Year& Semester : **II B.Tech., I SEM**

Faculty Name : **P MAHENDRA VARMA**

Program Name : **B.Tech- Computer Science and Engineering**

Name of the Course : **COMPUTER ORIENTED STATISTICAL METHODS PPT**

Course Code : COSM(MA2103BS)

Unit 1: Probability

- Events and Event spaces
- Random variables
- Joint probability distributions
  - Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
- Mean and Variance

Sample space and Events

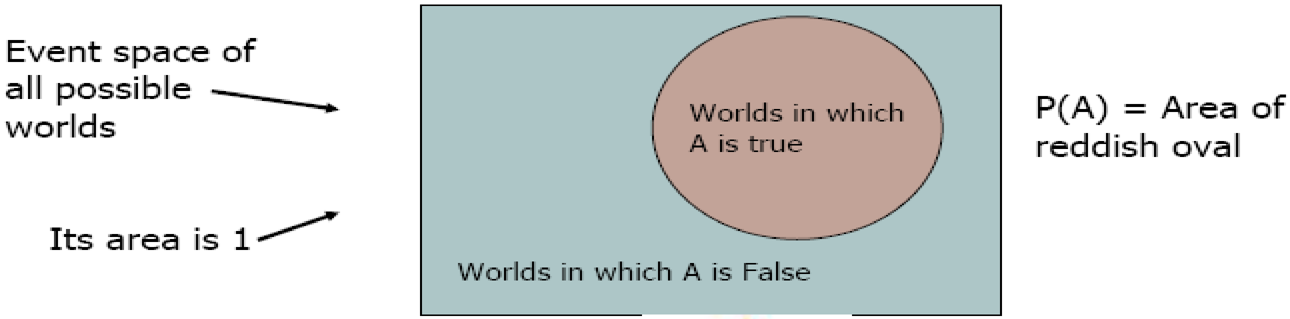
- Ω : Sample Space, result of an experiment
  - If you toss a coin twice $\Omega = \{HH,HT,TH,TT\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, etc.
  - Contains the empty event and Ω

Probability Measure

- Defined over (Ω, S) s.t.
  - $P(\alpha) >= 0$ for all α in S
  - $P(\Omega) = 1$
  - If α, β are disjoint, then
    - $P(\alpha \cup \beta) = p(\alpha) + p(\beta)$
- We can deduce other axioms from the above ones
  - Ex: $P(\alpha \cup \beta)$ for non-disjoint event

