ACDS Lecture Series

Lecture - 5

CSIR

Probability and Statistics

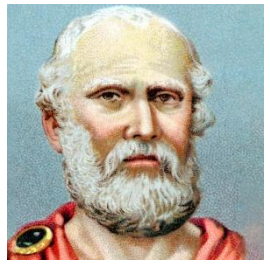
G. N. Sastry & Team

ADVANCED COMPUTATION AND DATA SCIENCES (ACDS) DIVISION

CSIR-North East Institute of Science and Technology, Jorhat, Assam, India

Ancient origins of the concept of probability

"Chance"

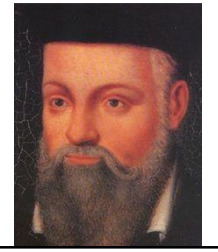


Plato (428-348 BC)

Plato (384-322 BC)

In 324 BC, a Greek, **Antimenes (530–510 BC)**, first developed a system of insurance which guaranteed a sum of money against wins or losses of certain events.

During 15th century the pragmatic approach to problems of games of chance with dice began in Italy



Gerolamo Cardano (1501–1576)

Famous physician, mathematician and gambler **Professor and chair of Mathematics at the University of Bologna in Italy**

Liber de Ludo Aleae (Games of Chance) - A short manual Contains first mathematical treatment of probability dealing with problems of mathematical expectation

Introduces the idea of probability p between 0 and 1 to an event whose outcome is random.

Concept of probability

Birth **Death** **Weather**

Uncertainties of everyday life

Health **Game**

↓

Concept of "Chance or Random variables"

Probability of an event is p, after a large **number of trials n**, the number of times it will occur is close to **np.**

Development of classical probability – 16th and 17th centuries

During the sixteenth and seventeenth centuries, in particular, a great deal of attention was given to games of chance, such as ***tossing coins, throwing dice or playing cards***, and to problems of gambling in general.

A problem of throwing Dice – First recorded in the history of mathematical probability theory.

Galileo Galilei
(1564–1642)

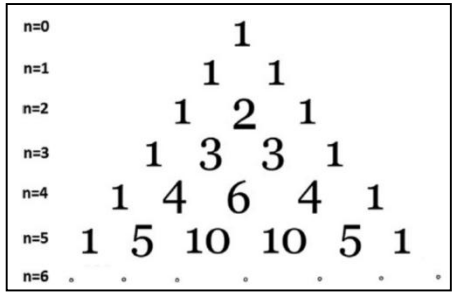
Introduced the concept of probability, mean (or expected) value and conditional probability

Baise Pascal
(1623–1662)

Pierre de Fermat
(1601–1665)

If a coin is tossed twice, four possible outcomes:
HH, HT, TH, TT
Binomial coefficients = $(H+T)^2 = H^2 + HT + HT + T^2 = H^2 + 2HT + T^2$

Similarly, If a coin is tossed thrice, eight possible outcomes:
HHH, HHT, HTH, THH, HTT, THT, TTH and TTT
Binomial coefficients = $(H + T)^3 = H^3 + 3H^2T + 3HT^2 + T^3$



Pascal triangle

Binomial coefficients = ${}^nC_r = \binom{n}{r}$ involved in the coefficient of arb $^{n-r}$ in the binomial expansion of $(a+b)^n$ for any integer $n \geq 0$ and $0 \leq r \leq n$

Development of classical probability – 16th and 17th centuries

Christian Huygens
(1629-1695)

Published the first treatise on probability theory *De Rationiis in Ludo Aleae* (*On Reasoning in Games of Dice*) in 1657.

Huygens introduced the definition of probability of an event as a quotient of ***favorable cases by all possible cases***. He also reintroduced ***the concept of mathematical expectation (or mean) and elaborated Cardano's idea of expected value (or mean) of random variables***.

Jacob Bernoulli
(1654–1705)

JACOBI BERNOULLI,
Profess. Basil. & utriusque Societ. Reg. Scientiar. Gall. & Prussi. Sodal.
Mathematici Celeberrimi,
ARS CONJECTANDI,
opus posthumum.
Accedit
TRACTATUS
DE SERIEBUS INFINITIS,
Et Epistola Gallicè scripta
DE LUDO PILÆ
RETICULARIS.
BASILEÆ,
Impensis THURNISIORUM, Fratrum.
cIↃ IↃcc xiii.

Bernoulli's Theorem - A name given by a celebrated French mathematician, Simeon Poisson (1781 – 1840)

In his book Bernoulli added the problem of repeated experiments in which a particular outcome is either a success or a failure.

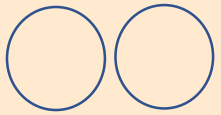
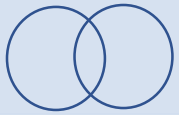
Using the general properties of binomial coefficients $\binom{n}{r}$ dealing with the number of ways in which r elements can be chosen from a population of (n≥r) elements, Bernoulli formulated the probability B(r;n,p) of r successes (and hence, n-r failures) in n dependent tails in the following form

$$B(r;n,p) = \binom{n}{r} p^r q^{n-r}$$

where p is the probability of a success and q = 1 – p is the probability of failure of an event.

Terminologies

- *Random Experiment:* An experiment whose outcome cannot be predicted
- *Sample Space:* The entire set of possible outcomes of a random experiment
- *Event*: One or more outcomes of an experiment, a subset of sample space

Disjoint Events	• Do not have any common outcome • Eg. A man cannot be dead and alive 
Non disjoint events	• Have common outcomes • Eg. Marks Maths and Chemistry can be 100 

Mathematical Backbone of Probability

Permutation and ***combination*** are used to determine the probabilities and hence can be considered the backbone

Both permutation and combination are built on the concept of **factorial!**

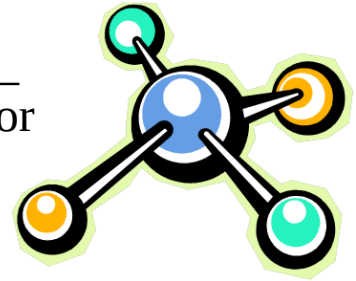
Factorial of any number say n denoted by "n!" can be computed by

$$n! = n(n-1)(n-2)(n-3) \ldots 1$$

For example, 5! = 5 x 4 x 3 x 2 x 1 = 120

To see how factorial is used in counting lets take another example,

"*The simplest protein molecule in biology is called vasopressin and is composed of 8 amino acids that are chemically bound together in a particular order. The order in which these amino acids occur is of vital importance to the proper functioning of vasopressin. If these 8 amino acids were placed in a hat and drawn out randomly one by one, how many different arrangements of these 8 amino acids are possible?*"



Solution: Let A,B,C,D,E,F,G,H symbolize the 8 amino acids. They must fill 8 slots: ___ ___ ___ ___ ___ ___ ___ ___ . There are 8 choices for the first position, leaving 7 choices for the second slot, 6 choices for the third slot and so on.

The number of different orderings is :**8(7)(6)(5)(4)(3)(2)(1) =8! = 40,320**

Hence, the probability that all 8 amino acids are in correct order is $\frac{1}{40,320}$ which is very unlikely to happen

Two problems illustrating combinations and permutations.

Problem 1: Consider the set { a , b , c}. How many two-letter "words" (including nonsense words) can be formed from the members of this set?
Answer: We will list all possibilities: ab, ba, bc, cb, ac, ca a total of 6.

Problem 2: Consider the set consisting of three males: {Paul, Ed, Nick}. How many two-man crews can be selected from this set?
Answer: (Paul, Ed), (Paul, Nick) and (Ed, Nick) and that is all!

The difference between the two problems is that:
Both problems involved counting the numbers of arrangements of the same set {a , b , c}, taken 2 elements at a time, without allowing repetition. However, in the first problem, the **order of the arrangements mattered** since ***ab*** and ***ba*** are two different "words". In the second problem, the **order did not matter** since (**Paul, Ed**) and (**Ed, Paul**) represented the same two-man crew. So we counted this only once.

The first example was concerned with counting the number of **permutations** of 3 objects taken 2 at a time.

The second example was concerned with the number of **combinations** of 3 objects taken 2 at a time

What is Permutation?

The notation P(n,r) represents the number of permutations (arrangements) of n objects taken r at a time when r is less than or equal to n. In a permutation, the **order is important.**
In general, P(n,r) = n(n-1)(n-2)(n-3)…(n-r+1) or

Permutation

The number of permutations of n objects taken r at a time is the **quotient** of $n!$ and $(n-r)!$

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

Example find P(5,3) here, n = 5 and r = 3 so we have P(5,3) = (5)(5-1)5-3+1) = 5(4)3 = 60

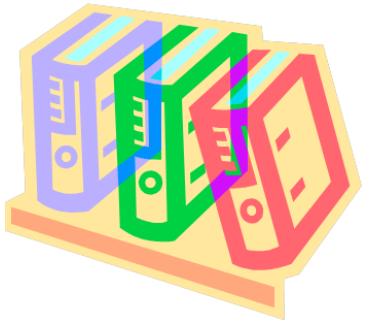
Or $\frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{5 \times 4 \times 3 \times 2!}{2!} = 60$ This means there are 60 arrangements of 5 items taken 3 at a time.

Applications of Permutation

Application:
How many ways can 5 people sit on a park bench if the bench can only seat 3 people?

Solution: Think of the bench as three slots ___ ___ ___ .
There are five people that can sit in the first slot, leaving four remaining people to sit in the second position and finally 3 people eligible for the third slot. Thus, there are 5(4)(3)=60 ways the people can sit. The answer could have been found using the permutations formula: P(5,3) = 60, since we are finding the number of ways of arranging 5 objects taken 3 at a time.

Application:
A bookshelf has space for exactly 5 books. How many different ways can 5 books be arranged on this bookshelf?



Solution: ___ ___ ___ ___ ___ Think of 5 slots, again. There are five choices for the first slot, 4 for the second and so on until there is only 1 choice for the final slot. The answer is 5(4)(3)(2)(1)
which is the same as P(5,5) = 120.

What is Combination?

An arrangement or listing in which order is not important is called a **combination.**
For example, if you are choosing 2 salad ingredients from a list of 10, the order in which you choose the ingredients does not matter.

Combination

Combination

The number of combinations of n objects taken r at a time is the **quotient** of $n!$ and $(n-r)! * r!$

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

Examples:

Find C(8,5)

Solution: C(8,5) = $\frac{P(8,5)}{5!} = \frac{8(7)(6)(5)(4)}{5(4)(3)(2)(1)} = \frac{8(7)(6)}{3(2)(1)} = 8(7) = 56$

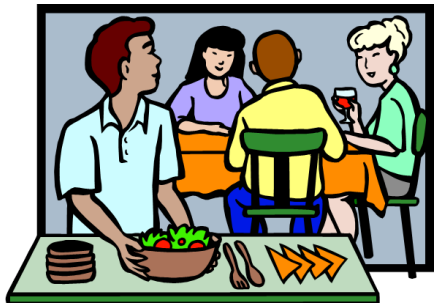
Find C(8,8)

Solution: C(8,8) = $\frac{P(8,8)}{8!} = \frac{8(7)(6)(5)(4)(3)(2)(1)}{8(7)(6)(5)(4)(3)(2)(1)} = 1$

Applications of Combination

Application:

In how many ways can you choose 5 out of 10 friends to invite to a dinner party?



Solution: Does the order of selection matter? If you choose friends in the order A,B,C,D,E or A,C,B,D,E the same set of 5 was chosen, so we conclude that the order of selection does not matter. We will use the formula for combinations since we are concerned with how many **subsets of size 5** we can select from a set of 10.

$$C(10,5) = \frac{P(10,5)}{5!} = \frac{10(9)(8)(7)(6)}{5(4)(3)(2)(1)} = \frac{10(9)(8)(7)}{(5)(4)} = 2(9)(2)(7) = 252$$

Application:

A certain state lottery consists of selecting a set of 6 numbers randomly from a set of 49 numbers. To win the lottery, you must select the correct set of six numbers. How many possible lottery tickets are there?

Solution: The order of the numbers is not important here as long as you have the correct set of six numbers. To determine the total number of lottery tickets, we will use the formula for combinations and find C(49, 6), the number of combinations of 49 items taken 6 at a time. Using our calculator, we find that C(49,6) = 13,983,816

Permutation and Probability

A computer program requires the user to enter a 7-digit registration code made up of the digits 1, 2, 4, 5, 6, 7, and 9. Each number has to be used, and no number can be used more than once.

a) How many different registration codes are possible?

Since the order of numbers in the code is important, this situation is a permutation of 7 digits taken 7 at a time.

$${}_nP_r = {}_7P_7 \qquad \text{n = 7, r = 7; recall that 0! = 1.}$$

$${}_7P_7 = \frac{7*6*5*4*3*2*1}{1} \qquad \text{or } 5040$$

There are 5040 possible codes with the digits 1, 2, 4, 5, 6, 7, and 9.

Probability first 3 digits even

$$P(\text{first 3 digits even}) = \frac{144}{5040} \quad \begin{matrix}\longleftarrow \text{favorable outcomes} \\ \longleftarrow \text{possible outcomes}\end{matrix}$$

The probability that the first three digits of the code are even is $\frac{1}{35}$ or about **3%**.

b) What is the probability that the first three digits of the code are even numbers?

Use the **Fundamental Counting Principle** to determine the number of ways for the first three digits to be even.

3 2 1 4 3 2 1

So, the number of favorable outcomes is **3 * 2 * 1 * 4 * 3 * 2 * 1** or **144**.

Permutation and Probability

When working with permutations and combinations, it is vital to be able to distinguish when the counting order is important, or not. This is only recognizable after a considerable amount of practice.

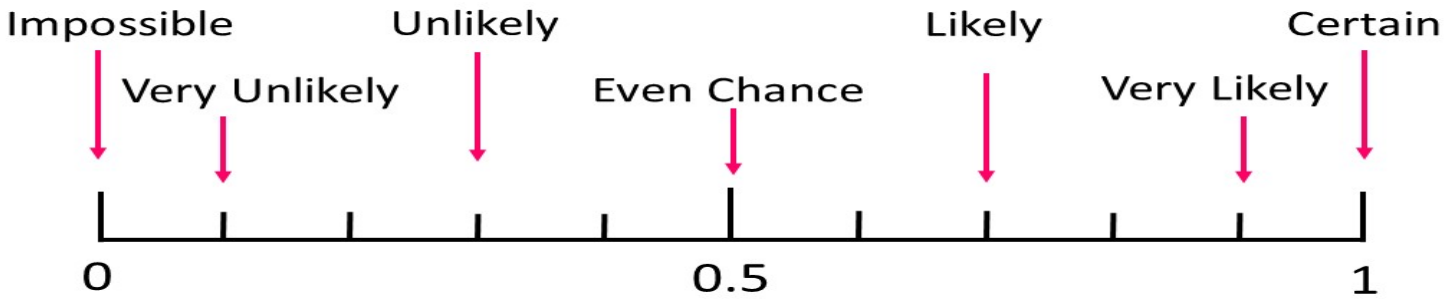
Example:

The students of Mr. Fant's Seminar class had to choose 4 out of the 7 people who were nominated to serve on the Student Council. How many different groups of students could be selected?

