Chapter 1

Marks 1

1) b) Discrete Random Variable
2) b) 0, 1, 2, 3, 4
3) d) chance variable
4) b) variable
5) a) Continuous variable
6) b) use of digital computers
7) a) Discrete Random variables
8) c) continuous Random variable
9) b) Random variable
10) b) Probability function.

11) 3men and 2women $= {}^7C_3 \times {}^6C_2 = 35 \times 15 = 525$
4men and 1 woman $= {}^7C_4 \times {}^6C_1 = 35 \times 6 = 210$
5men and 0 woman $= {}^7C_5 \times {}^6C_0 = 21 \times 1 = 21$

$\therefore$ Total number of ways to select atleast 3men from 7men and 6 women is

$525 + 210 + 21 = 756$



12) LEADING $\Rightarrow$ $\underbrace{LDNG}_{4 \text{ letters}}$ $\underbrace{EAI}_{1 \text{group}}$

all the vowels are considered as a group
all the vowels can arrange among themselves in 3! ways
So, the total no. of ways $= 5! \times 3! = 720$.

13) Probability that the first card is an ace = $4/52$
Probability that the second card is an ace = $3/51 \times 4/52$
$= 12/2652 = 1/221$

14) Probability of getting a 6 in 1st try is $1/6$.
Probability of getting two consecutive 6 = $(1/6)^2 = 1/36$

Probability of getting tail = $1/2$
Probability of getting two consecutive tail = $(1/2)^2 = 1/4$.

$\therefore$ Probability of two consecutive 6 and two consecutive tail is $1/36 \times 1/4 = 1/144$.

15) Prime Numbers = 2, 3, 5.
Probability of getting a prime number = $3/6 = 1/2$.
Probability of getting 6 consecutive prime number = $(1/2)^6 = 1/6$[illegible]

Marks 5

]

1) The random variable X can take any integer from 0 to 100. So range of X is

$$R_X = \{0, 1, 2, 3, \dots, 100\}$$

2) 9 am The random variable Y can take any Integer (Positive) so range of Y is

10 am

1 pm $R_Y = \{1, 2, 3 \dots\} =$ 4 pm $\mathbb{N}$

11 am

2 pm

5 pm

3) The random variable T can take any non negative real number. So range of T is

$R_T = [0, \infty)$.

2) Probability that a trainee will remain ~~with~~ $= 0.6 = P(R)$

Probability of salary greater than 10K $= 0.5 = P(10K)$

Probability of salary >10K or trainee remain $= 0.7 = P(10K \cup R)$

Probability of salary >10K and trainee remain $= P(10K \cap R)$

$\cancel{P(10K \cap R) = P(10K) + P(R) - R}$

$P(10K \cup R) = P(10K) + P(R) - P(10K \cap R)$

$\Rightarrow 0.7 = 0.5 + 0.6 - P(10K \cap R)$

$\Rightarrow 0.7 = 1.1 - P(10K \cap R)$

$\Rightarrow P(10K \cap R) = 0.4$

$$P(10K|R) = \frac{P(R \cap 10K)}{P(R)} = \frac{0.4}{0.6} = \frac{2}{3}.$$

$\boxed{3}$ $RGR = \frac{8}{12} * \frac{4}{11} * \frac{7}{10} = \cancel{501/330}\ 56/330$

$GRG = \frac{4}{12} * \frac{8}{11} * \frac{3}{10} = \cancel{400/330}\ 8/110$

$\cancel{\frac{571}{330}} + \cancel{\frac{449}{330}} = \cancel{\frac{1020}{330}}$

$\frac{56}{330} + \frac{8}{110} = \frac{80}{330} = \frac{8}{33}.$

9 am 12 noon 3 pm 10 am 1 pm 4 pm 11 am 2 pm 5 pm

4) 1st case : Red ball is drawn. White is put.

$3/10 * 8/10$

2nd Case : White ball is drawn. Blue is put.

$7/10 * 4/10$

$(3/10 * 8/10) + (7/10 * 4/10) = 0.34.$

5) i) Two success $= 1/15 * 1/15 = \frac{1}{225}$

ii) Exactly one success $= 1/15 * 1/15 = 1/225$

iii) Atleast one success $= (1/15 * 1/15) + (1/15 * 1/15) = 2/225$

iv) No success $= (1/15 * 1/15) = 1/225.$

15 marks

1) i) P(one from each category) $= \frac{{}^3C_1 * {}^4C_1 * {}^2C_1 * {}^1C_1}{{}^{10}C_4}$

$= 24/210 = 0.1143$

ii) P (atleast one from purchase department)

= 1 − P(none from purchase department)

$= 1 - \frac{{}^6C_4}{{}^{10}C_4} = 0.9286.$

iii) P(chartered accountant must in the community)

$$= \frac{{}^{9}C_3}{{}^{10}C_4} = 0.40$$

3) $P(A) = 0.4$
$P(B) = 0.7$
$P(A \cup B) = 0.8$
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$0.8 - 0.7 - 0.4 = -P(A \cap B) \Rightarrow -0.3 = -P(A \cap B)$

$\Rightarrow P(A \cap B) = 0.3$

~~4) $(1 - \frac{1}{20}) =$~~

~~4) Probability~~

~~4) originator has $(n-1) = 9$ choices to spread.~~

4)
a) Originator has to spread to some other.
Every other person can spread to everyone except himself and the originator

So every person has ~~19~~ 19 choices. [21-1-1]

So prob. that 1 person not return to the originator $= 19/20$

So " " 9 " " " " $= (19/20)^9$.

9 am 12 noon 3 pm

b) $\frac{20}{20} \times \frac{19}{20} \times \frac{18}{20} \times \cdots \times \frac{11}{20} = \frac{20 \times 19 \times 18 \times 17 \times 16 \times 15 \times 14 \times 13 \times 12 \times 11}{20^{10}}$.

