STATISTICS

Outline

Probability

Sample Space:

A set consisting of all possible outcomes that can results from a random experiments (real or conceptual), is defined to be a sample space for the experiment and is ~~defined~~ denoted by the letter S- Each possible outcome is a member of the sample space and is called a sample point in that space-

Example:

Experiment of tossing a coin results in either of the two possible outcomes : a head (H) or a tail (T); landing on its edge or rolling away is not considered- The sample space for this experiment is set notation as S = {H, T}-

Scanned with CamScanner

Scanned with CamScanner

Events =

An event is an individual outcome or any number of outcomes (sample points) of a random experiment or a trail. In set termimology, any subset of a sample space S of the experiment, is called an event. An event that contains exactly one sample point, is defined a sample event. A compound event contains more than one sample point and is produced by the union of sample events.

Counting Sample Events:

When the number of sample points in a sample space S is very large, it becomes very inconvinient and difficult to list them all and to count the number of points in the sample space S and in subsets of S. We then need some rules which helps us to count number of all sample points without actually listing them.

Scanned with CamScanner

Scanned with CamScanner

⇒ Rules / Methords :-

i) : Rule of Multiplication :-

If a compound experiment consists of two experiments such that the first experiment has exactly n distinct outcomes and, if corresponding to each outcome of the first experiment there can be n distinct outcomes of the second experiment, than the compound experiment has exactly mn outcomes.

Example :-

The compound experiment of tossing a coin and throwing a die together consist of two experiments - The total number of possible outcomes are $2 \times 6 = 12$.

ii) : Rule of Permutation :-

A permutation is any ordered subset from a set of n distinct objects - The number of permutation of r objects, selected in a definite order from n distinct objects is denoted by the symbol ${}^nP_r$.

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

Scanned with CamScanner

Scanned with CamScanner

iii)- Rules of Combination:

A combination is any subset of r objects, selected without regard to their order, from a set of n distinct objects. The total number of such combinations is denoted by the symbol ${}^nC_r$ or $\binom{n}{r}$

where $r \leq n$.

$$\binom{n}{r} = \frac{{}^nP_r}{r!}$$

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

Probility of an Event:

If an event A is defined in a sample space S, then its probility will be equal to the sum of the probilities of all sample points that are included in A,

ie; $P(A) = \Sigma P(E_i)$. When all n possible outcomes of a random experiment are eavually like to occur, then =

$$P(A) = \frac{m}{n} = \frac{\text{Number of sample points in A}}{\text{Number of sample points in S}} = \frac{n(A)}{n(S)}$$

$\Rightarrow \frac{1}{n}$ is the probility of each outcome

Scanned with CamScanner

Scanned with CamScanner

and its remains same-

Additive Law:

If A and B are any two events defined in a sample space S, then-

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

Proof:

Two events $A \cup B$ may be written as the union of two mutualy exclusive events A and $B \cap \bar{A}$. That is-

$$A \cup B = A \cup (B \cap \bar{A})$$

Then

$$P(A \cup B) = P(A) + P(B \cap \bar{A}) \quad \text{---- (i)}$$

Again the event B may also be decomposed into two mutualy exclusive events as-

$$B = (A \cap B) \cup (\bar{A} \cap B)$$

$$P(B) = P(A \cap B) + P(\bar{A} \cap B) \quad \text{---- (ii)}$$

By subtratting (i) and (ii)-

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

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

Hence proved-

Scanned with CamScanner

Scanned with CamScanner

Conditional Probility:

It is the probility of one event occuring with some relationship to one or more other events.

Example:

- Event A is that it is raining outside, and it has a 0.3 (30%) chances of raining today.
- Event B is that you will need to go outside, and that has a probility of 0.5 (50%).

A conditional probility would look at these two events in relationship with one another, such as the probility that it is both raining and you will need to go outside -

Formula:

$$P(A/B) = \frac{\text{number of sample points in } A \cap B}{\text{number of sample points in } B}$$

$$= \frac{n(A \cap B)}{n(B)}$$

denominator

Scanned with CamScanner

Scanned with CamScanner

we get

$$P(A/B) = \frac{n(A \cap B)}{n(S)} \cdot \frac{n(S)}{n(B)} = \frac{P(A \cap B)}{P(B)}$$

So, if $P(B) = 0$, the conditional probility $P(A/B)$ remains undefined.