The order in which the people are being chosen does not matter because the positions for which they are being chosen are the same. They are all going to be members of the student council, with the same duties. ***(Combination)***

However, if Mr. Fant's class was choosing 4 out of 7 students to be president, vice-president, secretary, and treasurer of the student council, then the order in which they are chosen would matter. **(Permutation)**

Probability Scale



A card is drawn from a pack of 52 playing cards. Find the probability that the card is (i) a diamond (ii) a queen (iii) a king or a queen

The probability that the card is a:

(i) Diamond: There are 13 diamonds in a pack of card $P(Diamond) = \frac{13}{52}$

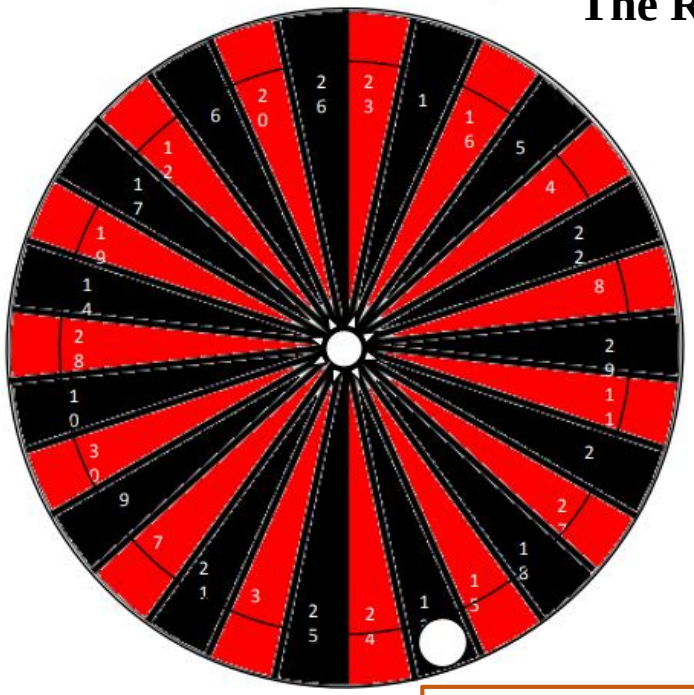
(ii) Queen: There are 4 Queen in a pack of card $P(Diamond) = \frac{4}{52}$

(iii) King or Queen: There are 8 Kings or Queens in a pack of card

$$P(King/Queen) = \frac{8}{52}$$

Probability and chance

The Roulette Wheel



ODD		EVEN
1	11	21
2	12	22
3	13	23
4	14	24
5	15	25
6	16	26
7	17	27
8	18	28
9	19	29
10	20	30
1 to 10	11 to 20	21 to 30
RED		BLACK

$$P(\text{odd number}) = \frac{15}{30} = 50\ \%$$

$$P(1 \text{ to } 10) = \frac{10}{30} = 33\ \%$$

$$P(\text{Black}) = \frac{15}{30} = 50\ \%$$

$$P(\text{number } 1) = \frac{1}{30} \quad 3.3\ \%$$

Flipping the Coin

When two coins are tossed possible outcomes:

- There could be two heads
- There could be two tails
- There could be a head and a tail
- There could be a tail and a head

The probability of outcome while tossing two coins:

(i) Two heads (HH): $P(HH) = \frac{1}{4}$

(ii) A head and a tail (HT): $P(HT) = \frac{2}{4}$

(iii) A tail and a head (TH): $P(TH) = \frac{2}{4}$

(iv) Two tails (TT) = $P(TT) = \frac{1}{4}$

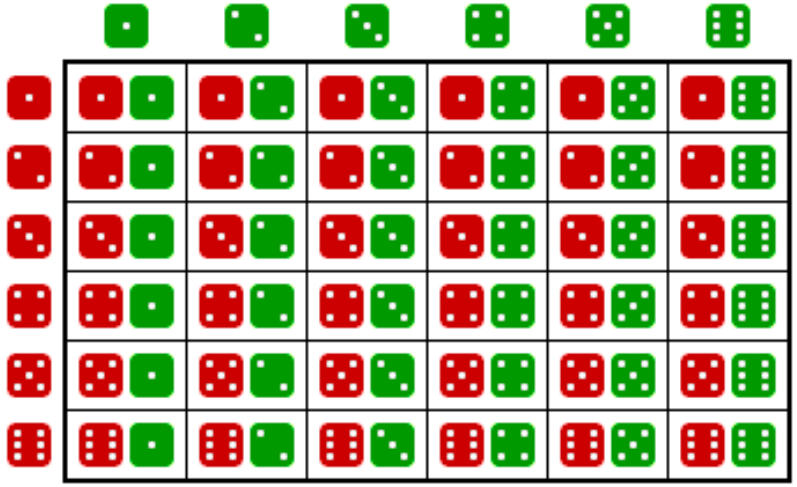
Sample Space HH, TH, HT, TT

Rolling a Dice

Each dice has 6 faces numbered 1 .. 6

- There are 6 possible outcomes for 1 dice
- There are 36 possible outcomes for 2 dice

Sample Space



The probability of outcome while rolling 1 dice:

(i) Getting a 5: $P(5) = \frac{1}{3}$

(ii) Getting an even number: $P(Even) = \frac{3}{6}$

The probability of outcome while rolling 2 dice:

(iii) Getting one dice even: $P(Even) = \frac{18}{36}$

(iv) Getting sum of 8 = $P(Sum\ 8) = \frac{5}{36}$

(2, 4), (3, 6) (4, 5) (1, 6) …..

(3, 5), (4, 4) (2, 6) (6, 1) (5, 3)

Sample Space

Introduction to Statistics

Statistics is the science of collecting, organizing, analysing and interpreting data.

"What is the weight of a mouse?"

Is this a statistical question – Yes, one would expect the weights of mice to vary.

"How many shares did my video get in YouTube?"

Is this a statistical question – Again, Yes, because the number of shares of a video will always vary

Why Statistics?

Statistics can help to take decisions in uncertain situations

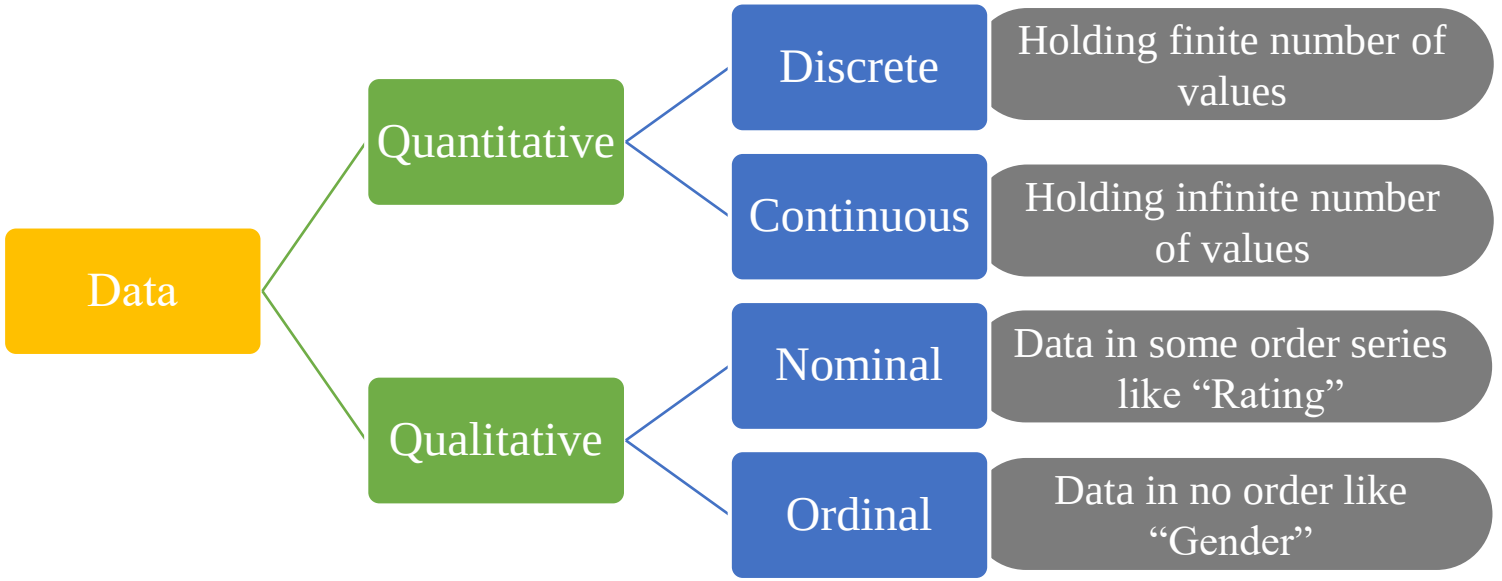
Statistics uses the term **"PROXY"** which is something which we want to measure but isn't exactly what we want to measure

- Data and its Types
- Mean, Variance, Standard Deviation
- Randomness and distributions
- Hypothesis Testing

"We measure whether a movie is good or bad based on reviews" – which is PROXY

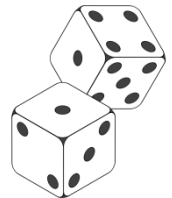
What is Data?

Data can be referred to facts and statistics gathered for analysis or reference.



Statistics is an area of applied mathematics which is focuses on data collection, analysis, interpretation and presentation. *Example: Amount of rain in an area, Increase in market prices etc.*

Probability is the measure on how likely an event will occur. It is the ratio of desired outcome and total outcome
Example: The number on the dice, the combination of card, winning chances of a baseball team



Expectations

Let X be a discrete random variable with range $R_X=\{x_1,x_2,x_3,\ldots\}$ (finite or countably infinite). The *expected* value of X, denoted by EX is defined as,

$$EX = \sum_{x_k \in R_X} x_k P_X(x_k)$$

Example: Consider the probability distribution given below to find the expected value.

x	10	11	12	13	14
P(x)	0.4	0.2	0.2	0.1	0.1

Solution: Expected value $EX = \sum_{i=1}^{5} x_i P(x_i)$ $= (10 \times 0.4) + (11 \times 0.2) + (12 \times 0.2) + (13 \times 0.1) + (14 \times 0.1)$

$= 11.3$ (which is the expected value)

With given expected value, we can find any missing value from a distribution table

Mean

- The mean is the arithmetic average for a set of data.
- It is denoted by x, and it is computed as sum of all the observed outcomes from the sample divided by the total number of events

Here is the formula for – x.

$$\bar{x} = \frac{\sum x}{n}$$

∑ is sigma and is used to represent a summation, so ∑x is the sum of all values. The letter n is the number of values in the sample and is referred to as the sample size.

Example:

A manager of a local restaurant is interested in the number of people who eat there on Fridays. Here are the totals for nine randomly selected Fridays:

712, 626, 600, 596, 655, 682, 642, 532, 526

Find the mean for this set of data.

We sum the values and get a total of 5571, and divide by 9 to find the mean.

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{5571}{9}$$

$$= 619$$

The mean for this set of data is 619 people

Median

- The median of a set of data is a value that divides the set of data into two equal groups, after the values have been put in order from lowest to highest.
- If there are an odd number of values in the set, there will be one value in the center of the data, and this value is the median.

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

Median = 23

- If there are an even number of values in the set of data, there is not a single value in the center. In this case, we take the mean of the two values in the center.

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

the middle numbers are **21 and 23**.

21+23=44
44 ÷ 2 = 22

Median = 22

Example:

A manager of a local restaurant is interested in the number of people who eat there on Fridays. Here are the totals for nine randomly selected Fridays: 712, 626, 600, 596, 655, 682, 642, 532, 526

Arrange the values in ascending order

526, 532, 596, 600, 626, 642, 655, 682, 71

Divide the set into two equally sized groups, the one value left over in the middle is the median for this set.

526	532	596	600	626	642	655	682	712

Median= 626

Example:

Tracy, a real estate agent, sold six homes last week. Here are their sale prices. Find the median sale price for these six homes.

88000,112000, 99000,106000, 95000, 2500000

The median is equal to the mean of 99,000 and 106,000

Median = 99,000 + 106,000/2

= 102,500

Mode

The mode of a set of data is the value(s) that occur most frequently in the set.

3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

In order these numbers are

3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56

Numbers that appear **most often**.

Mode = 23

Example:

Here are the ages of the 42 presidents of the United States at inauguration. Find the mode for this set of data.

57 61 57 57 58 57 61 54 68
51 49 64 50 48 65 52 56 46
54 49 **51** 47 55 55 54 42 **51**
56 55 **51** 54 **51** 60 62 43 55
56 61 52 69 64 46

Mode =51

Variance

- It is a measure of how spread out a data set is.
- It is calculated as the average squared deviation of each number from the mean of a data set.

Sample Variance

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

Population Variance

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

The symbol s^2 represent sample variance, and the symbol σ^2 represent population variance. σ is the lowercase Greek letter sigma, which is the Greek letter s.

For example, for the numbers 1, 2, and 3 the mean is 2 and the variance is 0.667.

$$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0.667$$

Variance tells us how are data are spread, larger the variance more the spread

Example :

A taxi dispatcher is interested in the number of fares for his drivers on Fridays. He randomly selects seven drivers, and then randomly selects one Friday for each of the drivers. Here are the number of fares for each.

32, 27, 30, 41, 29, 38, 34

Find the variance for these totals.

Since the dispatcher is interested in all Fridays, these data are a sample. For this calculation, just as with mean deviation, it is best to use a column approach.

x	$x - \bar{x}$	$(x - \bar{x})^2$
32	–1	1
27	–6	36
30	–3	9
41	8	64
29	–4	16
38	5	25
34	1	1
$\bar{x} = 33$		152

$$s^2 = \frac{152}{7 - 1}$$

$$= \frac{152}{6}$$

$$= 25.33$$

The variance is 25.33

Standard Deviation

- The standard deviation of a set of data is the square root of the variance.
- The measure of dispersion that we will use most often in this course is the standard deviation

$$\text{Standard deviation} = \sqrt{\text{variance}}$$

Here, s represent sample standard deviation, and σ to represent population standard deviation.

Sample Standard Deviation

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma(x - \mu)^2}{N}}$$

Standard deviation tells us, on average how far each score lies from the mean

Example :

- A taxi dispatcher is interested in the number of fares for his drivers on Fridays. He randomly selects seven drivers, and then randomly selects one Friday for each of the drivers. Here are the number of fares for each.

32, 27, 30, 41, 29, 38, 34

Find the standard deviation for these totals.

x	$x - \bar{x}$	$(x - \bar{x})^2$
32	–1	1
27	–6	36
30	–3	9
41	8	64
29	–4	16
38	5	25
34	1	1
$\bar{x} = 33$		152

$$s^2 = \frac{152}{7-1}$$

$$= \frac{152}{6}$$

$$= 25.33$$

$$s = \sqrt{25.33}$$

$$= 5.03$$

The sample standard deviation is 5.03 fares.

Covariance

Covariance is a measure of how much two random variable vary together. It's similar to variance, but where variance tells you how a *single* variable varies, co variance tells you how two variables vary together.

$$Cov(X,Y) = \Sigma\ E((X-\mu)E(Y-\nu)) / n-1$$

where:
X and Y are random variable
$E(X) = \mu$ is the expected value (the mean) of the random variable X and
$E(Y) = \nu$ is the expected value (the mean) of the random variable Y
n = the number of items in the data set

Example :

- The table below describes the rate of economic growth (xi) and the rate of return on the S&P 500 (y_i). Using the covariance formula, determine whether economic growth and S&P 500 returns have a positive or inverse relationship. Before you compute the covariance, calculate the mean of x and y.

Economic Growth % (x_i)	S&P 500 Returns % (y_i)
2.1	8
2.5	12
4.0	14
3.6	10

x = 2.1, 2.5, 4.0, and 3.6 (economic growth)
y = 8, 12, 14, and 10 (S&P 500 returns)

$\bar{x} = \sum xin$
$\bar{x} = 2.1+2.5+4+3.64$
$\bar{x} = 12.24$
$\bar{x} = 3.1$

$\bar{y} = \sum xin$
$\bar{y} = 8+12+14+10$
$\bar{y} = 44$
$\bar{y} = 11$

- Now, substitute these values into the covariance formula to determine the relationship between economic growth and S&P 500 returns.

xi	yi	$xi - \bar{x}$	$yi - \bar{y}$
2.1	8	-1	-3
2.5	12	-0.6	1
4.0	14	0.9	3
3.6	10	0.5	-1

$$Cov(x,y)=(-1)(-3)+(-0.6)1+(0.9)3+(0.5)(-1)/4-1$$

$$=3-0.6+2.7-0.5/3$$

$$=4.6/3$$

$$=1.533$$

Covariance can be positive or negative but is covariance is 0 then both variables are independent

Measures of Central Tendency

Data Set: 3,4,3,1,2,3,9,5,6,7,4,8

- Mean 4.583
- Median 1,2,3,3,3,4,4,5,6,7,8,9 Hence median = 4
- Mode The value 3 appears the most time so mode = 3

Measure of Dispersion

Data Set: 3,4,3,1,2,3,9,5,6,7,4,8

- Range (Max - Min): 9-1=8
- Inter Quartile Range: 3rd Quartile – 1st quartile = 75th percentile – 25th percentile = 6.5 -3 = 3.5
- Variance:
- Standard Deviation:

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

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Concept of Random Variables

A **random variable** is a function that associates a real number with each element in the sample space.

- Useful to denote outcomes of random experiments by number.
- Can be done even for categorical outcomes.

- Variable that associates a number with an outcome of a random experiment is random variable.
- **Notation:** Random variable is denoted by a X and its value is denoted by small letter (x)

  Example: The rainfall on a particular day is a random variable R

  Question: What is the probability that rainfall is greater than 10mm? i.e. P(R>10)=?

- For a uniform random variable (all outcomes are equally likely).
- For example, for an unbiased coin, unbiased dice.
- So the probability distribution is called a uniform distribution.

Probability Distribution

Probability Distribution: Tells us how likely a random variable is to take each of its possible states.

Discrete Random variable (DRV) → Probability Mass Function (PMF)

- Has finite or countably infinite range.
- Example – Rolling dice, Tossing coin, no of diagnostic error.
- Probability measured by Probability Mass Function (PMF).

Continuous Random variable (CRV) → Probability Density Function (PDF)

- Has real number interval range.
- Example – Height, Temperature, pressure etc.
- Probability measured by Probability Density Function (PDF).