10 am 1 pm 11 am 2 pm 5 pm

074-291 • Week 11 MARCH FRIDAY 5) A)

5) A)

_ _ _ 0 4 → 3! Total 5 digit number

_ _ _ 2 4 → 2×2!

_ _ _ 3 2 → 2×2! $= 4\times4\times3\times2\times1 = 4\times4!$

_ _ _ 1 2 → 2×2!

_ _ _ 4 0 → 3!

_ _ _ 2 0 → 3!

4 4 3

$$\frac{6+4+4+4+6+6}{4\times4!} = \frac{30}{4\times4!}$$

$$= \frac{15}{48} = \frac{5}{16}$$

B) First person has 4 option

Second " " 3 "

Third " " 2 "

∴ Total number of option without condition $= 4*4*4 = 64$.

∴ Probability $= \frac{4\times3\times2}{64} = 24/64 = 0.375$

9 am

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

Chapter - 2

Marks 1

1) $P(1) = 1k$ $\quad P(1) + P(2) + P(3) + \cdots + P(6) = 1$

$P(2) = 2k$ $\quad 1k + 2k + \cdots + 6k = 1$

$\vdots$ $\quad k(1+2+3+4+5+6) = 1$

$P(6) = 6k$ $\quad k = 1/21$

$\therefore P(4) = 4k = 4/21$

2) Mean of 5 innings $= (50+70+82+93+20)/5 = 63$

Standard Deviation, $\sigma = \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$

$\Sigma(x_i-\mu)^2 = (50-63)^2 + (70-63)^2 + (82-63)^2 + (93-63)^2 + (20-6$

$= (-13)^2 + (7)^2 + (19)^2 + (30)^2 + (-43)^2$

$= 169 + 49 + 361 + 900 + 1849$

$= 3328$

$\sigma = \sqrt{\frac{3328}{5}} = \sqrt{665.6} = 25.79$



3) 4, 5, 9, 11, 13, 14, 15, 18, 18

No. of terms = 9 = odd.

$\therefore$ Median $= \frac{9+1}{2} = 10/2 = 5^{th}$ term $= 13$

mode = 18 = occurance 2times

MARCH 2013

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 |

4) $V(x) = E(x^2) - (E(x))^2$

5) The area under cummulative density function = 1.

6) $\int_0^3 x^2 = \int 3^2 - \int 0^2 = \int 3^2 = \frac{3^{2+1}}{2+1} = \frac{3^3}{3} = \frac{27}{3}$
$= 9$

7) $Var(x) = 0.2$ $var(y) = 0.5$ $Z = 5X - 2Y$

$var(z) = Var(5x - 2y) = Var(5x)$

8) $E(x) = 2$ $E(z) = 4$ $E(z-x) = ? = 2.$

$E(z-x) = E(z) - E(x) = 4 - 2 = 2.$

9) $f(x) = \begin{cases} 0.2 & |x| < 1 \\ 0.1 & 1 < |x| < 4 \\ 0 & \text{otherwise} \end{cases}$

$\int_{0.5}^{1} 0.2 + \int_{1}^{4} 0.1 + \int_{4}^{5} 0 = [0.2 - 0.1] + [0.4 - 0.1] + 0$

$= 0.1 + 0.3 = 0.4$

9 am
10 am
11 am
12 noon
1 pm
2 pm
3 pm
4 pm
5 pm

10) $\int_{-\infty}^{\infty} f(x) = 1$

Marks 5

1) i) Mean of binomial distribution $= np = (10) \times (60\%) = 10 \times 0.6 = 6$

Variance of binomial distribution $= npq = 10 \times 0.6 \times 0.4 = 2.4$

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

$P(1) = \frac{e^{-\lambda} \lambda}{1!} = e^{-\lambda} \lambda$

$P(3) = \frac{e^{-\lambda} \lambda^{3}}{3!}$

$e^{-\lambda} \lambda = \frac{e^{-\lambda} \lambda^{3}}{3!}$

$\Rightarrow \lambda = \frac{\lambda^{3}}{3!}$

$\Rightarrow 3! = \lambda^{2} \Rightarrow 6 = \lambda^{2} \Rightarrow \lambda = \sqrt{6}$

iii) Mean $= np = 8 \times \frac{1}{2} = 4$

2) i) P (choosing boy) $= 1/2$

P (choosing girl) $= 1/2$

P (choosing maths by boy) $= 40/100$

P (choosing maths by girl) $= 60/100$

P (maths) $= \left(1/2 \times 40/100\right) + \left(1/2 \times 60/100\right) = 1/2 = 0.5$.

If the no. of boys = no. of girls.

ii) $1/3 \times \frac{10}{100} + 1/3 \times \frac{20}{100} + 1/3 \times \frac{5}{100}$

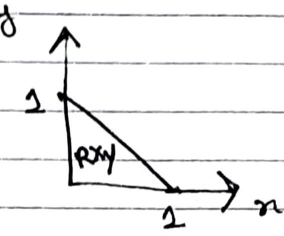
$= 0.116 \approx 0.12$.

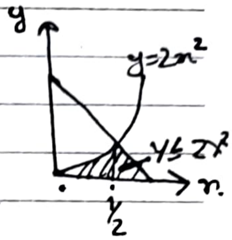
9 am 12 noon 3 pm 10 am 1 pm 4 pm 11 am 2 pm 5 pm

3) i) $\left(\frac{1}{2}\times\frac{5}{100}\right)+\left(\frac{1}{2}\times\frac{10}{250}\right)=0.045$

ii) $\frac{1}{5}\times\left(\frac{1}{2}+\frac{1}{3}+\frac{2}{3}+\frac{1}{5}+\frac{1}{6}\right)=0.37$

4) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,dy = 1$




Therefore.