Similarly,

$$P(B/A) = \frac{P(A \cap B)}{P(A)} \quad \text{where } P(A) > 0.$$

Baye's Theoram=

$$P(Ai/B) = \frac{P(Ai)\,P(B/Ai)}{\sum_{i=1}^{k} P(Ai)\,P(B/Ai)}$$

for $i = 1, 2, 3, \ldots k$.

Proof=

By multiplicative law of probilities, we have=

$$P(B \cap Ai) = P(B)\,P(Ai/B)$$

$$P(B \cap Ai) = P(Ai)\,P(B/Ai)$$

Equating the equivalent relations of $P(B \cap Ai)$ and

Scanned with CamScanner

Scanned with CamScanner

Solving for $P(A_i/B)$ we get

$$P(B)P(A_i/B) = P(A_i)P(B/A_i)$$

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(B)}$$

We may write B as $B = S \cap B =$

$$B = (A_1 \cup A_2 \cup A_3 \cup \dots \cup A_k) \cap B$$

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup (A_3 \cap B) \cup \dots \cup (A_k \cap B)$$

NOW $P(B) =$

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \dots + P(A_k \cap B)$$

$$P(B) = \sum_{i=1}^{k} P(A_i \cap B)$$

Using the multiplicative law of probility, we may express each term $P(A_i \cap B)$ as $P(A_i)P(B/A_i)$. Then:

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_k)(B/A_k).$$

$$P(B) = \sum_{i=1}^{k} P(A_i)P(B/A_i)$$

So

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{\sum_{i=1}^{k} P(A_i)P(B/A_i)}$$

Hence proved.

Scanned with CamScanner

Scanned with CamScanner

Random Variables

Outline

Concept of Random Variables:

Such a numerical quantity whose value is determined by the outcome of a random experiment is called a random variables.

Mathematically we assign a single value / real number to each outcome of the sample space, and hence we state that a random variable is a real valued function defined on a sample space.

A random variable is also called a chance variable, a stochastic variable or simply a variate and is abbriviated as rv. The random variables are usally denoted by capital Latin letters such as X, Y, Z; while the values taken by them are represented by the corresponding small letters such as x, y, z. There are two types of random variables:

i):- Discrete.

ii):- Continous.

Scanned with CamScanner

Scanned with CamScanner

Discrete Probility Distribution:

A discrete distribution describes the probility of occurence of each value of a discrete random variable- A discrete random variable is a random variable that has a countable values, such as a list of non-negative integers-

With a discrete probility distribution each possible value of the discrete random variables can be assosiated with a non-zero probility- Thus a discrete probility distribution is often presented in tabular form-

Scanned with CamScanner

Scanned with CamScanner

CONTINOUS PROBILITY DISTRIBUTION:

A continous distribution describes the probility of the possible values of a continous random variable. A continous random variable is a random variable with a set of possible values that is infinite and uncountable.

Probilities of a continous random variable (x) are defined as the area under the curve of its PDF. Thus only ranges of values can have a nonzero probility. The probility that a continous random variable equals some value is always zero.

Probility Density Function: A probility density function (p.d.f) has the following properties:

i) $f(x) \geq 0$, for all x.

ii) $\int_{-\infty}^{\infty} f(x)\,dx = 1$

iii) The probility that x takes on a value in the interval $[c, d]$, $c \leq d$ is given by:

$$P(c < x \leq d) = F(d) - F(c)$$

$$= \int_{-\infty}^{d} f(x)\,dx - \int_{-\infty}^{c} f(x)\,dx$$

$$= \int_{c}^{d} f(x)\,dx$$

Which is the area under curve.

Scanned with CamScanner

Scanned with CamScanner

Joint Distributions:

The distribution of two or more random variables which are observed simulataneously when an experiment is performed is called thier joint distribution. It is neccessary to call the distribution of a single r.v as univarative. Likewise, a distribution involving two, three or many r.v.'s simulatanious ly is refered to be a bivariative, trivariate or multivariate.

=> Bivariate Distribution Function:

Let X and Y be two r.v's defined on the same sample space S. Then the function $f(x,y)$ defined by $F(x,y) = P(X \leq x \text{ and } Y \leq y)$ where $F(x,y)$ gives the probility that X will take on a value less than or eqval to x and, at the same time, Y will take on a value less

Scanned with CamScanner

Scanned with CamScanner

than or equal to y is called a bivariate or joint distribution function of x and y.