  $P(\alpha \cup \beta) = p(\alpha) + p(\beta) - p(\alpha \cap \beta)$

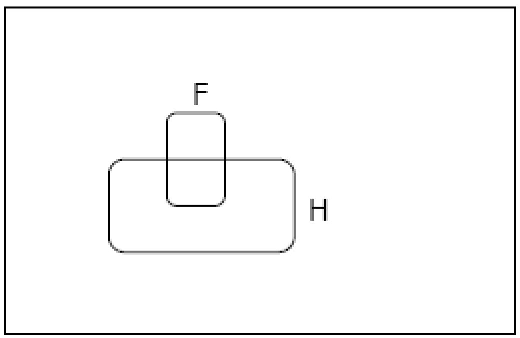
Visualization



- We can go on and define conditional probability, using the above visualization

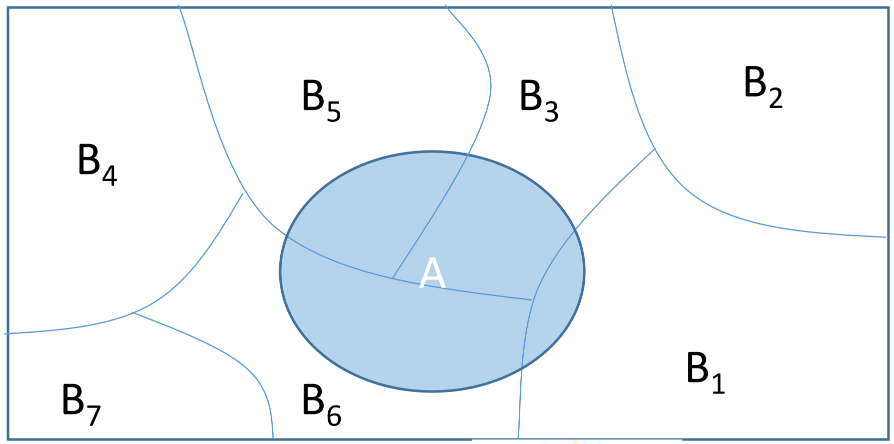
Conditional Probability

P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability

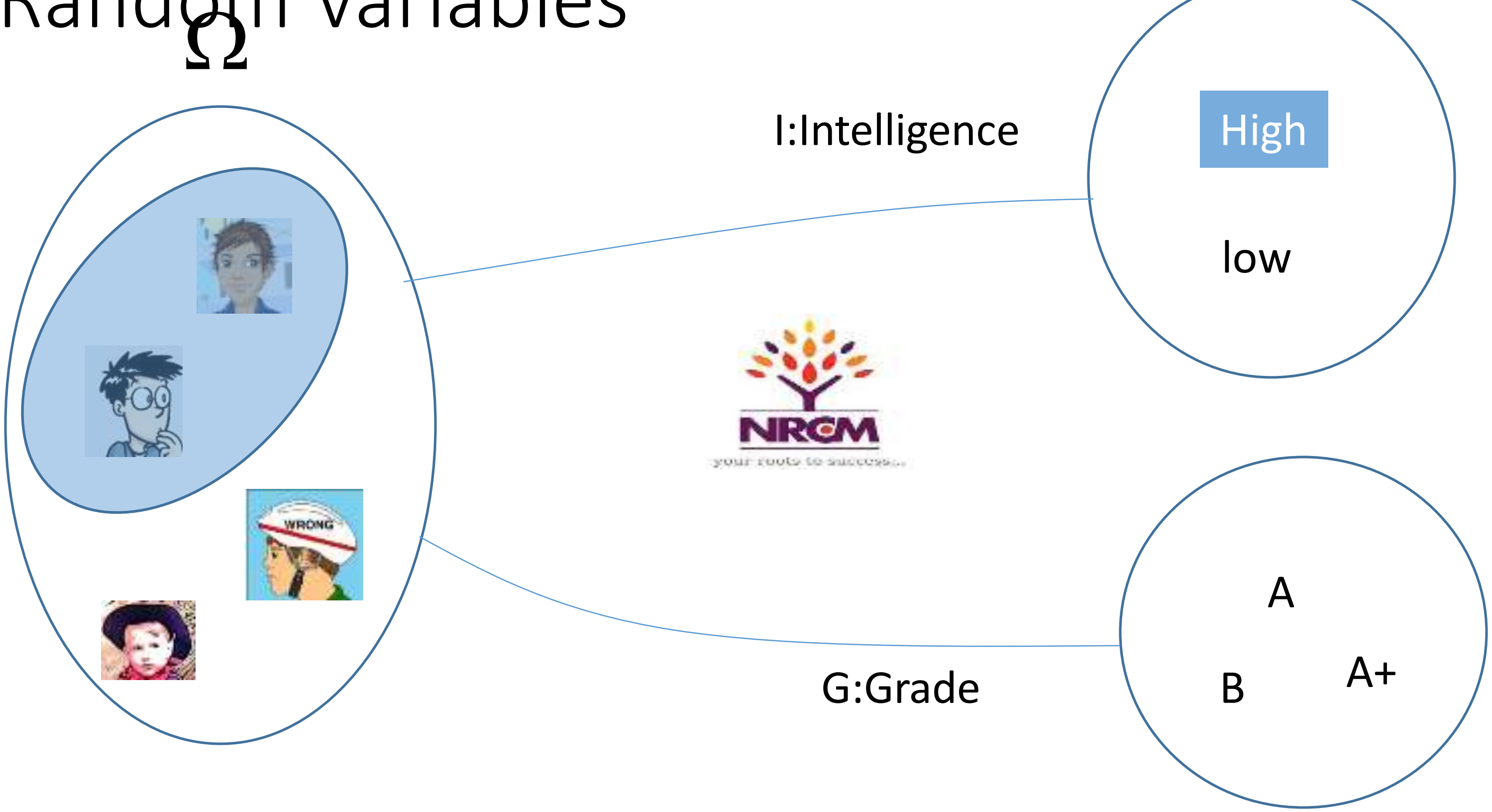


$$p(A) = \sum P(B_i) P(A \mid B_i)$$

Random Variable

- Almost all the semester we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\Omega =$ all possible students
  - What are events
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need "functions" that maps from Ω to an attribute space.
  - $P(G = A) = P(\{student \in \Omega : G(student) = A\})$

Random Variables



P(I = high) = P({all students whose intelligence is high})

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2, …, 100

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

Unit 2:
Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x.