$\cancel{\int_{0}^{1}\int_{-\infty}^{\infty} cx+1\,dy\,dx}$

$\cancel{\int_{0}^{1}\int_{0}^{1-x} cx+1\,dy\,dx} = \cancel{\int_0^1\left(c\int_0^{1-x} x\,dy + \int_0^{1-x} 1\,dy\right)dx}$

$\cancel{=\int_0^1 c} = \int_0^1\int_0^{1-x} cx+1\,dy\,dx = \int_0^1\left((cx+1)\int_0^{1-x} 1\,dy\right)dx$

$= \int_0^1 (cx+1)(1-x)\,dx = \frac{1}{2}+\frac{1}{6}\frac{c}{\cancel{x}} = 1 \quad \therefore \underline{c = 3.}$

9 am

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

5) $f_{X,Y}(x,y) = \begin{cases} 6e^{-(2x+3y)} & x,y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$

i) $f_X(x) = 2e^{-2x}\, u(x)$ $f_Y(y) = 3e^{-3y}\, u(y)$.

Thus X and Y are independent.

ii) $E[Y \mid X>2] = E[Y]$ $Y \sim$ Expontential (3) $\therefore E[Y] = 1/3$

iii) $\displaystyle\int_0^\infty 3e^{-3y}\, dy.$

9 am

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | | | | 1 | 2 | 3 |
| 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 11 | 12 | 13 | 14 | 15 | 16 | 17 |
| 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 25 | 26 | 27 | 28 | 29 | 30 | 31 |

Marks 15

J PDF

1) i) A joint probability distribution shows a probability distributi for two or more random variables.

The formal definition is

$$f(x, y) = P(X = x,\ Y = y)$$

The whole point of joint probability distribution is to look for a relationship between two variables.

The table below shows probabilities of X and Y happening at the same time.

| Y \ X | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0 | 1/6 | 1/6 |
| 2 | 1/6 | 0 | $\boxed{1/6}$ |
| 3 | 1/6 | 1/6 | 0 |

There, the probability for X = 3 and Y = 2 will be 1/6.

Cummulative Distributio

A cummulative distribution function is used to describe the probability distribution of a random variable. It can be used to describe the probability for a discrete, continuous or a mixed variable. The formal definition is

9 am

$$F_X(x) = P(X \le x),\ \forall x \in R$$

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

Probability of getting an outcome on rolling a dice.

~~$P(X \le x)$ 1 2 3~~

| $X = x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $P(x)$ | $1/6$ | $2/6$ | $3/6$ | $4/6$ | $5/6$ | $6/6$ |

The probability value always lies between 0 and 1. It is non decreasing and right continuous in nature.

ii) Expectation

The expected value of a random variable with finite number of outcome is a weighted average of all possible outcomes. In case of continuous random variable, the expectation is defined as integration. It is denoted by $E(x)$

The expectation of when a die is rolled one time with 6 possible outcome is.



$$E[x] = 1 \cdot 1/6 + 2 \cdot 1/6 + 3 \cdot 1/6 + 4 \cdot 1/6 + 5 \cdot 1/6 + 6 \cdot 1/6$$
$$= 3.5$$

$$E(x) = \sum x_i P(x_i)$$

## Variance

It is the measure of dispersion, meaning it is a measure of how ~~much~~ far a set of number is spread out from their average value.

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

$$Var(X) = Cov(X, X)$$

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

variance of a fair six-sided die canbe calculated as

$$\sum_{i=1}^{6} \frac{1}{6}\left(i - \frac{7}{2}\right)^2 \qquad E(X) = \frac{7}{2}.$$

$$\Rightarrow \frac{35}{12} \approx 2.92$$

## Standard Deviation

It is the measurement of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values should be close to mean. while a high standard deviation indicates the values are spread out over a wider range.

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

9 am 12 noon 3 pm 10 am 1 pm 4 pm 11 am 2 pm 5 pm

If variance in marks of a group of students is 75.6 then the standard deviation will be $\sqrt{75.6} = 8$

iii) Significance of Expectation Value

Suppose there is a insurance company giving term life insurance for ~~20~~ 10 years. The insurance premium per year is ~~Rs~~ Rs 50,000 and the insurance cover is: Rs 50,00,000.

Let say there are two medical plans.

Plan A: Costs ~~Rs~~ Rs 80,000/year and have to pay first 10,000 ruppees of any medical cost.

Plan B: costs Rs 60,000/year and have to pay first 25,000 ruppees of any medical cost.

Let for a person his probababilty of medical expense is given:-

| Medical cost | Probabilily |
|---|---|
| Rs 0 | 30%. |
| Rs 10,000 | 25%. |
| Rs 40,000 | 20%. |
| Rs 70,000 | 20%. |
| Rs 1,50,000 | 5%. |

~~X = expected~~ cost.

~~Plan A~~:

~~$E(X) = 80000 + 0 \times (0.3) + 10,000 \times (0.25) + 40,000 \times (0.2) + 70,000 \times (0.2) + 1,50,000 \times (0.05)$~~

9 am 12 noon 10 am 1 pm 4 pm 11 am 2 pm 5 pm

X = expected cost

Plan A : $E(x) = 80{,}000 + 0\times(0.3) + 10000(0.25) + 10000(0.2) + 10000(0.2) + 10000(0.05)$

$= 80000 + 2500 + 2000 + 2000 + 500$

$= 87000$

Plan B : $E(x) = 60000 + 0\times(0.3) + 10000(0.25) + 25000(0.2) + 25000(0.2) + 25000(0.05)$

$= 60000 + 2500 + 5000 + 5000 + 12500$

$= 1{,}05{,}000$

∴ The expected expense in plan B is more than plan A. So plan B should be choosen

Significanof Variance

Data Scientists often use variance to better understand the distribution of a database set. Machine Learning uses variance to make generalization about a dataset adding in ~~...~~ neural network's understanding of the database. ~~It~~ often used in conjunction with probability distribution.

