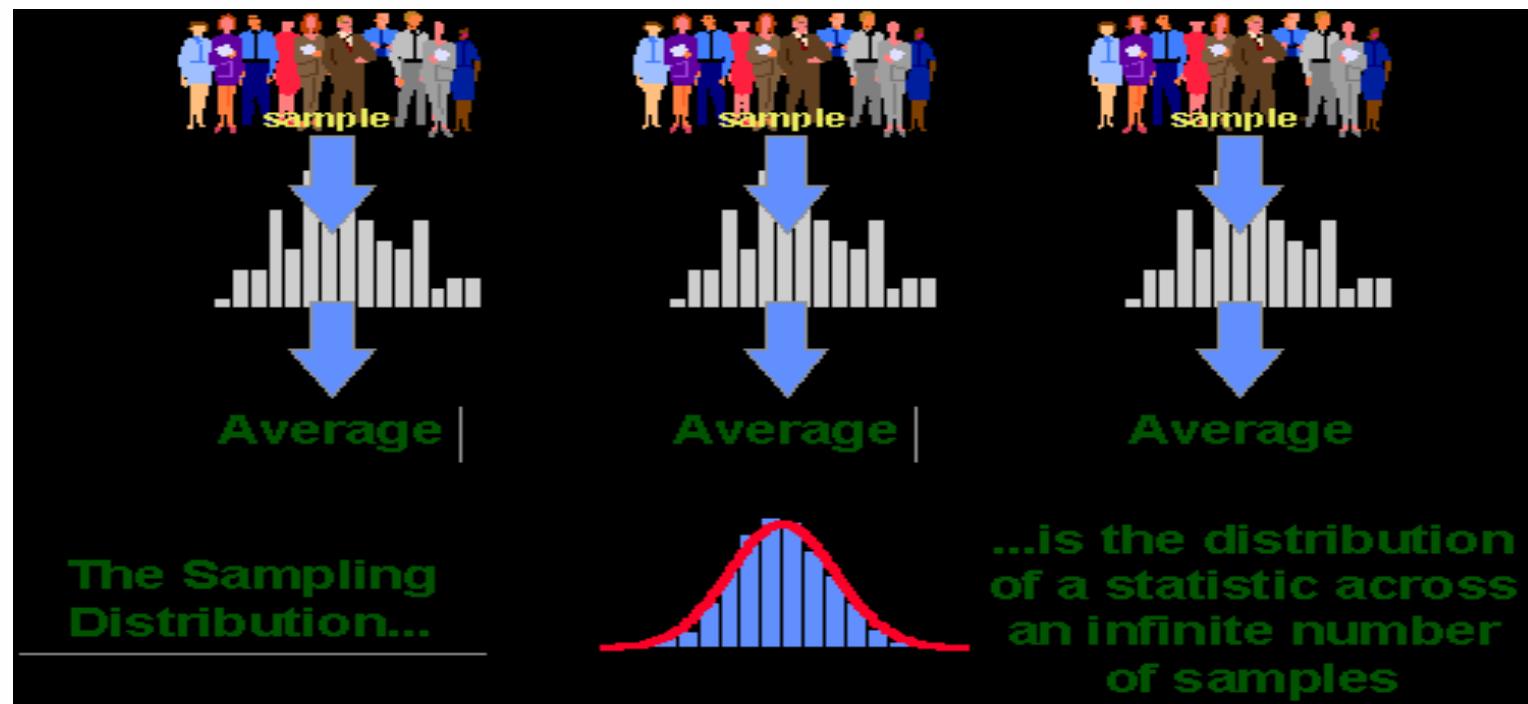
Sample from  
person D:  
Mean = 8.77

Each Sample  
consists of 10  
observations

Hence we see that if the experiment is performed by different people we get different values for the sample mean.

# So we implies that...

- The values from different samples namely by persons A, B, C, D etc. also follow a random pattern.
- This pattern of randomness is the sampling distribution.



# Population and Sample Notations

	Population	Sample
Mean	$\mu$	$\bar{x}$
Variance	$\sigma^2$	$s^2$
Standard Deviation	$\sigma$	$s$

Population  
Parameters

Statistics

A Statistic is said to be an estimator of a population parameter.

# Central Limit Theorem

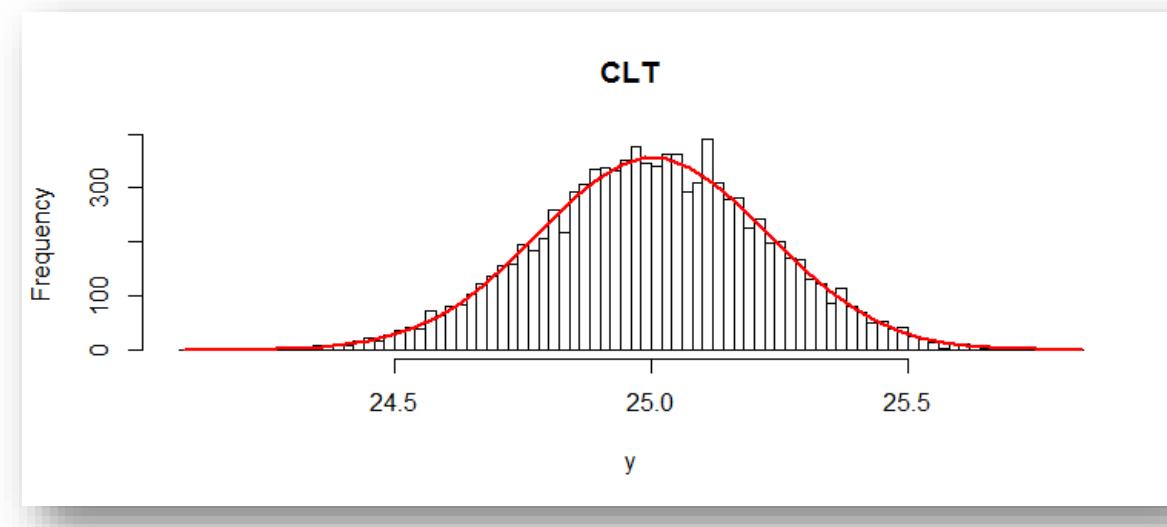
- The central limit theorem states that the sampling distribution of the mean of any independent, random variable will be normal or nearly normal, if the sample size is large enough.
- In practice, some statistics practitioners say that a sample size of 30 is large enough for the population distribution to be roughly bell-shaped.
- Hence, we can state population parameters for distribution of means as:

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

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

# Illustration of CLT

- Suppose a sample of 1000 observations is taken 10,000 times.
- Each time the sum of the sample of 1000 is calculated and we have 10,000 such means
- Then we draw the following histogram for the distribution means
- Hence we see here that the shape of the histogram tends to be a bell shaped symmetric curve



# Estimation Terminology

- Point Estimate
- Estimation Error
- Standard Error of the estimate
- Interval Estimate

# Point Estimate

- Point estimate is calculated as being a “best guess” of the population parameter. e.g. Sample mean is point estimate of population mean.
- Points estimates can never be equal to the population parameter. The difference between the point estimate and the true value of the population parameter is called estimation error or sampling error.
- As the sample size increases point estimates come closer to their corresponding population parameters
- Also we have seen that distribution of point estimates is Sampling Distribution

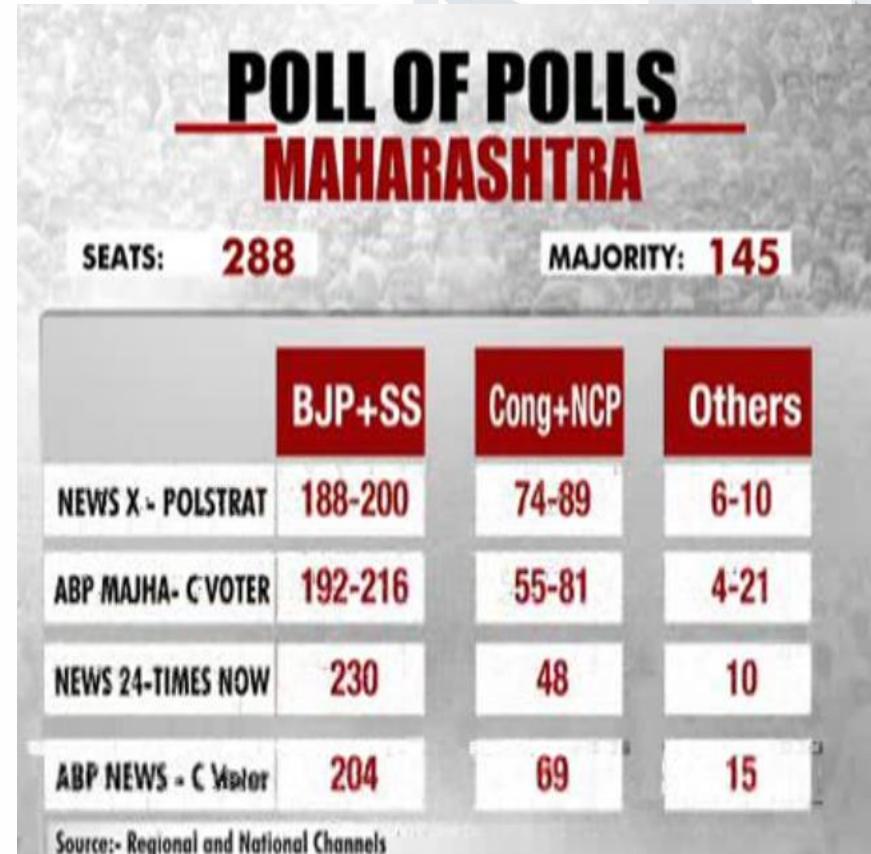
# Standard Error of the Estimate

- It is a measure of uncertainty associated with the point estimate
- For Population,  $SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- As  $\sigma$  is unknown for any population we use the following formula:  
 $SE = \frac{s}{\sqrt{n}}$  where n is sample size

```
> mtcars$mpg  
[1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4  
[20] 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4  
> SE <- sd(mtcars$mpg)/sqrt(nrow(mtcars))  
> SE  
[1] 1.065424
```

# Interval Estimation

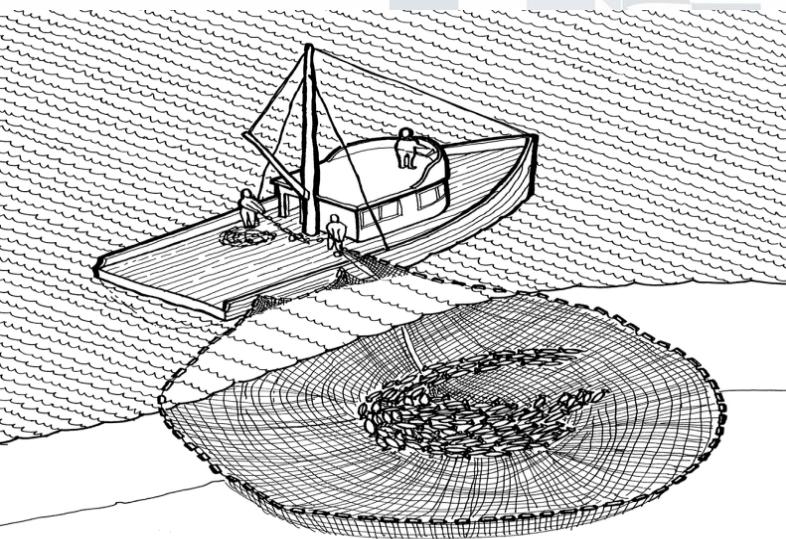
- A confidence interval is an interval around the point estimate calculated from the sample data, where it is strongly believed that the true value of the population parameter lies.
- Here we get the lower bound value and upper bound value
- e.g. [12.3, 41.5] C.I. indicates that the true population parameter value might be between 12.3 and 41.5



# Difference in Point and Interval Estimation



Point Estimation



Interval Estimation

# Confidence Interval

- Interval  $(x_1, x_2)$  is said to be 95% confidence interval of population parameter  $\mu$ ,
  - If  $P(x_1 \leq \mu \leq x_2) = 0.95$
- Similarly,
- Interval  $(x_1, x_2)$  is said to be 99% confidence interval of population parameter  $\mu$ ,
  - If  $P(x_1 \leq \mu \leq x_2) = 0.99$

# C.I. of $\mu$

- Assuming Normal Distribution, C.I. of mean  $\mu$  with known standard deviation is given by the formula(Sample size = n):

$$(\bar{x} - z.value \frac{\sigma}{\sqrt{n}}, \bar{x} + z.value \frac{\sigma}{\sqrt{n}})$$

- Assuming Normal Distribution, C.I. of mean  $\mu$  with unknown standard deviation is given by the formula(Sample size = n):

$$(\bar{x} - t.value \frac{s}{\sqrt{n}}, \bar{x} + t.value \frac{s}{\sqrt{n}})$$

# Margin of Error

- The quantity  $t\text{.value}^* s / \sqrt{n}$  is margin of error.
- More is the margin of error wider is the C.I.
- If its 95% C.I. then its confidence coefficient is 0.95.
- If its 99% C.I. then its confidence coefficient is 0.99.
- Confidence coefficient is denoted by  $(1 - \alpha)$
- More the confidence coefficient wider is the C.I.
- Also greater is the sample size  $n$ , lesser would be the margin of error.

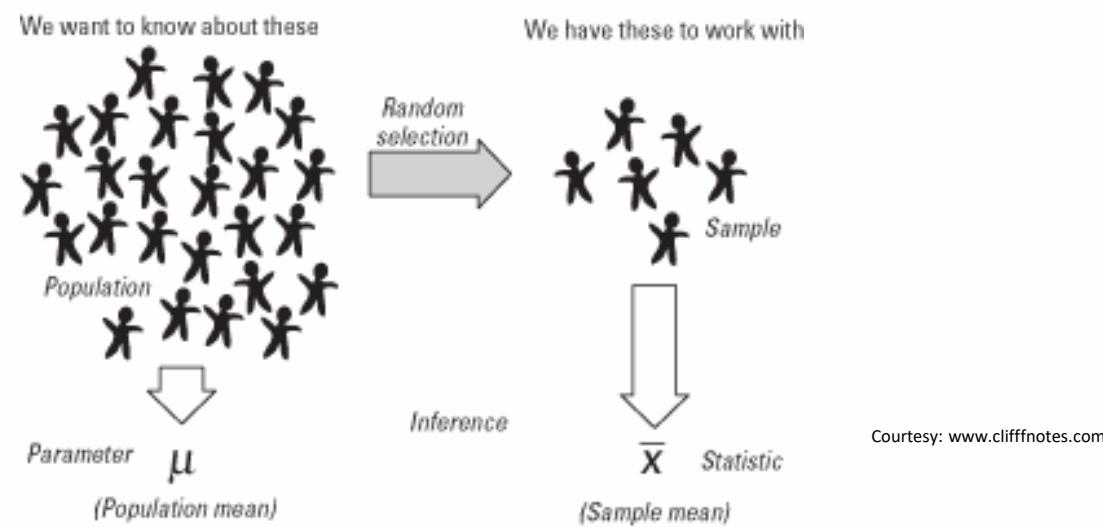
# Hypothesis Testing

# Statistical Hypothesis

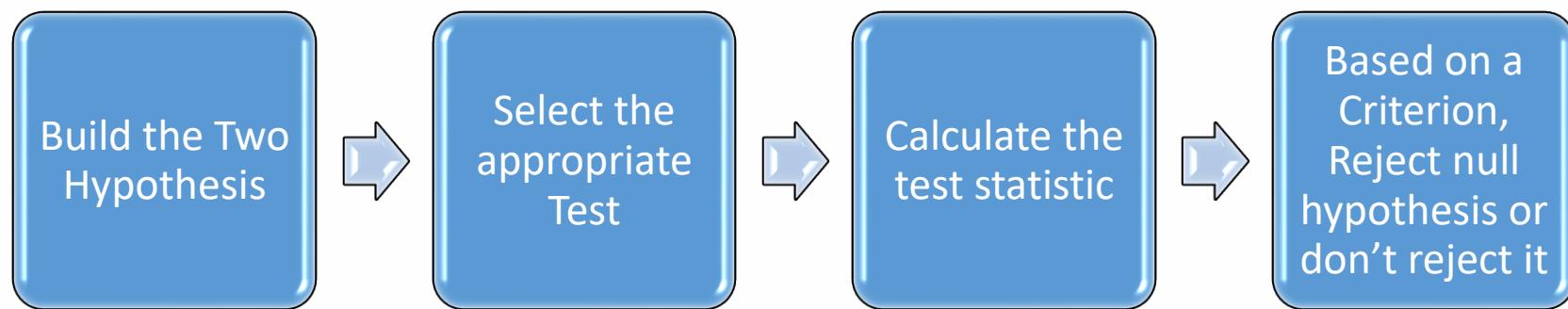
- A hypothesis is a statement or assertion about the state of nature. e.g. The campaign was effective.
- Every hypothesis implies its contradiction or alternative. e.g. The campaign was not effective.
- Either of the hypothesis statement can be true or false.

# Testing the Hypothesis

- For testing any hypothesis we make use of the sample data.
- Based on the sample data, we reject either of the hypothesis statements.
- The decision to reject or failing to reject any of the hypothesis can be either correct or wrong.



# Flow for Hypothesis Testing



# Actual and Findings

- Found fact would be based on sample whereas actual fact would be based on population.
- We won't be able to directly find the actual fact due to the large size of population.
- Hence we analyze the sample and try to find the fact.
- It may happen that:
  - Found fact from the sample is same as actual fact
  - Found fact from the sample is exactly opposite of the actual fact.

# Actual and Findings

- When found fact from sample is not same as the actual fact from the population then we would be wrong.
- We are interested in knowing the probability of us getting wrong.

# Statistical Hypothesis

- Null Hypothesis: A **null hypothesis**, denoted by  $H_0$ , is an assertive statement about one or more population parameters. This is the statement we hold to be true until we have sufficient statistical evidence to reject it.
- The **alternative hypothesis**, denoted by  $H_1$ , is the assertive statement of all situations *not* covered by the null hypothesis.

# Example

- H<sub>0</sub>: The mean sales are 1400.
- H<sub>1</sub>: The mean sales are not 1400.
- OR
- H<sub>0</sub>: The mean sales are less than or equal to 1400.
- H<sub>1</sub>: The mean sales are greater than 1400.

# Example

- Consider the example:
  - $H_0$ : The mean sales are less than or equal to 1400.
  - $H_1$ : The mean sales are greater than 1400.
- Suppose we draw a random sample for testing the above hypotheses.
- We can be wrong in either of the cases:
  - The mean sales for population are less than or equal to 1400 and our sample suggests it as greater than 1400.(Rejecting true  $H_0$ )
  - The mean sales for population are greater than 1400 and our sample suggests it as less than or equal to 1400.(Failing to reject the false  $H_0$ )

# Errors

- Type I: Rejecting true  $H_0$
- Type II: Failing to reject the false  $H_0$
- $P(\text{Type I Error}) = \alpha$
- $P(\text{Type II Error}) = \beta$
- If we try to reduce  $\alpha$ , then  $\beta$  increases and if we try to reduce  $\beta$ , then  $\alpha$  increases.
- More serious is the error  $\alpha$ . Hence a level for  $\alpha$  is maintained. This level is called level of significance. It is denoted by  $\alpha$ . Usually its is maintained as 0.1 or 0.05 or 0.01.
- More often maintained as 0.05.

# Types of Errors

Decision	H <sub>0</sub> True	H <sub>0</sub> False
Reject H <sub>0</sub>	Type I Error ( $\alpha$ ) Producer's Risk	Correct
Fail to reject H <sub>0</sub>	Correct	Type II Error ( $\beta$ ) Consumer's Risk

# P - Value

- The *p-value* is the probability of getting a value of the test statistic as extreme as, or more extreme than, the actual value obtained, when the null hypothesis is true.
- The p-value is the smallest level of significance,  $\alpha$ , at which the null hypothesis may be rejected using the obtained value of the test statistic.
- **Policy to be followed: When the *p*-value is less than  $\alpha$  , reject  $H_0$ , otherwise we do not reject  $H_0$ .**

# Power of a Test

- The **power** of a statistical hypothesis test is the probability of rejecting the null hypothesis when the null hypothesis is false.

$$\text{Power} = (1 - \beta)$$

- Power is the probability that your test will reject the null hypothesis when the null hypothesis is false, or the probability that you will detect a difference when a difference actually exists.