Discrete Probability Distribution

Discrete variable → Probability Mass Function (PMF)

- PMF – list of possible values along with their probabilities

  Example

  X: Number that comes up on throwing a biased dice

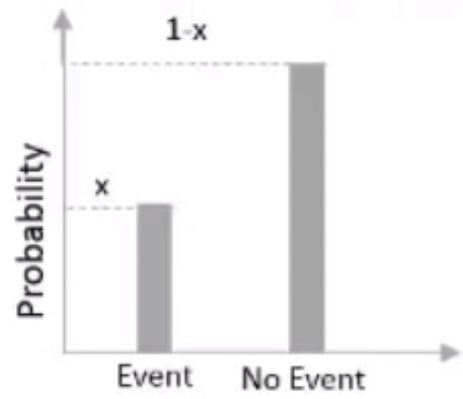
  P(X=1) = 0.1 P(X=2) = 0.1 P(X=3) = 0.2

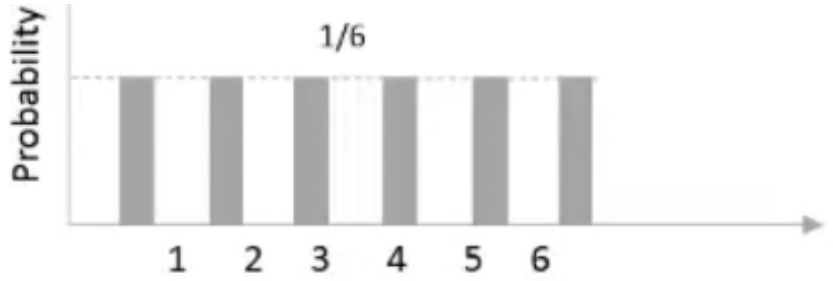
  P(X=4) = 0.2 P(X=5) = 0.2 P(X=6) = 0.2

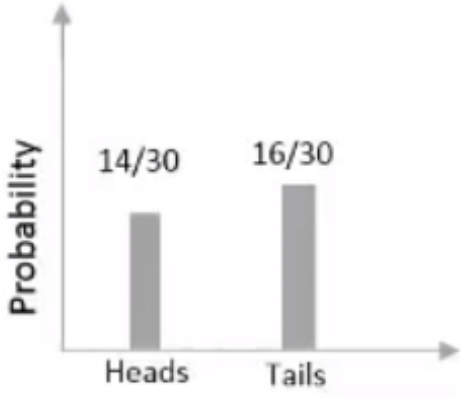
- To be a PMF for a random variable X, a function P satisfies:

  Domain of P should be all possible states of X

- Uniform random variable: P(X=x) = 1/k



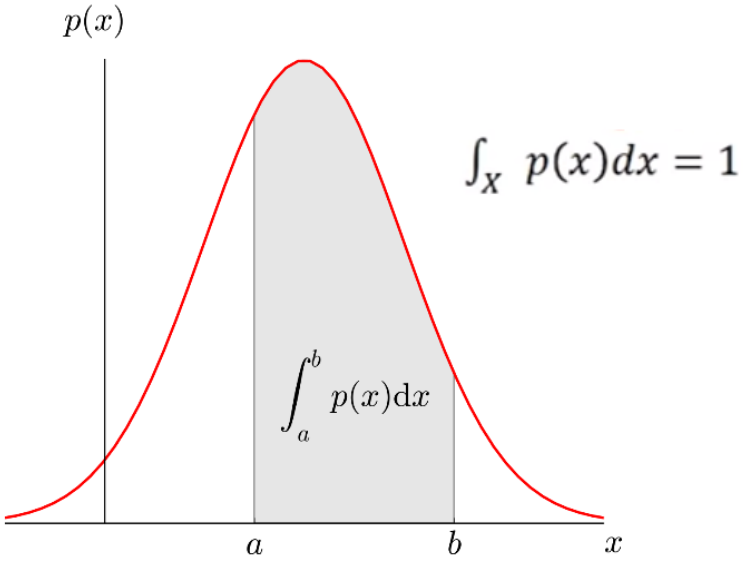




Probability Density Function (PDF) → Continuous variable

R: Amount of rainfall

$P(10<X<20) = P(10 \leq X \leq 20) = \int_{10}^{20} p(x)dx$



- To be a PDF for a continuous variable X, a function P satisfies:

  Domain of P should be all possible states of X

Properties of Probability Density Function

Property 1 Graph of a PDF will be continuous over a range

Property 2 Area bounded by the curve of density function and the x-axis is equal to 1

Property 3 Probability that a random variable assumes a value between a & b is equal to the area under the PDF of a & b

Common Probability Distribution

Discrete

- Uniform Discrete
- Binomial
- Poison
- Geometric

Continuous

- Uniform Continuous
- Normal
- T
- Chi-square

- **Binomial Distribution**

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!\,x!} p^x q^{n-x}$$

Where,
n = number of trials
x = number of success desired
P = probability of getting a success in one trial
q =1-p = probability of getting a failure in one trial

- **Poisson Distribution**

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

Where,
x = 0,1,2,3,….
λ = mean number of occurrences in the interval
e = Euler's constant = 2.71

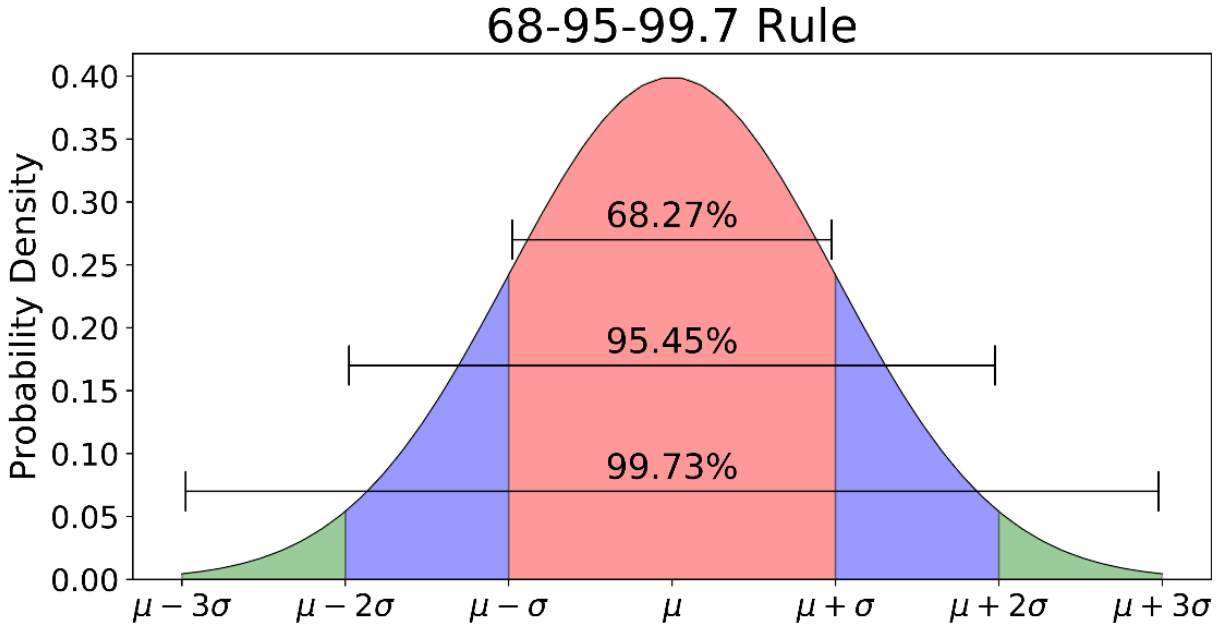
- **Normal Distribution**

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

Where,
μ = mean
σ = standard deviation
e = Euler's constant = 2.71
Π = 3.14

Normal Distribution

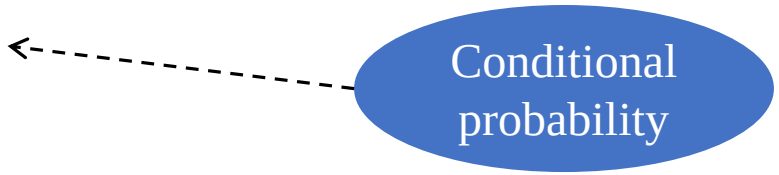
- The percentage of values that lie within a band around the mean in a normal distribution with a width of two, four and six standard deviations, respectively.
- More accurately, 68.27%, 95.45% and 99.73% of the values lie within one, two and three standard deviations of the mean, respectively.



Conditional Probability and Bayes' Theorem

Probability of A given probability of B denoted by P(A|B) is given by –

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



Multiplication Rule

For Dependent events (Probability of one event depend upon another event)

$$P(A \cap B) = P(B) \cdot P(A|B) \qquad P(A \cap B) = P(A) \cdot P(B|A)$$

For Independent events (Probability of one event does not depend upon another event)

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

$$P(A \cap B \cap C) = P(A) \cdot P(B|A) \cdot P(C \mid A \cap B)$$
$$P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$$

One of the most widely used theorem in conditional probability is Bayes' Theorem.

Classification algorithm "Naïve Bayes" is based on Bayes' Theorem

Conditional Probability and Bayes' Theorem

Finding reverse Probability/ Posterior Probability P(B|A) Given P(A|B) and P(B)

We know, Product Rule → $P(A \cap B) = P(A|B)\,P(B)$

$= P(B|A)\,P(A)$

So, $P(A|B) / P(A) = P(B|A)\,P(B)$

Bayes' Theorem → $P(B|A) = P(B) \,.\, P(A|B) / P(A)$

where, P(A) is not given which is total probability

That means, $P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3) + \ldots\ldots$

If the events $B_1, B_2, B_3, \ldots\ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ for i=1,2,3,….,k then for any event A for S such that $P(A) \neq 0$

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} \quad \text{for } r = 1, 2, \ldots, k$$

Chain Rule

The chain rule states that if $A_1, A_2,.., A_n$ are events with probability $P(A_1 \cap A_2 \ldots\ldots \cap A_n) > 0$ then by chain rule,

$$P(A_1 \cap A_2 \ldots\ldots \cap A_n) = P(A_1)\, P(A_2|\, A_1)\, P(A_3|A_1 \cap A_2) \ldots P(A_{n-1}|A_1 \cap A_2 \ldots A_{n-2})$$

Recalling the product rule: p(x, y) = p(x | y) p(y) let us consider 3 variables (x, y, z)
p(x, y, z) = p(x, a) where a is the event (y, z)

$$\Rightarrow p(x, y, z) = p(x, a) = p(x|a)p(a)$$
$$= p(x|a)\, p(y, z)$$
$$= p(x|a)\, p(y|z)\, p(z)$$
$$= p(x|\, y, z)\, p(y|z)\, p(z)$$
$$= p(z)\, p(y|z)\, p(x|y, z)$$

Chain rule of three variables

In general,

$$P(x^{(1)}, x^{(2)}, x^{(3)},.., x^{(n)}) = P(x^{(1)})\, P(x^{(2)}|\, x^{(1)}) \ldots P(\, x^{(n)} \mid x^{(1)}, .. , x^{(n-1)})$$

Chain rule of conditional probability

Example: If we want to find the probability that the image belongs to a cat, we consider the image as 60 × 60 matrix with 3600 pixels and find the joint probability which will be the probability that the image belongs to a cat. Computation of joint probability can be done using chain rule.

Why do we need Regression Analysis?

- To establish relation between the variables.
- To reduce insignificant many to vital few variables.
- To better understand the processes
- To forecast the variables
- To identify the factors effecting the response
- To provide baseline for process performance

Identify Variables...

The important component of statistical analysis is to establish the relationship between the output variable(Y) and input variable(X).

$$\mathbf{Y = f(x)}$$

Where, f is the functional relationship

To understand an output Y we need to focus on the x

Establish the relationship between Y and X's

y = f(x)

Dependent	Independent
Output	Input
Response	Variable
Effect	Cause
Symptom	Problem

Identify the X's which influence Y more effectively

Correlation

- The correlation is a statistical tool which studies the relationship between two variables.
- Correlation first developed by Sir Francis Galton (1822-1911) and then reformulated by Karl Pearson(1857-1936)
- Note: The degree of relationship or association is known as the degree of relationship.

Types of Correlation:-

(i) Positive and Negative correlation

(ii) Simple, Partial and multiple Correlation

(iii) Linear and Non-linear correlation

Correlation Analysis

- Two variables are said to be correlated if change in variable results in a corresponding change in other variable.
- Correlation analysis gives an idea of degree and direction of the relationship between two variables under study.
- The correlation between two variables can be studied by
    i. Scatter diagram
    ii. Karl Pearson's coefficient of correlation
    iii. Pearson's Rank Correlation
    iv. Correlation Ratio

Positive and Negative Correlation

- If the values of the two variables deviate in the same direction i.e. if an increase(or decrease) in the values of one variable results, on an average, in a corresponding increase (or decrease) in the values of the other variable the correlation is said to be positive.
- Examples: Height and weights

  Household income and expenditure

  Price and supply of commodities

  Amount of rainfall and yield of crops
- Correlation between two variables is said to be negative or inverse if the variables deviate in opposite direction. That is, if increase(or decrease) in the values of one variable results on an average, in corresponding decrease(or increase) in the values of other variable.
- Examples:- Volume and pressure of perfect gas

  Price and demand of goods

Linear and Non-linear Correlation

- The correlation between two variables is said to be linear if the change of one unit in one variable result in the corresponding change in the other variable over the entire range of values.
- Example:-

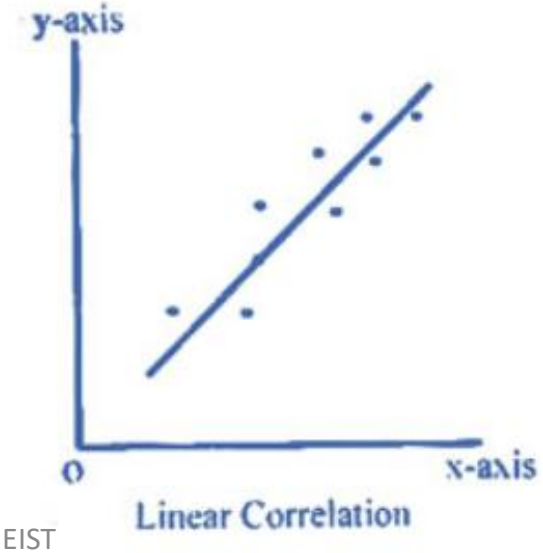
X	2	4	6	8
Y	7	13	19	25

- For a unit change in the value of X, there is a constant change in the corresponding values of Y and the above data can be expressed by the relation;

$$y = 3x+1$$

In general;

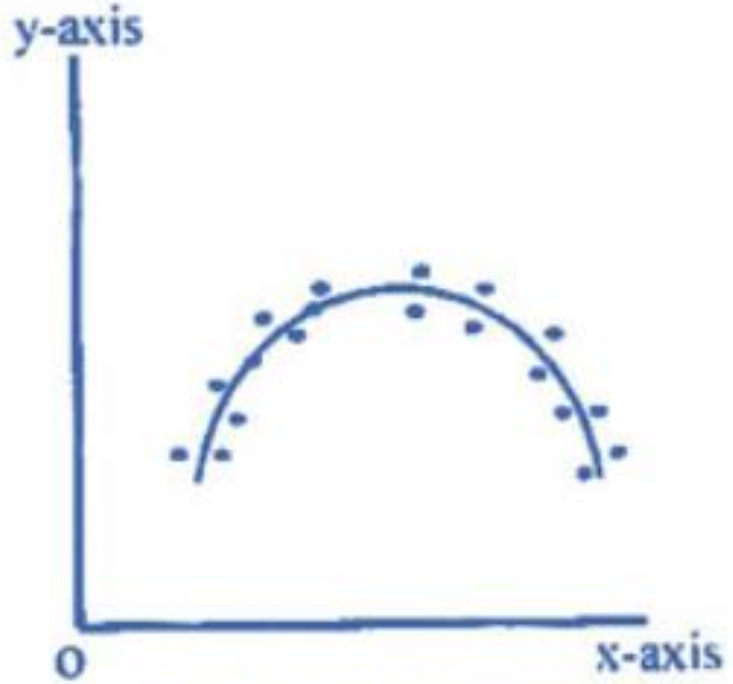
$$y = a + bx$$



Linear Correlation

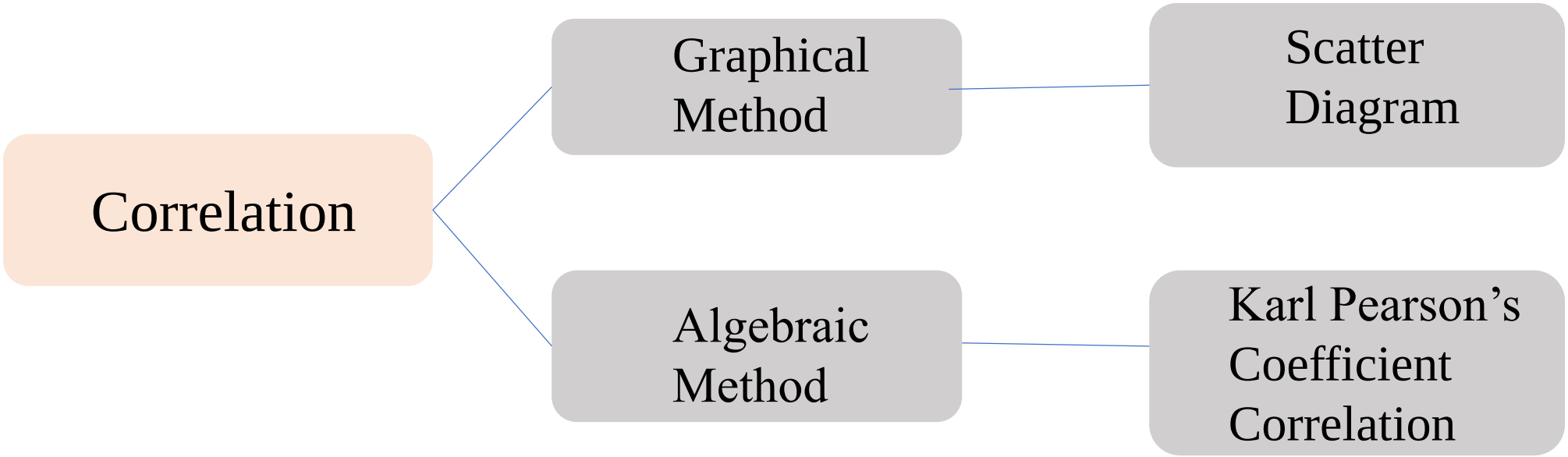
Non-Linear Correlation

- The relationship between two variables is said to be non-linear if corresponding to a unit change in one variable, the other variable does not change at a constant rate but changes at a fluctuating rate.
- In such cases if the data is plotted on a graph sheet we will not get a straight line curve.
- Example: y = a + bx + cx2



Non Linear Correlation

Methods of studying Correlation

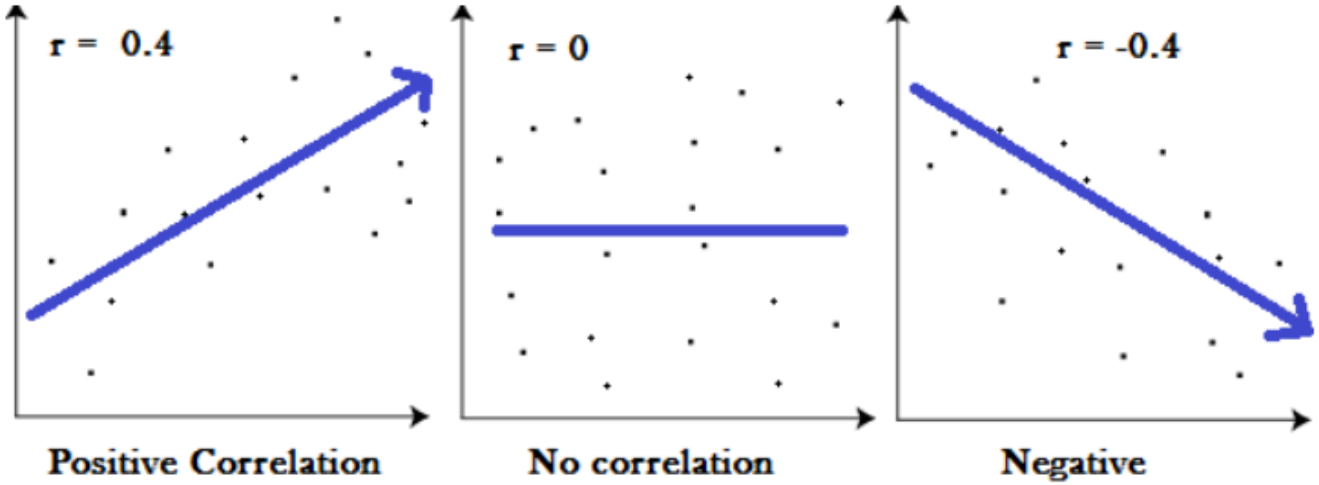


The Coefficient of Correlation

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

- Correlation of coefficient ‘r’ measures the degree of association between the two values of related variables given in the dataset.
- Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

  - 1 indicates a strong positive relationship.
  - -1 indicates a strong negative relationship.
  - A result of zero indicates no relationship at all.