9 am 12 noon 10 am 1 pm 4 pm 11 am 2 pm 5 pm

## 2)i) Moment

A moment is a quantitative measurement for the shape of a function.

$$s^{th} \text{ moment} = (x_1 s + x_2 s + \dots + x_n s)/n$$

First moment is function's mean.
Second moment is central moment or function's variance.
Third moment is function's skewness or to which extent the probability distribution is offset from mean.
Fourth moment is function's Kurtosis which describes the shape of tail of a probability ~~function~~ distribution.

The ~~...~~ third and fourth moments are normalized ~~that~~ meaning the moment function is divided by an expression of standard deviation. This allows the third and fourth moment to be scale invariant.

### Significance in Data Science

~~...~~ Calculationg moments are simple so they provide a first insight (quantitative) into the data. It provides the data scientist with a good understanding of the dataset before any training of ML models.

9 am
10 am 1 pm 4 pm
11 am 2 pm 5 pm

ii) Quartile

A quartile is a statistical term that splits data into quarters ~~of~~ or four defined intervals. It is an extended version of median. Median divides the dataset into two equal parts whereas quartiles divides it into four equal parts. Before ~~arranging~~ dividing into quartile the data must be arranged in ascending order in the number.

1st quartile basically seperates the lowest 25% from the highest 75%. It is also called lower quartile

2nd quartile is same as median. It divides the data into two parts. It is also called middle quartile

3rd quartile or the upper quartile seperates the highest 25% of the data from the lowest 25%.

Lower Quartile (Q1) = $\frac{(N+1)}{4}$

Middle Quartile (Q2) = $\frac{N+1}{2}$

Upper Quartile (Q3) = $N+1 \times \frac{3}{4}$

$IQR = Q_3 - Q_1$

9 am 12 noon 3 pm
10 am 1 pm 4 pm
11 am 2 pm 5 pm

Quartiles are often used in neural networks and machine learning functions. For example, if one is using a neural network to analyse a set of data, it is ~~integral~~ ~~to~~ ~~needed~~ to understand ~~the boundary~~ of the probability density function. integral in the creation of a probability density function to define and understand the boundary quartile that separate the quantiles containing the possible ~~outcome~~ probabilities of continuous random variable occuring.

3) i) Mean $(x) = \dfrac{\sum f_i x_i}{N} = \sum \left( \dfrac{f_i x_i}{N} \right) = \sum \dfrac{f_i}{N} x_i$

$= \sum p_i x_i$ = Expectation.

ii) $\mathrm{Var}(X+c) = E((X+c - E(X+c))^2)$

$= E((X+c - E(X) - c)^2)$

$= E((X - E(X))^2)$

$= \mathrm{Var}(X)$

iii) $\mathrm{Var}(aX+b) = E((aX+b - E(aX+b))^2)$

$= E((aX+b - E(aX) - b)^2)$

$= E((aX - aE(X))^2$

$= a^2 E(X - E(X))^2 = a^2 \mathrm{Var}(X).$

iv) $\mathrm{Cov}(aX, Y) = E(aXY) - E(aX)E(Y)$

$= aE(XY) - aE(X)E(Y)$

$= a(E(XY) - E(X)E(Y))$

$= a\,\mathrm{Cov}(X,Y)$

v) Standard deviation measures how ~~p~~ far apart numbers are in a dataset whereas variance measures how much a data is away from the mean. SD is measured as the square root of variance. SD is the indicator of the observation whereas variance is indicator of individual spread out in a group.

9 am

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

## Conditional Probability

The likelihood of an event or outcome occurring based on the occurrence of a previous event or outcome

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{OR} \quad P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$P(A|B)$ is the probability of A given B.

**Example** ·

A red card is drawn so what is the probability it is 4

$$P(\text{Four}|\text{Red}) = 2/26 = 1/13.$$

## Density Function

PDF defines the probability function representing the density of a continuous random variable lying between a specific range of values. It produces the likelihood of continuous random variable.

$$P(a<x<b) = P(a \le x \le b) = P(a \le x < b) = P(a < x \le b) = \int_a^b f(x)\,dx$$

It is non negative everywhere

9 am | 12 noon | 3 pm
10 am | 1 pm | 4 pm
11 am | 2 pm | 5 pm

4) iii) $P(A|C) = P(B|C) = 1/2$

$P(A \cap B|C) = 1/2 \cdot 1/2 = 1/4$

$P(A) = P(A|C)P(C) + P(A|\bar{C})P(\bar{C})$

$= 1/2 \cdot 1/2 + 1 \cdot 1/2 = 3/4.$

$P(B) = 3/2.$

$P(A \cap B) = P(A \cap B|C)P(C) + P(A \cap B|\bar{C})P(\bar{C})$

$= P(A|C)P(B|C)P(C) + P(A|\bar{C})P(B|\bar{C})P(\bar{C})$

$= 1/2 \cdot 1/2 \cdot 1/2 + 1 \cdot 1 \cdot 1/2$

$= 5/8.$

$P(A \cap B) = 5/8 \neq P(A) \cdot P(B) = 9/16$ which means A and B are not independent.

9 am

12 noon

3 pm

10 am

1 pm

4 pm

11 am

2 pm

5 pm

CHAPTER 3

Marks 1

1) Expected no. of women is ~~$E(x) = \sum x_i P_i(x)$~~

$= E(x) = \sum_i x_i P_i(x)$

$= \left[0 \times \frac{2}{11} + 1 \times 5/11 + 2 \times 4/11\right] = 13/11 \approx 1$ women

2) P (getting blue) = 1/4.
E (blue) = ~~1/4~~ $np = 30 \times 1/4 = 7.50$.

Therefore out of 30 experiments 7 to 8 balls will appear blue.

3) $E(\text{success}) = (0 \times 6/11) + (1 \times 9/22) + (2 \times 1/22)$
$= 11/22 = 1/2$.

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

where $\lambda$ is the mean.
e is the euler constant 2.71

5) Mean = np
variance = npq.

9 am

12 noon

3 pm

~~6) Correlation~~

10 am

1 pm

4 pm

11 am

2 pm

5 pm

6) Correlation

It refers to the process of establishing the relationship between two variables.