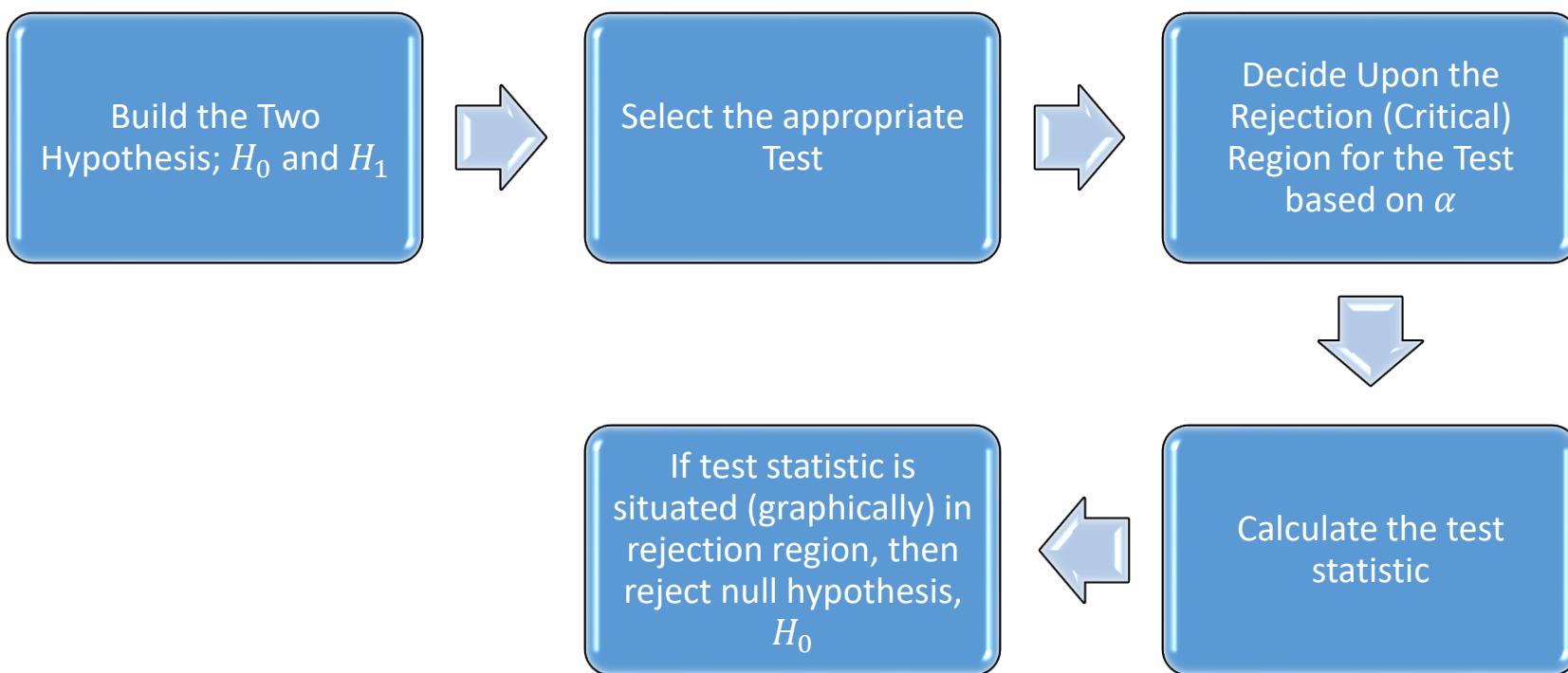
# Tail of Test

- H<sub>0</sub>: The mean sales are 1400. i.e.  $\mu = 1400$
- H<sub>1</sub>: The mean sales are not 1400. i.e.  $\mu \neq 1400$ 
  - This is **two tailed** test as in H<sub>1</sub>, mean sales can be greater than or less than 1400.
- H<sub>0</sub>: The mean sales are less than or equal to 1400. i.e.  $\mu \leq 1400$
- H<sub>1</sub>: The mean sales are greater than 1400. i.e.  $\mu > 1400$ 
  - This is **right tailed** or **upper tailed** test as in H<sub>1</sub>, mean sales are greater than 1400.
- H<sub>0</sub>: The mean sales are greater than or equal to 1400. i.e.  $\mu \geq 1400$
- H<sub>1</sub>: The mean sales are less than 1400. i.e.  $\mu < 1400$ 
  - This is **left tailed** or **lower tailed** test as in H<sub>1</sub>, mean sales are less than 1400.

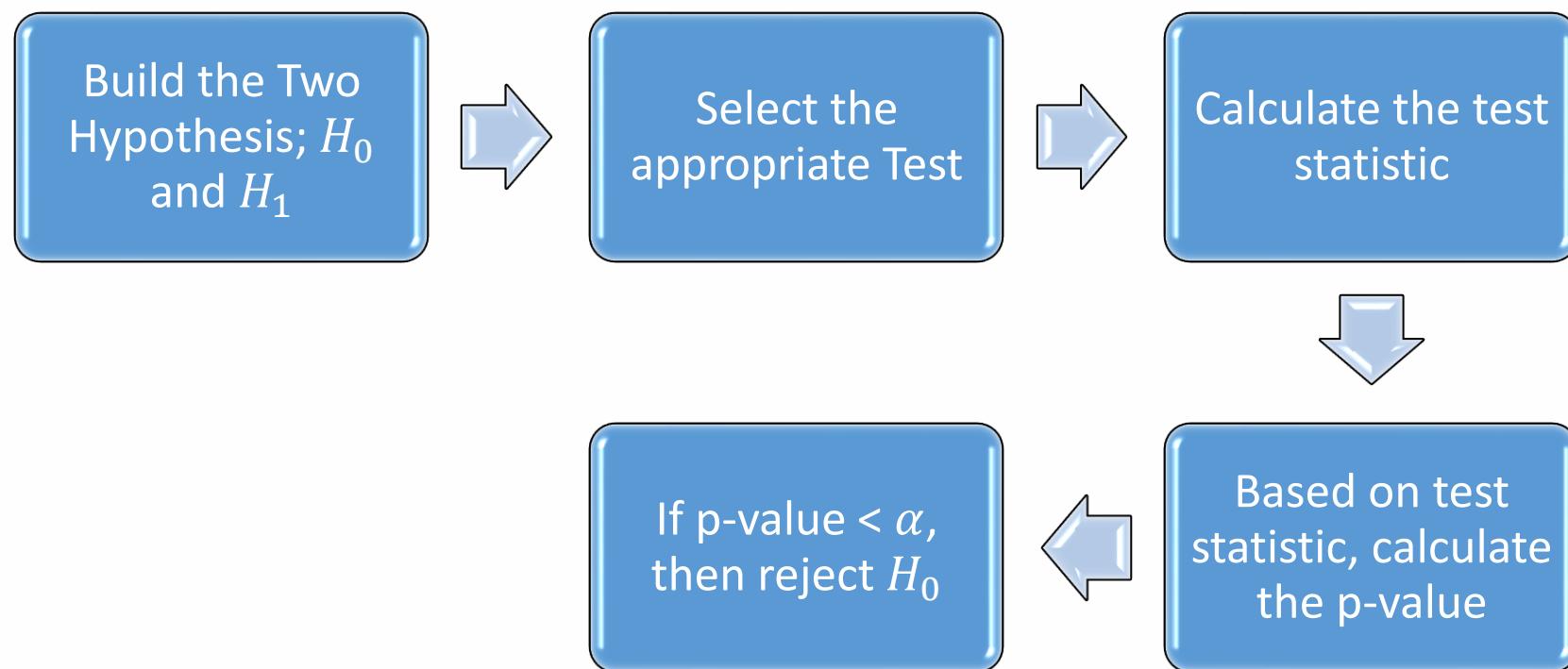
# Testing of Hypothesis Approaches

- Critical Value Approach
- P-value Approach

# Critical Value Approach



# P-value Approach



# Types of Hypothesis Tests

- Parametric: Tests which assume a particular distribution(Normal) of the population
  - t-test
  - F-test
  - Chi-square test for variance
- Non-Parametric: Tests which assume no particular distribution of the population
  - Median Test
  - Wilcoxon's Rank Sum Test
  - Mann-Whitney Test
  - Kruskal Wallis Test



A large, abstract graphic element occupies the upper right portion of the slide. It consists of a grid of light gray squares of varying sizes, some of which are filled with a darker shade of gray. This creates a pattern that resembles a stylized tree or a network of interconnected paths.

Thank You



# Hypothesis Tests

Population Mean for one sample

Population Standard Deviation for one sample

# Parametric Tests for Means

- Test for means is done under two assumptions:
  - Population Standard Deviation is known
  - Population Standard Deviation is unknown
- For Known Standard Deviation: Z-test
- For Unknown Std Deviation: t-test
- Assumption: Sample is drawn from a population which follows Normal Distribution



Parametric Test

## **T-TEST FOR ONE SAMPLE MEAN**

# One Sample t-test

- One Sample t-test is test for mean of single population
- **Assumption:** Sample has been drawn from population which is Normal
- Suppose that we want to test whether population mean of the population from which sample is drawn is a particular value, say  $\mu_0$

$$H_0: \mu = \mu_0 \text{ against } H_1: \mu \neq \mu_0$$

# Test Statistic

- The test statistic of the t-test can be proved to be following t distribution with  $(n-1)$  degrees of freedom
- The test statistic is given by:

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

Where

$\bar{x}$  : Sample Mean

$\mu_0$  : Population mean to be tested

$s$  : Sample Standard Deviation

$n$  : Sample Size

# Example

- Given data on plant growth contains weights of dried plants for three different treatments
- We want to test the hypothesis whether the mean weight of the dried plants is 6 for the population

# Solution

$H_0: \mu = 6$  against  $H_1: \mu \neq 6$

$$t = \frac{\bar{x} - 6}{\frac{s}{\sqrt{n}}}$$

Where

$\bar{x}$  : Sample Mean

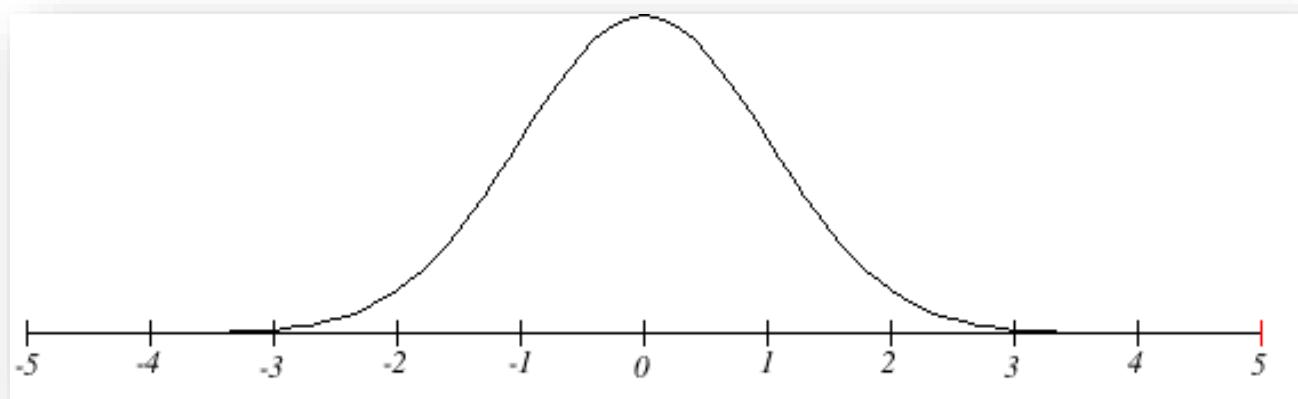
$\mu_0$  : Population mean to be tested

$s$  : Sample Standard Deviation

$n$  : Sample Size

# Calculations

$$t = \frac{\bar{x} - 6}{s / \sqrt{n}} = -7.241$$



- We see here that area under the curve at both sides is less than 0.01. Hence, we reject null hypothesis  $H_0$

# One Sample t-test in Python

Syntax:

```
scipy.stats.ttest_1samp(a, popmean, ...)
```

where

a : Numerical vector

popmean: Population mean to be tested ( $\mu_0$ )

# Python Program

```
In [1]: import pandas as pd
        from scipy import stats
        PlantGrowth = pd.read_csv("G:/Statistics (Python)/Datasets/PlantGrowth.csv")
        stats.ttest_1samp(PlantGrowth.weight, popmean=6.0)
```

```
Out[1]: Ttest_1sampResult(statistic=-7.241082682752039, pvalue=5.666151490495602e-08)
```

- We observe here that p-value is less than 0.05 and even 0.01. Hence, we reject  $H_0$ .
- Conclusion: Population mean of the weight of plants may not be 6 at 1% level of significance



A large, abstract graphic in the upper right corner consists of a grid of light gray and white squares of varying sizes, creating a tessellated or woven pattern.

# Questions?

# **Comparison of Two Population**

*TESTS OF MEANS*

# Comparing Means

- For comparing means of two populations, we can use the following two alternatives assuming that the distribution of population is Normal:
  - Paired t-test : Matched Samples
  - Two Samples t-test : Independent Samples



Test for Matched Samples

## **PAIRED T-TEST**

# Paired-samples Scenario

- We apply this test when we have data with matched samples

Prewt	Postwt
80.7	80.2
89.4	80.1
91.8	86.4
74.0	86.3
78.1	76.1
88.3	78.1

- In the given example, we have Weight of the patient before treatment in Prewt and weight of the same patient after treatment in Postwt.
- The data in both the columns is matched samples data. Here, we will be interested in knowing whether the average weight before treatment is significantly different from that after treatment.
- In other words, we want to analyze as : **Did treatment make any impact on weight of the patients?**

# Paired t-test

- Let  $x_i$  and  $y_i$  be the paired observations under study with  $n$  as the sample size
- Let  $d_i = x_i - y_i$  be the difference in corresponding paired observations
- Let  $s_d$  be the sample standard deviation for the difference and  $\bar{d}$  be the mean of differences in samples
- Let  $D$  be the population mean for difference
- The two tailed hypotheses for the test can be written as:

$$H_0 : D = 0 \text{ against } H_1 : D \neq 0$$

- The test statistic of paired t-test is

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

The test statistic has  $n-1$  degrees of freedom

# Paired t-test in Python

`scipy.stats.ttest_rel(a, b, axis=0, nan_policy='propagate')`

[\[source\]](#)

Calculate the T-test on TWO RELATED samples of scores, a and b.

This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values.

Parameters: `a, b : array_like`

The arrays must have the same shape.

`axis : int or None, optional`

Axis along which to compute test. If `None`, compute over the whole arrays, `a`, and `b`.

`nan_policy : {'propagate', 'raise', 'omit'}, optional`

Defines how to handle when input contains `nan`. `'propagate'` returns `nan`, `'raise'` throws an error, `'omit'` performs the calculations ignoring `nan` values. Default is `'propagate'`.

Returns: `statistic : float or array`

t-statistic

`pvalue : float or array`

two-tailed p-value

# Example : Paired t-test

- Consider the a subset of the dataset *anorexia* from the package **MASS** with Treat = Cont.

```
In [8]: anoCont = anorexia[anorexia.Treat == "Cont"]
...: anoCont.head()
Out[8]:
   Treat  Prewt  Postwt
0  Cont    80.7    80.2
1  Cont    89.4    80.1
2  Cont    91.8    86.4
3  Cont    74.0    86.3
4  Cont    78.1    76.1
```

- We find here, whether there is any significant difference between Prewt and Postwt.

# Example : Paired t-test

- Hypothesis:

$H_0 : D = 0$  i.e. There is no difference in weights before and after treatment. Hence, treatment **Cont** is not effective against

$H_1 : D \neq 0$  i.e. There is some difference in weights before and after treatment. Hence, treatment **Cont** may be effective

# R Program and Output

```
In [9]: stats.ttest_rel(anoCont.Prewt,anoCont.Postwt)
Out[9]: Ttest_relResult(statistic=0.2872253910150255,
pvalue=0.7763070622194167)
```

- We observe here that the p-value is greater than 0.05, hence we are inclined to not reject  $H_0$ .
- Conclusion: The treatment **Cont** might not be effective

# Example : Paired t-test

- Let us consider some other treatment in the data namely, **FT** and perform the similar test on the data

```
In [10]: anoFT = anorexia[anorexia.Treat == "FT"]
....: anoFT.head()
Out[10]:
   Treat  Prewt  Postwt
55     FT    83.8    95.2
56     FT    83.3    94.3
57     FT    86.0    91.5
58     FT    82.5    91.9
59     FT    86.7   100.3
```

# R Program and Output

```
In [11]: stats.ttest_rel(anoFT.Prewt,anoFT.Postwt)
Out[11]: Ttest_relResult(statistic=-4.184908135290033,
pvalue=0.0007002531056005393)
```

- We observe that p-value is less than 0.05 and even less than 0.01. Hence we reject  $H_0$  at 1% level of significance.
- Conclusion: Treatment **FT** might be effective.

# One-Tailed Test

- We can also consider here the hypothesis as

$$H_0: D \geq 0 \text{ against } H_1 : D < 0$$

```
In [11]: stats.ttest_rel(anoFT.Prewt,anoFT.Postwt)
Out[11]: Ttest_relResult(statistic=-4.184908135290033,
pvalue=0.0007002531056005393)
```

```
In [16]: pvalue2tailed = stats.ttest_rel(anoFT.Prewt,anoFT.Postwt)[1]
....: pvaluetailed = stats.ttest_rel(anoFT.Prewt,anoFT.Postwt)[1]/2
```

```
In [17]: pvalue2tailed
Out[17]: 0.0007002531056005393
```

```
In [18]: pvaluetailed
Out[18]: 0.00035012655280026967
```

- We observe that p-value is less than 0.05 and even less than 0.01. Hence we reject  $H_0$  at 1% level of significance.
- Conclusion: Treatment FT might be effective in increasing weight.



# 2 INDEPENDENT SAMPLES TESTS

# Two Sample Tests

- Used to test whether there is a significant difference between the means of two samples.
- Here, the two samples are independent
- Two-tailed Hypotheses for variances can be stated as:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ against } H_1: \sigma_1^2 \neq \sigma_2^2$$

- Two-tailed Hypotheses for means can be stated as:

$$H_0: \mu_1 = \mu_2 \text{ against } H_1: \mu_1 \neq \mu_2$$

## Two Sample Test for Variance : Bartlett's test

- This test checks the equality of variances

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 \text{ against } H_1: \sigma_i^2 \neq \sigma_j^2 \text{ for atleast one pair } (i,j)$$

Bartlett's test is used to test the null hypothesis,  $H_0$  that all  $k$  population variances are equal against the alternative that at least two are different.

If there are  $k$  samples with sizes  $n_i$  and sample variances  $S_i^2$  then Bartlett's test statistic is

$$\chi^2 = \frac{(N - k) \ln(S_p^2) - \sum_{i=1}^k (n_i - 1) \ln(S_i^2)}{1 + \frac{1}{3(k-1)} \left( \sum_{i=1}^k \left( \frac{1}{n_i - 1} \right) - \frac{1}{N-k} \right)}$$

where  $N = \sum_{i=1}^k n_i$  and  $S_p^2 = \frac{1}{N-k} \sum_i (n_i - 1) S_i^2$  is the pooled estimate for the variance.

The test statistic has approximately a  $\chi^2_{k-1}$  distribution. Thus the null hypothesis is rejected if  $\chi^2 > \chi^2_{k-1,\alpha}$  (where  $\chi^2_{k-1,\alpha}$  is the upper tail critical value for the  $\chi^2_{k-1}$  distribution).

Source: [Bartlett's test - Wikipedia](#)

# Example: Two Samples Test

- The CO<sub>2</sub> uptake of six plants from Quebec and six plants from Mississippi was measured at several levels of ambient CO<sub>2</sub> concentration.
- Half the plants of each type were chilled overnight before the experiment was conducted.
- We will see whether there is a significant difference in the variances and means of Uptake in plants

# Dataset: CO2

```
In [20]: co2.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 84 entries, 0 to 83
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Plant        84 non-null    object  
 1   Type         84 non-null    object  
 2   Treatment    84 non-null    object  
 3   conc          84 non-null    int64   
 4   uptake        84 non-null    float64 
dtypes: float64(1), int64(1), object(3)
memory usage: 3.4+ KB
```

# Program and Output

$$H_0: \sigma_{chilled}^2 = \sigma_{non-chilled}^2$$

$$H_1: \sigma_{chilled}^2 \neq \sigma_{non-chilled}^2$$

```
In [5]: import pandas as pd
....: from scipy import stats
....:
....: co2 = pd.read_csv("CO2.csv")
....: co2_chill = co2[co2.Treatment == "chilled"]
....: co2_nonchill = co2[co2.Treatment == "nonchilled"]
....:
....: uptake_chill = co2_chill.uptake
....: uptake_nonchill = co2_nonchill.uptake
....: stats.bartlett(uptake_chill,uptake_nonchill)
Out[5]: BartlettResult(statistic=0.5315695885641828, pvalue=0.46594771841246396)
```

- We observe that p-value is greater than 0.05. Hence we cannot reject  $H_0$  at 5 % level of significance
- Conclusion : Variances of uptakes of two treatments might be same

# T-test for two Samples

- The t-test for comparison of means between two samples can be applied under two cases:
  - Two Samples having same variance
  - Two Samples having different variance
- It has got different statistics under these two cases

# Two Samples With Equal Variances

- The test statistic in this case is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$\bar{x}_1$  : Sample mean of sample from population 1

$\bar{x}_2$  : Sample mean of sample from population 2

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

# Two Samples With Equal Variances

- The t statistic has degrees of freedom as

$$(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

# Two Samples With Unequal Variances

- The test statistic in this case is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where

$\bar{x}_1$  : Sample mean of sample from population 1

$\bar{x}_2$  : Sample mean of sample from population 2

$s_1^2$  : Sample variance of sample from population 1

$s_2^2$ : Sample variance of sample from population 2

# Two Samples With Unequal Variances

- The t statistic has degrees of freedom as

$$df = \frac{\left( \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

# Example : Mean Uptake

- We had seen in the example of CO<sub>2</sub>, that the variances of the uptake values are equal.
- Let us examine the means of uptake and compare them for variable Treatment

$$H_0: \mu_{chilled} = \mu_{non-chilled} \quad H_1: \mu_{chilled} \neq \mu_{non-chilled}$$

```
In [9]: stats.ttest_ind(uptake_chill,uptake_nonchill,equal_var=True)
Out[9]: Ttest_indResult(statistic=-3.0484611149819503,
pvalue=0.0030957332525416484)
```

Conclusion : The means of uptake values may not be equal for two treatments at 5% level of significance

# Example : CO2

- We can also test for the following hypothesis:

$$H_0: \mu_{chilled} \geq \mu_{non-chilled} \quad H_1: \mu_{chilled} < \mu_{non-chilled}$$

```
In [16]: stats.ttest_ind(uptake_chill,uptake_nonchill,equal_var=True)
Out[16]: Ttest_indResult(statistic=-3.0484611149819503,
pvalue=0.0030957332525416484)

In [17]: pvalue2tailed =
stats.ttest_ind(uptake_chill,uptake_nonchill,equal_var=False)[1]

In [18]: pvalue1tailed =
stats.ttest_ind(uptake_chill,uptake_nonchill,equal_var=False)[1]/2

In [19]: pvalue1tailed
Out[19]: 0.0015534684495496483
```

Conclusion : The mean of uptake values of chilled plants may be lesser than mean of uptake values of non-chilled plants at 5% level of significance



A large, abstract graphic in the upper right corner consists of a grid of light gray and white squares of varying sizes, creating a tessellated or woven pattern.