Example

A study was conducted to find whether there is any relationship between the weight and blood pressure of an individual. The set of data was arrived at from a clinical study. Determine the coefficient of correlation for this set of data. The first column represents the serial number and the second and the third columns represent the weight and blood pressure of each patient.

S. No	Weight	Blood Pressure
1	78	140
2	86	160
3	72	134
4	82	144
5	80	180
6	86	176
7	84	174
8	89	178
9	68	128
10	71	132

Solution

X	Y	X^2	Y^2	XY
78	140	6084	19600	10920
86	160	7396	25600	13760
72	134	5184	17956	9648
82	144	6724	20736	11808
80	180	6400	32400	14400
86	176	7396	30976	15136
84	174	7056	30276	14616
89	178	7921	31684	15842
68	128	4624	16384	8704
71	132	5041	17424	9372
796	1546	63,776	243036	1242069

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Regression

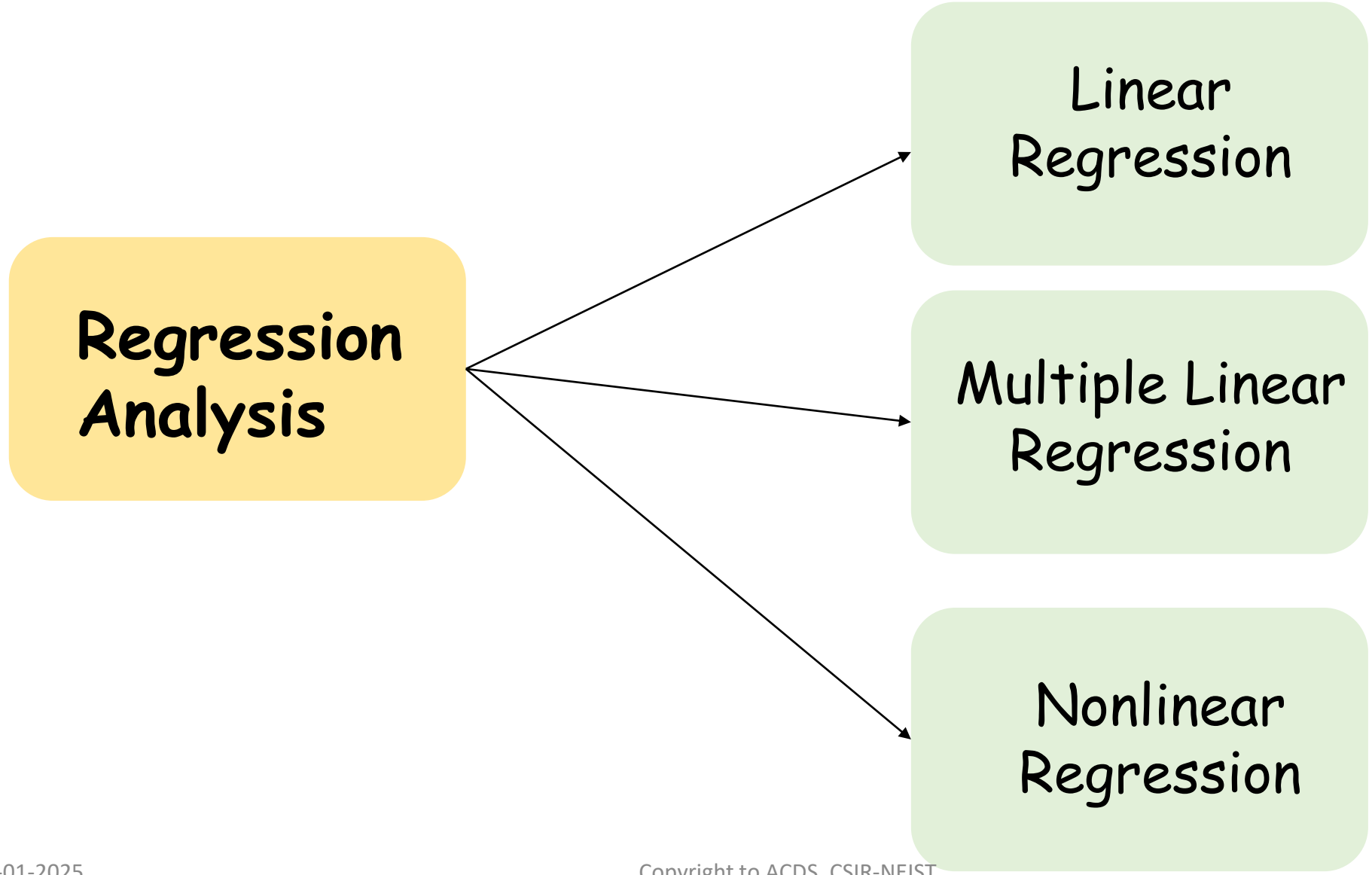
- Regression means stepping back or returning to the average value.
- Regression analysis is a mathematical measure of average relationship between two or more variables.
- The regression analysis equation is the same as the equation for a line which is

$$y = MX + b$$

Where,

- Y= the dependent variable of the regression equation
- M= slope of the regression equation
- x=dependent variable of the regression equation
- B= constant of the equation

Regression Analysis



Regression Analysis–Linear model assumptions

Linear regression analysis is based on six fundamental assumptions:

- The dependent and independent variables show a linear relationship between the slope and the intercept.
- The independent variable is not random.
- The value of the residual (error) is zero.
- The value of the residual (error) is constant across all observations.
- The value of the residual (error) is not correlated across all observations.
- The residual (error) values follow the normal distribution.

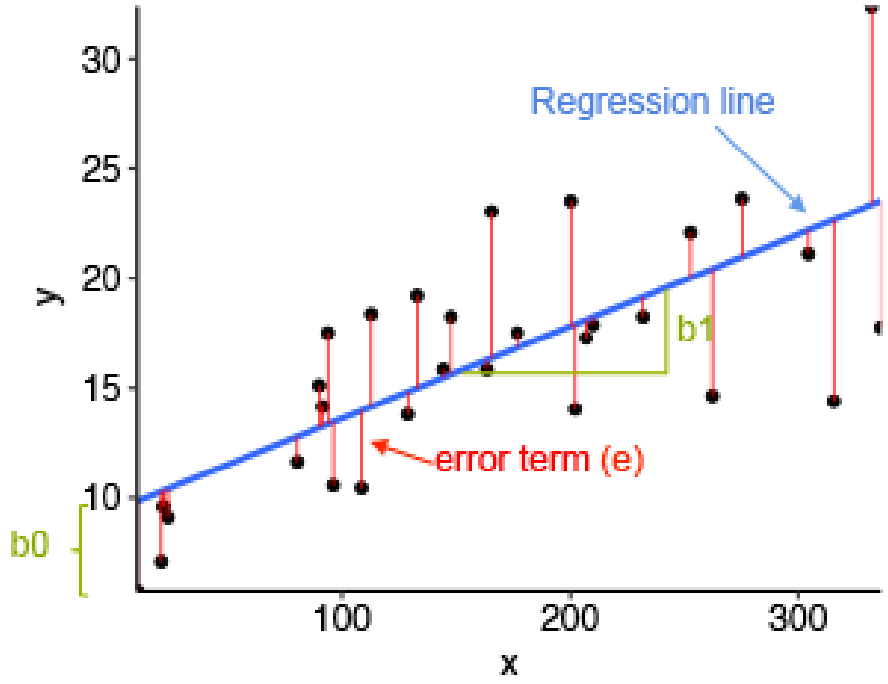
Linear Regression

- Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \epsilon$$

Where:

- **Y** – Dependent variable
- **X** – Independent (explanatory) variable
- **a** – Intercept
- **b** – Slope
- **ϵ** – Residual (error)



Example

Find linear regression equation for the following two sets of data:

X	2	4	6	8
Y	3	7	5	10

Solution:

Construct the following table:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
∑x = 20	∑y = 25	$\sum x^2 = 120$	∑xy = 144

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

$$b = \frac{4\times 144 - 20\times 25}{4\times 120 - 400}$$

$$b = 0.95$$

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

$$a = \frac{25\times 120 - 20\times 144}{4(120) - 400}$$

$$a = 1.5$$

Linear regression is given by:

$y = a + bx$

$Y = 1.5 + 0.95\ x$

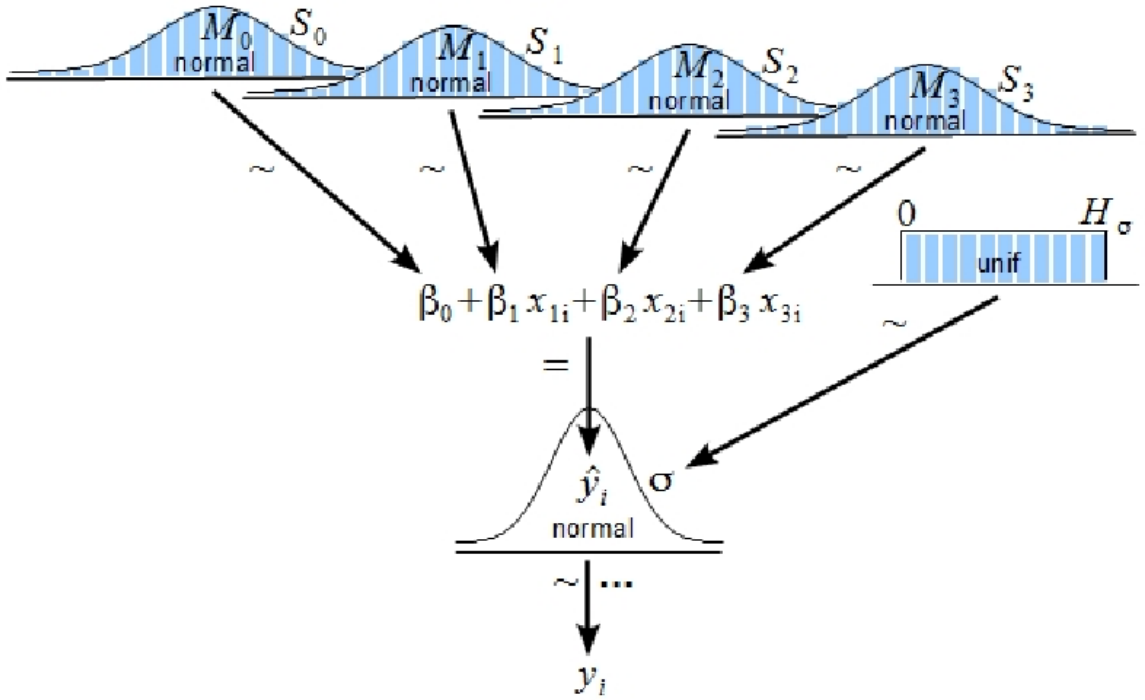
Multiple Linear Regression

- Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \epsilon$$

Where:

- **Y** – Dependent variable
- **X_1, X_2, X_3** – Independent (explanatory) variabl
- **a** – Intercept
- **b, c, d** – Slopes
- **ϵ** – Residual (error)



To find the best-fit line for each independent variable, multiple linear regression calculates three things:

- The regression coefficients that lead to the smallest overall model error.
- The *t*-statistic of the overall model.
- The associated *p*-value (how likely it is that the t-statistic would have occurred by chance if the null hypothesis of no relationship between the independent and dependent variables was true).

It then calculates the *t*-statistic and *p*-value for each regression coefficient in the model.

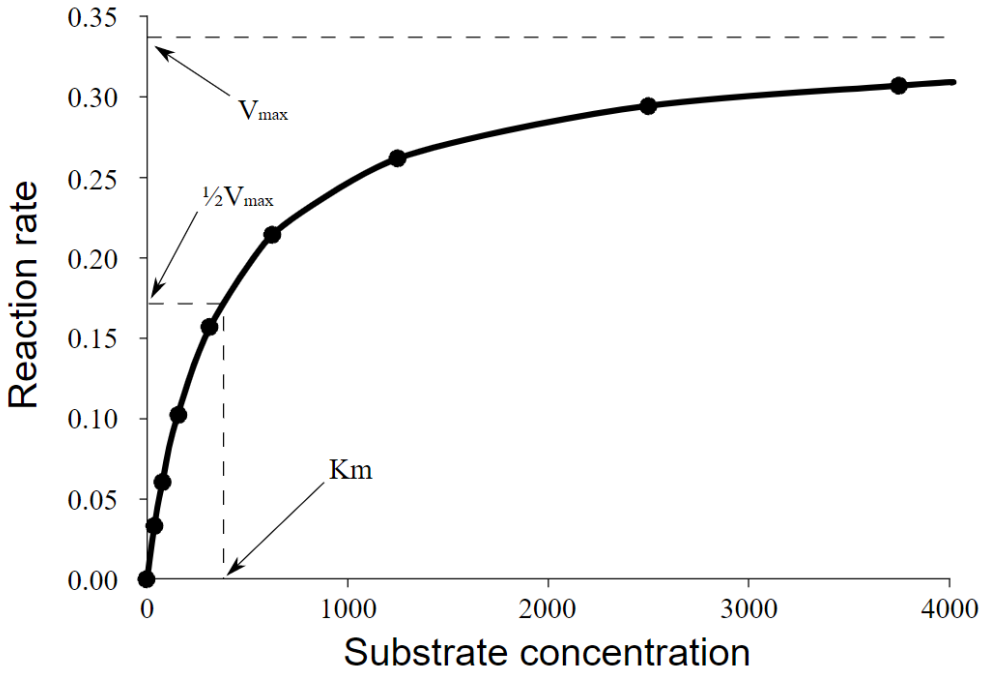
Non linear Regression

- Nonlinear regression is a regression in which the dependent or criterion variables are modeled as a non-linear function of model parameters and one or more independent variables.
- **Data:-**The dependent and independent variables should be quantitative. Categorical variables, such as religion, major, or region of residence, need to be recoded to binary (dummy) variables or other types of contrast variables.

$$Y = f(X,\beta) + \varepsilon$$

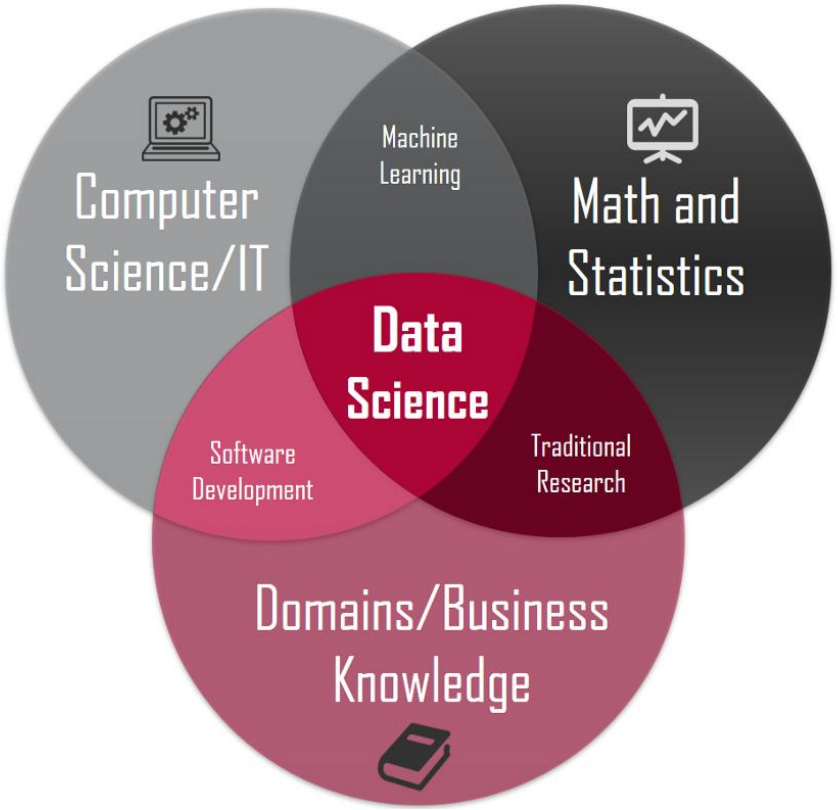
Where:

- X = a vector of *p* predictors,
- β = a vector of *k* parameters,
- f(-) = a known regression function,
- ε = an error term.