A correlation coefficient very close to 0 but either positive or negative implies little or no relation between two variables.

A positive correlation indicates that increase in one variable will lead to increase in other variable whereas a negative correlation implies that increase in one variable will lead to decrease in another variable.

7) Covariance

Covariance is a measure of the relationship between two random variables and to what extent they will change together.

Population Covariance

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N}$$

Sample Covariance

9 am

12 noon

3 pm

$$Cov(x,y) = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{N-1}.$$

1 pm

4 pm

10 am

5 pm

2 pm

11 am

8) $P(H) = \frac{1}{2} \quad P(T) = \frac{1}{2}. \quad N = 12.$

$B(r) = {}^nC_r \cdot p^r \cdot (1-p)^{n-r}.$

$B(7) = {}^nC_7 \cdot (\frac{1}{2})^7 \cdot (\frac{1}{2})^{12-7}$

$= {}^{12}C_7 \times (\frac{1}{2})^7 (\frac{1}{2})^5$

$= 0.193$

9) $p = 3/4 \quad q = 1/4 \quad n = 5$

At least thrice.

$P(X=3) + P(X=4) + P(X=5)$

$= {}^5C_3 \cdot (3/4)^3 (1/4)^2 + {}^5C_4 \cdot (3/4)^4 (1/4)^1 + {}^5C_5 (3/4)^5 (1/4)^0$

$= 459/512.$

10) ${}^nC_6 (\frac{1}{2})^6 (\frac{1}{2})^{n-6} = {}^nC_8 (\frac{1}{2})^8 (\frac{1}{2})^{n-8}$



$\Rightarrow {}^nC_6 (\frac{1}{2})^n = {}^nC_8 (\frac{1}{2})^n$

$\Rightarrow {}^nC_6 = {}^nC_8 = \cancel{\frac{n!}{6!(n-6)!}} = \cancel{\frac{n!}{8!(n-8)!}}$

$= \cancel{56(n-7)(n-8) = 1} \Rightarrow n = 14$

$\Rightarrow 56 = (n-7)(n-8) \Rightarrow n = 14, -1 \quad \therefore n = 14.$

Marks 5

Mean = $(1\times0.15) + (2\times0.10) + (3\times0.10) + (4\times0.01) + (5\times0.08) + (6\times0.01) + (7\times0.05) + (8\times0.02) + (9\times0.28) + (10\times0.20)$

$E(x) = 6.56.$

$E(x^2) = (1^2\times0.15) + (2^2\times0.10) + (3^2\times0.10) + (4^2\times0.01) + (5^2\times0.08) + (6^2\times0.01) + (7^2\times0.05) + (8^2\times0.02) + (9^2\times0.28) + (10^2\times0.20)$

$= 50.38.$

$Var(x) = 50.38 - (6.56)^2 = 7.35.$

2) Normal Distribution / Gaussian Distribution

The normal distribution is a continuous probability distribution that is symmetrical around its mean. It produces a bell shaped curved.

Standard Normal Variate (SNV)

SNV method performs a normalization of the spectra, that consists in subtracting each spectrum by its own mean and dividing it by its own standard deviation. After SNV, each spectrum will have a mean of 0 and standard deviation of 1.

## Advantage of Normal Distribution

The mean, median, mode of the distribution are equal.

The entire population canbe explained with the help of only mean and standard deviation.

3. $f(x) = \begin{cases} 4x^3 & \text{if } 0<x<1 \\ 0 & \text{otherwise.} \end{cases}$

$E(x) = \int_0^1 x \cdot 4x^3 + 0 = 4 \cdot \frac{x^5}{5} = 4 \cdot \frac{1}{5}$

$= 4/5.$

$var(x) = \int_0^1 x^2 \cdot 4x^3 = \int_0^1 4x^5 = 4 \cdot \frac{x^6}{6} = \frac{4}{6} = \frac{2}{3}.$

4. $f(x) = ke^{-2x}, \; 0 \le x \le \infty$

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

$\int_0^{\infty} ke^{-2x} = 1 = k\int_0^{\infty} e^{-2x}\,dx = 1$

$\Rightarrow k\left[\frac{e^{-2x}}{-2}\right]_0^{\infty} = 1 \Rightarrow k = 2.$

9 am 12 noon 3 pm

10 am 1 pm 4 pm

11 am 2 pm 5 pm

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

$$= \int_0^{\infty} xke^{-2x}\, dx = 2\int_0^{\infty} xe^{-2x}\, dx.$$

$$= 2\int_0^{\infty} xe^{-2x}\, dx = \int_0^{\infty} e^{-2x}\, dx = 1/2.$$

Mar ko15

Monte Carlo Estimation

It is is a ~~mod~~ model that is used to predict the probability of a variety of outcomes when the potential for random variables is present. It is also called multiple probability simulation.

characteristics

- It must generate random sample.
- It's input distribution must be known.
- It's result must be known while performing the experiment.