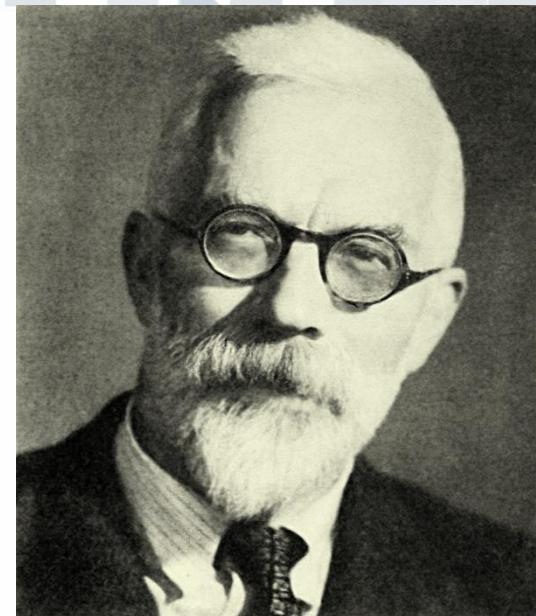
Thank You



# Analysis of Variance

# What is ANOVA?

- **ANOVA** (ANalysis Of VAriance) is a statistical method for testing the equality of several population means.
  - ANOVA is designed to detect differences among means from populations subject to different groups often called as *treatments*
  - ANOVA tests for the equality of several population means by calculating and analyzing the two estimators of the population variance. Hence, the name *analysis of variance*.
- This technique was developed by Statistician Prof. Ronald Fisher



Ronald Fisher



# One-Way

ANOVA

# 1-way ANOVA Model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$


Overall Effect      Treatment Effect      Error

- In 1-way, we think of any observation value(univariate) to be comprised of
  - An overall effect
  - Group or treatment effect
  - Error

# Example

- Consider an agricultural experiment, in which we check the yield of a crop planted on a plot of land.
- Suppose that we divide the plot in 4 parts in the interest of applying 4 different treatments (fertilizers) to the parts.
- In the four parts, suppose that we are able to plant 6, 7, 5 and 6 plants respectively.

I	II	III	IV
$y_{11}$	$y_{21}$	$y_{31}$	$y_{41}$
$y_{12}$	$y_{22}$	$y_{32}$	$y_{42}$
$y_{13}$	$y_{23}$	$y_{33}$	$y_{43}$
$y_{14}$	$y_{24}$	$y_{34}$	$y_{44}$
$y_{15}$	$y_{25}$	$y_{35}$	$y_{45}$
$y_{16}$	$y_{26}$		$y_{46}$
	$y_{27}$		

where ,  $y_{ij}$ : yield (kg) of  $j^{th}$  plant from  $i^{th}$  part of the plot

# Example

- After a certain period (an year), we note down the yields of all the plants as follows:

I	II	III	IV
23.4	34.2	23.8	36.7
24.1	45.2	24.5	39.5
19.6	24.9	29.3	43.2
23.9	40.3	18.3	50.2
29.4	39.4	19.4	47.2
21.9	35.3		34.1
	38.4		

# Statements of Hypothesis

- The hypothesis test of analysis of variance:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

$H_1$ : Not all  $\mu_i$  ( $i = 1, \dots, r$ ) are equal

- In our example,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_1$ : Not all  $\mu_i$  ( $i = 1, 2, 3, 4$ ) are equal

# Hypothesis Test of ANOVA

- In an analysis of variance:
  - We have  $r$  independent random samples, each one corresponding to a population subject to a different treatment.
  - We have:
    - $n = n_1 + n_2 + n_3 + \dots + n_r$  total observations.
    - $r$  sample means:  $x_1, x_2, x_3, \dots, x_r$
    - $r$  sample variances:  $s_{12}, s_{22}, s_{32}, \dots, s_{r2}$ 
      - These sample variances can be used to find a pooled estimator of the population variance.

# Result of ANOVA

Sources of Variation	Sums of Squares	Degrees of freedom	Mean Square	F Ratio	P-Value
Treatment	SSTR	r - 1	MSTR=SSTR / (r - 1)	MSTR/MSE	
Error	SSE	n - r	MSE = SSE / (n - r)		
Total	SST	n - 1			

$$SSTR = \sum_i \frac{(\sum_j y_{ij})^2}{n_i} - \frac{(\sum_j \sum_i y_{ij})^2}{n}$$

$$SSE = \sum_j \sum_i y_{ij}^2 - \sum_i \frac{(\sum_j y_{ij})^2}{n_i}$$

$$SST = \sum_j \sum_i y_{ij}^2 - \frac{(\sum_j \sum_i y_{ij})^2}{n}$$

# Example

I	II	III	IV
23.4	34.2	23.8	36.7
24.1	45.2	24.5	39.5
19.6	24.9	29.3	43.2
23.9	40.3	18.3	50.2
29.4	39.4	19.4	47.2
21.9	35.3		34.1
	38.4		

- In our example,  $r = 4$  ,  $n = 6+7+5+6 = 24$
- Our Python, function `anova_lm()` calculates not only the means and variances but also all the sums of squares

# ANOVA in Python

Syntax :

```
anova_lm(*args, **kwargs)
```

## Where

args : fitted linear model results instance

One or more fitted linear models

scale : float

Estimate of variance, If None, will be estimated from the largest model. Default is None.

test : str {"F", "Chisq", "Cp"} or None

Test statistics to provide. Default is "F".

typ : str or int {"I", "II", "III"} or {1,2,3}

The type of ANOVA test to perform.

# R Program and Output

```
In [39]: import pandas as pd
.....
....: from statsmodels.stats.anova import anova_lm
....: from statsmodels.formula.api import ols
....: #####Example 1#####
....: agr = pd.read_csv("G:/Statistics (Python)/Datasets/Yield.csv")
....: agrYield = ols('Yield ~ Treatments', data=agr).fit()
....: table = anova_lm(agrYield, typ=2)
....: print(table)
      sum_sq      df          F    PR(>F)
Treatments  1551.607762   3.0  18.293252  0.000006
Residual     565.457238  20.0        NaN      NaN
```

As p-value < 0.01, we can reject  $H_0$  at 1% level of significance. Hence, we conclude that the yields are significantly different for all the 4 treatments.

# Assumptions

- We assume *independent random sampling* from each of the  $r$  populations
- We assume that the  $r$  populations under study:
  - are *normally distributed*,
  - with means  $\mu_i$  that may or may not be equal,
  - but with *equal variances*,  $\sigma_i^2$ .

# Statements of Hypothesis

- The hypothesis test of analysis of variance:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

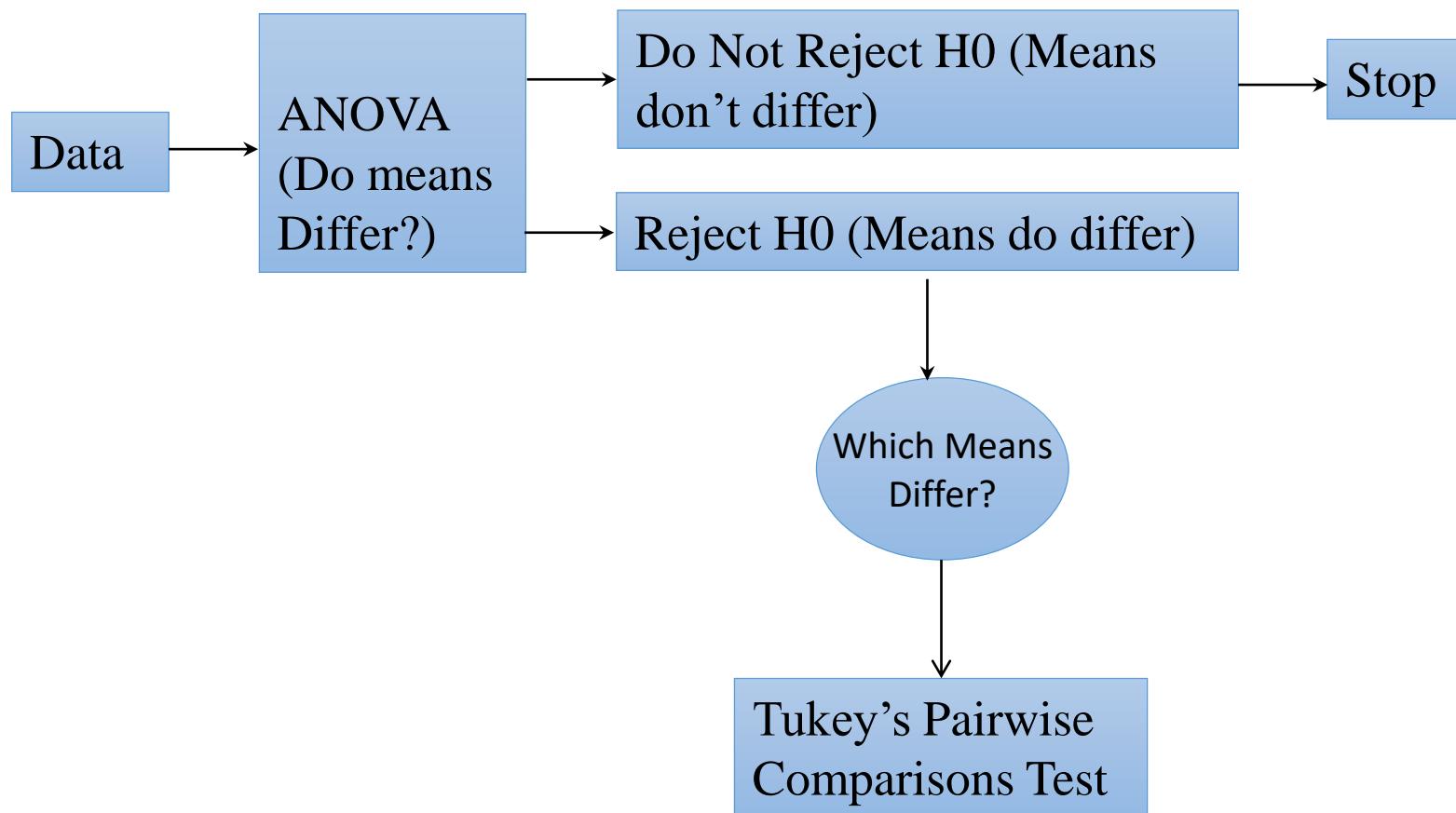
$H_1$ : Not all  $\mu_i$  ( $i = 1, \dots, r$ ) are equal

- In our example,

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$H_1$ : Not all  $\mu_i$  ( $i = 1,2,3,4$ ) are equal

# Further Analysis



# Tukey's Test in Python

In [9]:

```
##### Post Hoc Tukey HSD #####
from statsmodels.stats.multicomp import pairwise_tukeyhsd

compare = pairwise_tukeyhsd(agr.Yield, agr.Treatments, alpha=0.05)
pd.DataFrame(compare._results_table.data)
```

Out[9]:

	0	1	2	3	4	5
0	group1	group2	meandiff	lower	upper	reject
1	I	II	13.0976	4.8174	21.3779	True
2	I	III	-0.6567	-9.6689	8.3556	False
3	I	IV	18.1	9.5072	26.6928	True
4	II	III	-13.7543	-22.469	-5.0396	True
5	II	IV	5.0024	-3.2779	13.2826	False
6	III	IV	18.7567	9.7444	27.7689	True

The p-values for pair-wise comparisons namely II & I , IV & I , III & II , IV & III indicate that they have significant differences.

# Further Studies

- Two way ANOVA : The way we analyzed the effect of one factor variable, we can also analyze the effects of two factor variables with interaction or without interactions.
- The design we saw is called Completely Randomized Design
- There are also following designs in this field of study of statistics:
  - Factorial Design
  - Lattice Design
  - Split Plot Design
  - Repeated Measures Design
  - Multivariate Analysis of Variance

## Case: Funds

- *A magazine reports percentage returns and expense ratios for stock and bond funds.* The data FUNDS.csv are the expense ratios for 10 midcap stock funds, 10 small-cap stock funds, 10 hybrid stock funds, and 10 specialty stock funds.
- Test for any significant difference in the mean expense ratio among the four types of stock funds.



# Thank You

# Chi-Square Test

Independence of Attributes

# Contingency Table Analysis: A Chi-Square Test for Independence

		First Classification Category					Row Total
		1	2	3	4	5	
Second Classification Category	1	O <sub>11</sub>	O <sub>12</sub>	O <sub>13</sub>	O <sub>14</sub>	O <sub>15</sub>	
	2	O <sub>21</sub>	O <sub>22</sub>	O <sub>23</sub>	O <sub>24</sub>	O <sub>25</sub>	
	3	O <sub>31</sub>	O <sub>32</sub>	O <sub>33</sub>	O <sub>34</sub>	O <sub>35</sub>	
	4	O <sub>41</sub>	O <sub>42</sub>	O <sub>43</sub>	O <sub>44</sub>	O <sub>45</sub>	
	5	O <sub>51</sub>	O <sub>52</sub>	O <sub>53</sub>	O <sub>54</sub>	O <sub>55</sub>	
Column Total		C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>	C <sub>4</sub>	C <sub>5</sub>	n

# Contingency Table Analysis: A Chi-Square Test for Independence

A and B are independent if:  $P(A \cap B) = P(A) \times P(B)$ .

If the first and second classification categories are independent:  $E_{ij} = (R_i)(C_j)/n$

Null and alternative hypotheses:

H<sub>0</sub>: The two classification variables are independent of each other

H<sub>1</sub>: The two classification variables are not independent

Chi-square test statistic for independence:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Degrees of freedom:  $df = (r-1)(c-1)$

Expected cell count:  $E_{ij} = \frac{R_i C_j}{n}$

# Correlation

# Need for correlation

- To find the association between the variables
- To find the degree of the association

# What is correlation?

- The relationship between two variables is called their correlation.
- Positive Correlation: As one variable becomes large, the other also becomes large, and vice versa.
- Negative Correlation: As one variable becomes small, the other becomes large, and vice versa.

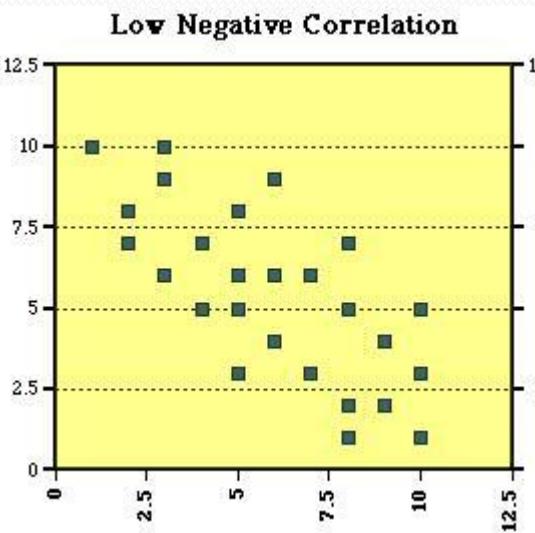
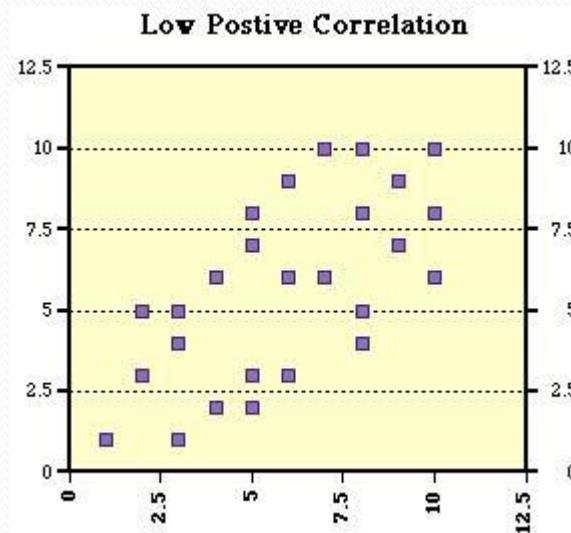
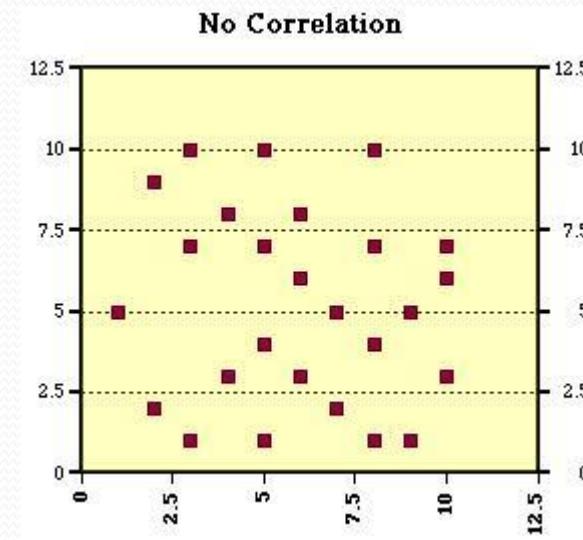
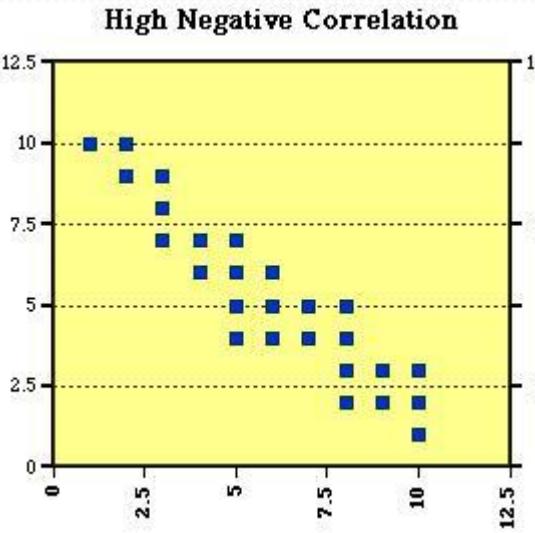
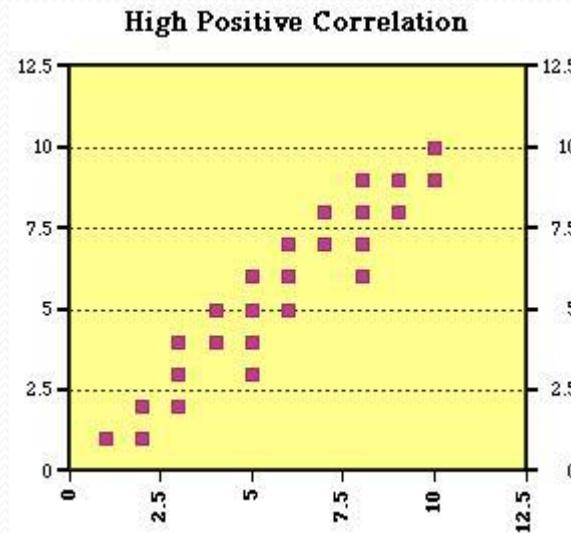
# Other types of correlation

- Linear: Corresponding to a unit change in one variable, there is a constant change in the other variable.
- Non-Linear: Corresponding to a unit change in one variable, the other variable doesn't change at a constant rate but it changes at a fluctuating rate.