NRCM
your roots to success...

Probability of Continuous RV

- Properties of pdf
  - $$f(x) \geq 0, \forall x$$
  - $$\int_{-\infty}^{+\infty} f(x) = 1$$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

$$P(0 \leq X \leq 1) = \int_{0}^{1} f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \le v)$
- Discrete RVs
  - $F_X(v) = \Sigma_{vi} P(X = v_i)$
- Continuous RVs
  -
  - $$F_X(v) = \int_{-\infty}^{v} f(x)dx$$

$$\frac{d}{dx} F_x(x) = f(x)$$

Joint Probability Distribution

- Random variables encodes attributes
- Not all possible combination of attributes are equally likely
  - Joint probability distributions quantify this
- P(X= x, Y= y) = P(x, y)
  - Generalizes to N-RVs
  - $\sum_x \sum_y P(X = x, Y = y) = 1$
  - $\int_x \int_y f_{X,Y}(x, y)\,dx\,dy = 1$

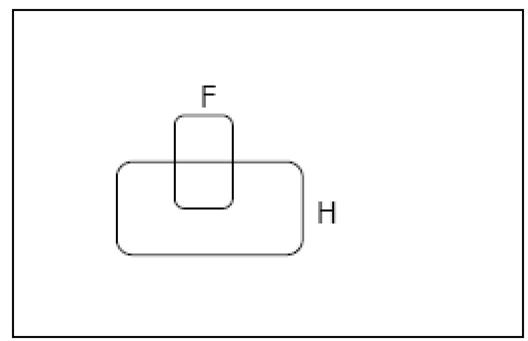
Conditional Probability

events

$$P\left(X = x \middle| Y = y\right) = \frac{P\left(X = x \cap Y = y\right)}{P\left(Y = y\right)}$$

But we will always write it this way:

$$P(x \mid y) = \frac{p(x, y)}{p(y)}$$



Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z)P(y \mid x, z)}{P(y \mid z)}$$

Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - $P(X|Y=y) = P(X)$
  - $P(X=x, Y=y) = P(X=x)\ P(Y=y)$
  - The above should hold for all x, y
  - It is symmetric and written as $X \perp Y$

Independence

- $X_1, \ldots, X_n$ are independent if and only if

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P\left(X_i \in A_i\right)$$

- If $X_1, \ldots, X_n$ are independent and identically distributed we say they are *iid* (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P$$

CI: Conditional Independence

- RV are rarely independent but we can still leverage local structural properties like Conditional Independence.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
  - P(rain $\perp$ sprinkler's on | cloudy)
  - P(rain $\not\perp$ sprinkler's on | wet grass)

NRCM
your roots to success...

Conditional Independence

- $P(X=x \mid Z=z, Y=y) = P(X=x \mid Z=z)$
- $P(Y=y \mid Z=z, X=x) = P(Y=y \mid Z=z)$
- $P(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z)\, P(Y=y \mid Z=z)$

Ve call these factors : very useful concept !!

Mean and Variance

- Mean (Expectation): $$\mu = E(\mathrm{X})$$
  - Discrete RVs: $$E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$$

$$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs:

$$E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x) dx$$

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Mean and Variance

- Variance:

$$Var(X) = E((X-\mu)^2)$$

$$Var(X) = E(X^2) - \mu^2$$

  - Discrete RVs:
  - Continuous RVs:

$$V(X) = \sum_{v_i} (v_i - \mu)^2 P(X = v_i)$$

- Covariance:

$$V(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx$$

- Covariance:

$$Cov(X,Y) = E((X-\mu_x)(Y-\mu_y)) = E(XY) - \mu_x\mu_y$$

Properties

- Mean
  - $$E\left(\mathrm{X}+\mathrm{Y}\right)=E\left(\mathrm{X}\right)+E\left(\mathrm{Y}\right)$$
  - $$E\left(a\mathrm{X}\right)=aE\left(\mathrm{X}\right)$$
  - If X and Y are independent,

$$E\left(\mathrm{XY}\right)=E\left(\mathrm{X}\right)\cdot E\left(\mathrm{Y}\right)$$