Importance of Statistics in Data Sciences

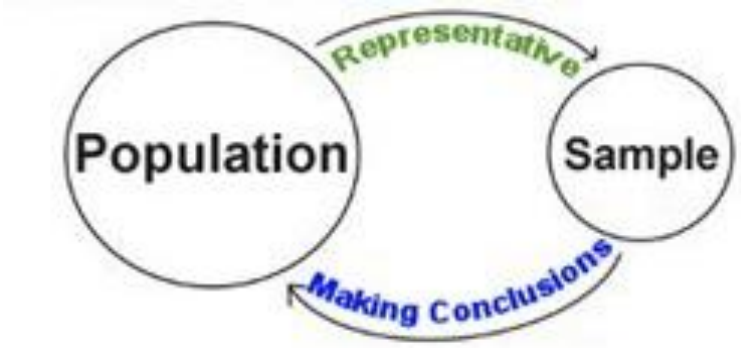
- Data Science is that sweet spot that sits perfectly amidst computer programming, statistics and the domain on which the analysis is performed. Let us see how.



Significance in Data Science

- Making inferences about the population from the sample.
  - Whether a sample is significantly different from the population?
  - If adding or removing a feature from a model will really help to improve the model?
  - If one model is significantly better than the other?

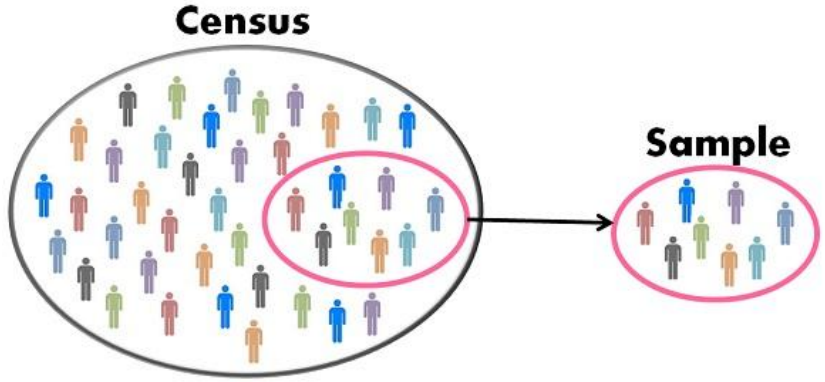
- Aim is on hypothesis testing in general.



Population vs. Sample

Population (Donated by N)

- Contains all members of a specified group (the entire list of possible data values).
- All the Data Science students in India.



Sample (Denoted by n)

- A sample data set contains a part or a subset of a population.
- The size of a sample is always less than the size of the population from which it is taken.
- Sampling is the main technique employed for data selection.
- It is often used for both the preliminary investigation of the data and the final data analysis.
- It has approximately the same property (of interest) as the original set of data.
- Sampling used because –
  - Obtaining entire set of data of interest is too expensive or time consuming.
  - Processing the entire set of data of interest is too expensive or time consuming.

Hypothesis Test

- A rule that specifies whether to accept or reject a claim about a population depending upon the evidence provided by a sample.

- A hypothesis test examines two opposing hypotheses about a population:
  - Null hypothesis - It is the statement being tested - "no effect" or "no difference“

$$H_0: \mu = \mu_0$$

  - Alternative hypothesis - It is the statement that conclude true based on evidence from sample.

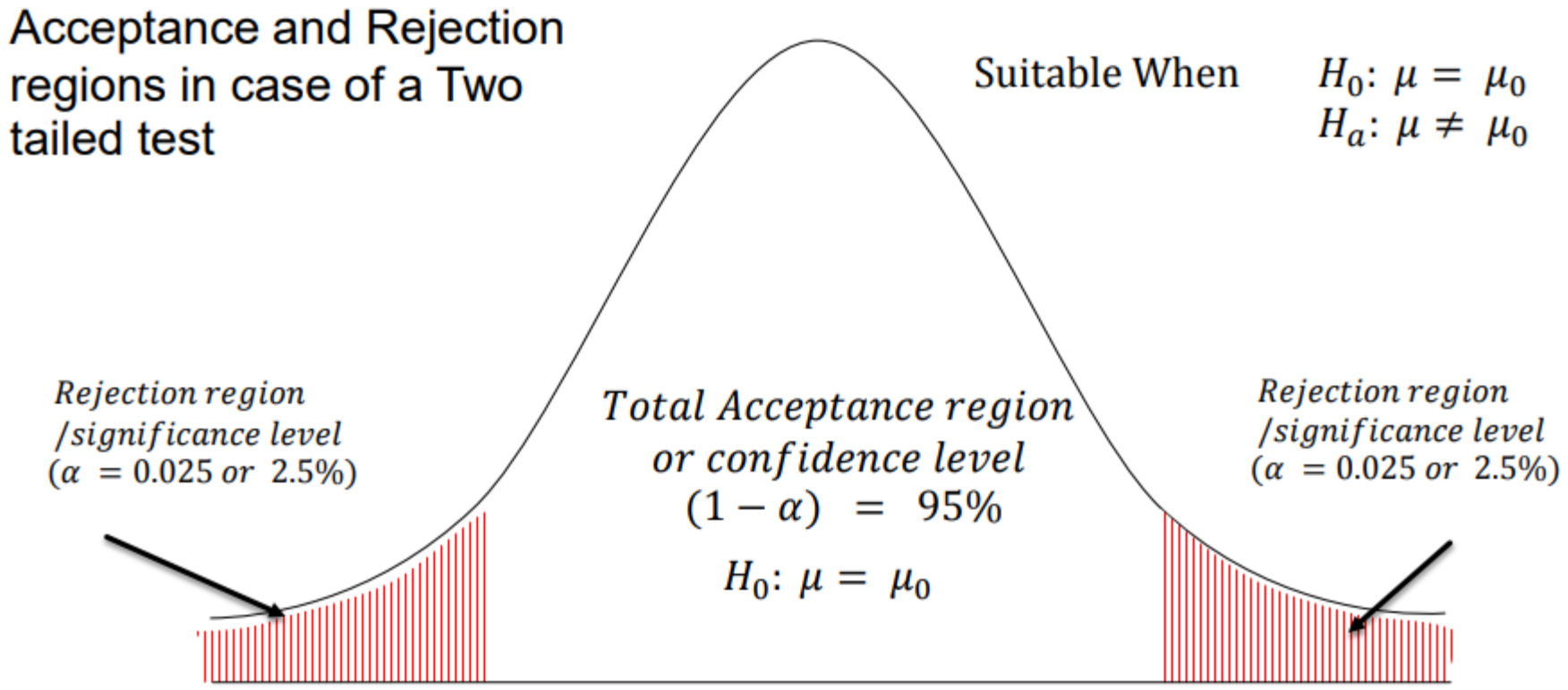
$$H_a: \mu \neq \mu_0$$

$$H_a: \mu > \mu_0$$

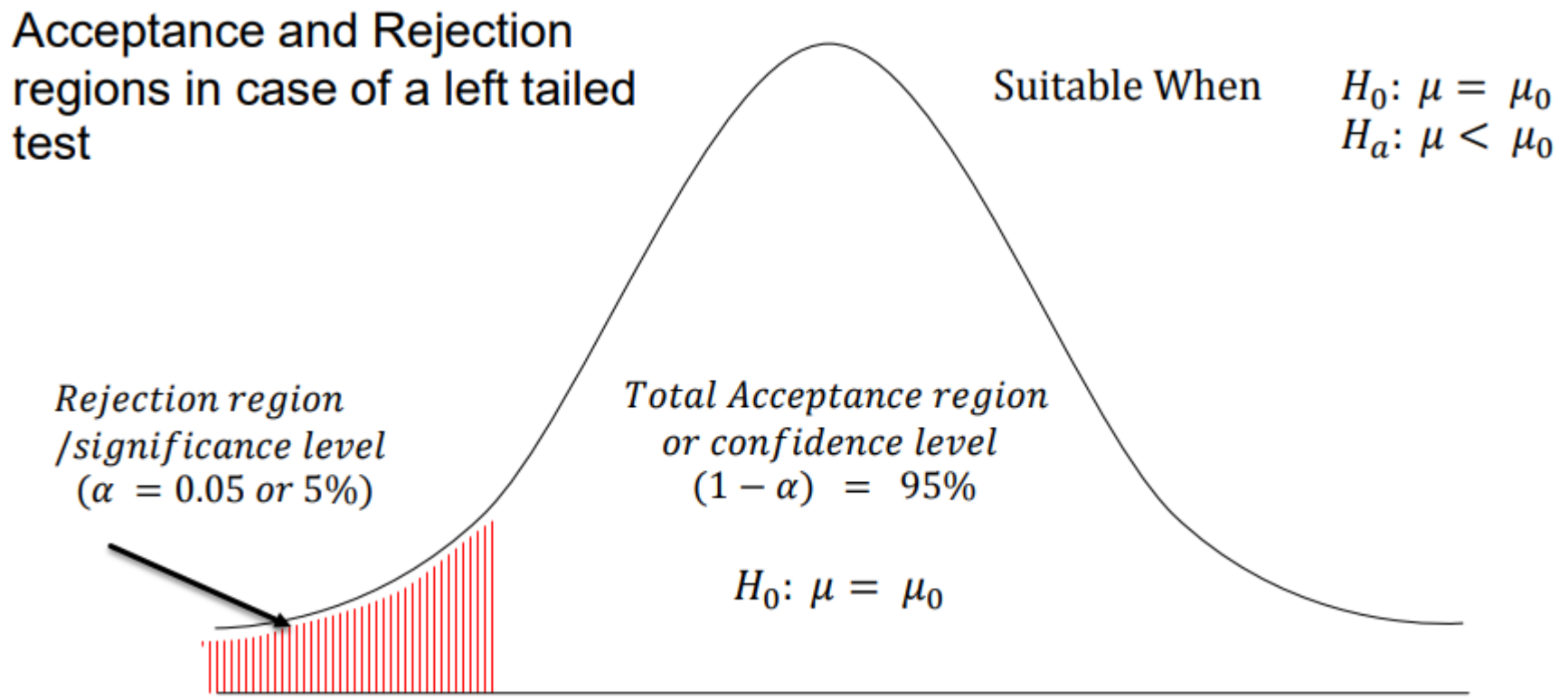
$$H_a: \mu < \mu_0$$

- Significance means the percentage risk to reject a null hypothesis when it is true and it is denoted by α. Generally taken as 1%, 5%, 10%
- $(1 - \alpha)$ is the confidence interval in which the null hypothesis will exist when it is true.

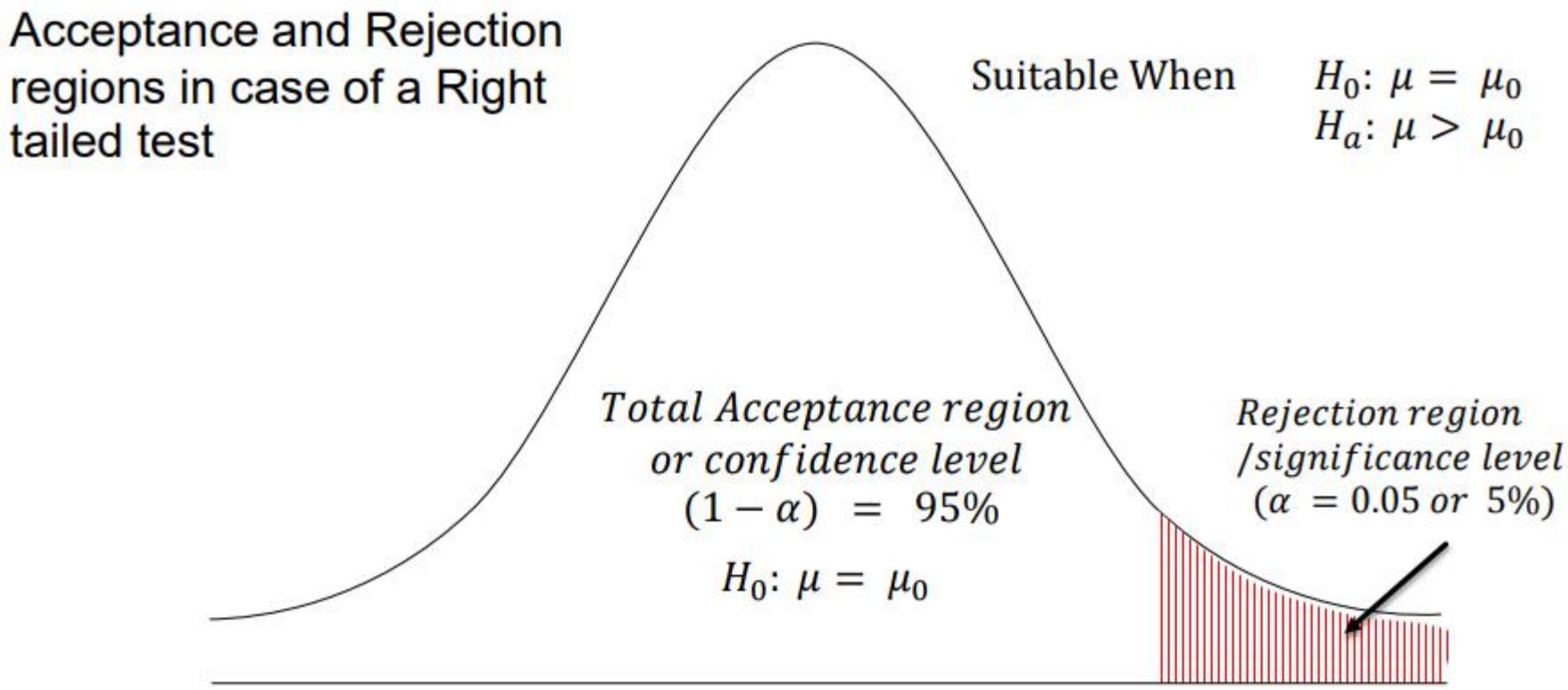
Two tailed test at Significance level of 5%



Left tailed test at Significance level of 5%



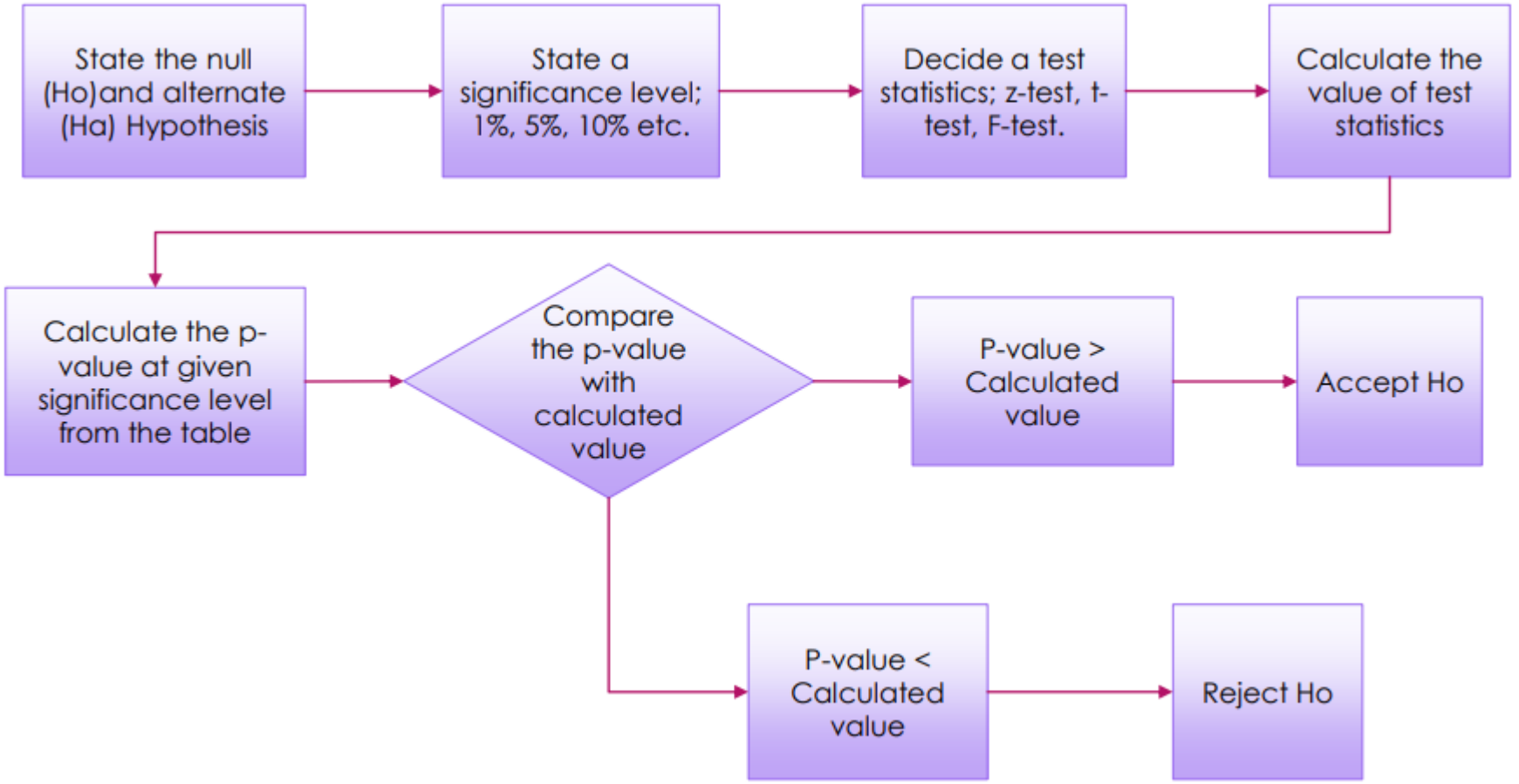
Right tailed test at Significance level of 5%



Type I and Type II errors

Situation	Decision	
	Accept Null	Reject Null
Null is true	Correct	Type I error (α *error*)
Null is false	Type II error (β *error*)	Correct

Procedure for Hypothesis Testing



Z Test

- Used to validate a hypothesis that the sample drawn belongs to the same population.