# How to find the correlation?

- ▶ Scatter plots show how much one variable is affected by another.
- ▶ Correlation Coefficients give the degree of correlation.

# Example – Numerical Data (Two variables) Scatter Plot



# Karl Pearson's Coefficient

- It is calculated a formula involving variance and covariance values.

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad -1 \leq \rho \leq 1$$



# Simple Linear Regression

# Example

A firm has a chain of pizza restaurants around the country. To see the effectiveness of its advertising activities, it has collected the data from 19 randomly selected metropolitan regions. There are two variables in the data:

- Promote: Promotional Expenditure in thousand rupees
- Sales in thousand rupees

We are interested in building the relationship between the two variables.

# Simple Linear Regression Model

$$\text{Sales} = \beta_0 + \beta_1 \text{ Promote} + \varepsilon$$

Sales is a linear function of Promote plus  $\varepsilon$

$\beta_0$  and  $\beta_1$  are parameters of the model,  
 $\varepsilon$  is a random variable.

The linear term of  $\beta_0 + \beta_1 \text{ Promote}$  is the **variations in Sales that can be explained by Promote**

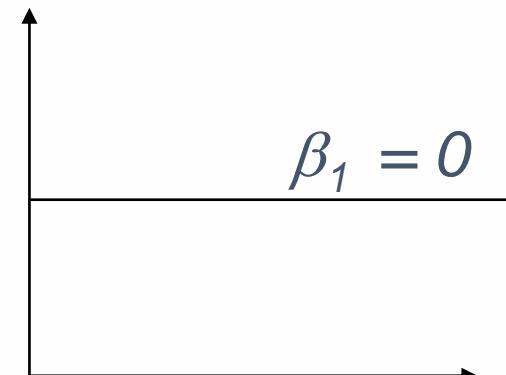
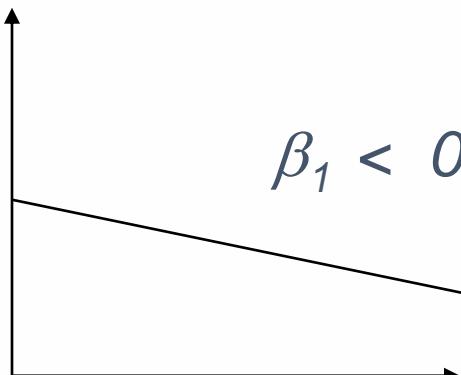
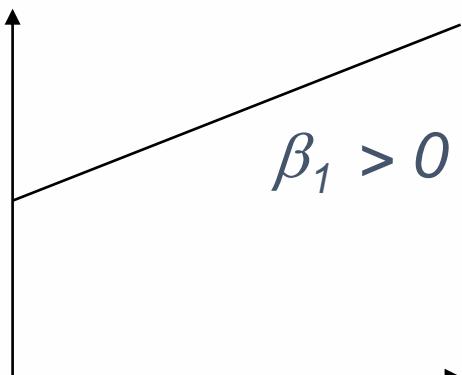
The error term of  $\varepsilon$  is variations in Sales that **can not be explained by the liner relationship between Promote and Sales.**

# Simple Linear Regression Equation

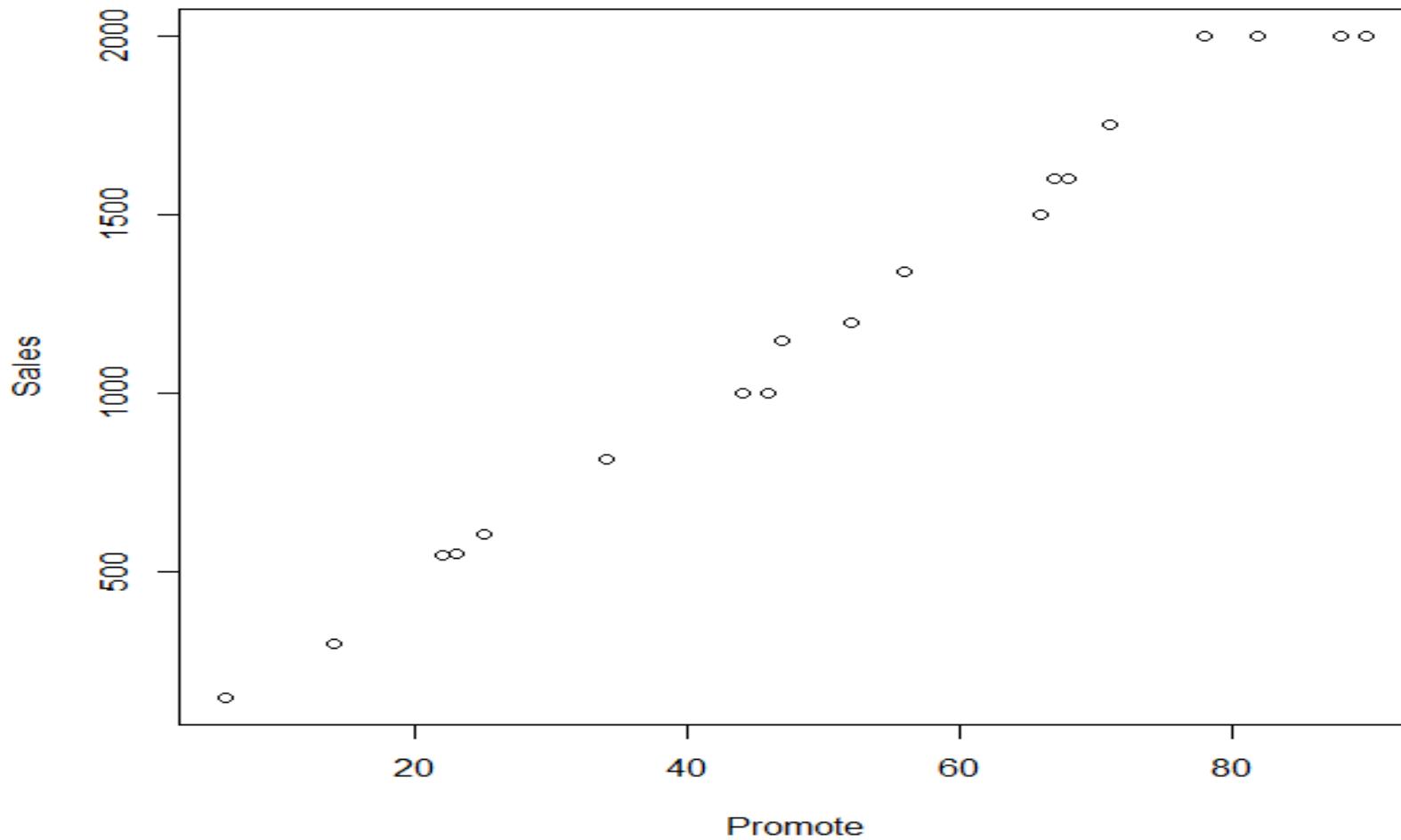
For the time being let us forget  $\varepsilon$ . The following equation describes **how the mean value of Sales is related to Promote**.

$$\text{Expected Sales} = \beta_0 + \beta_1 \text{ Promote}$$

$\beta_0$  is the intersection with y axis,  $\beta_1$  is the slope.

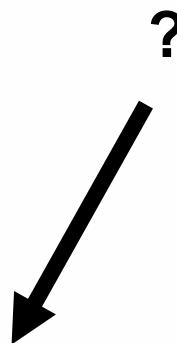


# Scatter Diagram

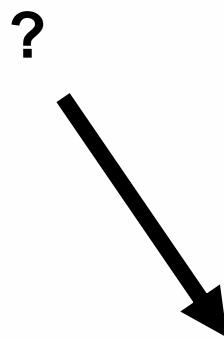


# Estimated Linear Regression Equation

We want to estimate the relationship between



Promotional Expenditure

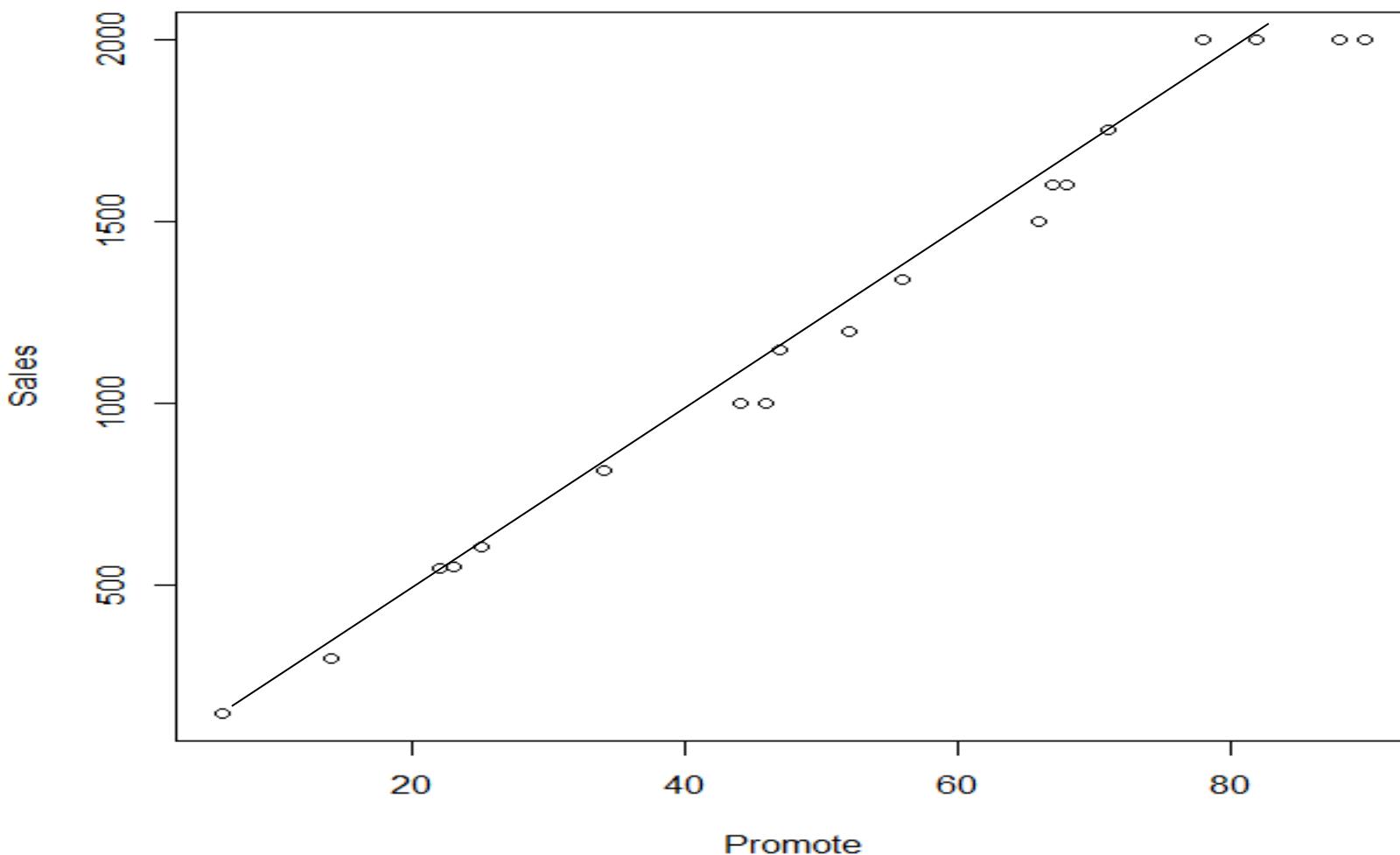


Mean value of sales

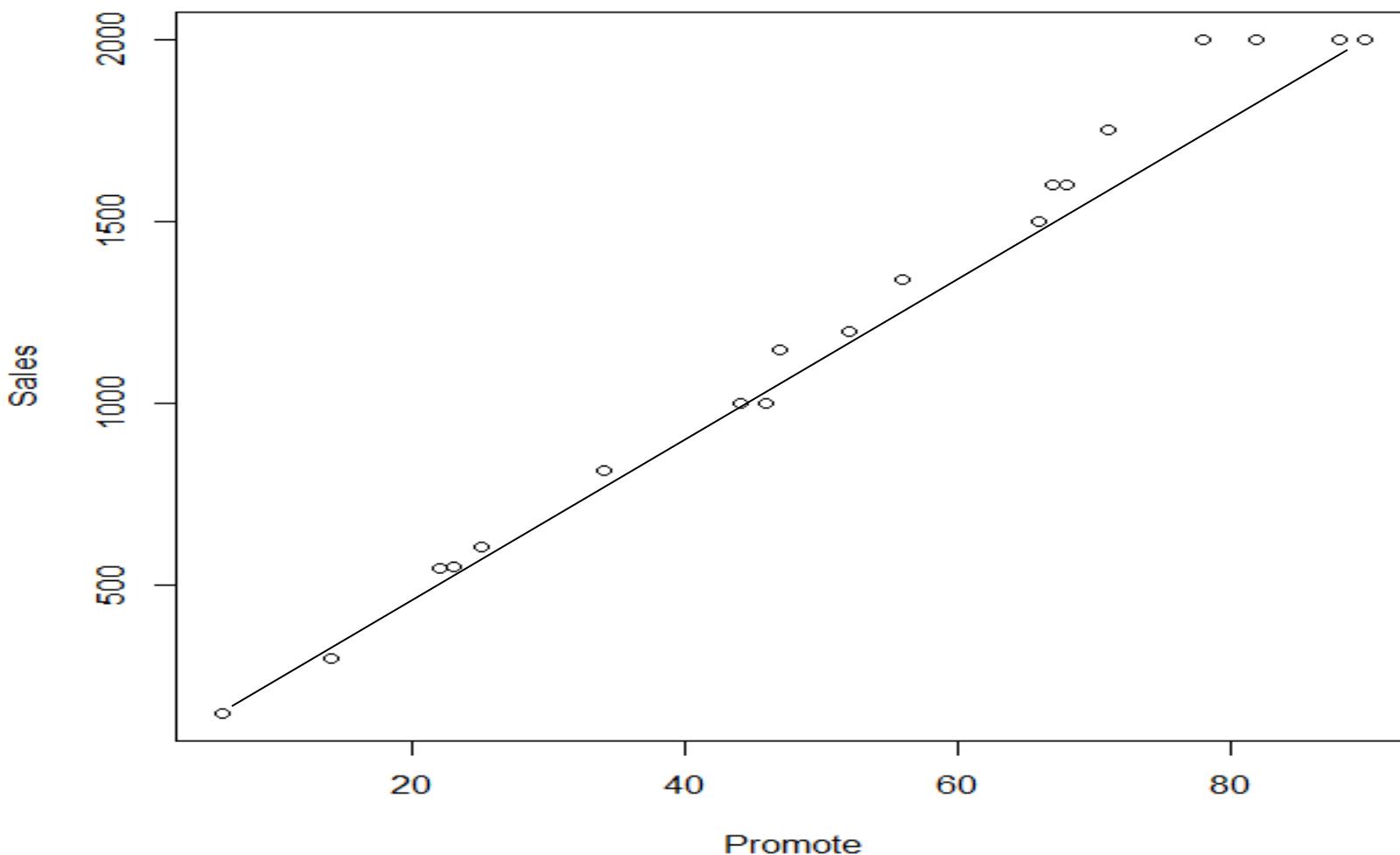
We may rely on own judgement, and draw a line to fit them.

Then we measure the intersection with y axis and that is  $b_0$ , and the slope is  $b_1$

# Judgmental Solution 1



# Judgmental Solution 2

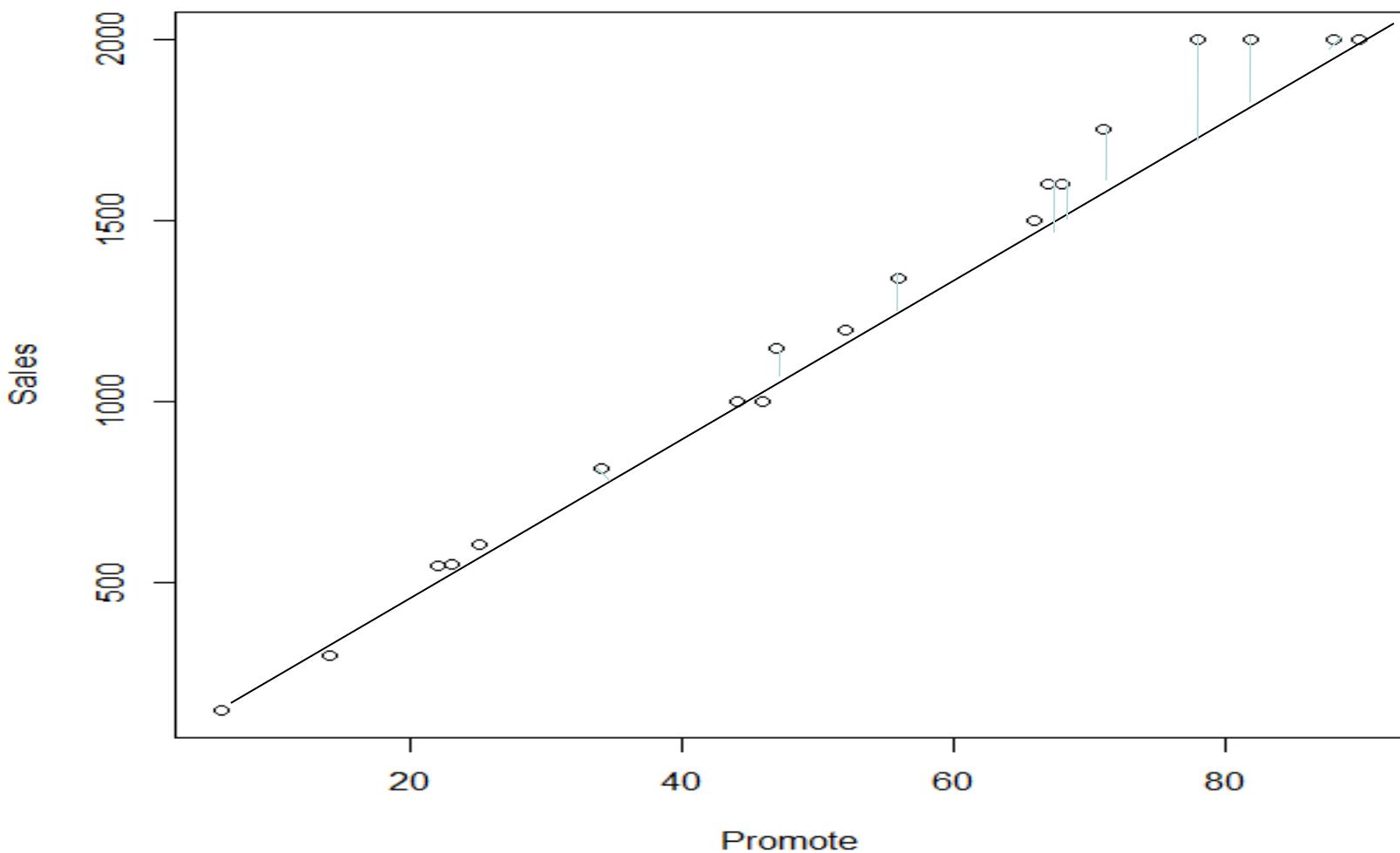


# The Least Square Method

Judgmental Solution can't be a standard approach as it may be person dependent.

We need to use algebra and calculus for correctly calculating the optimal line.

Hence we follow **The Least Square Method** approach.



# The Least Square Method

$y_i$	$x_i$	$\hat{y}_i$
$y_1$	$x_1$	$b_0 + b_1 x_1$
$y_2$	$x_2$	$b_0 + b_1 x_2$
$y_3$	$x_3$	$b_0 + b_1 x_3$
.	.	.
.	.	.
$y_n$	$x_n$	$b_0 + b_1 x_n$

$$\text{Min} \quad Z = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{Min} \quad Z = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

# Classic Minimization

$$\text{Min} \quad Z = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

We want to minimize this function with respect to  $b_0$  and  $b_1$

This is an optimization problem.

We may remember from high school algebra that to find the minimum value we should get the derivative and set it equal to zero.