- Variance
  - $$V\left(a\mathrm{X}+b\right)=a^{2}V\left(\mathrm{X}\right)$$
  - If X and Y are independent,

$$V\left(\mathrm{X}+\mathrm{Y}\right)=V(\mathrm{X})+V(\mathrm{Y})$$

Some more properties

- The conditional expectation of Y given X when the value of X = x is:

- The Law of Total Expectation or Law of Iterated Expectation:

$$E(Y \mid X = x) = \int y * p(y \mid x) dy$$

$$E(Y) = E[E(Y \mid X)] = \int E(Y \mid X = x) p_X(x) dx$$

Some more properties

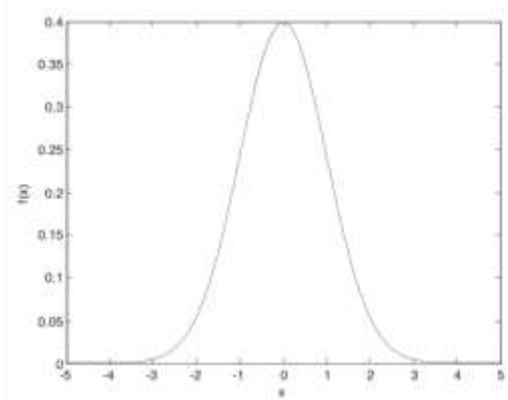
- The law of Total Variance:

$$Var(Y) = Var\left[E(Y \mid X)\right] + E\left[Var(Y \mid X)\right]$$

Unit 3 : Distributions

- Normal distribution: X is continuous R,V
- $N(\mu, \sigma^2)$
  - E.g. the height of the entire population

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



Common Distributions

- Uniform X $U[1, \ldots, N]$
  - X takes values 1, 2, … N
  - $P(X = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial X $Bin(n, p)$
  - X takes values 0, 1, …, n
  -
  - E.g. coin flips

$$p(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$$

UNIT 4
Test of hypothesis

- Given observations from a model
  - What (conditional) independence assumptions hold?
    - Structure learning
  - If you know the family of the model (ex, multinomial), What are the value of the parameters: MLE, Bayesian 1.
    - Parameter learning

Chi Square Test for Independence (Example)

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- State the **hypotheses**

H_0: Gender and voting preferences are independent.
H_a: Gender and voting preferences are not independent

- Choose **significance level**

Say, 0.05

Chi Square Test for Independence

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- Analyze sample data
  - **Degrees of freedom** =
    $|g|-1 * |v|-1 = (2-1) * (3-1) = 2$
  - **Expected frequency count** =
    $E_{g,v} = (n_g * n_v) / n$

    $E_{m,r} = (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,d} = (400 * 450) / 1000 = 180000/1000 = 180$
    $E_{m,i} = (400 * 100) / 1000 = 40000/1000 = 40$
    $E_{f,r} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,d} = (600 * 450) / 1000 = 270000/1000 = 270$
    $E_{f,i} = (600 * 100) / 1000 = 60000/1000 = 60$

Chi Square Test for Independence

	Republican	Democrat	Independent	Total
Male	200	150	50	400
Female	250	300	50	600
Total	450	450	100	1000

- Chi-square **test statistic**

$$X^2 = \left[\sum \frac{(O_{g,v} - E_{g,v})^2}{E_{g,v}} \right]$$

- $X^2 = (200 - 180)^2/180 + (150 - 180)^2/180 + (50 - 40)^2/40 + (250 - 270)^2/270 + (300 - 270)^2/270 + (50 - 60)^2/40$
- $X^2 = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60$
- $X^2 = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2$

Chi Square Test for Independence

- **P-value**
  - Probability of observing a sample statistic as extreme as the test statistic
  - $P(X^2 \geq 16.2) = 0.0003$
- Since **P-value** (0.0003) is less than the significance level (0.05), we cannot accept the null hypothesis
- There is a relationship between gender and voting preference

UNIT 5:

CORRELATION AND REGRESSION

Definition Correlation is a statistical tool which studies the relationship b/w 2 variables & correlation analysis involves various methods & techniques used for studying & measuring the extent of the relationship b/w them.