⇒ Bivariate Probility Function:

Let X and Y be two discrete r.v's defined on the same samplae space S, X taking the values $x_1, x_2, ---, x_n$ and Y takes on the value $y_1, y_2, ---, y_n$. Then the probility that X takes on the values x_i and Y takes on the values y_j, denoted by $f(x_i, y_j)$ of p_{ij} is defined to be Bivariate probility Function or joint probility function and its values at point (x, y) is given by

$$f(x_i, y_j) = (X = x_i \text{ and } Y = y_j)$$

⇒ Conditional Probility Functions:

Let x and Y be two discrete r.v's with joint probility function $f(x, y)$. Then the condition probility for X given $Y = y$, denoted as $f(x|y)$, is defined by:

$$f(x_i|y_j) = P(X = x_i | Y = y_j)$$

Scanned with CamScanner

Scanned with CamScanner

$$= \frac{P(X = x_i \text{ and } Y = y_j)}{P(Y = y_j)}$$

$$= \frac{f(x_i, y_j)}{h(y_j)} \quad \text{for}$$

$i = 1, 2, \ldots$

$j = 1, 2, \ldots$

where $h(y_j)$ is the marginal probility and $h(y_j) > 0$ -

Scanned with CamScanner

Scanned with CamScanner

Continous Probility Distribution:

Outline #6

Continous Uniform Distribution:

The density function of a continous r.v X is called a uniform distribution when between the end points any two subintervals of the same length containing X, have the same probability.

Alternatively, a r.v X is said to be uniformly distributed if its density function is defined as:

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b$$

$$= 0 \quad \text{elsewhere.}$$

The distribution derives its name from the fact that its density is constant or uniform over the interval [a,b] and is 0 elsewhere.

Properties:

i. Let X be the uniform distribution over [a, b]. Then its mean is $a+b/2$ and variance is $(b-a)^2/2$.

Scanned with CamScanner

Scanned with CamScanner

Now $$E[U] = \int_{-\infty}^{\infty} f(u)\,du = \frac{1}{b-a}\int_{a}^{b} u\,du$$

$$= \frac{1}{b-a}\left[\frac{u^2}{2}\right]_a^b$$

$$= \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{a+b}{2} \Rightarrow \text{mid-point}$$

And

$$Var[X] = E[X^2] - [E[U]]^2$$

$$E[X^2] = \int_a^b u^2 \cdot \frac{1}{b-a}\,du$$

$$= \frac{1}{b-a}\left[\frac{u^3}{3}\right]_a^b$$

$$= \frac{b^3-a^3}{3(b-a)}$$

$$= \frac{(a^2+ab+b^2)(b-a)}{3(b-a)}$$

$$= \frac{a^2+ab+b^2}{3}$$

$$\therefore Var[X] = E[X^2] - [E[X]]^2$$

$$= \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4}$$

$$= \frac{(b-a)^2}{12}$$

Scanned with CamScanner

2:- The shape of distribution is rectangular-

NORMAL DISTRIBUTION

The normal distribution is a continous probability distribution that is symmetrical on both sides of the ~~center is 0~~ mean, go the right side of the center is a mirror image of the left side - The area under the curve of normal distribution represents probability and the total area under the curve sums to one-

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

where,

$f(x)$ = probability density function-

σ = Standard deviation-

μ = mean-

Area Under the Normal Curve:

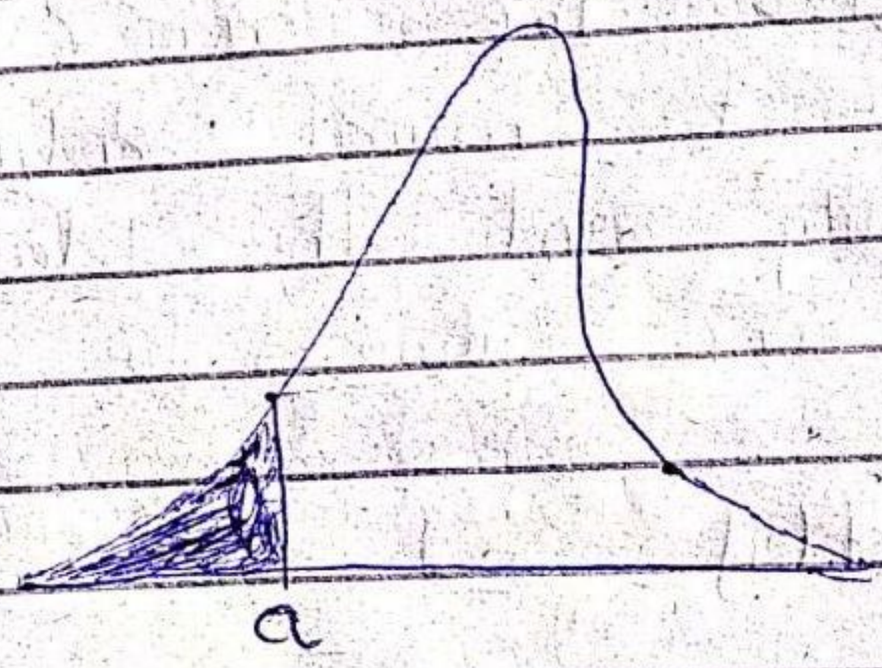
1):- The total area under the normal curve is equal to 1-

2):- The probability that a normal random variable x equals any particular valve is 0-

Scanned with CamScanner

i) The probability that X is greater than a equals the area under the normal curve bounded by a and minus infinity (indicated by non-shaded area).

ii) The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (indicated by shaded area).



Applications of Normal Distribution:

OR

Properties:

The main properties of normal distribution are:

1) The function f(x) defining the normal distribution is a proper pdf ie $f(x) \geq 0$

Scanned with CamScanner

and the total area under the normal curve is unity.

2J= The mean and variance of the normal distribution are μ and σ^2.

3J= The median and mode of the normal distribution are each equal to μ, the mean of distribution.

4J= The mean deviation of the normal distribution is approximately 4/5 of its standard deviation.

5J= The normal curve has points of inflection which are equidistant from the mean.

6J= For the normal distribution, the odd order moments about the mean are all zero and even order moments are given by =

$$\mu_{2n} = (2n-1)(2n-3) \cdots 5, 3, 1, \sigma^{2n}$$

7J= If X is $N(\mu, \sigma^2)$ and if $Y = a+bX$ than $Y = N(a+b\mu, b^2\sigma^2)$

8J= The sum of independent normal variables is a normal variable.

Scanned with CamScanner

Normal Approximation to the Binomial Distribution:

The binomial distribution $b(x; n, p)$ can be closely approximated by the normal distribution when n is sufficiently large and neither p nor q is close to zero. As a rule of thumb, normal distribution provides a responsible approximation to the binomial distribution if both np and nq are equal to or greater than 5.

Now, the probability for a binomial random variable X to the value x is:

$$f(x) = \binom{n}{x} p^x q^{n-x}$$

For $0 \leq x \leq n$ and $q = 1-p$:

Scanned with CamScanner

Gamma Distribution:

A conlinous r.v X is said to have a gamma distribution with parameter $m > 0$, if its p.d.f is defined by

$$f(x) = \begin{cases} \frac{1}{\Gamma(m)} x^{m-1} \cdot e^{-x} & \text{for } 0 \le x \le \infty \\ 0, & \end{cases}$$

A gamma variable with parameter m is actvally / usally denoted by $\gamma(m)$. A straightforward inlegration shows that $\int_0^\infty f(x)\,dx = 1$ and hence it represents a p.d.f.

The distribution funclion $F(x)$ is

$$F(x) = \int_0^x \frac{1}{\Gamma(m)} x^{m-1} e^{-x}\,dx \qquad x \ge 0$$

$$= 0 \qquad x < 0$$

which is also called incomplete gamma distribution.