  Null: Sample mean is same as the population mean

  Alternate: Sample mean is not same as the population mean

- In a z-test, the sample is assumed to be normally distributed.

- The statistics for this hypothesis testing is called z-statistic, the score for which is calculated as

  $$\mathbf{z = (x - \mu) / (\sigma/\sqrt{n})}$$

  where

  x = sample mean

  μ = population mean

  $\sigma/\sqrt{n}$ = population standard deviation

- If the test statistic is lower than the critical value, accept the hypothesis or else reject.

TTest

- Used to compare the mean of two given samples.
- Like a z-test, a t-test also assumes a normal distribution of the sample.
- A t-test is used when the population parameters (mean and standard deviation) are not known.

- The statistic for this hypothesis testing is called t-statistic, the score for which is calculated as

$$t = (x1 - x2) / (\sigma/\sqrt{n1} + \sigma/\sqrt{n2})$$

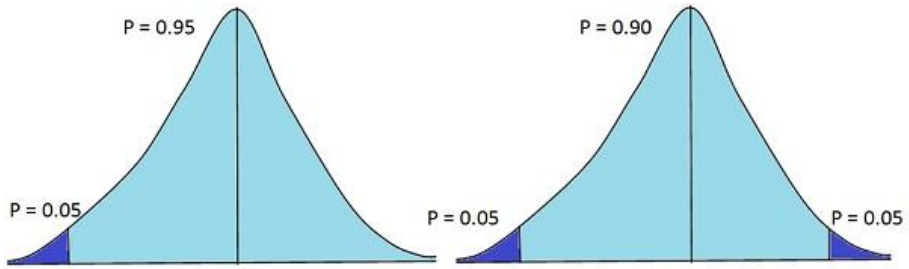
where

x1 = mean of sample 1

x2 = mean of sample 2

n1 = size of sample 1

n2 = size of sample 2



One-tailed Test Vs Two-tailed Test

P and Alpha Value

P-value

- During hypothesis testing *p*-value helps in determining the significance of results.
- All hypothesis tests ultimately use a *p*-value to weigh the strength of the evidence.

Alpha-value

- Alpha (critical) value is a point beyond which we reject the null hypothesis.
- Typically alpha-value is 0.05

- Both *p*-value and alpha-value is a number between 0 and 1 and interpreted in the following way:
  - *p*-value > alpha-value says weak evidence against null hypothesis so fail to reject null hypothesis.
  - *p*-value ≤ alpha-value says strong evidence against null hypothesis, so reject null hypothesis.
  - *p*-values close to cutoff alpha-value are considered to be marginal (could go either way).

Example of Hypothesis Testing

Q. A principal at a certain school claims that the students in his school are above average intelligence. A random sample of thirty students IQ scores have a mean score of 112. Is there sufficient evidence to support the principal's claim? The mean population IQ is 100 with a standard deviation of 15.

Solution:

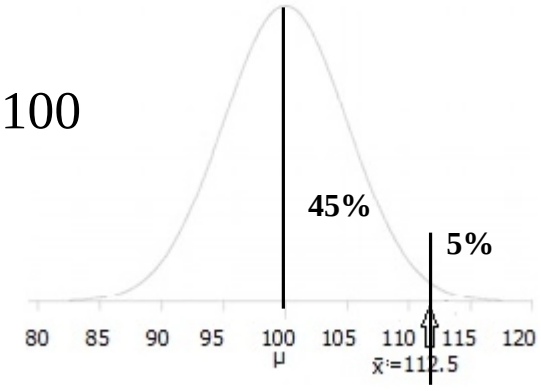
Step 1: Null hypothesis – The fact that mean population IQ = 100, H_0 : $\mu = 100$

Step 2: Alternate hypothesis – The claim that students have average IQ, H_1 : $\mu > 100$

Step 3: Since α value is not stated, we take the general value $\alpha = 0.05$ (5%)

Step 4: Find the rejection region from the z table which is 1.645 for $\alpha = 0.05$

Step 5: Use z test to find the test statistics with $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$



$Z = (112.5 - 100) / (15/\sqrt{30}) =$ **4.56**

If Step 5 is greater than Step 4, reject the null hypothesis, else null hypothesis cannot be rejected.

In this case, it is greater (4.56 > 1.645), so you can reject the null.

Example of Hypothesis Testing

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545

Statistical Modelling and Graphical Diagnostics

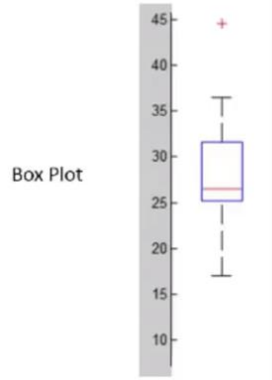
Graphical analysis in statistics makes us visualize the data/numbers

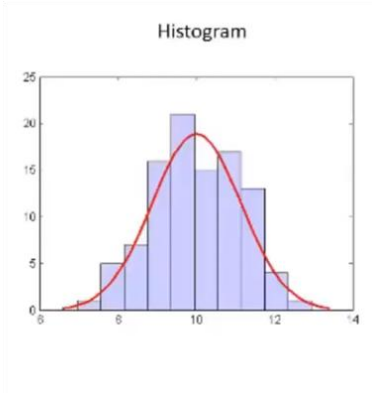
Benefits of Graphical analysis

- ❖ Allows to learn about the nature of the process.
- ❖ Enables clarity of communication.
- ❖ Helps understanding sources of variation in the data.

Single variable

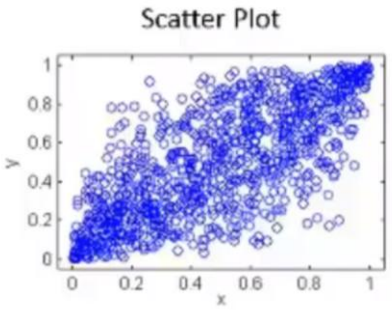
- ***Bar graph, Pie Chart:*** For categorical Variables
- ***Box Plot, Histogram:*** For quantitative Variables

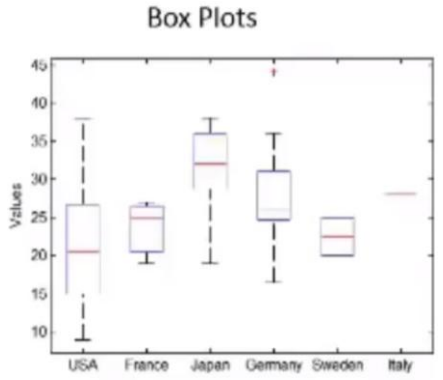




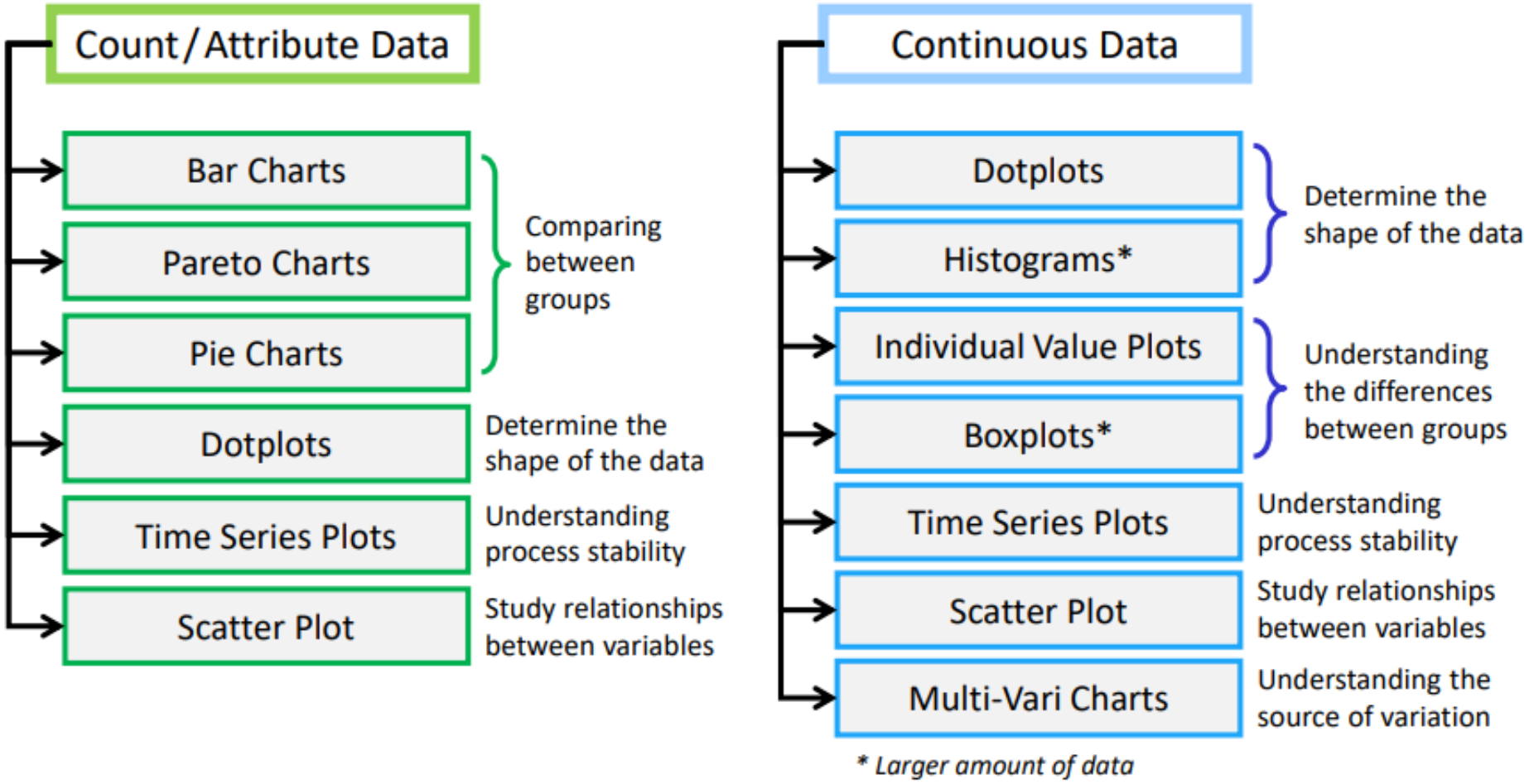
Multiple variable

- ***Scatter plot:*** Two quantitative variable
- ***Box plot:*** One categorical with one quantitative variable



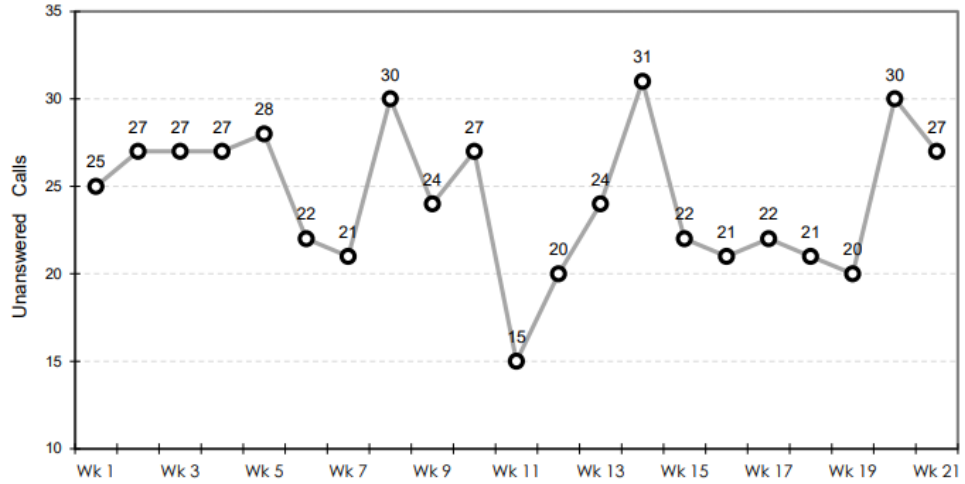


Graph Selection



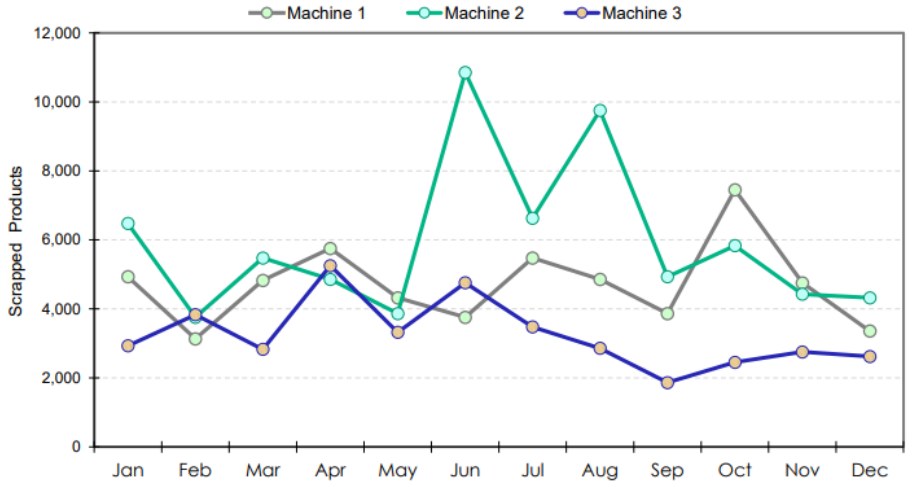
Graphical Representation

Example – The number of unanswered calls in a call center:



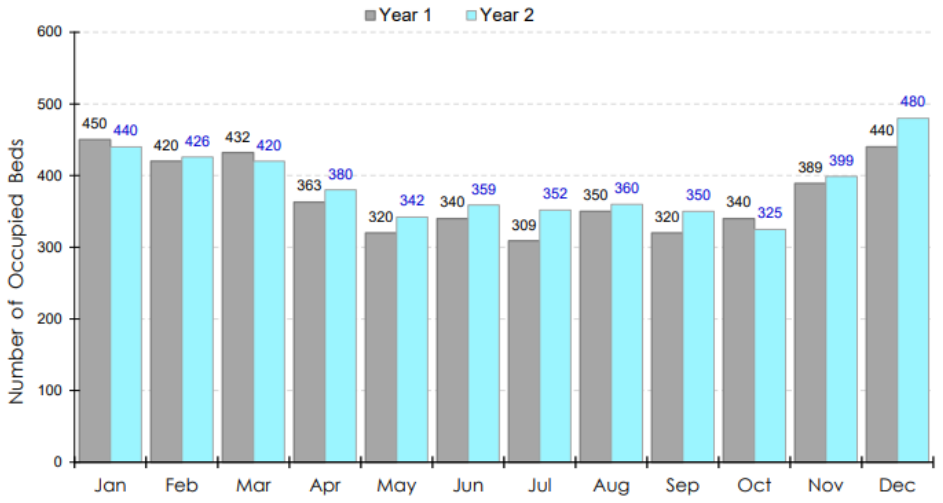
A time series plot for evaluating **count data**

Example – The number of scrapped products generated from three machines:

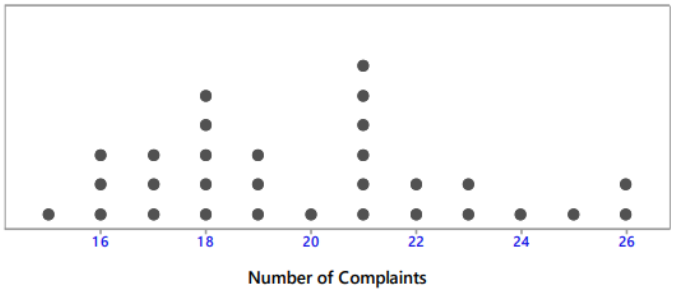


Graphical Representation

Example – A grouped bar chart displaying the number of occupied beds in a hospital in two consecutive years.