Advantage

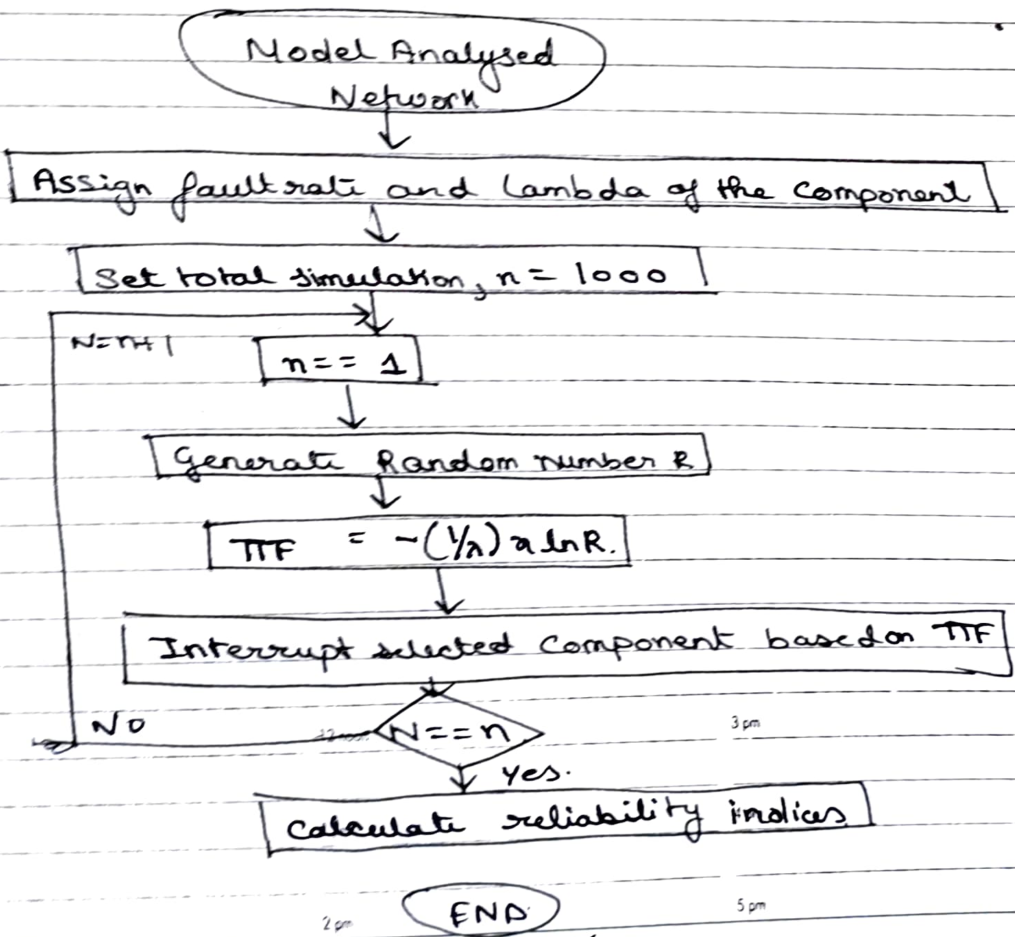
- Easy to implement
- Provides statistical sampling for numerical experiments using computer.
- Provides approximate sol^n to mathematical problem.
- can be used for both stochastic and deterministic problem.

Dis Advantage

1) Time consuming as there is a need to generate large number of sampling to get desired output.

2) The result of this method are only the approximation of true values and not the exact.

Flow Diagram



i) Conditional Expectation

Let X and Y be two random variables. The conditional expectation of X given Y = y is the weighted average of the values that X can take on, where each possible value is weighted weighted by its respective conditional probability (conditional on the information that Y = y).

The expectation of a random variable X conditional on Y = y is denoted by $E[X|Y=y]$.

Convergence in Probability

Convergence in probability is stronger than convergence in distribution.
A sequence of random variable $X_1, X_2, X_3, \ldots$ converges in probability to a random variable X, is shown by

$$X_n \rightarrow X, \text{ if}$$

$$\lim_{n\to\infty} P(|x_n - X| \geq \epsilon) = 0, \ \forall \ \epsilon > 0.$$

Application of Convergence in Probability in Data Science

9 am

10 am

11 am

12 noon

1 pm

2 pm

3 pm

4 pm

5 pm

## Generating Function

For a sequence $a_0, a_1, a_2, \ldots, a_i, \ldots$ the generating function $f(x)$ is the series:-

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n + \cdots = \sum_{i=0}^{\infty} a_i x^i$$

So, $a_n$, the $n^{th}$ term in the sequence is the coefficient of $x^n$ in $f(x)$.

## Binomial Distribution

For any random variable $X$, the binomial distribution is given by:

$$P(X) = {}^nC_r\, p^r\, (1-p)^{n-r}$$

where $n$ is the number of independent trials.
$p$ is the probability of success.
$(1-p)$ is the probability of failure.



### Example

Suppose a game is there that you can either win or lose. The probability of winning is 55% and the probability of losing is 45%. Then if you play 20 rounds, then probability of winning 15 times is

$$P(X=15) = {}^{20}C_{15} \times (0.55)^{15} (0.45)^5.$$

| APRIL - 2013 | | | | | | |
|---|---|---|---|---|---|---|
| M | T | W | T | F | S | S |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | | | | | |

## Applications

i) Binomial Trials are done to see effectiveness of a drug.

ii) Binomial distribution is used to detect defective goods in a manufacturing company.

3)

| $X = x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $P(x)$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ | $1/6$ |

mean $= \mu = \sum x_i p_i$

$= (1\times 1/6) + (2\times 1/6) + (3\times 1/6) + (4\times 1/6) + (5\times 1/6) + (6\times 1/6)$

$E(x) = 3.5$.

$E(x^2) = (1^2\times 1/6) + (2^2\times 1/6) + (3^2\times 1/6) + (4^2\times 1/6) + (5^2\times 1/6) + (6^2\times 1/6)$

$= 1/6 + 4/6 + 9/6 + 16/6 + 25/6 + 36/6$

$= 91/6$

Varian $= 91/6 - (3.5)^2 = 2.9166$

9 am
10 am
11 am
12 noon
1 pm
2 pm
3 pm
4 pm
5 pm

$\mu = 90 \quad \sigma = 10 \qquad P(x>100) = ?$

chapter-5

1) The data used in calculating a chi-square static must be random raw, mutually exclusive, drawn from independent variables and drawn from a large enough sample. Chi square tests are often used to test hypothesis. The result of tossing a fair coin meets this idea.

2) Stochastic Process

A stochastic process, also known as random process, is a collection of random variables that are indexed by some mathematical set. Each random variable in the collection of values is taken from a the same mathematical space known as state space.