# The Least Square Method

**Note :** Our unknowns are  $b_0$  and  $b_1$ .  
 $x_i$  and  $y_i$  are known. They are our data.

$y_i$	$x_i$	$\hat{y}_i$
$y_1$	$x_1$	$b_0 + b_1 x_1$
$y_2$	$x_2$	$b_0 + b_1 x_2$
$y_3$	$x_3$	$b_0 + b_1 x_3$
.	.	.
.	.	.
$y_n$	$x_n$	$b_0 + b_1 x_n$

$$Z = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

Find the derivative of  $Z$  with respect to  $b_0$  and  $b_1$  and set them equal to zero

# Derivatives

$$Z = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2$$

$$\frac{\partial Z}{\partial b_0} = \sum_{i=1}^n 2(-1)(y_i - b_0 - b_1 x_i) = 0$$

$$\frac{\partial Z}{\partial b_1} = \sum_{i=1}^n 2(-x_i)(y_i - b_0 - b_1 x_i) = 0$$

# **b<sub>0</sub>** and **b<sub>1</sub>**

$$b_1 = \frac{\sum xy - (\sum x \sum y) / n}{\sum x^2 - (\sum x)^2 / n}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

# Example

Promote(X)	Sales (Y)	XY	X square	Y Square
23	554	12742	529	306916
56	1339	74984	3136	1792921
34	815	27710	1156	664225
25	609	15225	625	370881
67	1600	107200	4489	2560000
82	2000	164000	6724	4000000
46	1000	46000	2116	1000000
14	300	4200	196	90000
6	150	900	36	22500
47	1150	54050	2209	1322500
52	1200	62400	2704	1440000
88	2000	176000	7744	4000000
71	1750	124250	5041	3062500
78	2000	156000	6084	4000000
66	1500	99000	4356	2250000
44	1000	44000	1936	1000000
68	1600	108800	4624	2560000
90	2000	180000	8100	4000000
22	550	12100	484	302500
<b>Totals</b>		<b>979</b>	<b>23117</b>	<b>1469561</b>
			<b>62289</b>	<b>34744943</b>

**b<sub>1</sub>**

$$b_1 = \frac{\sum xy - (\sum x \sum y) / n}{\sum x^2 - (\sum x)^2 / n}$$

$$b_1 = 23.506$$

**b<sub>0</sub>**

$$\bar{y} = b_0 + b_1 \bar{x}$$

$$\bar{y} = \frac{23117}{20} = 1155.85$$

$$\bar{x} = \frac{979}{20} = 48.95$$

$$1155.85 = b_0 + 23.506(48.95)$$

$$b_0 = 5.48$$

# Estimated Regression Equation

$$y = 5.48 + 23.51x$$

Now we can predict.

For example, if one of restaurants of this Pizza Chain is having an expenditure of 72

We predict the mean of its quarterly sales is

$$y = 5.48 + 23.51(72)$$

$$y = 1697.95 \quad \textit{thousand rupees}$$

# Summary : The Simple Linear Regression Model

- Simple Linear Regression Model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Simple Linear Regression Equation

$$E(y) = \beta_0 + \beta_1 x$$

- Estimated Simple Linear Regression Equation

$$\hat{y} = b_0 + b_1 x$$

# Summary : The Least Square Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where

$y_i$  = observed value of the dependent variable  
for the  $i$  th observation

$\hat{y}_i$  = estimated value of the dependent variable  
for the  $i$  th observation

# Summary : The Least Square Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum x_i y_i - (\sum x_i \sum y_i) / n}{\sum x_i^2 - (\sum x_i)^2 / n}$$

- $y$ -Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x}$$

$x_i$  = value of independent variable for  $i$  th observation

$y_i$  = value of dependent variable for  $i$  th observation

$\bar{x}$  = mean value for independent variable

$\bar{y}$  = mean value for dependent variable

$n$  = total number of observations

# Coefficient of Determination (on training set)

- It is the fraction of variation of the dependent variable explained by the regression line.
- It is another measure of goodness of fit
- Bigger the  $R^2$ , better is the model fit
- Its formula is

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} \quad 0 \leq R^2 \leq 1$$

# Dummy Variables

- Categorical Variables can be converted into indicators called dummy variables.
- e.g.
  - Gender having values 1 and 0
  - Quarter: four dummy variables Q1-Q4 with 1s and 0s

# Multiple linear regression

# Multiple Linear Regression

- Instead of fitting a line we fit a plane.
- General Form is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

# Assumptions of Linear Regression

- Normality: Errors are Normally Distributed with mean zero
- Independence: Errors are independent
- Linearity: Mean of dependent variable Y is linearly related to Xis
- Homoscedasticity: Errors have constant variance

# Decision Trees

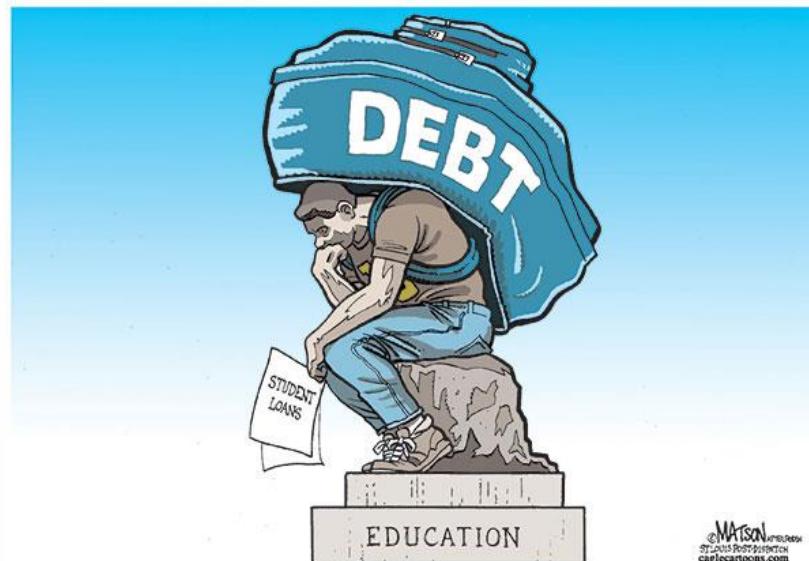
# Decision Trees

- Decision Trees create a set of binary splits on the predictor variables
- These splits are used to classify new observations into one of the two groups
- There are two categories of decision trees:
  - Classification Tree for Categorical Response Variable
  - Regression Tree for Numerical Response Variable

# Classification Trees

# Example 1: Loan Defaulter

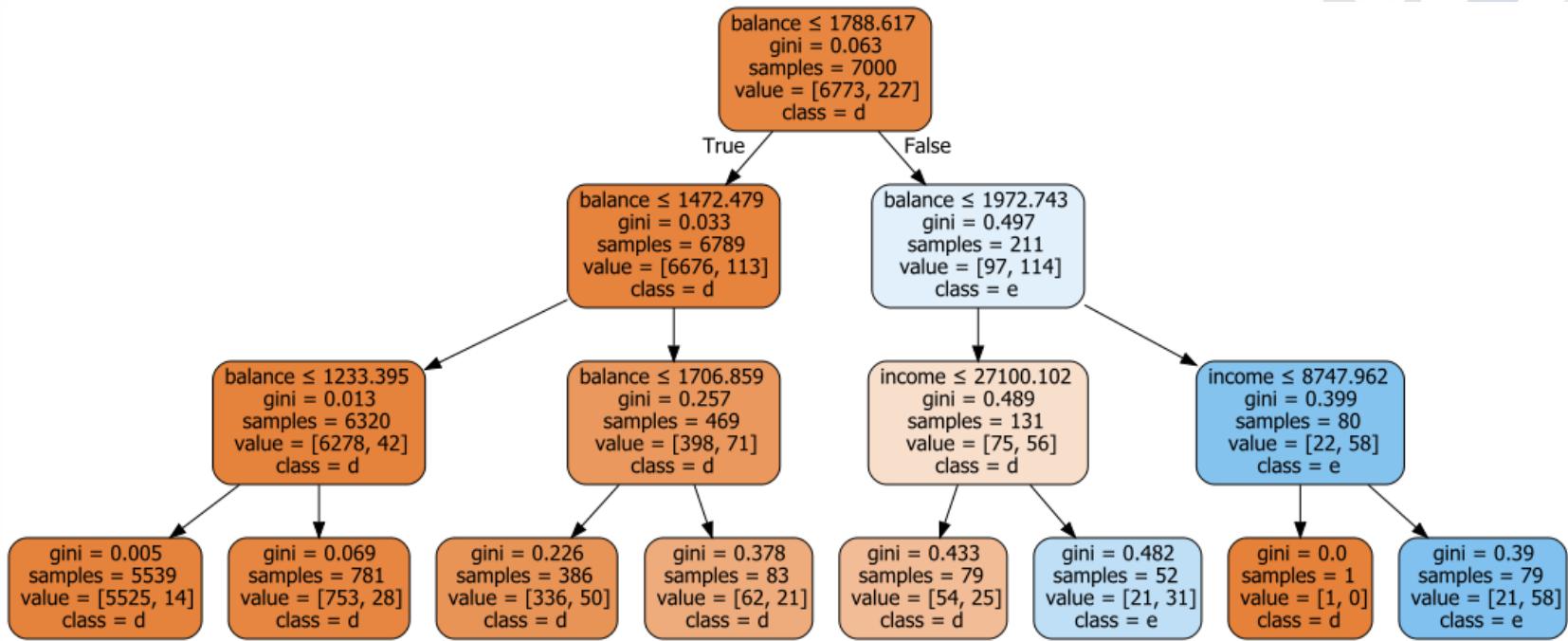
- Consider the data on loan defaulters.
- For analysis, we have taken non-defaulter + defaulters
- Data has been recorded in the file Default.csv



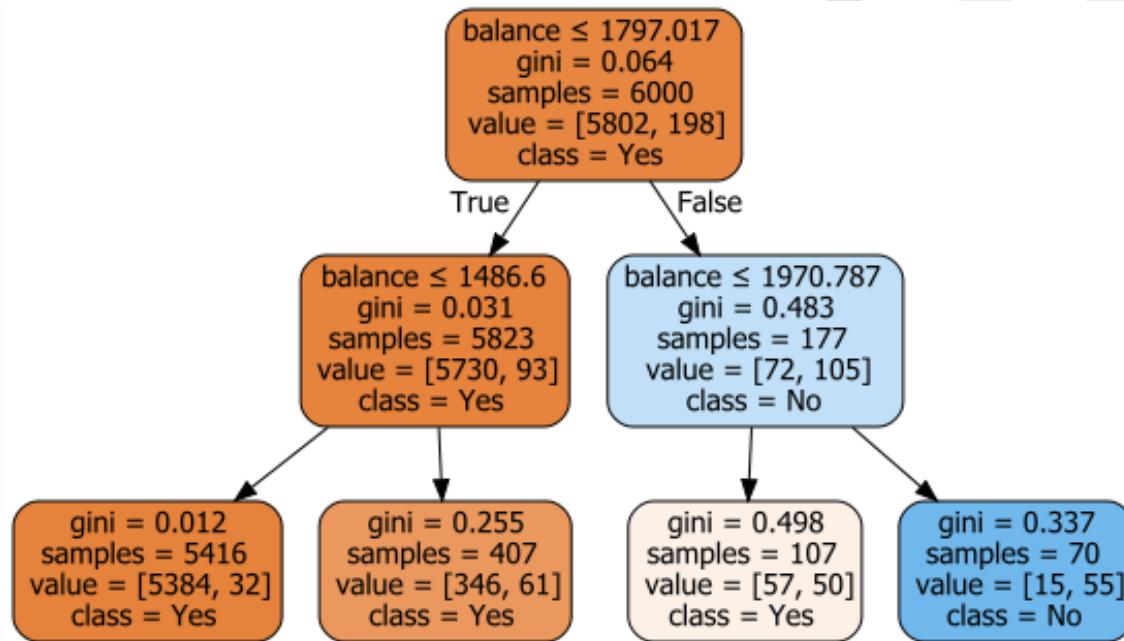
# Data

Index	default	student	balance	income
0	No	No	729.526	44361.6
1	No	Yes	817.18	12106.1
2	No	No	1073.55	31767.1
3	No	No	529.251	35704.5
4	No	No	785.656	38463.5
5	No	Yes	919.589	7491.56
6	No	No	825.513	24905.2
7	No	Yes	808.668	17600.5
8	No	No	1161.06	37468.5
9	No	No	0	29275.3
10	No	Yes	0	21871.1
11	No	Yes	1220.58	13268.6
12	No	No	237.045	28251.7
13	No	No	606.742	44994.6

# Classification Tree



# What did we gain from classification?



- For the whole training data, we had  $n(\text{Yes})=198$  and  $n(\text{No})=5802$  with proportions as 0.033 and 0.967 respectively
- By splitting on balance cut-off as 1797.017, we got two partitions with proportions as  $[ P(\text{Yes})=0.016, P(\text{No})=0.98 ]$  and  $[ P(\text{Yes})=0.41, P(\text{No})=0.59 ]$  respectively
- Hence we gained a **purity** or **homogeneity** in one case and lost it in other case
- The splits get proceeded for increasing purity

# Types of Decision Tree Algorithms

- There are many specific decision-tree algorithms. Notable ones include:
  - ID3 (Iterative Dichotomiser 3)
  - C4.5 (successor of ID3)
  - CART (Classification And Regression Tree)
  - CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
  - MARS: extends decision trees to handle numerical data better.
  - Conditional Inference Trees: Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid overfitting. This approach results in unbiased predictor selection and does not require pruning.
- We will be covering CART

# Homogeneity Measures

- Two homogeneity measures are considered for optimization in classification tree algorithms:
  - Gini's Impurity Index
  - Entropy

# Gini's Impurity Index

- Gini impurity can be computed by summing the probability of each item being chosen times the probability of a mistake in categorizing that item.
- It reaches its minimum (zero) when all cases in the node fall into a single target category.

$$I_G(f) = \sum_{i=1}^m f_i(1 - f_i) = \sum_{i=1}^m (f_i - f_i^2) = \sum_{i=1}^m f_i - \sum_{i=1}^m f_i^2 = 1 - \sum_{i=1}^m f_i^2 = \sum_{i \neq k} f_i f_k$$

# Entropy

- Entropy can be computed by summing the term of product of prevalence probability and its log
- It reaches its minimum (zero) when all cases in the node fall into a single target category.

$$H_i = - \sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^n p_{i,k} \log(p_{i,k})$$

Where

$p_{i,k}$ : Probability of Prevalence of class k for the  $i^{\text{th}}$  node

# How does CART calculate the threshold for splitting ?

- CART is a popular among all the tree algorithms
- Used in libraries like rpart ( R ) and scikit-learn ( Python )
- The algorithm goes on splitting the node with a feature k and a threshold  $t_k$  for which the following cost function is minimized

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

Where

$m_{left}$ : number of observations in the left sub-setted node

$G_{left}$ : Impurity of the left sub-setted node

$m_{right}$ : number of observations in the right sub-setted node

$G_{right}$ : Impurity of the right sub-setted node

# Classical Decision Trees Algorithm

- Response : Categorical outcome variable
- Predictors : Categorical and/or Continuous Variables.
- Steps are as follows:
  1. Predictor variable gets chosen in such a way that it best splits the data into two groups with maximized purity.
    - a) If the predictor is continuous, then cut-point is chosen for maximizing the purity
    - b) If the predictor is categorical ( Not for in sklearn ), then the categories are combined together to obtain two groups with maximized purity
  2. The data is separated into two groups and the process for each subgroup is continued
  3. Steps 1 and 2 are repeated until a subgroup is obtained containing fewer number of observations than a minimum number specified or the algorithm may terminate if further splitting doesn't increase purity beyond a specific threshold

# Classical Decision Trees Algorithm

- The subgroups in the lowest branch of the tree are called terminal nodes or leaf nodes.
- Each leaf node is classified as one single category of the outcome
- For classifying any observation in the validation / test data set, that observation is traversed through the branches of tree and leaf node at which the traversal stops is predicted as the outcome of the observation.

# Package tree from scikit-learn

- We need to instantiate the DecisionTreeClassifier first
- Then, call fit function on training data
- Finally, we predict on test data

```
clf = tree.DecisionTreeClassifier(max_depth=2)
clf2 = clf.fit(X_train, y_train)
```

```
y_pred = clf2.predict(X_test)
```

# DecisionTreeClassifier( )

Syntax :

```
sklearn.tree.DecisionTreeClassifier(criterion='gini', max_depth=None, ...)
```

Where

criterion : split criterion; “gini” or “entropy”

max\_depth : The maximum depth of the tree

# Program and Output

```
# Create training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4,
                                                    random_state=42)

clf = tree.DecisionTreeClassifier(max_depth=2)
clf2 = clf.fit(X_train, y_train)

y_pred = clf2.predict(X_test)

print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
In [111]: y_pred = clf2.predict(X_test)
.....
....: print(confusion_matrix(y_test, y_pred))
....: print(classification_report(y_test, y_pred))
[[3850  15]
 [ 100  35]]
          precision    recall   f1-score   support
      0       0.97     1.00     0.99    3865
      1       0.70     0.26     0.38     135
avg / total       0.97     0.97     0.96    4000
```

# Regularization Hyper-parameters

- If the tree is made fully to grow unconstrained, then it specializes itself only to the training data. In other words, it overfits.
- Hence, we need to put some constraints to its growth to minimize the risk of overfitting.
- The following can be some of the regularization parameters for controlling the growth:
  - Maximum depth of the tree - `max_depth`
  - Minimum number of observations a node must have before a split is applied on it - `min_samples_split`
  - Minimum number of observations a leaf node must have – `min_samples_leaf`

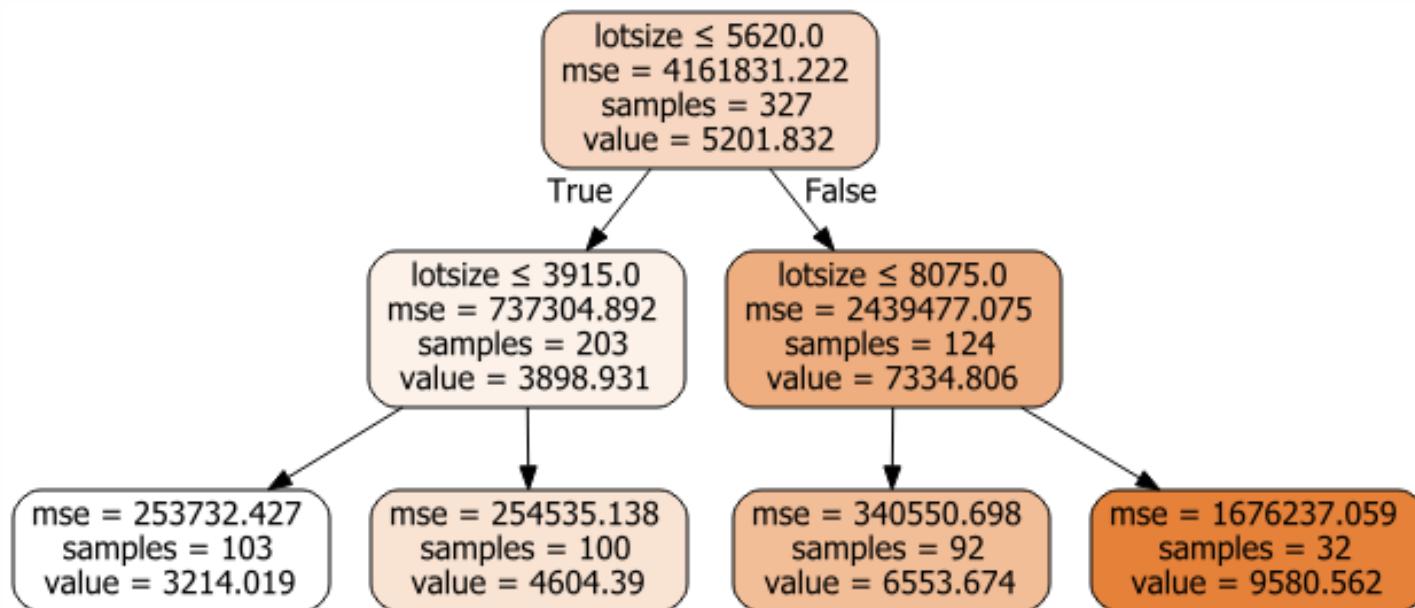


Questions?

# Regression Tree

# Typical Regression Tree Output

- In case of regression trees, the difference is that on the leaf nodes we have the means of the response variable values.



# Regression Tree

- The data gets divided into two parts in the interest of decreasing the variation of response variable
- The child nodes have lesser variation than their respective parent nodes for response variable

# Comparison in types of trees

	Classification	Regression
Response Variable Type	Categorical	Numerical
Measuring Homogeneity	Gini, Entropy	Squared Error, MSE
Prediction	Majority Class in the leaf node	Mean of response variable in the data in leaf node
Evaluation	Confusion Matrix metrics, ROC, LogLoss	MSE, MAE, $R^2$

# Regression Tree in Python

- From scikit-learn, we import package tree
- We instantiate the class of tree.DecisionTreeRegressor
- Call the method fit() on it
- On the built model, we call predict()

```
clf = tree.DecisionTreeRegressor(max_depth=2)
clf2 = clf.fit(x_train, y_train)

y_pred = clf2.predict(x_test)
```

# Example : Sales Prices of Houses in the City of Windsor

- Description
  - a cross-section from 1987
  - number of observations : 546
  - country : Canada
- A dataframe containing :
  - price : sale price of a house
  - lotsize : the lot size of a property in square feet
  - bedrooms : number of bedrooms
  - bathrms : number of full bathrooms
  - stories : number of stories excluding basement
  - driveway : does the house has a driveway ?
  - recroom : does the house has a recreational room ?
  - fullbase : does the house has a full finished basement ?
  - gashw : does the house uses gas for hot water heating ?
  - airco : does the house has central air conditioning ?
  - garagepl : number of garage places
  - prefarea : is the house located in the preferred neighbourhood of the city ?