Two variables are said to be correlated if the change in one variable results in a corresponding change in the other.

The Types of Correlation

1) Positive and Negative Correlation: If the values of the 2 variables deviate in the same direction

i.e., if the increase in the values of one variable results in a corresponding increase in the values of othervariable (or) if the decrease in the values of one variable results in a corresponding decrease in the values of other variable is called Positive Correlation.

e.g. Heights & weights of the individuals If the increase (decrease) in the values of one variable results in a corresponding decrease (increase) in the values of other variable is called Negative Correlation.

e.g, Price and demand of a commodity.

2) Linear and Non-linear Correlation: The correlation betweentwo variables is said to be Linear if the corresponding to a unit change in one variable there is a constant change in the other variable over the entire range of the values (or) two variables x, y are said to be linearly related if there exists a relationship of the form y = a + bx.

e.g when the amount of output in a factory is doubled by doubling the number of workers.

Two variables are said to be Non-linear or curvilinear if corresponding to a unit change in one variable the other variable doesnot change at a constant rate but at fluctuating rate. i.eCorrelation is said to be non-linear if the ratio of change is not constant. In other words, when all the points on the scatter diagram tend to lie near a smooth curve, the correlation is saidto be non-linear (curvilinear).

1) **Partial and Total correlation:** The study of two variables excluding some other variablesis called Partial Correlation.

e.g. We study price and demand eliminating the supply. In Total correlation all the facts are taken into account.

e.g Price, demand &supply ,all are taken into account.

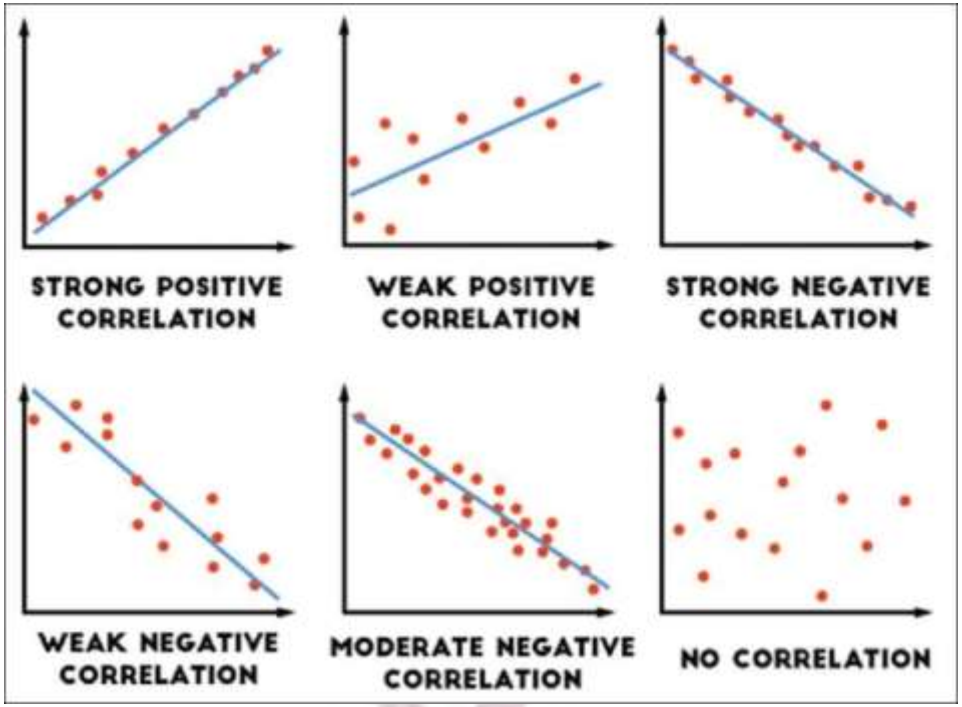
1) **Simple and Multiple correlation:**When we study only two variables, the relationship is described as Simple correlation.

E.g quantity of money and price level, demand and price.