3) Markov Chain

A markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules. A markov chain is also known as discrete time markov chain (DTMC) or markov process.

9 am 12 noon 3 pm

It is used to predict future state of a variable based on its past state

10 am 1 pm 4 pm

11 am 2 pm 5 pm

Markov chain is used in pagerank algorithm implemented by google. An user's behaviour can be predicted based on ~~past~~ previous preferences and interaction with it.

4. Cluster Sample

Cluster sampling is a probability sampling technique where researchers divide the population into multiple group (or clusters) for research. Researchers then select random groups with a simple random ~~variable~~ or systematic random sampling technique for data collection and data analysis.

Types:-

1) Single stage clusters.
2) Two stage clusters
3) Multi-stage clusters.

5. 4 types of Stochastic Process

i) Markov Process
ii) Poisson Process
iii) Birth Death Process
iv) Discrete or Continuous State Process.

9 am
10 am
11 am
1 pm
2 pm
3 pm
4 pm
5 pm

6. Need of Stochastic Process

Stochastic Processes provide quantitative study through mathematical model.

Stochastic Processes are commonly used in game theory.

Stochastic process involves studying and measuring patterns over a period of time for continuous phenomenon.

Marks 5

~~Am~~ Markov Chain

A markov chain is a discrete time process for which future behaviour given the past and present only depends on the present and not the past.

Markov Process

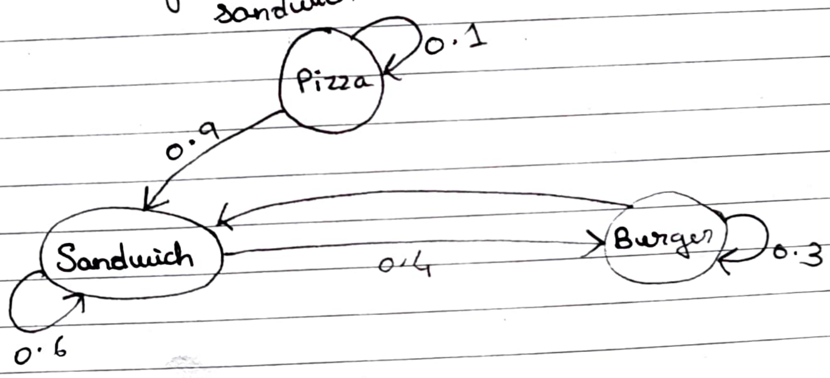
Markov Process is a continuous time version of a markov chain. Many queueing models are infact Markov process.

9 am
12 noon
3 pm
10 am
1 pm
4 pm
11 am
2 pm
5 pm

| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| 15 | 16 | 17 | 18 | 19 | 20 | 21 |
| 22 | 23 | 24 | 25 | 26 | 27 | 28 |
| 29 | 30 | | | | | |

## Example

let us assume that there is a restaurant which has only three items: Pizza, Burger, Sandwich. If the restaurant serves pizza on a day then probability of serving pizza next day is 0.1 and that of serving sandwich is 0.9. If the day is a sandwich day then probability of sandwich next day is 0.6 and probability of burger is 0.4. If the day is the burger day then probability of serving burger next day is 0.3 and that of ~~pizza~~ sandwich is 0.7.




2. Poisson Process

A poisson process is a model for a series of discrete events where the average time between event is known but the exact time of event is random.

9 am 12 noon 3 pm

10 am 1 pm 4 pm

11 am 2 pm 5 pm

Example

Suppose, there is a website which goes down on average once per 60 days, but one failure doesnot affect probability of next. Therefore we know average time between failures.

Criteria for Poisson Process

i) Events are independent of each other.
ii) Average & rate (event per time period) is constant.
iii) Two events cannot occur at the same time.

3. ~~What~~ Time Series Data.

Time series data also referred to as time-stamped data is a sequence of data point indexed in time order.

These data points typically consists of successive measurement made from the same source over a fixed time interval and are used to track change over time.



Time series data is everywhere, since time is the constituent of everything that is observable. As our world gets increasingly instrumented, sensors and systems are constantly emitting relentless stream of time series data. These data has various applications across numerous industries → stock, monthly subs.

4. Condition under which stochastic process becomes counting process

A counting process is a stochastic process $\{N(t), t \geq 0\}$ with values that are non-negative, integer and nondecreasing. $N(t) \geq 0$. $N(t)$ is an integer.
If $s \leq t$ then $N(s) \leq N(t)$.

5) State Space in Stochastic Process

The range (or possible values) of random variables in a stochastic process is called the state space of the process

6) Done

7) The stochastic process is a model for analysis of time series. The stochastic process is considered to generate the infinite collection (called the ensemble) of all possible time series that has been observed. Every member of the ensemble is a possible ~~real~~ realization of stochastic process.

9 am
10 am
12 noon
1 pm
2 pm
3 pm
4 pm
5 pm

Harks 15

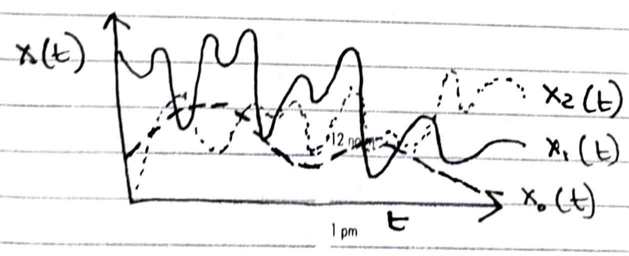
1) Stochastic means non-determinisistic or unpredictable.

Random generally means unrecognizable, not adhering to a pattern.

A random variable is also called a stochastic variable.

2) In general stochastic is a synonym for random. For example, a stochastic variable is a random variable. A stochastic process is a random process. Typically random is used to refer to a lack of dependence between observations in a sequence.

3) A stochastic function is a many valued numerical function of an independent argument t, whose value for any fixed value t∈T (where T is the domain of the argument) is a random variable called a cutset.