# Program and Output

```
Housing = pd.read_csv("F:/Python Material/ML with Python/Cases/Real Estate/Housing.csv")
dum_Housing = pd.get_dummies(Housing.iloc[:,1:11], drop_first=True)

from sklearn.model_selection import train_test_split
from sklearn import tree
X = dum_Housing
y = Housing.iloc[:,1]

# Create training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4,
                                                    random_state=42)

clf = tree.DecisionTreeRegressor(max_depth=2)
clf2 = clf.fit(X_train, y_train)

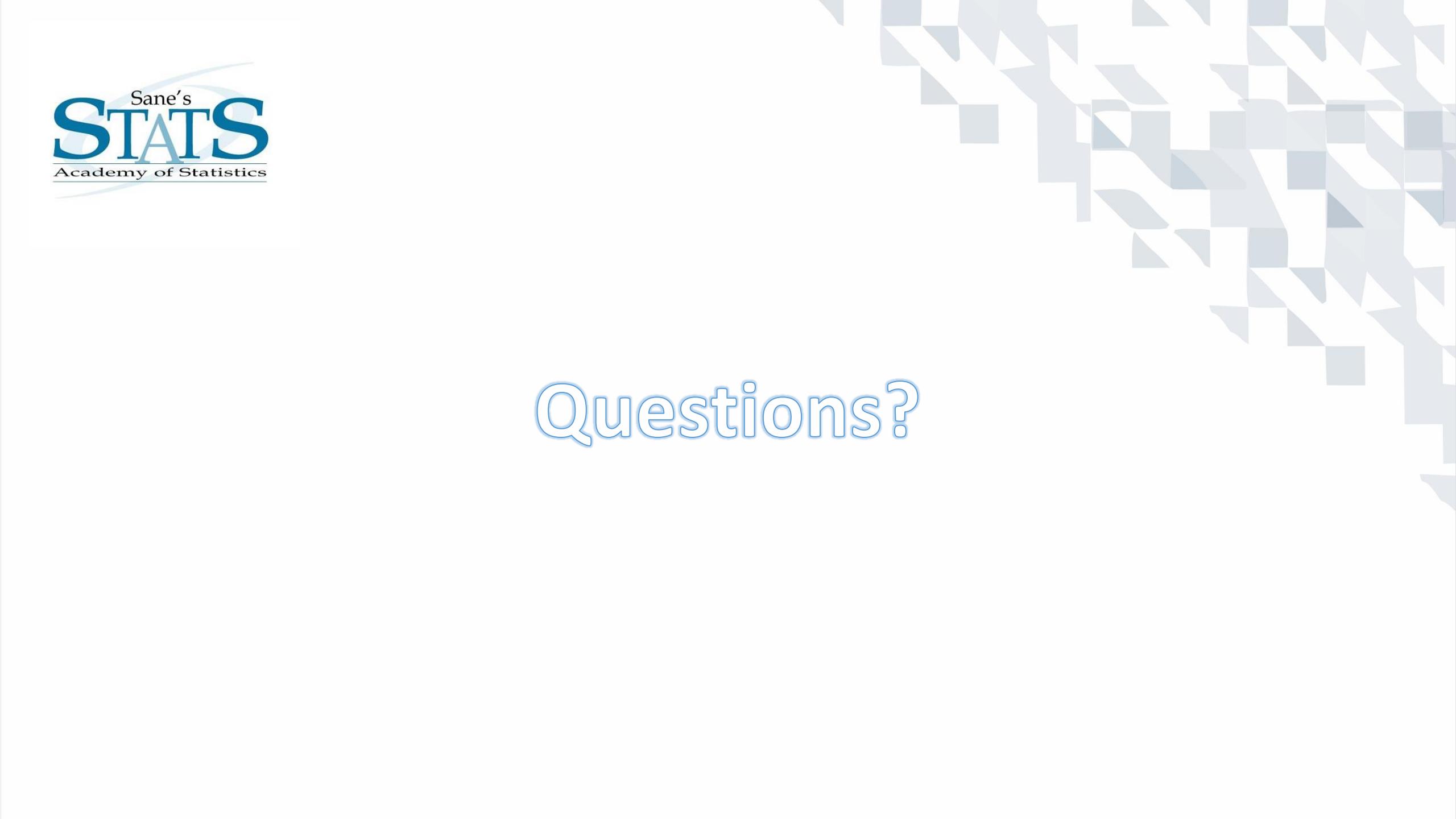
y_pred = clf2.predict(X_test)

from sklearn.metrics import mean_squared_error,mean_absolute_error,r2_score
```

```
In [138]: mean_squared_error(y_test, y_pred)
Out[138]: 909968.9687529949
```

```
In [139]: mean_absolute_error(y_test, y_pred)
Out[139]: 601.9324009754998
```

```
In [140]: r2_score(y_test, y_pred)
Out[140]: 0.8337797533789042
```



The background of the slide features a subtle, abstract geometric pattern composed of light gray squares and triangles, creating a sense of depth and texture.

# Questions?

# Bayes' Theorem

# Example : Telecom Customers

- A telecom firm has many customers. Each customer either talks for the duration of more than 100 minutes or less than 100 minutes. The firm has launched a plan for the customers who talk more specially to optimize the amount spent by them on bills.
- Call Centre staff had been instructed to call some customers. In that operation, some customers bought the new plan and others didn't.
- In this case each customer is a record, and the response of interest,  $Y = \{\text{Bought ,Not Bought}\}$ , has two classes: C1 = Bought and C2 = Not Bought.

# Conditional Probabilities

- A conditional probability of event A given event B [denoted by  $P(A|B)$ ] represents the chances of event A occurring only under the scenario that event B occurs.
- In the response example, we may be interested in  $P(\text{bought} | \text{Talk Time} \geq 100, \text{gender}=\text{Male})$ , also  $P(\text{bought} | \text{Talk Time} \geq 100, \text{gender}=\text{Female})$ , as we have gender as additional feature of the customers

# BAYES FORMULA

- The Bayes theorem gives us the following formula to compute the probability that the record belongs to class Ci:

$$P(C_i|X_1, \dots, X_p) = \frac{P(X_1, \dots, X_p|C_i)P(C_i)}{P(X_1, \dots, X_p|C_1)P(C_1) + \dots + P(X_1, \dots, X_p|C_m)P(C_m)}.$$

Where

Ci : classes of interest

X<sub>1</sub>,X<sub>2</sub>,...X<sub>p</sub> : Variables which co-exist with Classes of interest

# Example

Talks for more than 100 min? (TT >= 100)	Gender	Response
y	male	not bought
n	male	not bought
n	female	not bought
n	female	not bought
n	male	not bought
n	male	not bought
y	male	bought
y	female	bought
n	female	bought
y	female	bought

# Bayes' Formula Calculations

$$P(Buy|Male, TT \geq 100)$$

$$= \frac{P(Male, TT \geq 100 | Buy) P(Buy)}{P(Male, TT \geq 100 | Buy) P(Buy) + P(Male, TT \geq 100 | Not Buy) P(Not Buy)}$$

$$= \frac{P(Male|Buy)P(TT \geq 100|Buy) P(Buy)}{P(Male|Buy)P(TT \geq 100|Buy)P(Buy) + P(Male|Not Buy)P(TT \geq 100|Not Buy)P(Not Buy)}$$

$$\begin{aligned} &= \frac{\frac{1}{4} \times \frac{3}{4} \times \frac{4}{10}}{\frac{1}{4} \times \frac{3}{4} \times \frac{4}{10} + \frac{4}{6} \times \frac{1}{6} \times \frac{6}{10}} \\ &= 0.529 \end{aligned}$$

(TT >= 100)	Gender	Response
y	male	not bought
n	male	not bought
n	female	not bought
n	female	not bought
n	male	not bought
n	male	not bought
y	male	bought
y	female	bought
n	female	bought
y	female	bought

# Bayes Probabilities

- For the conditional probability of bought behaviors given ( $TT \geq 100$ ) = y, gender = male , the numerator is a multiplication of the proportion of ( $TT \geq 100$ ) = y instances among the bought customers, times the proportion of gender = male instances among the bought customers, times the proportion of bought customers:  $(3/4)(1/4)(4/10) = 0.075$ .
- To get the actual probabilities, we must also compute the numerator for the conditional probability of not bought given ( $TT \geq 100$ ) = y, gender = male :  $(1/6)(4/6)(6/10) = 0.067$ .
- The denominator is then the sum of these two conditional probabilities ( $0.075 + 0.067 = 0.14$ ).

# Bayes Probabilities

- The conditional probability of bought behaviors given ( $TT \geq 100$ ) = y, gender = male is therefore  $0.075/0.14 = 0.53$ .
- Similarly,
  - $P(\text{bought} | (TT \geq 100) = y, \text{gender} = \text{female}) = 0.87$ ,
  - $P(\text{bought} | (TT \geq 100) = n, \text{gender} = \text{male}) = 0.07$ ,
  - $P(\text{bought} | (TT \geq 100) = n, \text{gender} = \text{female}) = 0.31$ .

# Naïve Bayes Algorithm

# Naïve Bayes

- Naïve Bayes is a classification algorithm
- There are two types of Naïve Bayes Algorithms:
  - Discrete Naïve Bayes: For categorical predictors
  - Kernel Naïve Bayes: For numerical predictors

# Discrete Naive Bayes

- In this algorithm, the probability of a record belonging to a certain class is evaluated on the basis of conditional probability calculated using Bayes theorem
- Discrete Naive Bayes works only with predictors that are categorical.
- Numerical predictors must be binned and converted to categorical variables before the application Naive Bayes algorithm
- In Python, package `sklearn.naive_bayes` supports Descrete Naïve Bayes (`MultinomialNB`)

# Kernel Naïve Bayes

- Kernel Naïve Bayes works with numeric predictors assuming some distribution of the predictors
- It can assume Normal Distribution (Gaussian Naïve Bayes ) or any other distribution
- On assuming the distribution, the prior probabilities are calculated
- In Python, package `sklearn.naive_bayes` supports Gaussian (GaussianNB) Naïve Bayes (assumes Normality of predictors)

# Example 1 : Telecom Customers (Discrete NB)

- A telecom firm has many customers. Each customer either talks for the duration of more than 100 minutes or less than 100 minutes. The firm has launched a plan for the customers who talk more specially to optimize the amount spent by them on bills. In this case each customer is a record, and the response of interest,  $Y = \{\text{Bought}, \text{Not Bought}\}$ , has two classes that a company can be classified into:  $C_1 = \text{Bought}$  and  $C_2 = \text{Not Bought}$ .
- Apart from talk time we also have information about the gender of the customer

# Data

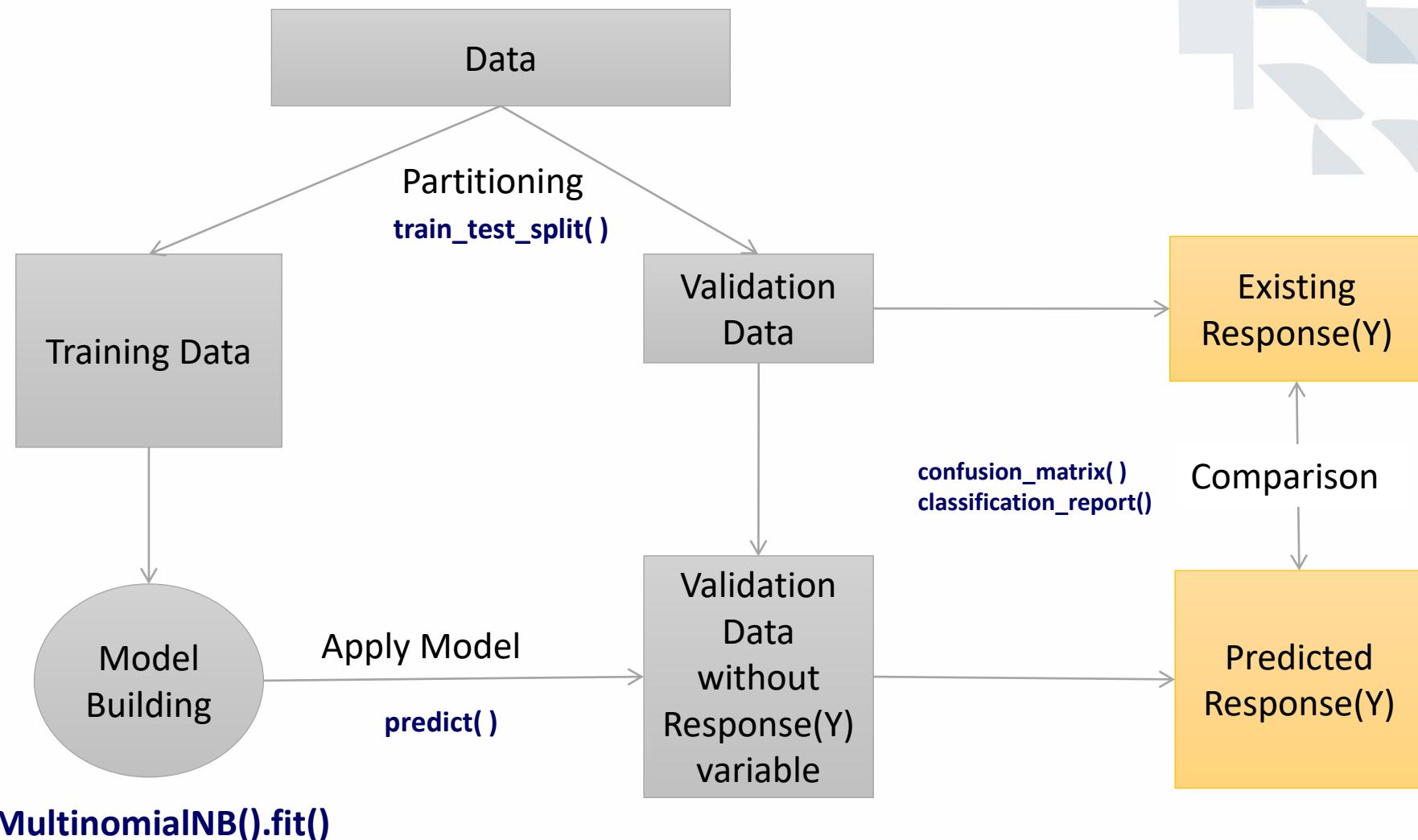
- 150 Observations , 3 variables
- First 8 observations:

telecom - DataFrame

Index	Gender	TT_gt_100	Response
0	F	Y	N
1	M	N	N
2	M	N	N
3	F	Y	Y
4	F	N	N
5	F	N	N
6	F	Y	Y
7	M	Y	Y
8	M	Y	N

Format Resize  Background color  Column min/max OK Cancel

# Naïve Bayes Classifier



# Program and Output

```
In [1]: import pandas as pd
.....
....: telecom = pd.read_csv("G:/Statistics (Python)/Cases/Telecom/Telecom.csv")
.....
....: dum_telecom = pd.get_dummies(telecom, drop_first=True)
.....
....: from sklearn.model_selection import train_test_split
....: from sklearn.metrics import confusion_matrix
....: from sklearn.metrics import classification_report, accuracy_score
....: from sklearn.naive_bayes import MultinomialNB
.....
....: X = dum_telecom.iloc[:,0:2]
....: y = dum_telecom.iloc[:,2]
.....
....: # Create training and test sets
....: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3,
....:                                                 random_state=42,
....:                                                 stratify=y)
.....
....: multinomial = MultinomialNB()
....: multinomial.fit(X_train, y_train) # Model Building
.....
....: y_probs = multinomial.predict_proba(X_test)
....: y_pred = multinomial.predict(X_test) # Applying built on test data
.....
....: print(confusion_matrix(y_test, y_pred))
[[18  4]
 [ 2 21]]
```

# Evaluation

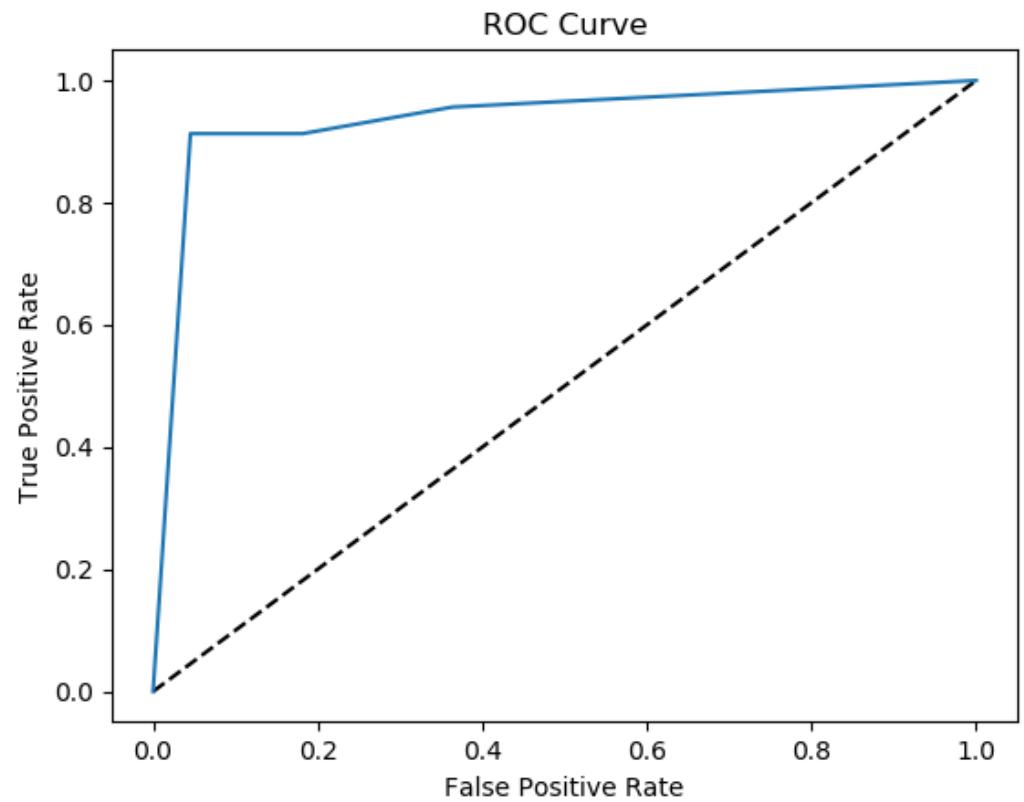
```
In [2]: print(classification_report(y_test, y_pred))
      precision    recall  f1-score   support

          0       0.90      0.82      0.86      22
          1       0.84      0.91      0.87      23

   accuracy                           0.87      45
  macro avg       0.87      0.87      0.87      45
weighted avg       0.87      0.87      0.87      45
```

```
In [3]: print(accuracy_score(y_test, y_pred))
0.8666666666666667
```

```
In [5]: roc_auc_score(y_test, y_pred_prob)
Out[5]: 0.9377470355731224
```



# Example 2: Predicting Defaulters (Gaussian NB)

- Data Set Details:
  - **default** : A categorical variable with levels No and Yes indicating whether the customer defaulted on their debt
  - **student** : A categorical variable with levels No and Yes indicating whether the customer is a student
  - **balance** : The average balance that the customer has remaining on their credit card after making their monthly payment (Numeric Variable)
  - **income** : Income of customer (Numeric Variable)
- Source: <http://www-bcf.usc.edu/~gareth/ISL/>

# Data

- 4 variables and 10,000 observations

Default - DataFrame

Index	default	student	balance	income
0	No	No	729.526	44361.6
1	No	Yes	817.18	12106.1
2	No	No	1073.55	31767.1
3	No	No	529.251	35704.5
4	No	No	785.656	38463.5
5	No	Yes	919.589	7491.56
6	No	No	825.513	24905.2
7	No	Yes	808.668	17600.5
8	No	No	1161.06	37468.5
9	No	No	0	29275.3
10	No	Yes	0	21871.1
11	No	Yes	1220.58	13268.6
12	No	No	237.045	28251.7
13	No	No	606.742	44994.6

# Program and Output

```
In [10]: import pandas as pd
.....
....: Default = pd.read_csv("F:/Python Material/ML with Python/Datasets/Default.csv")
....: dum_Default = pd.get_dummies(Default, drop_first=True)
.....
....: from sklearn.model_selection import train_test_split
....: from sklearn.metrics import confusion_matrix, classification_report
....: from sklearn.naive_bayes import GaussianNB
.....
....: X = dum_Default.iloc[:,[0,1,3]]
....: y = dum_Default.iloc[:,2]

In [11]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state=42)
.....
....: gaussian = GaussianNB()
....: y_pred = gaussian.fit(X_train, y_train).predict(X_test)
.....
....: print(confusion_matrix(y_test, y_pred))
....: print(classification_report(y_test, y_pred))
[[3840  25]
 [ 102  33]]
          precision    recall  f1-score   support

           0       0.97      0.99      0.98     3865
           1       0.57      0.24      0.34      135

avg / total       0.96      0.97      0.96     4000
```



The background of the slide features a subtle, abstract geometric pattern composed of light gray squares and triangles, creating a sense of depth and texture.