⇒ Properties:

1): The mean and variance of gamma distribution is eqval to its parameter m.

as $\mu = m$

and $\sigma^2 = m$

Scanned with CamScanner

3)- The sum of two independent Gamma distributions with parameter m and n is Gamma distribution with parameter (m+n)-

Exponential Distribution:

A random variable X is said to have an exponential distribution with parameter λ, if its p.d.f is defined by:-

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x > 0$$
$$= 0$$

where $\lambda > 0$. The p.d.f may also written as

$$f(x) = \frac{1}{\beta} e^{-x/\beta} \quad \text{for } x > 0.$$
$$= 0$$

The function $f(x)$ is a proper p.d.f so

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\,dx = -\left[-e^{\lambda x}\right]_0^{\infty} = 1$$

The distribution function of the exponential r.v. X is given by:-

$$F(x) = P(X \le x)$$
$$= \int_0^x \lambda e^{-\lambda x}\,dt$$

Scanned with CamScanner

$= \left[-e^{-\lambda t}\right]_0^x$

$= 1 - e^{-\lambda x}$, for $x_i > 0$

$= 0$

and hence $P(X > x) = e^{-\lambda x}$

⇒ Properties:

1)= The mean and standard deviation of the exponential distribution are equal-

2)= The distribution is extremely skewed and thus there does not exisy any mode-

BETA DISTRIBUTION:

A continous r.v, X is said to be bela distribution with two parameters m and n, if its p.d.f is defined by:

$$f(x) = \begin{cases} \frac{1}{B(m,n)} x^{m-1}(1-x)^{n-1}, & 0 \le x \le 1; \\ 0; & m, n > 0 \end{cases}$$

This is called bela distribution of first kind and refered as $\beta_1(m,n)$-

Scanned with CamScanner

⇒ Properties of $\beta_1(m,n)$:-

•]:- The mean and variance of this distribution are $\frac{m}{m+n}$ and $\frac{mn}{(m+n)^2(m+n+1)}$ respectively.

Beta Distribution of Second kind:

A continous r.v X is said to have a beta continous distribution of second kind with the parameters m and n, if its pdf is defined by:-

$$f(x) = \begin{cases} \frac{1}{B(m,n)} \cdot \frac{x^{m-1}}{(1+x)^{m+n}} & 0 \le x \le \infty \\ & m,n > 0 \\ 0 ; & \end{cases}$$

It is generally denoted by $\beta_2(m,n)$:-

Scanned with CamScanner

Chi-Squvare Distribution:

In probability theory and statistics the chi-squvare distribution with k- degree of freedom is the distribution of a sum of the squvares of k independent standard normal r.v - The chi-squvare is a special case of the gamma distribution.

Its parameter is k.

The p.d.f of chi-squvare is:-

$$f(u) = \begin{cases} \frac{1}{2^{k/2}\,\Gamma(k/2)}\, u^{k/2-1} e^{-u/2} & \text{for } 0 \le u < \infty \\ 0; & \text{and } 0 < t < \infty \end{cases}$$

Scanned with CamScanner

Introduction To Statistics AND Data Analysis:

Outline #1

Statistical Interface:

o:- It is the branch of Statistics which is concerned with using probability concept to deal with uncertainly in decision making-

:- It refers to the process of selecting and using a sample to draw inferance about population from which sample is drawn-

Samples:

A sample is a part or a subset of population. Generally it contains of some of the observations but in certain situations, it may include the whole of the population. The number of observations included in a sample is called the size of the sample and is denoted by letter n-

Scanned with CamScanner

Scanned with CamScanner

Populations:-

A population or a statistical population is the collection or set of all possible observations whether finite or infinite, relevent to some charactrestics of intrest. A statistical population may be real such as the hieght of all college students or hypothetical such as the possible outcomes from the toss of a coin.

Sampling Procedures:-

1. Choosing part of a population to use to test hypotheses about the entire population.
2. Used to choose the number of participants, interviews, or work samples to use in the assessment process.

Collection of Data:-

The most important part of statistical work is perhaps the collection of data. There are two types of collection of data.

1. Primary Collections
2. Secondary Collections

Scanned with CamScanner

Scanned with CamScanner

⇒ Collection of Primary Data:

One or more of the following methords are employed to collect primary data-

i)= Direct Personal Investigation
ii)= Indirect Investigation or Personal Interviews-
iii)= Collection through Questionnaries-
iv)= Collection through Enumerators-
v)= Collection through local sources-
vi)= Computer Interviews-

⇒ Collection of Secondary Data:

The secondary data may be obtained from following sources-

i)= Internal secondary data-
ii)= External secondary data-
•)= Offical
•)= Semi-offical

Measure of Location:

The three most common measure of location are the mean, the median and the mode-

Scanned with CamScanner

Scanned with CamScanner

Mean:

The mean of a data set is found by adding all numbers in a data set and then dividing by the numbers of value in a set-

Median:

The median is the middle value when a data set is ordered from least to greatest

Mode:

The mode is a number that occurs most often in a data set-

Scanned with CamScanner

Scanned with CamScanner

Measures of Variability:

The most common measures of variability are-

i: Range:

The range R is defined as the diffrence b/w the largest and the smallest observations in a set of data- Symbolically,

$$R = x_n - x_0$$

ii: Interquatile Range:

The interquatile range is a measure of disperssion, defined by the diffrence between the third and first quartilies- Symbolically:

$$Q.D = \frac{Q_3 - Q_1}{2}$$

iii: Mean Deviation:

The mean deviation of a set of data is defined as the arithmetic mean of the deviations measured either from the mean or from the median, all deviations being counted as positive-

$$M.D = \frac{\Sigma|x_i - \bar{x}|}{n}$$ for sample data,

$$M.D = \frac{\Sigma|x_i - \mu|}{n}$$ For population data

Scanned with CamScanner

Scanned with CamScanner

iv) Variance:-

The variance of the set of observations is defined as the mean of squares of deviations of all the observations from thier mean- Symbolically:-

$$\sigma^2 = \frac{\sum (xi - \mu)^2}{N} \quad \text{for population data}$$

$$S^2 = \frac{\sum (xi - \bar{x})^2}{n} \quad \text{for sample data-}$$

v): Standard Deviation:-

The positive squcare root of the variance is called the Standard deviation- Symbolically-

$$\sigma = \sqrt{\frac{\sum (xi - \mu)^2}{N}}, \quad \text{for population data.}$$

$$S = \sqrt{\frac{\sum (xi - \bar{x})^2}{n}}, \quad \text{for sample data.}$$

DISCRETE AND CONTINOUS DATA:

"Discrete data" is information that can only take certain values-

"Continous data" is data that ccun take any value-

Scanned with CamScanner

Scanned with CamScanner

i)- Define Probability?

Ans- The word probability has two basic meanings: (i) a quantitative measure of uncertainty and (ii) a measure of degree of belief in a particular statement or problem-

ii)- Mutually and non-mutually exclusive events?

=> Mutually exclusive events-

Two events A and B of a single experiment are said to be mutually exclusive or disjoint if and only if they cannot both occur at the same time. That is they have no points in common-

=> Non-mutually exclusive events-

Two events A and B are said to be non-mutually exclusive if both events can occur at same time- eg; if we draw a card from an ordinary deck of 52 playing cards, it can be both a king and a diamond- Therefore,

Scanned with CamScanner

Scanned with CamScanner

kings and diamonds are not mutually exclusive-

=> Independent and Dependent Events=

Independent Events=

Two events A and B in the same sample space S, are defined to be independent, if the probability that one event occurs, is not affected by weather the other event has or has not occured that is;

$$P(A/B) = P(A)$$

and

$$P(B/A) = P(B)$$

It then follows that two events A and B are independent if and only if =

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

Dependent Events =

Two events A and B are defined to be dependent if $P(A \cap B) \neq P(A) \times P(B)$. This means that the occurance of one of the events in some way affects the probability of the occurance of the other events.

Scanned with CamScanner

Scanned with CamScanner

=> Statement of general law of Addition:-

If A and B are any two events defined in a sample space, then

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

This is the statement for general law of addition.

=> Defination of expectation:-

Scanned with CamScanner

Scanned with CamScanner

⇒ Properties of Normal Distribution:-

The main properties of Normal distribution are given below:-

1) :- The function $f(x)$ defining the normal distribution is a proper p.d.f ie, $f(x) \geq 0$ and the total area under the normal curve is unity.

2) :- The mean and variance of the normal distribution are μ and σ^2 respectively.

3) :- The median and the mode of the normal distribution are each equal to μ, the mean of the distribution.

4) :- The mean deviation of the normal distribution is approximately $\frac{4}{5}$ of its standard deviation.

5) :- The normal curve has points of inflection which are equidistant from the mean.

6) :- For the normal distribution, the odd number movements about mean are all zero and the even order movements, are $\Rightarrow \mu_{2n} = (2n-1)(2n-3) \cdots 5, 3, 1$

Scanned with CamScanner

Scanned with CamScanner