9 am 10 am 11 am 12 noon 1 pm 2 pm 3 pm 4 pm 5 pm

MAY - 2013

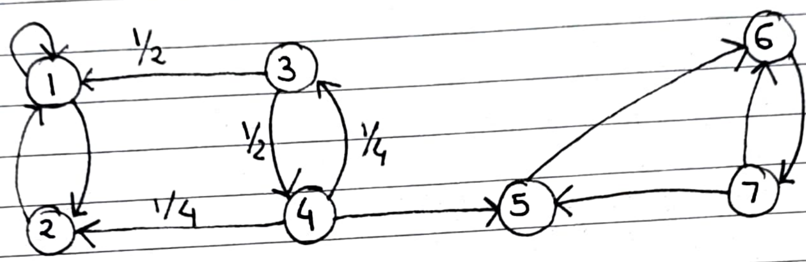
| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| 20 | 21 | 22 | 23 | 24 | 25 | 26 |
| 27 | 28 | 29 | 30 | 31 | | |

4. Primary Component of stochastic Process

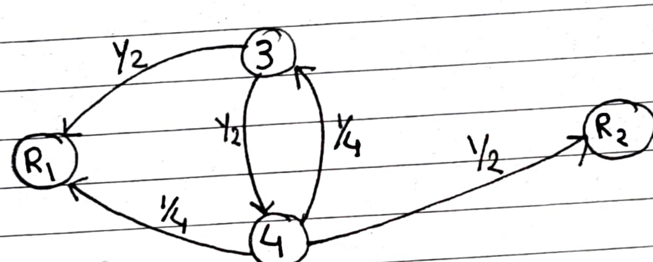
i) observation at certain time

ii) the outcome, that is the observed value at each time is a random variable.

5.




$R_1 = \{1, 2\}$ $R_2 = \{5, 6, 7\}$.




| | $R_1$ | 3 | 4 | $R_2$ |
|---|---|---|---|---|
| $R_1$ | 0 | 0 | 0 | 0 |
| 3 | ½ | 0 | ½ | 0 |
| 4 | ¼ | ¼ | 0 | ½ |
| $R_2$ | 0 | 0 | 0 | 0 |

9 am
noon
3 pm
10 am
1 pm
4 pm
11 am
2 pm
5 pm

$R_1 = [1 \quad 0 \quad 0 \quad 0]$

$$[1 \quad 0 \quad 0 \quad 0]\begin{bmatrix} 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 1/4 & 1/4 & 0 & 1/2 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$= [0 \quad 0 \quad 0 \quad 0]$

Therefore it is a reducible Markov chain.

So, $a_{R_1} = 1$ and $a_{R_2} = 0$ ($\because$ finding absorption in $R_1$).

$a_3 =$ all outward edges

$= \frac{1}{2} a_{R_1} + \frac{1}{2} a_4 = \frac{1}{2} + \frac{1}{2} a_4$

$a_4 =$ all outward edge

$= \frac{1}{4} a_{R_1} + \frac{1}{2} a_{R_2} + \frac{1}{4} a_3 = \frac{1}{4} + 0 + \frac{1}{4} a_3.$

$= \frac{1}{4} + \frac{1}{4} a_3$

$a_3 = \frac{1}{2} + \frac{1}{2} a_4 \qquad a_4 = \frac{1}{4} + \frac{1}{4} a_3$

9 am $2a_3 = 1 + a_4 \qquad 4a_4 = 1 + a_3$

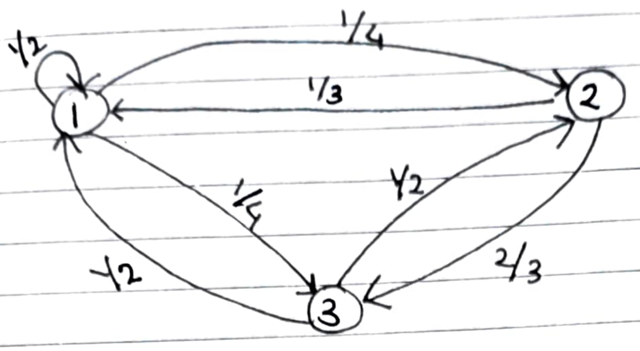
$\Rightarrow 8a_4 = 2 + 2a_3$

10 am $\Rightarrow 8a_4 - 2 = 1 + a_4 \qquad \Rightarrow 2a_3 = 8a_4 - 2$

$\Rightarrow 7a_4 = 3 \Rightarrow a_4 = 3/4$ and $a_3 = 5/7.$

11 am 2 pm 5 pm

6.




i) Yes, the chain is irreducible Since we can go from any state to other state in finite number of steps.

ii) Yes, the chain is aperiodic since there is a self transition.

iii) Stationary distribution

$\pi_i$ = all ~~out~~ in ward edge of i

$\pi_1 = \frac{1}{2}\pi_1 + \frac{1}{3}\pi_2 + \frac{1}{2}\pi_3$

$\pi_2 = \frac{1}{4}\pi_1 + \frac{1}{2}\pi_3$ Solving we get

$\pi_3 = \frac{1}{4}\pi_1 + \frac{2}{3}\pi_2$

$\pi_1 = 0.457$

$\pi_2 = 0.257$

9 am $\pi_1 + \pi_2 + \pi_3 = 1$ 12 noon $\pi_3 = 0.286$. 3 pm

10 am iv) Stationary distribution is limiting distribution for the chain as it is irreducible and aperiodic. 1 pm 4 pm

11 am 2 pm 5 pm

## Inherited Attribute

An attribute is said to be inherited attribute if its parse tree nodes values is determined by the attribute value at its parent or sibling nodes.

The production must have a non-terminal as a symbol in its body.

It is defined only in terms of its parent, itself and its siblings.

It can be evaluated during single top-down or sideways traversal of a parse tree.

It can only be contained by non terminals.

It can be used by only L-attributed SDT.

E.val = F.val



E val

↓

F val.