# Questions?

# Optimization Techniques

# Optimization Techniques

- ▶ Need for Optimization
- ▶ Setting up Optimization Problem
  - Formulation
- ▶ Different Optimization Techniques:
  - Linear Programming
  - Transportation Problem
  - Assignment Problem
  - Queuing Systems

# Need for Optimization

- ▶ Controlling the excess usage of resources
- ▶ Maximising the profit
- ▶ Minimising the cost

# Setting up Optimization Problem

- ▶ Formulation
- ▶ Construction of a mathematical model
- ▶ Acquiring the data

# Different Optimization Techniques – Linear Programming

- ▶ LPP is a mathematical technique for allotting the limited resources of a firm in an optimum manner.

# Linear Programming (LP) Problem

- ▶ The maximization or minimization of some quantity is the objective in all linear programming problems.
- ▶ All LP problems have constraints that limit the degree to which the objective can be pursued.
- ▶ A feasible solution satisfies all the problem's constraints.
- ▶ An optimal solution is a feasible solution that results in the largest possible objective function value when maximizing (or smallest when minimizing).

# Linear Programming (LP) Problem

- ▶ If both the objective function and the constraints are linear, the problem is referred to as a linear programming problem.
- ▶ Linear functions are functions in which each variable appears in a separate term raised to the first power and is multiplied by a constant (which could be 0).
- ▶ Linear constraints are linear functions that are restricted to be "less than or equal to", "equal to", or "greater than or equal to" a constant.

# Problem Formulation

- ▶ Problem formulation or modeling is the process of translating a verbal statement of a problem into a mathematical statement.

# Guidelines for Model Formulation

- ▶ Understand the problem thoroughly.
- ▶ Describe the objective.
- ▶ Describe each constraint.
- ▶ Define the decision variables.
- ▶ Write the objective in terms of the decision variables.
- ▶ Write the constraints in terms of the decision variables.

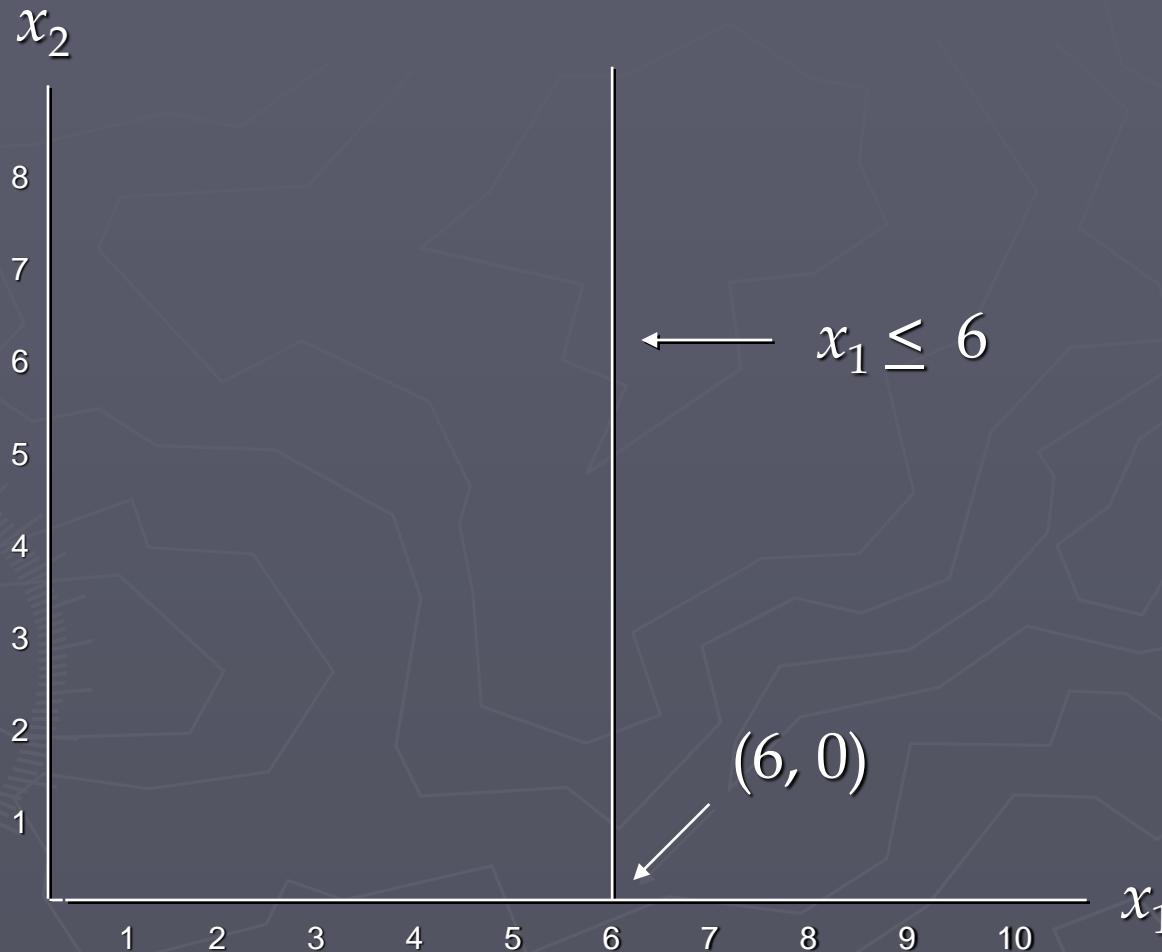
# Example 1: A Maximization Problem

## ► LP Formulation

$$\begin{aligned} \text{Max } & 5x_1 + 7x_2 \\ \text{s.t. } & x_1 \leq 6 \\ & 2x_1 + 3x_2 \leq 19 \\ & x_1 + x_2 \leq 8 \\ & x_1, x_2 \geq 0 \end{aligned}$$