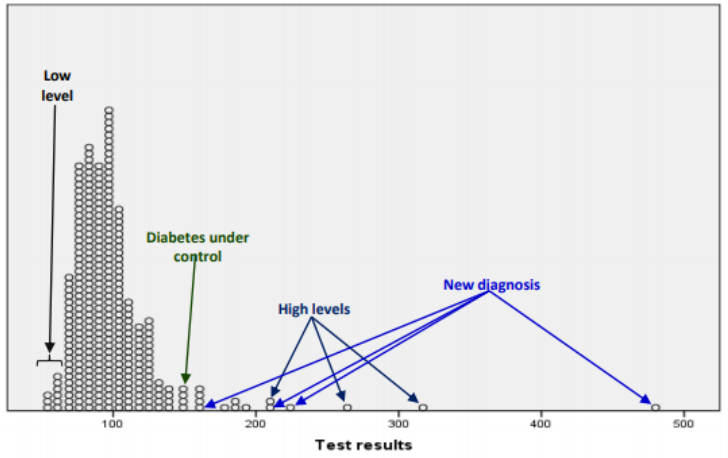


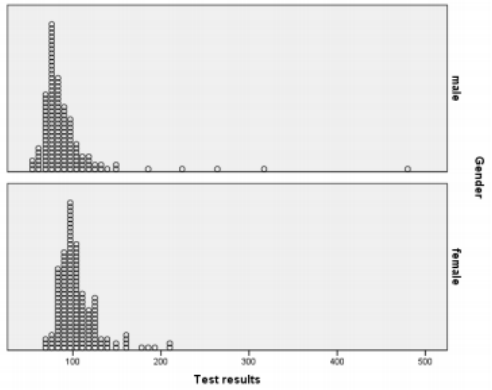
Example – A dotplot that displays the number of complaints made by customers in a given period of time.



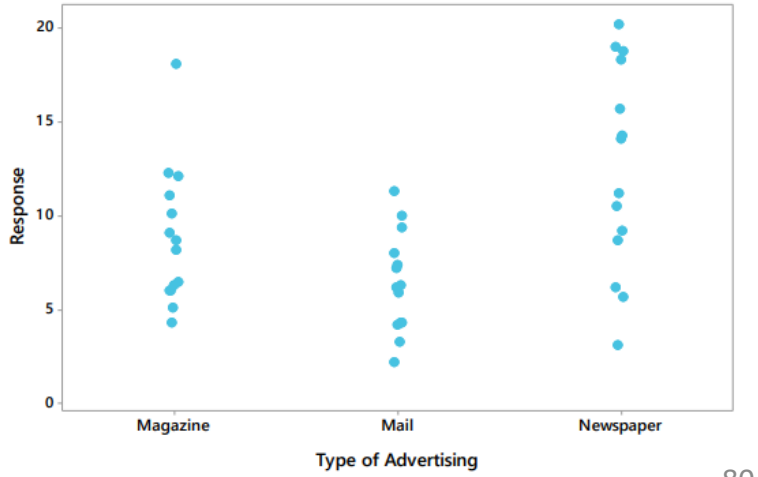
Graphical Representation

Example – An analysis that was conducted for diagnosing the presence of diabetes at a workplace.





Example – An individual value plot showing the responses of a particular marketing campaign that uses multiple advertising methods.



Basic Questions on Probability and Statistics

1. A college president wants to find out which courses are popular with students. What procedure would be most appropriate for obtaining an unbiased sample of students?

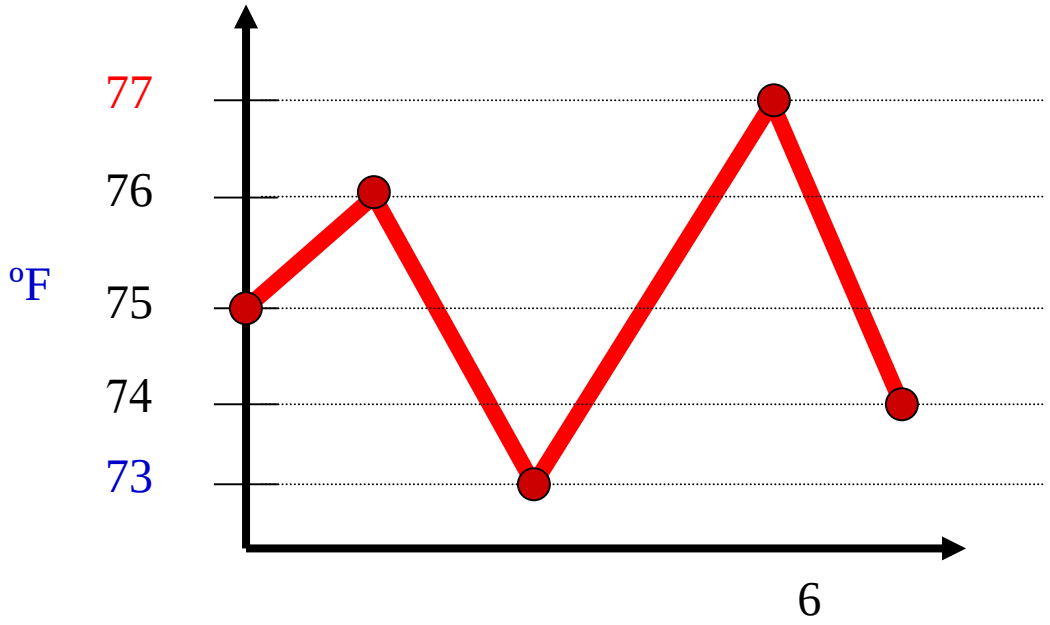
A. Survey a random sample of students from the English Department.

B. Survey the first hundred students from an alphabetical listing.

C. Survey random sample of students from list of entire student body.

2. The graph shows the yearly average temperature from 1980 to 1985. What is the difference between the highest and lowest?

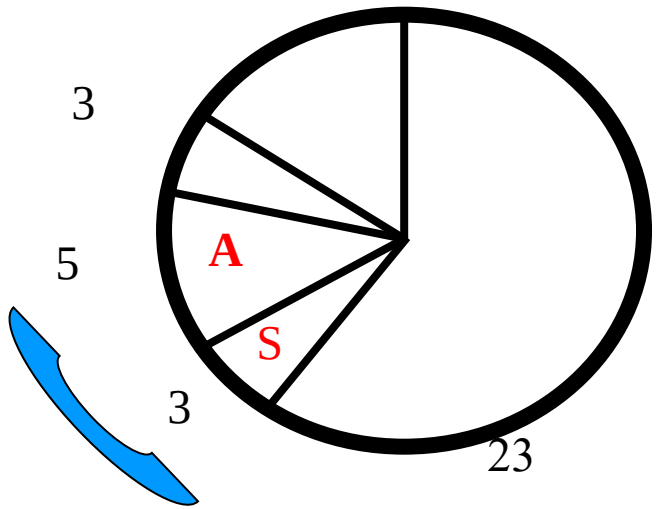
77 - 73 = 4



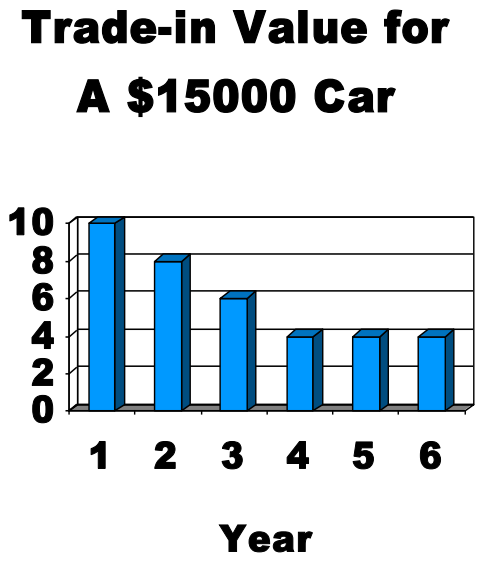
3. The number of people employed in different work areas in a manufacturing plant are represented by the circle graph. What **percent** are represented in **Sales and Administration combined?**

Total = 40

$$\frac{8}{40} = \frac{p}{100}, \quad 40p = 800, \quad p = 20$$



4. Consider the following graph showing the value of a $15,000 car after 1, 2, 3, 4, 5 and 6 years. In what year did the price of the car begin to stabilize?



5. Find mean, median & mode of the data in this sample: 6, 15, 24, 23, 29, 22, 21, 29, 29

Arrange in order:

6, 15, 21, 22, 23, 24, 29, 29, 29

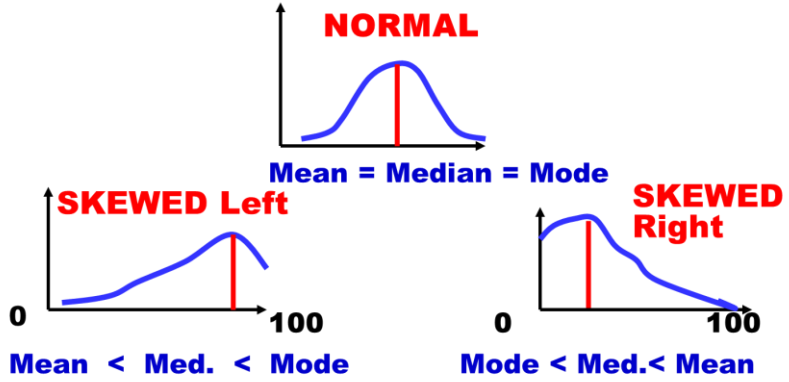
Median is 23 (middle)

Mode is 29 (most frequent)

(6+15+21+22+23+24+29+29+29)/9

198/9=22 the mean

But, (29 + 15)/2 = 22 !!!



Basic Question on Probability and Statistics

6. Two common sources of protein for US adults are beans & meat. If **75%** of US adults eat **meat**, **80%** eat **beans** and 70% **eat both** meat & beans, what is the probability that a randomly selected adult eats meat **or** beans?

P(meat or beans)
=P(meat) or P(beans) - P(both)
=75% + 80% - 70% = 85% $= \frac{85}{100} = \frac{17}{20}$

Worked Examples

1. If a coin is tossed 3 times, find the probability that no successive tosses show the same face?

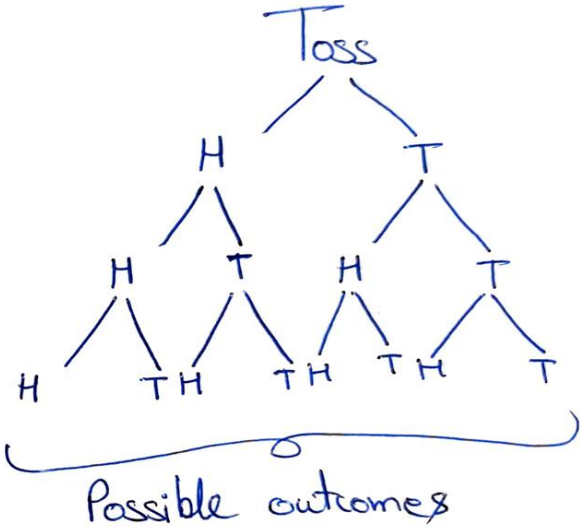
Solution:

∴ Possible outcomes are :-

{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

∴ Total outcomes = 8

Favourable outcomes = {HTH, THT}



$$\therefore \text{Probability} = \frac{\text{Favourable outcomes}}{\text{Total outcomes}}$$

$$= \frac{2}{8}$$

$$= \frac{1}{4}$$

2. If 8 boys are arranged in a row, what is the probability that 3 particular boys will sit together?

Solution:

Total numbers of boys = 8

Total numbers of possible arrangement = 8!

Now, 3 boy will be together,

Hence, just group these three boys together and consider as a single person.

∴ we can take number of boys as 6.

∴ These 6 boys can be arrangement = 6!

Also, we had grouped three boys together. These three boys can be arranged among themselves in 3!.

Hence, required number of ways to arrange boys is 6! × 3!

Therefore the required probability $= \frac{6! \times 3!}{8!} = \frac{6! \times 3 \times 2}{8 \times 7 \times 6!} = \frac{6}{28}$

Hence, the probability that 3 particular boys always sit together is $\frac{3}{28}$

Worked Examples

3. If a 5 digit number is formed using digits 1, 2, ..., 9 (without repetition), then what is the probability that the number is even?

Solution:

Total possible 5 digit number (without repetitive) $= {}^{9}P_5 = \frac{9!}{(9-5)!} = \frac{9!}{4!}$

Hence, Even no., $\underline{\quad}\ \underline{\quad}\ \underline{\quad}\ \underline{\quad}\ \underline{\quad}$ ⇝ even place (2, 4, 6, 8)

$\therefore 5 \times 6 \times 7 \times 8 \times 4 =$ Total Even Number.

$\therefore$ Probability $= \frac{8 \times 7 \times 6 \times 5 \times 4}{\frac{9!}{4!}} = \frac{8 \times 7 \times 6 \times 5 \times 4 \times 4!}{9 \times 8 \times 7 \times 6 \times 5 \times 4!} = \frac{4}{9}$

Worked Examples

4. The average man drinks 2L of water when active outdoors (with standard deviation of 0.7 L). A planning for a full day nature trip for 50 men and will bring 110L of water is being made. What is the probability that you will run out?

Solution:

The standard deviation, $\sigma = 0.7L$

Mean, $\mu = 2L$

$n = 50$

P(average water use per man is $> 2.2L$ Per man)

$\bar{x} = \frac{110L}{50} = 2.2L$

Standard distribution of $\bar{x}$ = $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$

$\therefore \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.7}{\sqrt{50}} = 0.099$

We just need to figure out how many standard deviations 2.2L is away from the mean (known as the Z-score).

$Z = \frac{2.2 - \mu}{\sigma} = \frac{2.2 - 2}{0.099} = 2.02$

Now, we can use a Z-table to figure out that probability:

Z-table of 2.02 = 0.9783

$\therefore$ P (running out) = 1 − (.9783) = 0.0217

Worked Examples

5. A study on the current material status of women aged 25 to 29 years old conducted on a large group of women between the ages 25 and 29. The table below summarizes the findings of this study.

Marital Status	Never married	Married	Divorced	Widowed
Proportion	0.468	0.459	0.06	0.013

What is the probability that a randomly selected woman aged 25 to 29 years is currently either never married or divorced? Also what is the probability that the first woman has never been married and the second women is married?

Solution:

Probability (Never Married or divorced)

Outcomes 1 :- never married
Outcomes 2 :- divorced

Hence, never married & divorced are disjoint.

Never Married — divorced.

∴ Probability (never married or divorced) = P(never Married) + P(divorced)

∵ disjoint

= 0.468 + 0.06

= 0.528

Outcomes 1 :- First woman has never married

Outcomes 2 :- Second women is married.

∴ P(first woman N.M and Second woman married)

= P(never married) × P(Married)

= 0.468 × 0.459

= 0.2148

6. An automobile manufacturer buys computer chips from a supplier. Each chip chosen from the shipment has probability of 0.05 of being defective. Each automobile uses 7 computer chips that are both selected independently or work independently of each other. What is the probability that all 7 chips in an automobile will work properly? Also what is the probability that at least 1 of the 7 chips in an automobile are defective?

Solution:

Probability (defective) = 0.05

∴ P(7 chip work properly)

Outcome 1 ⇒ chip 1 work properly
Outcome 2 ⇒ chip 2 work properly
⋮
Outcome 7 ⇒ chip 7 work properly

} 7 outcomes are independent

∴ (P(all 7 chips work properly) = P(chip 1) × P(chip 2) × ... × P(chip 7)

= 0.95 × 0.95 × ... × 0.95

= 0.6983

P(1 chip defective) = can be any one 1 chip

∴ atleast = 1, ∴ may other chip can be defective.

∴ may possible outcomes of defective

∴ P(atleast 1 chip defective) = 1 − P(no. of chip defective)

= 1 − 0.6983

= 0.3017 #

7. The bottom 30% of the students failed an end semester examination. The mean of the test was 120 and the standard deviation was 17. What was the passing score?

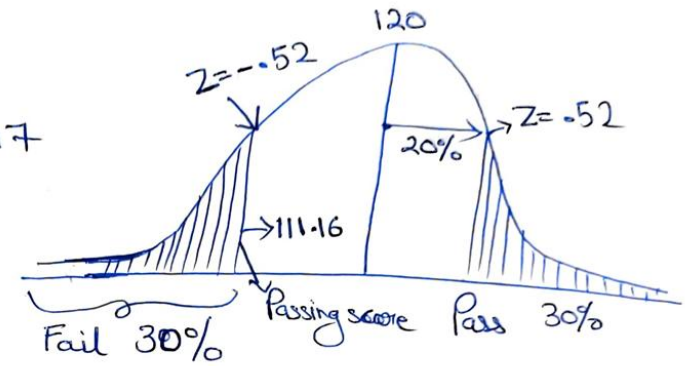
Solution:

Normal distribution

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

Standard deviation, $\sigma = 17$

Test, Mean, $\mu = 120$



Z-table of 20% is .52

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

$$X = Z\cdot\sigma + \mu$$

$$= (-0.52) \times (17) + 120$$

$$= 111.16$$

∴ The passing score for the test = 111.16 .

8. The manufacturer of a certain make of LED bulb claims that his bulbs have a mean life of 20 months. A random sample of 7 such bulbs gave the following values: Life of bulbs – 19, 21, 25, 16, 17, 14, 21.
Can the producer's claim be regard as valid at 1% level of significance?

Solution:

Given data,

Bulb have a mean life, $(\mu) = 20$ month.

Life of Bulb (in month) = 19, 21, 25, 16, 17, 14, 21

Level of significance = 1%

null hypothesis, $H_0 : \mu = \bar{x}$

median hypothesis, $H_a : \mu \neq \bar{x}$

Here, $\bar{x}$ is standard mean.

Test Significant,

$$t = \frac{|\bar{x} - \mu|}{s} \times \sqrt{n} \; ; \text{ Standard deviation, } s = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Calculation of $\bar{x}$ and s

x	$x - \bar{x}$	$(x-\bar{x})^2$
19	0	0
21	2	4
25	6	36
16	-3	9
17	-2	4
14	-5	25
21	2	4
$\sum = 133$		$\sum = 82$

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{133}{7} = 19$$

Worked Examples

Standard deviation,

$$S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{82}{7-1}}$$

$$= \sqrt{\frac{82}{6}} = \sqrt{13.67} = 3.7$$

test Significant, $t = \frac{|19-20|}{3.7} \times \sqrt{7}$

$$= \frac{1}{3.7} \times 2.65$$

$$t = 0.716$$

Degree of freedom, $\nu = n-1$

$$= 7-1$$

$$= 6$$

$t_{0.001}$ (tabulated value)

$t_{0.001} = 3.707$

H_0 is passed and accepted. There is no difference between the sample mean and population mean life of bulb.