NRCM

your roots to success...

The following are scatter diagrams of Correlation.



Rank Correlation Coefficient

Charles Edward Spearman found out the method of finding the Coefficient of correlation by ranks.

This method is based on rank & is useful in dealing with qualitative characteristics such as morality, character,

intelligence and beauty. Rank correlation is applicable to only to the individual observations.

formula: $\rho = 6 \sum D2$

$N(N2–1)$

where :ρ - Rank Coefficient of correlation

D^2- Sum of the squares of the differences of two ranks

N- Number of paired observations.

Properties

- The value of ρ lies between $+1$ and -1.
- If $\rho = 1$, then there is complete agreement in the order of the ranks & the direction of the rank is same.
- If $\rho = -1$, then there is complete disagreement in the order of the ranks & they are in opposite directions.

Equal or Repeated ranks

$$\text{Formula: } \rho = 1 - 6\left\{\frac{\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \ldots}{N^3 - N}\right\}$$

where m = the number of items whose ranks are common.

N - Number of paired observations.

D^2 - Sum of the squares of the differences

If any 2 or more items are with same value the in that case common ranks are given to repeated items.

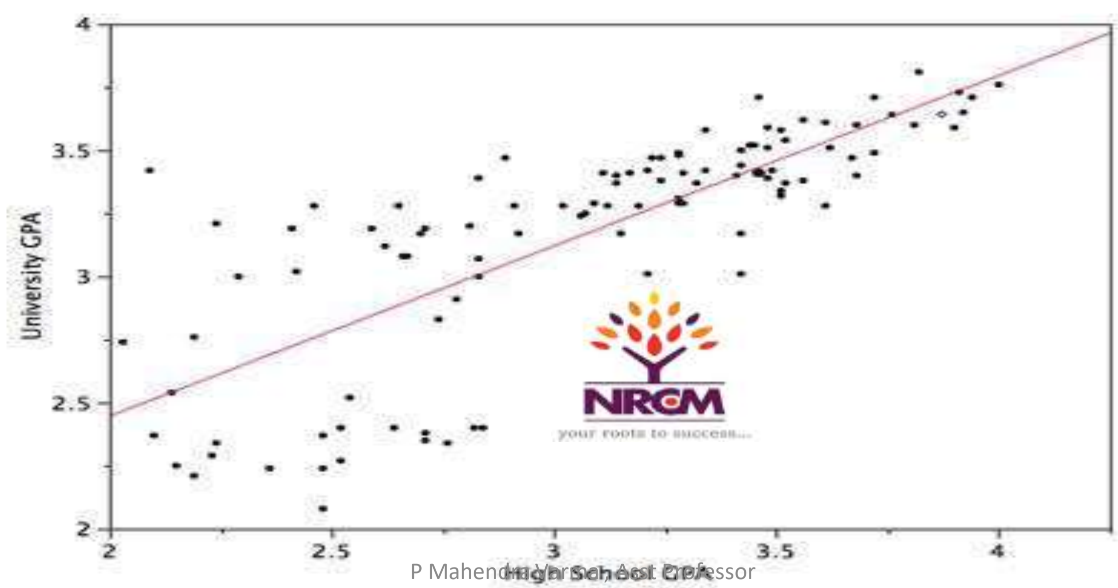
The common rank is the average of the ranks which these items would have assumed,

if they were different from each other and the next item will get the rank next to ranks already assumed.

REGRESSION

In regression we can estimate value of one variable with the value of the other variable which is known. The statistical method which helps us to estimate the unknown value of one variable from the known value of the related variable is called ‘Regression’.

The line described in the average relationship b/w 2 variables is known as <u>Line of Regression.</u>



Acknowledgment

your roots to success...

- Carlos Guestrin recitation slides: http://www.cs.cmu.edu/~guestrin/Class/10708/recitations/r1/Probability_and_Statistics_Review.ppt
- Andrew Moore Tutorial: http://www.autonlab.org/tutorials/prob.html
- Monty hall problem: http://en.wikipedia.org/wiki/Monty_Hall_problem
- http://www.cs.cmu.edu/~guestrin/Class/10701-F07/recitation_schedule.html
- Chi-square test for independence

  http://stattrek.com/chi-square-test/independence.aspx