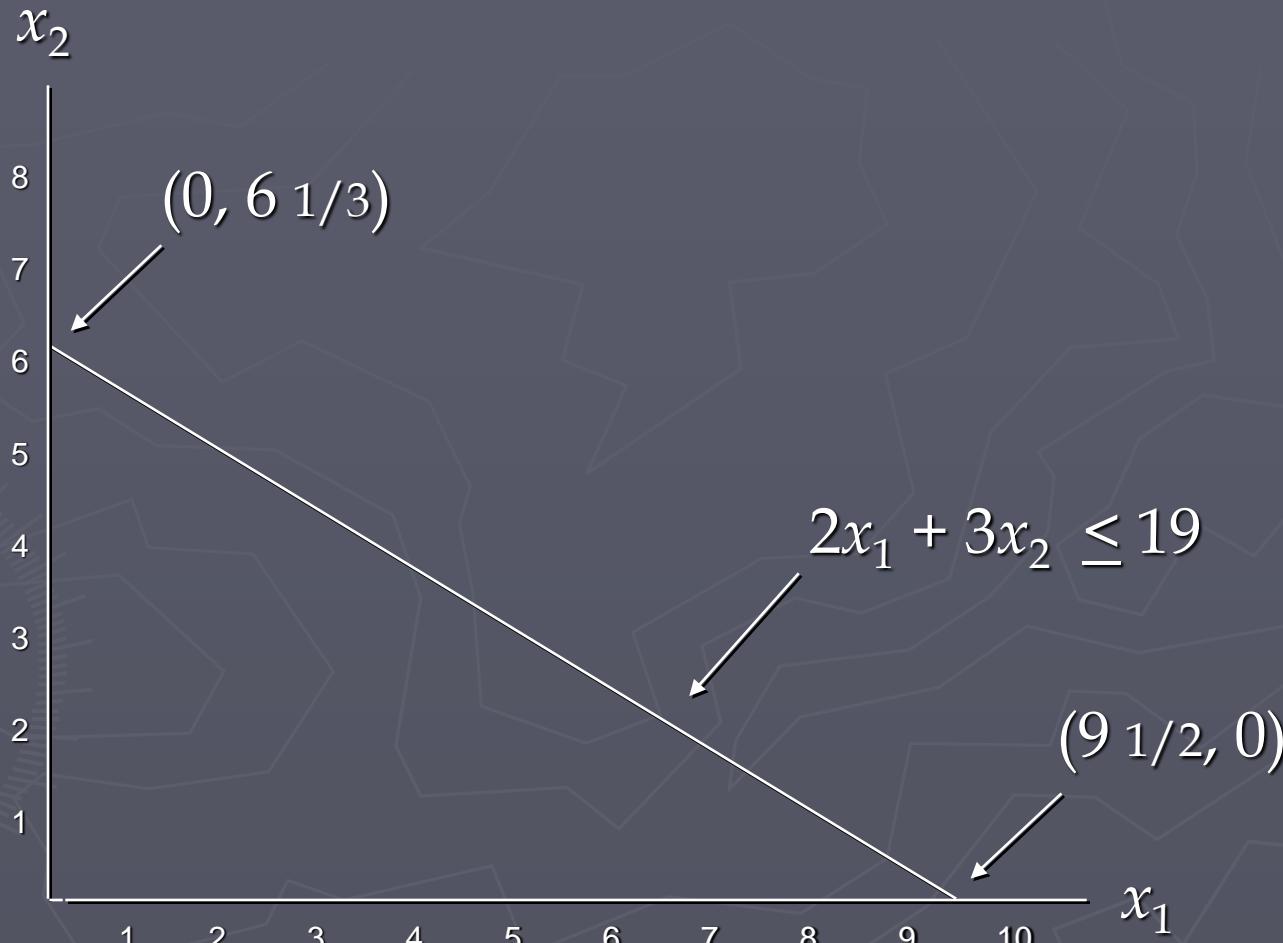
# Example 1: Graphical Solution

- Constraint #1 Graphed



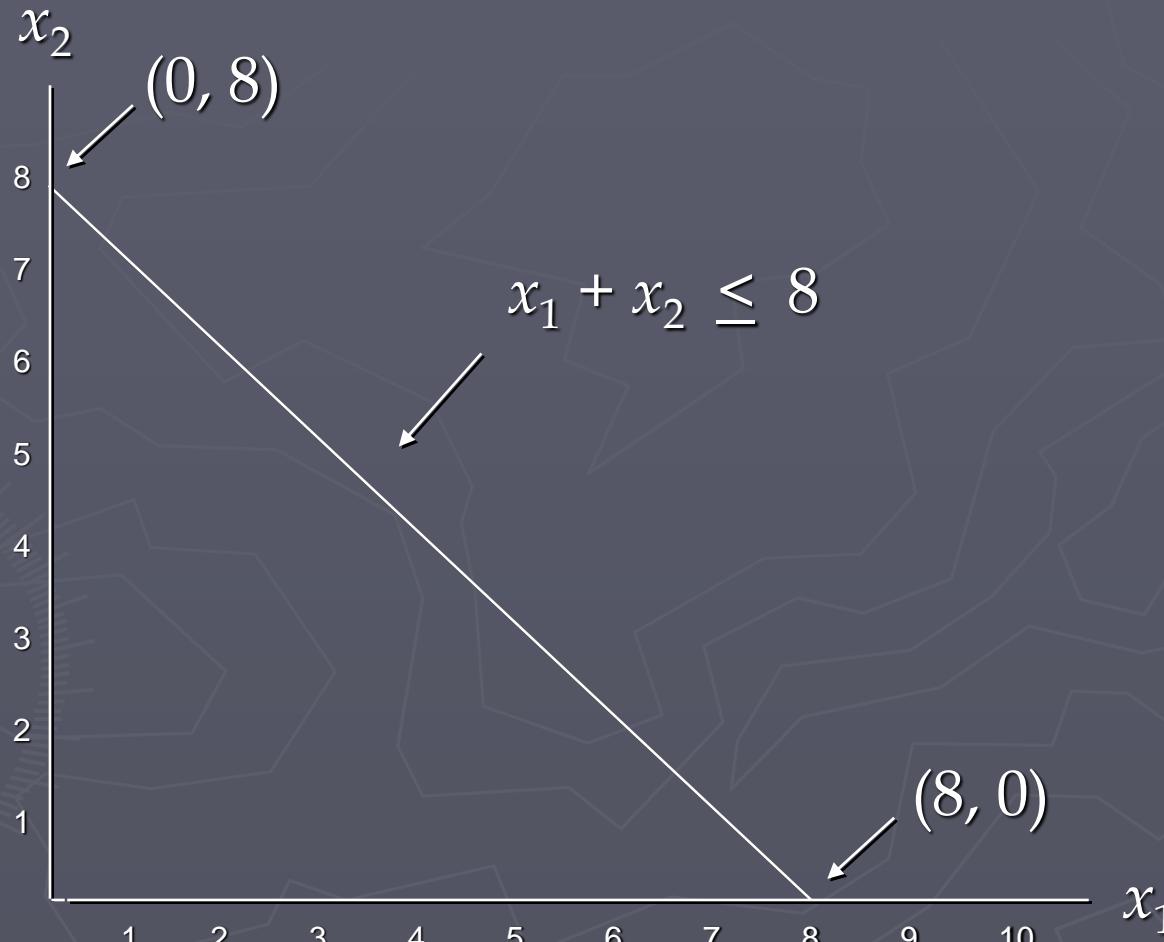
# Example 1: Graphical Solution

## ► Constraint #2 Graphed



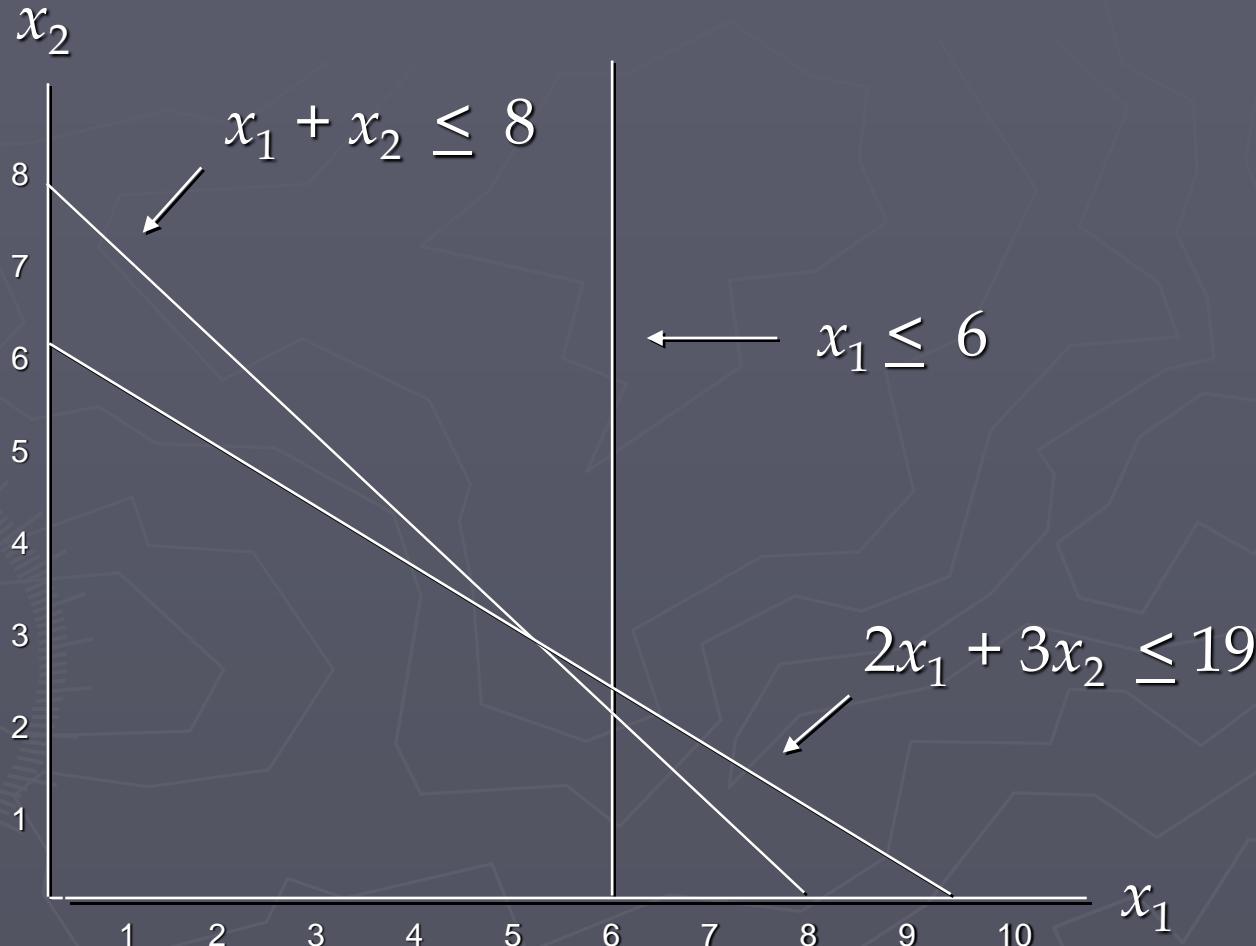
# Example 1: Graphical Solution

## ► Constraint #3 Graphed



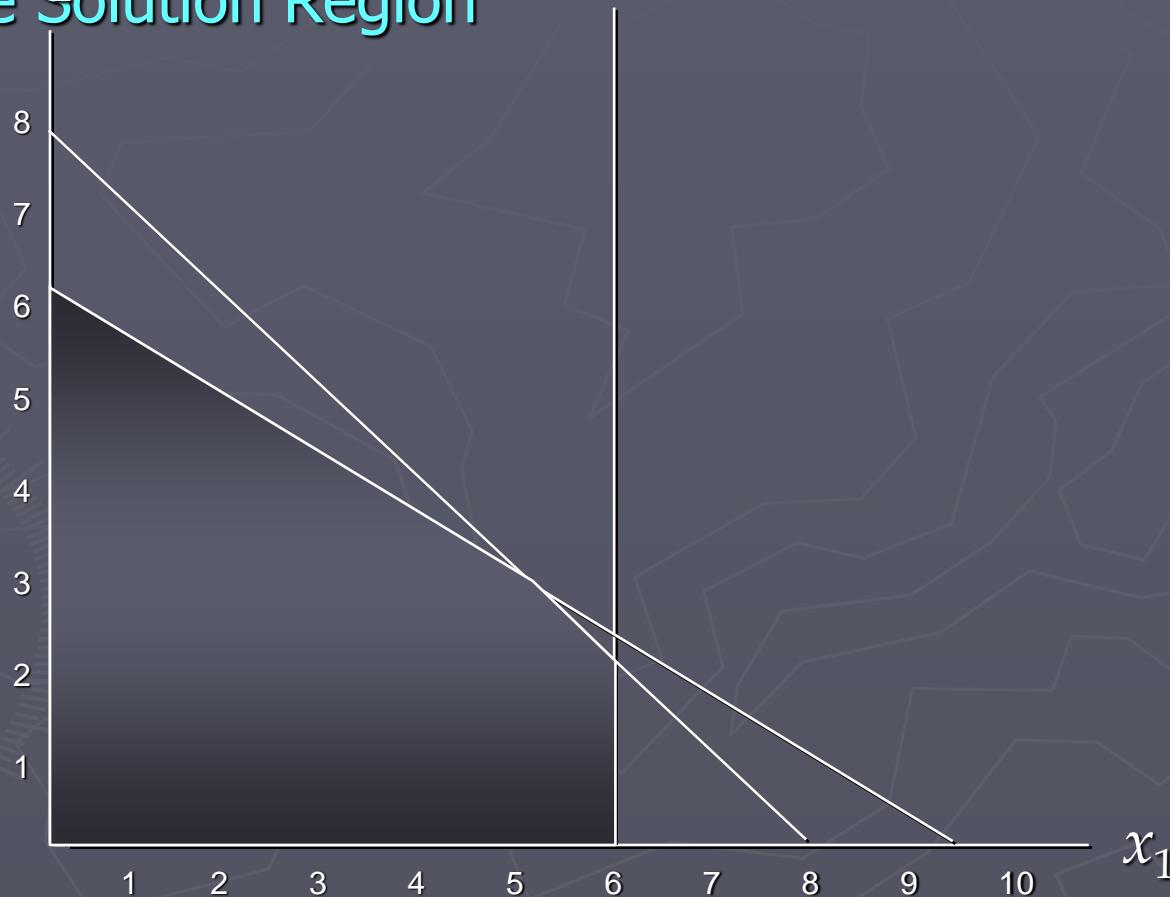
# Example 1: Graphical Solution

## ► Combined-Constraint Graph



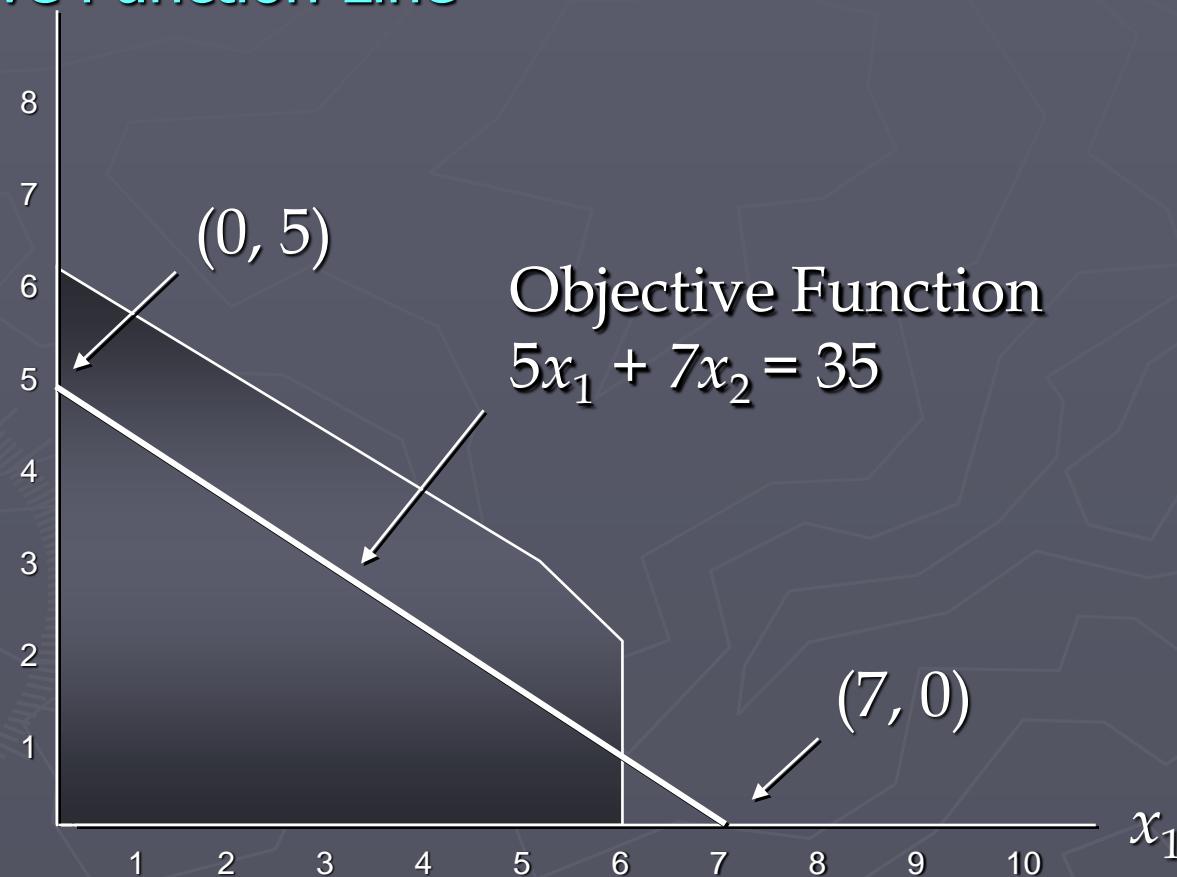
# Example 1: Graphical Solution

- Feasible Solution Region



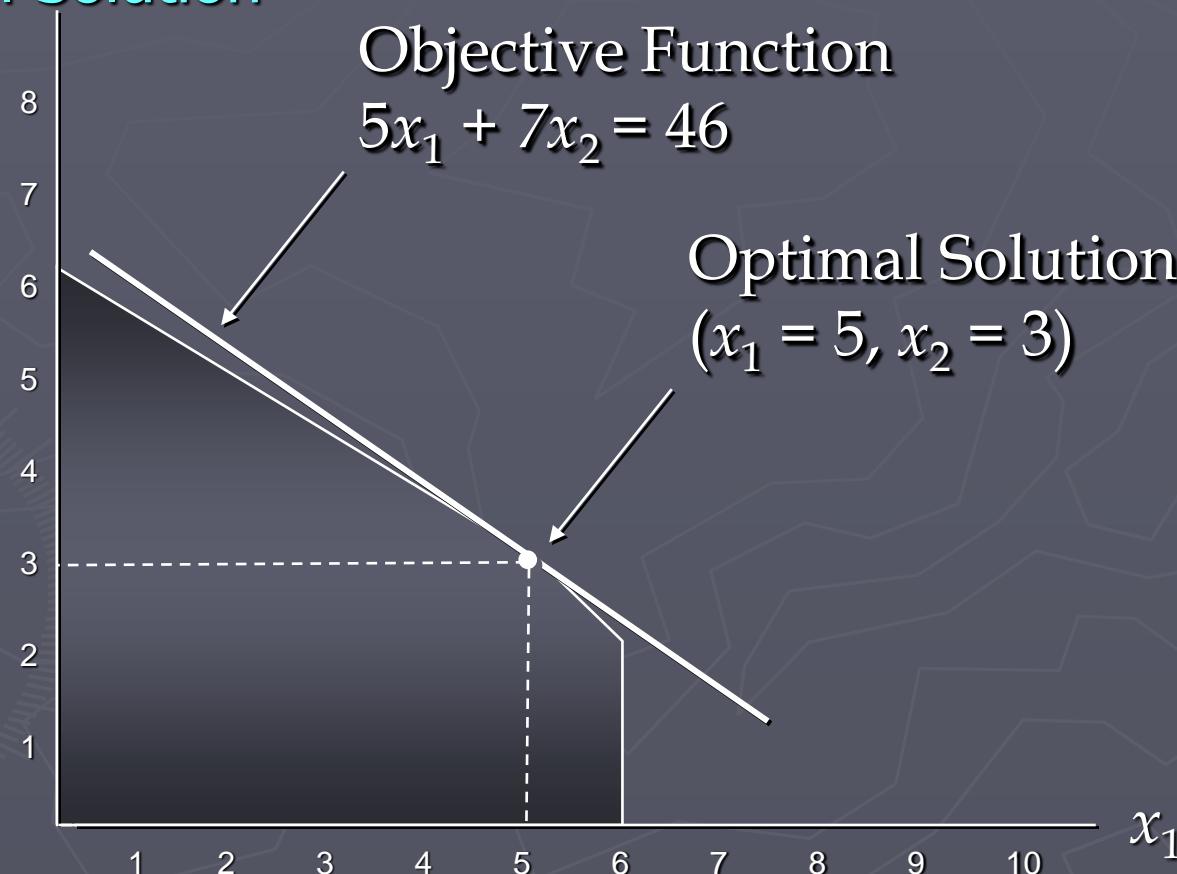
# Example 1: Graphical Solution

## ► Objective Function Line



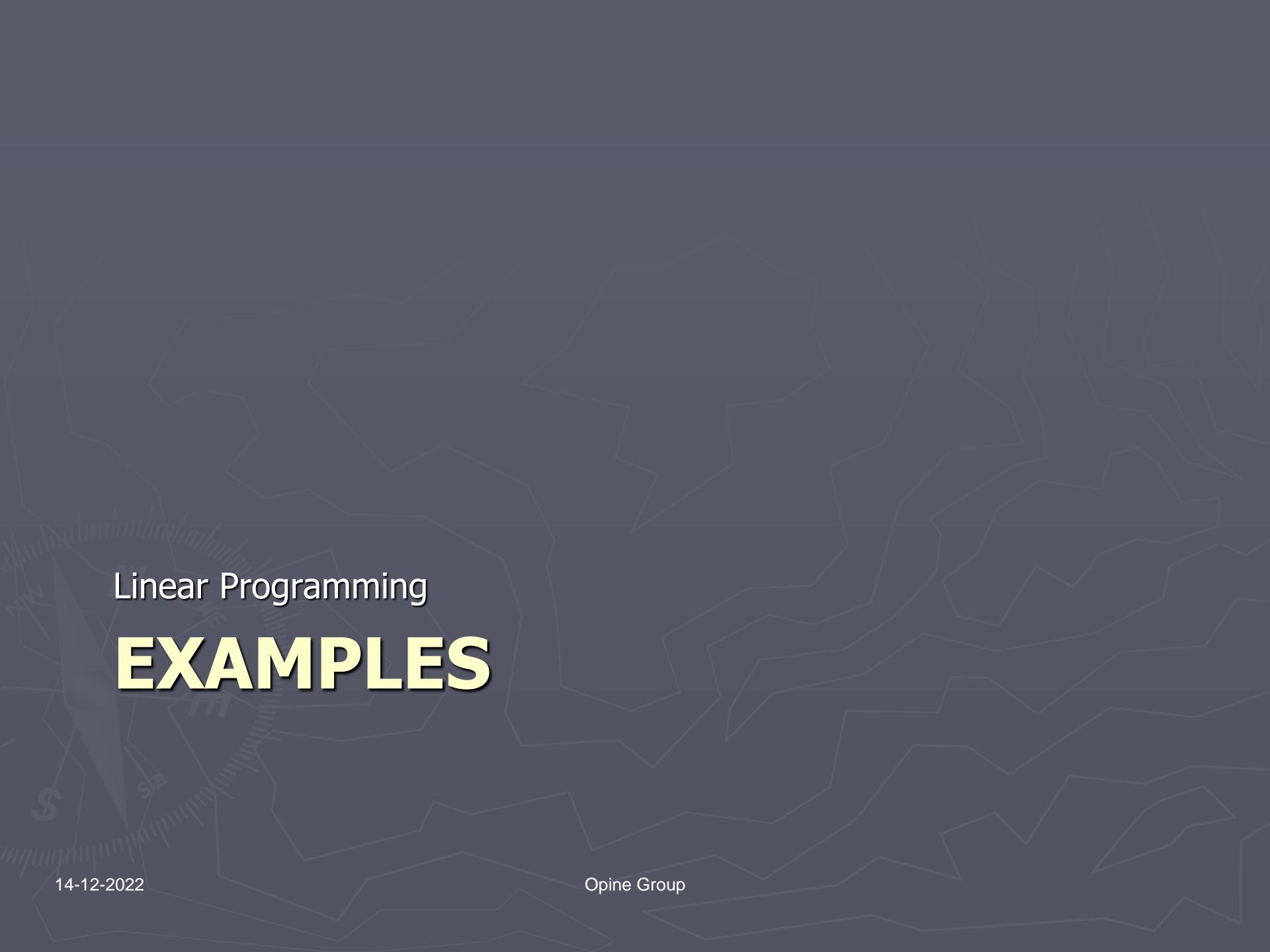
# Example 1: Graphical Solution

## ► Optimal Solution



# Feasible Region

- ▶ The feasible region for a two-variable linear programming problem can be nonexistent, a single point, a line, a polygon, or an unbounded area.
- ▶ Any linear program falls in one of three categories:
  - is infeasible
  - has a unique optimal solution or alternate optimal solutions
  - has an objective function that can be increased without bound
- ▶ A feasible region may be unbounded and yet there may be optimal solutions. This is common in minimization problems and is possible in maximization problems.



# Linear Programming **EXAMPLES**

A company that operates 10 hours a day manufactures two products on three sequential processes. The following table summarizes the data of the problem

Product	Minutes per unit			Unit Profit
	Process 1	Process 2	Process 3	
1	10	6	8	Rs. 2/-
2	5	20	10	Rs. 3/-

Determine the optimal mix of the two products.

## **Problem 2:**

Solve this linear programming problem.

**Maximize**

$$Z = 20x_1 + 10x_2 + 15x_3$$

**Subject to:**

$$3x_1 + 2x_2 + 5x_3 \leq 55$$

$$2x_1 + x_2 + x_3 \leq 26$$

$$x_1 + x_2 + 3x_3 \leq 30$$

$$5x_1 + 2x_2 + 4x_3 \leq 57$$

$$x_1, x_2, x_3 \geq 0$$

### Problem 3:

A company makes three models of desks, an executive model, an office model and a student model. Each desk spends time in the cabinet shop, the finishing shop and the crating shop as shown in the table:

Type of desk	Cabinet shop (in hrs.)	Finishing shop (in hrs.)	Crating shop (in hrs.)	Profit (in Rs.)
Executive	2	1	1	150
Office	1	2	1	125
Student	1	1	.5	50
Available hours	16	16	10	

How many of each type of model should be made to maximize profits?

# Directional Data Analysis

An Introduction to Circular Statistics

# What is Circular Statistics?

- **Directional statistics** (also **circular statistics** or **spherical statistics**) is the subdiscipline of **statistics** that deals with directions, axes or rotations.

# Angles

- The fact that 0 degrees and 360 degrees are identical angles, so that
- for example 180 degrees is not a sensible mean of 2 degrees and 358 degrees

# Time Periods

- Statistics involving
  - temporal periods (e.g. time of day, week, month, year, etc.),
  - compass directions,
  - dihedral angles in molecules, orientations, rotations and so on.

# Conversion

Time can be converted to an angular measurement using the equation:

$$a = \frac{(360^\circ)(X)}{k}$$

where  $a$  is the angular measurement,  $X$  is the time period, and  $k$  is the number of time units on the circular measurement scale.

What is the angular measurement  
of 6:15 a.m. (6.25a.m.)?  
(Remember to use a 24hr clock...)

$$a = \frac{(360^\circ)(6.25\text{hr})}{24\text{hrs}} = 93.75^\circ$$

What is the angular measurement  
of February 14<sup>th</sup>?  
(Remember to use total days...)

$$a = \frac{(360^\circ)(45\text{th day})}{365\text{ days}} = 44.38^\circ$$

# Calculation

To analyze directional data they must first be transformed into rectangular polar coordinates.

- First, we specify a ‘unit circle’ that has a radius of 1.
- The polar location is then defined as the angular measurement and its intersection with the unit circle.
- The cosine and sine functions are then used to place this location (based on the angle and unit distance) into a standardized Cartesian space.

$$\cos \alpha = \frac{x}{r} \quad \sin \alpha = \frac{y}{r}$$

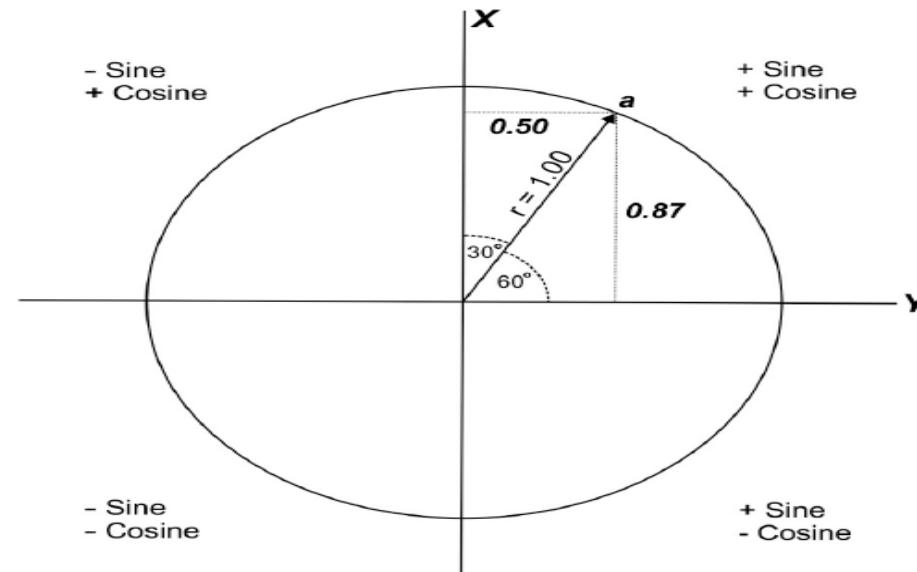
$$\cos 30^\circ = 0.50$$

$$\cos 60^\circ = 0.50$$

$$\sin 30^\circ = 0.87$$

$$\sin 60^\circ = 0.87$$

Note that the coordinates of opposite angles are identical. Also note that the x and y axes are opposite of the typical Cartesian plane.



# Mean Angle

The mean angle can not simply be the sum of the angles divided by the sample size, because the mean angle of 359° and 1° (north) would be 180° (south)! Therefore we use the following equations:

$$Y = \frac{\sum_{i=1}^n \sin_a}{n} \quad X = \frac{\sum_{i=1}^n \cos_a}{n}$$

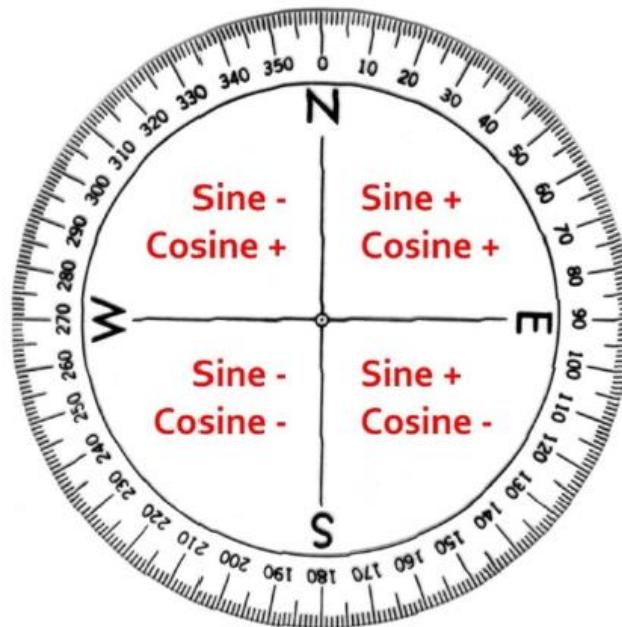
$$r = \sqrt{X^2 + Y^2}$$

$$\cos \bar{a} = \frac{X}{r} \quad \sin \bar{a} = \frac{Y}{r} \quad \theta_r = \arctan\left(\frac{\sin \bar{a}}{\cos \bar{a}}\right)$$

where  $X$  and  $Y$  are the rectangular coordinates of the mean angle, and  $r$  is the mean vector.

# Determining the Quadrant

- Sin +, Cos + : the mean angle is computed directly.
- Sin +, Cos - : the mean angle =  $180 - \theta_r$ .
- Sin -, Cos - : the mean angle =  $180 + \theta_r$ .
- Sin -, Cos + : the mean angle =  $360 - \theta_r$ .



# Calculations

Rocks Vectors	Sin (Azimuth)	Cos (Azimuth)
341	-0.32557	0.94552
330	-0.50000	0.86603
301	-0.85717	0.51504
299	-0.87462	0.48481
9	0.15643	0.98769
7	0.12187	0.99255
359	-0.01745	0.99985
334	-0.43837	0.89879
353	-0.12187	0.99255
15	0.25882	0.96593
27	0.45399	0.89101
28	0.46947	0.88295
25	0.42262	0.90631
23	0.39073	0.92050
350	-0.17365	0.98481
30	0.50000	0.86603
26	0.43837	0.89879
22	0.37461	0.92718
8	0.13917	0.99027
356	<u>-0.06976</u>	<u>0.99756</u>
$\Sigma$	0.34763	17.91415

- First take the sine and cosine of the angles (azimuths) and sum them.
- In Excel the formula is:  
 $=\sin(\text{radians}(\text{cell } \#))$  and  
 $=\cos(\text{radians}(\text{cell } \#))$

# Calculations

$$n = 20$$

$$\sum \sin_a = 0.34763 \quad \sum \cos_a = 17.91415$$

$$Y = \frac{0.34763}{20} = 0.01738$$

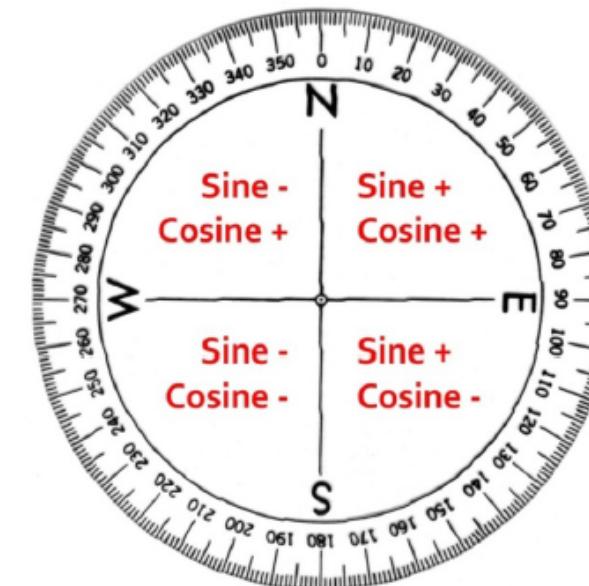
$$X = \frac{17.91415}{20} = 0.89571$$

$$r = \sqrt{0.01738^2 + 0.89571^2} = \sqrt{0.00030 + 0.80229} = 0.8959$$

$$\sin \bar{a} = \frac{0.01738}{0.8959} = 0.0194$$

$$\cos \bar{a} = \frac{0.89571}{0.8959} = 0.9998$$

$$\theta_r = \arctan\left(\frac{0.0194}{0.9998}\right) = 1.11 \quad \text{← Ignore the sign.}$$



# Angular Dispersion

The value of  $r$  is also a measure of angular dispersion, similar to the standard deviation with a few exceptions:

- Unlike the standard deviation it ranges from 0 – 1.
- A value of 0 means uniform dispersion.
- A value of 1 means complete concentration in one direction.

# Some Circular probability distributions

- von Mises circular distribution
- Circular uniform distribution
- Wrapped normal distribution
- Wrapped Cauchy distribution
- Wrapped Lévy distribution

For more info: [https://en.wikipedia.org/wiki/Circular\\_distribution](https://en.wikipedia.org/wiki/Circular_distribution)

# Some Directional Hypothesis Tests

- Rayleigh z Test
- Watson's  $U^2$  test.