Claim of the producer is correct.

9. Compute the effective coefficient of correlation and equation of lines of regression for the data given below:

X	1	2	3	4	5	6	7
Y	9	8	10	12	11	13	14

Solution:

Formula of co-efficient of co-relation :-

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

r = co-relation co-efficient.

x_i = Values of the x-variable in a sample.

$\bar{x}$ = mean of the value of the x-variable

y_i = Value of the y-variable in a sample.

$\bar{y}$ = mean of the value of the y-variable.

S.No	x	y	x^2	y^2	x·y
1	1	9	1	81	9
2	2	8	4	64	16
3	3	10	9	100	30
4	4	12	16	144	48
5	5	11	25	121	55
6	6	13	36	169	78
7	7	14	42	196	98
	$\Sigma x = 28$	$\Sigma y = 77$	$\Sigma x^2 = 133$	$\Sigma y^2 = 875$	$\Sigma xy = 334$

$$r = \frac{7(334) - (28 \times 77)}{\sqrt{[7(133) - (28)^2] \times [7(875) - (77)^2}}$$

$r = 0.9285$

Linear of regression :-

$\bar{x} = \frac{28}{7} = 4$, $\bar{y} = \frac{77}{7} = 11$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7(334) - 28 \times 77}{7(133) - (28)^2}$$

$$= \frac{2338 - 2156}{931 - 784} = \frac{182}{147} = \frac{26}{21}$$

Regression line y on x

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$(y - 11) = \frac{26}{21}(x - 4)$$

$$\Rightarrow 26x - 21y = 127 \#$$

Again

$$b_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma y^2 - (\Sigma y)^2} = \frac{7(334) - 28 \times 77}{7(875) - (77)^2}$$

$$= \frac{2338 - 2156}{6125 - 5929} = \frac{182}{196} = \frac{13}{14}$$

Regression line x on y

$$(x - \bar{x}) = b_{xy}(y - \bar{y})$$

$$= (x - 4) = \frac{13}{14}(y - 11)$$

$$\Rightarrow 14x - 13y = 23$$

Worked Examples

10. In a manufacturing factory, machines A, B, C are produce 25%, 35% and 40% bolts respectively. Out of total 5%, 4%, 2% are defective bolts. A bolt is drawn at random from product. If bolt drawn is found to be defective, what is the probability that it has been produced by B?

Solution:

E_1 = Bolt by A, $P(E_1) = \frac{25}{100}$

E_2 = Bolt by B, $P(E_2) = \frac{35}{100}$

E_3 = Bolt by C, $P(E_3) = \frac{40}{100}$

Defective bolts =

$P(\text{def. from A}) = \frac{5}{100} = P\left(\frac{A}{E_1}\right)$ → defective

$P(\text{def. from B}) = \frac{4}{100} = P\left(\frac{A}{E_2}\right)$

$P(\text{def. from C}) = \frac{2}{100} = P\left(\frac{A}{E_3}\right)$

∴ We have to find defective bolt taken from B,

So, we need $P\left(\frac{E_2}{A}\right)$

By, Bayes Theorem, $P\left(\frac{E_2}{A}\right) = \frac{P(E_2) \cdot P\left(\frac{A}{E_2}\right)}{P(E_1)P\left(\frac{A}{E_1}\right) + P(E_2)P\left(\frac{A}{E_2}\right) + P(E_3)P\left(\frac{A}{E_3}\right)}$

$= \frac{\frac{35}{100} \times \frac{4}{100}}{\frac{25}{100} \times \frac{5}{100} + \frac{35}{100} \times \frac{4}{100} + \frac{40}{100} \times \frac{2}{100}}$

$= \frac{140}{345}$

11. Determine the mean, median, mode, range, sample standard deviation and population standard deviation for the following data set.

Hour	Sample 1	Sample 2	Sample 3	Sample 4
1	10	9	25	20
2	7	9	10	20
3	8	12	15	15
4	7	17	18	15
5	5	20	12	14
6	15	15	15	8
7	20	15	10	9
8	14	7	12	16
9	8	8	10	13
10	10	5	9	11
Mean	10.4	11.7	13.6	14.1

Solution:

Hour	Sample 1	Sample 2	Sample 3	Sample 4
1	10	9	25	20
2	7	9	10	20
3	8	12	15	15
4	7	17	18	15
5	5	20	12	14
6	15	15	15	8
7	20	15	10	9
8	14	7	12	16
9	8	8	10	13
10	10	5	9	11
Mean	10.4	11.7	13.6	14.1

Total mean of the dataset = $\frac{10.4 + 11.7 + 13.6 + 14.1}{4}$ = 12.45

Median of sample 1 = 10

Median of sample 2 = 17.5

Median of sample 3 = 13.5

Median of sample 4 = 11

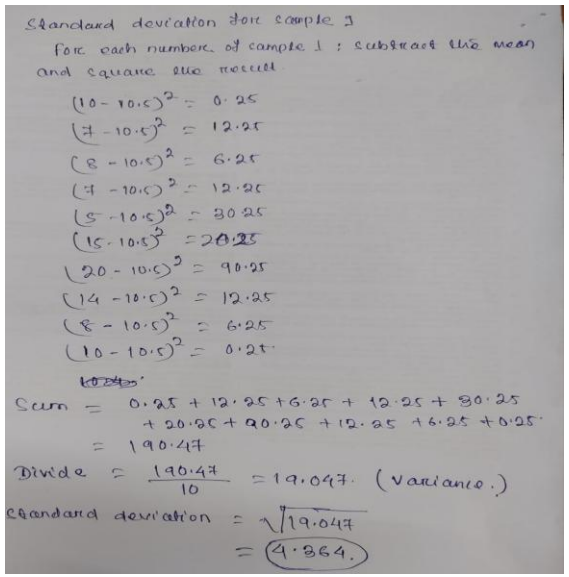
Mode of sample 1 = 7,8,10

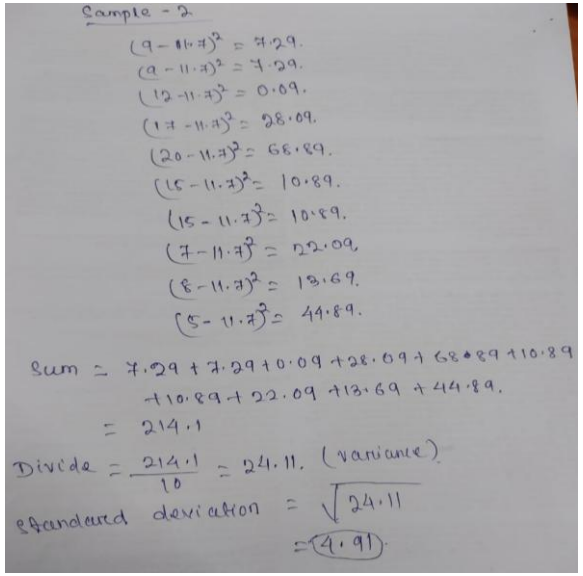
Mode of sample 2 = 9,15

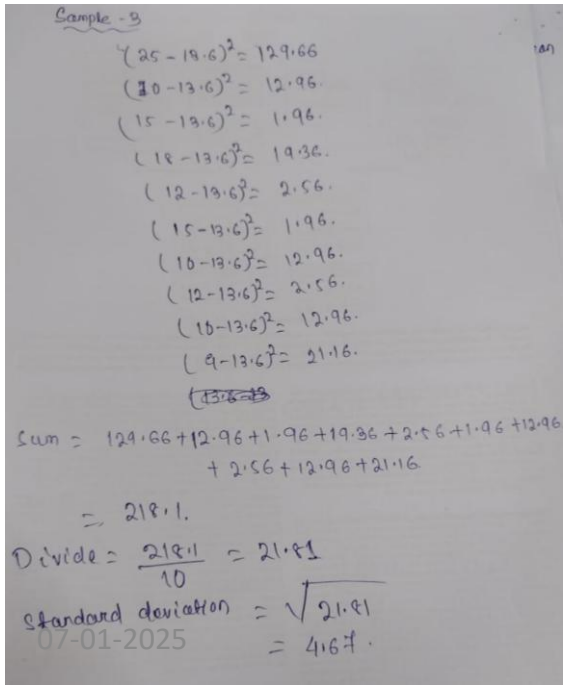
Mode of sample 3 = 15,10,12

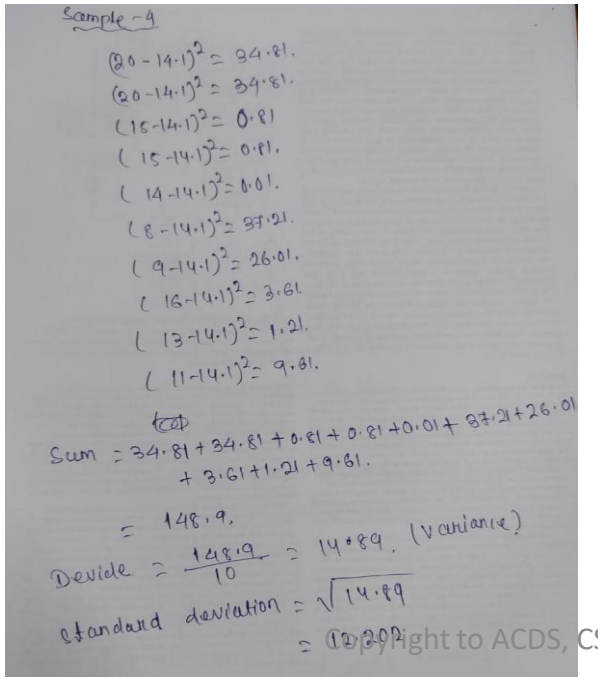
Mode of sample 4 = 15,20

Contd...









Standard deviation of sample 1 = 4.364

Standard deviation of sample 2 = 4.91

Standard deviation of sample 3 = 4.67

Standard deviation of sample 4 = 12.202

12. In a group of 40 people, 10 are healthy and every person of the remaining 30 has either high blood pressure, a high level of cholesterol or both. If 15 have high blood pressure and 25 have high level of cholesterol,

a) How many people have high blood pressure and a high cholesterol?

If a person is selected randomly from this group, what is the probability that he/she

b) Has high blood pressure?

c) Has high level of cholesterol?

d) Has high blood pressure and high level of cholesterol?

Solution:

a) Let x be the number of people with both high blood pressure and high level of cholesterol.

Hence (15 -x) will be the number of people with high blood pressure only and (25 - x) will be the number of people with high level of cholesterol only. We now express the fact that the total number of people with high blood pressure only, with high level of cholesterol only and with both is equal to 30.

(15 - x) + (25 - x) + x = 30

Solve for x: x = 10

b) 15 have high blood pressure, hence P(A) = 15/40= 0.375

c) 25 have high level of cholesterol, hence P(B) = 25/40= 0.625

d) 10 have both, hence P(A and B) = 10/40= 0.25

Exercises

1. An agricultural research organisation tested a particular chemical fertilizer to try to find out whether an increase in the amount of fertilizer used would lead to corresponding increase in the food supply.

Fertilizer	2	1	3	2	4	5	3
Yield of Beans	4	3	4	3	6	5	5

2. A card player is dealt a 13 card hand from a well-shuffled, standard deck of cards. What is the probability that the hand is void in at least one suit ("void in a suit" means having no cards of that suit)?

3. It is claimed that 90% of men cannot tell the difference between two different brands of Cheddar cheese, but of the members of a random sample of 500 men, 72 could distinguish between them. Is the claim justified?

4. Two unbiased dice are thrown simultaneously, and the sum of the scores on their uppermost faces is recorded. What is the distribution of this quantity, and what are its mean and variance?

5. Two coins are in a hat. The coins look alike, but one coin is fair, while other coin is biased, with probability ¼ of heads. One of the coins is randomly pulled from the hat, without knowing which of the two it is, Call the chosen coin "Coin c"

a) Coin c is tossed twice, showing heads both time. Given this information, what is the probability that Coin c is a fair coin?

b) Are the events "first toss of coin c is heads" and "second toss of coin c is heads" independent? Explain

c) Find the probability that in 10 flips of coin C, there will be exactly B heads.

6. Four cards are dealt from a standard pack of 52 cards. Find

(i) the probability that all four are spades;

(ii) the probability that two or fewer are spades;

(iii) the probability that all four are spades, given that the first two are spades;

(iv) the probability that spades and hearts alternate.

7. An urn contains three red and five white balls. A ball is drawn at random, its colour is noted, and it is replaced along with another ball of the same colour. This process is repeated until three balls have been drawn. Find the mean and standard deviation of the number of red balls drawn.

8. The following table shows the number of candidates who scored 0, 1, . . . , 10 marks for a particular question in an examination.

Mark	0	1	2	3	4	5	6	7	8	9	10
No.of Candidates	8	10	49	112	98	86	54	37	28	12	6

Calculate the mean, median and mode of the distribution of marks. What feature of the distribution is suggested by the fact that the mean is greater than the median?

9. A seed manufacturer claims that in a particular variety sold by him there will be one white flower for every three pink flowers. You buy a packet and plant the contents, obtaining 21 pink and 3 white flowers. Do you accept the manufacturer's claim?

10. The stock of a warehouse consists of boxes of high, medium and low quality light bulbs in respective proportions 1 : 2 : 2. The probabilities of bulbs of the three types being unsatisfactory are 0.0, 0.1 and 0.2 respectively. IF a box is chosen at random and two bulbs in it are tested and found to be satisfactory, what is the probability that it contains bulbs

(i) of high quality; (ii) of medium quality; (iii) of low quality?

11. Three quarters of the members of a sports club are adults, and one quarter are children. Three quarters of the adults, and three fifths of the children, are male. Half the adult males, and a third of the adult females, use the swimming pool at the club; the corresponding proportion for children of either sex is four fifths.

(a) Find the probability that a member of the club uses the swimming pool.

(b) Find the probability that a member of the club who uses the swimming pool is male.

(c) Find the probability that a member of the club is female.

(d) Find the probability that a member of the club who uses the swimming pool is female.

(e) Find the probability that a male user of the swimming pool is a child.

(f) Find the probability that a member of the club who does not use the swimming pool is either female or an adult.

12. Each Sunday a fisherman visits one of three possible locations near his home: he goes to the sea with probability +, to a river with probability a, or to a lake with probability i. If he goes to the sea there is an 80% chance that he will catch fish; corresponding figures for the river and the lake are 40% and 60% respectively.
(a) Find the probability that, on a given Sunday, he catches fish.
(b) Find the probability that he catches fish on at least two of three consecutive Sundays.
(c) If, on a particular Sunday, he comes home without catching anything, where is it most likely that he has been?
(d) His friend, who also goes fishing every Sunday, chooses among the three locations with equal probabilities.

Find the probability that the two fishermen will meet at least once in the next two weekends. (Any assumptions you make in solving this problem should be clearly stated.)

13. Blood glucose levels for obese patients have a mean of 100 with a standard deviation of 15. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample of 30 patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.

14. (a) On a particular day a random sample of 12 tins of peas is taken from the output of a canning factory, and their contents are weighed. The mean and standard deviation of weight for the sample are 301.8gm and 1.8gm respectively. Find 99% confidence limits for the mean weight of peas in tins produced by the factory on the day in question.

(b) On the following day a further random sample of 12 tins is taken, and the mean and standard deviation of contents for this sample are 302.1 gm and 1.6 gm respectively. Assuming that the variances of the weights are the same on the two days, show that a 95% confidence interval for the difference between mean weights on the two days includes zero.

(c) Assume now that the samples on both days are from the same population. Find a 99% confidence interval for the mean weight of tins in that population, based on both samples.

15. Measurements of I.Q. were made for a random sample of 200 grammar school children, with the results given below. Test whether a normal distribution gives a satisfactory fit to the data.

IQ	No of Children	IQ	No of Children
80-84	1	125-129	8
85-89	3	130-134	2
90-94	16	135-139	2
95-99	33	140-144	1
100-104	44	145-149	2
105-109	31	150-154	0
110-1 14	26	155-159	2
115-1 19	20	160-164	1
120-124	8		

References

- Deisenroth, M.P., Faisal, A.A., and Ong, C.S. (2020). *Mathematics for Machine Learning*. Cambridge University Press.
- Walpole, R.E., Myers R.H., Myers, S.L., and Ye, K. (2016). *Probability & Statistics for Engineers & Scientists*. Pearson.
- Soong, T.T. (2004). *Probability and Statistics for Engineers*. John Wiley & Sons, Ltd.
- *Montgomery, D.C., Runger J.C. (2005). Applied Statistics and Probability for Engineers*. Wiley.
- *Devore, J. (2012). Probability and Statistics for Engineering and the Sciences*. Cengage Learning.

3.5.2 Video Lectures

- Statistics 110: Probability | Harvard University

https://www.youtube.com/watch?v=KbB0FjPg0mw&list=PL2SOU6wwxB0uwwH80KTQ6ht66KWxbzTIo

- MIT 18.650 Statistics for Applications, Fall 2016 | MIT

https://www.youtube.com/watch?v=VPZD_aij8H0&list=PLUl4u3cNGP60uVBMaoNERc6knT_MgPKS0

Thank You