Random or stochastic process :-

→ in essence these processes whose outcomes are random within a fixed set of or range of values.

Probability :-

→ can be defined as the measure of certainity or uncertainity that a certain event or outcome will occur given a certain stochastic or random process.

population and sample :-

→ population refers to whole or complete data.

→ A collection of some randomly selected data from population that means a subset of population and which can serve as a good approximation of the population's characteristics and closely represent population is called sample.

pop probability distribution :-

→ refers to how the data is distributed among the set or range of values it contains.

Random variables :-

→ Random variables are variables that represent the collection of outcomes of a stochastic process.

Discrete & continuous random variables:-

→ Random variables that represent discrek & non continuous data are called Discrete random variable.

→ where as those which represent continuous data ~~are represent~~ are called continuous random variable.

Statistics:-

→ science of techniques & methodologies that are used for the collection, presentation and analysis of quantitative data for decision making.

→ Descriptive statistics (summarizing data)
→ Inferential statistics (drawing conclusions)

* Measures of centrality →

→ is a number that summarises the typical value or "tendency" of a random variable. These measures give the general indication as to where most values in a distribution fall.

MEAN → average value of random variable

MEDIAN → Middle value or average of middle two values in "SORTED" Random variable.

MODE → value with the Maximum frequency of occurance in random variable.

NOTE:- data with no repetation have no mode.

MEASURES OF DISPERSION :-

→ Measure of dispersion is a number or a range that summarises the "SCATTER" or "SPREAD" of a random variable with "RESPECT TO ITS MEAN".

MEAN OBSOLUTE DEVIATION :-

→ average of the obsolute deviation of all the values of a random variable with respect to mean.

$$\text{Mean obsolute deviation} = \frac{1}{n}\sum_{i=1}^{n}\left(|x_i - \text{mean}(x)|\right)$$

VARIANCE :-

→ average squared distance of the values of a r.v with respect to its mean. (squared to eliminate -ve values)

$$\text{variance} = 1/n \sum_{i=1}^{n} (x_i - \mu)^2$$

standard deviation :-

→ square root of the variance. it modifies the variance to provide measure of dispersion that is of same units as that of R.V.

Median absolute deviation :-

$$\text{Median}(\text{abs}(\mu - x_i)) \text{ for all } i)$$

$$= \text{Median}\left(|x_i - \text{~~median~~ mean}|\right)_{i=1}^{n}$$

IQR - Inter quartile range

→ $Q_3 - Q_1$ → gives mid 50% of dataset.

→ Q_2 - median

parameters:-

→ The statistical concepts of centrality, dispersion and shape when used in context to a population are called parameters.

statistics:-

→ concepts of centrality, dispersion & shape when used in context of sample are called statistics.

probability functions:-

→ mathematical charecterization of a probability distribution is called probability function. gives distribution type and its parameters we can completely define the probabity distribution, of a r.v.

probability Mass function:-

→ probability functions of discrete random variables are called PMF. They mathematically define the probability of occurance of each value or state within a discrete random variable.

Example: Binomial distribution.

Probability Density function :-

→ PDF used to express the frequency distributions of continuous r.v. maps the values of r.v to its probabili densities.

→ probability densities can have values greater than 1.

→ area under the curve defined by PDF represents the probability.

→ probabilites of continuous r.v expressed w.r.t ranges $P(x1 \leq X \leq x2)$.

→ PDF is non-negative everywhere and its integral over the entire space is equal to 1.

CUMULATIVE DENSITY FUNCTION :-

→ gives us the probability of any value in the random variable X being less than or equal to a particular value X.

$$CDF(x) = P(X \leq x)$$

→ CDF for both continuous & discrete r.v.

KERNAL DENSITY ESTIMATION :-

→ KDE is a statistical tool that allows us to create a smooth curve that approximates the PDF of a continuous r.v.

(i) replace each datapoint with PDF of normal distribution whose mean is equal to value of data point & standard deviation as function of bandwidth chosen.

→ Those PDF are called Kernals.
→ Band width is the range of values on either side of data point which we want to consider.
→ PDF of the dataset as a whole is approximated by summing up the values of individual kernals. at suitably chosen intervals.
→ The smoothness depends on size of the bandwidths chosen. larger bandwidths - less responsive to the sparsity of the data.

PROBABILITY DISTRIBUTIONS:-

THE BERNOULLI DISTRIBUTION:-

→ Discrete distribution, used to represent or model stochastic process that can have only two possible outcomes. success or failure

p — probability of success
1-p — probability of failure

→ Bernoulli is the basic building block for other discrete distributions like binomial, hypergeometric & poission.

→ Bernoulli(p) ~ X

mean — p
variance — pq

p Estimation — $\frac{n(1)}{n(0)+n(1)}$

The Binomial distribution:-

→ Binomial distribution is extension of the Bernoulli distribution. That is a binomial distribution or trail is the sequence of bernoulli events.

→ $X \sim B(n, p)$ — n: number of trials or events per outcome; p: probability of success.

PMF → $P(X = K) = P(K) = {}^{n}C_{K}\, p^{K}(1-p)^{n-K}$

CDF → $P(X \leq K) = \sum_{i=0}^{n} {}^{n}C_{K}\, p^{i}(1-p)^{n-i}$

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

$\text{var}(X) = npq = np(1-p)$

THE HYPER GEOMETRIC DISTRIBUTIONS:-

→ similar to Binomial distribution, but not same, it is picking without replacement, where Binomial is picking with replacement. in balls experiment if the # of balls is large relative to # of draws the distribution seem very similar to binomial distribution because the chance of success changes less.

→ $X \sim H(N, K, n)$ — N: population size; n: # of trials per outcome; K: number of success states in population

The poission distribution :-

→ deals with frequency with which an event occurs in a specfic interval of time, given that the probability of the event occuring is constant.

→ $X \sim Po(\lambda)$ (λ → rate parameter) expected or average frequency of success

→ can be used to model situations where the general rate of occurance of event is known and when we want to know the probability of same event occuring at the different rate of occurance.

$$PMF = P(X = K) = F(K) = \frac{\lambda^K e^{-\lambda}}{K!}$$

$$CDF = P(X \leq K) = F(K) = e^{-\lambda} \sum_{i=0}^{\lfloor K \rfloor} \frac{\lambda^i}{i!}$$

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

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

THE EXPONENTIAL DISTRIBUTION :-

→ exponential closely related to poission, while poission dist represents or models # of succesfull events per unit time, then exponential dist models the time period b/w 2 consecutive succesfull events., that follows a poission dist.

$$X \sim E(\beta) \rightarrow \beta = 1/\lambda$$

PDF = $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$

$E(x)$ = mean $= \beta = 1/\lambda$

CDF = $F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$

$Var(x) = \beta^2 = 1/\lambda^2$

→ memoryless distribution → p(event occur in 40 min | waited 30 min) = p(event occur in 10 min).

NORMAL OR GAUSSIAN DISTRIBUTION:-

→ used to represent r.v whose outcomes are continuous and exhibit 'centrality', the concentration of the frequency distribution is maximum near the central or mean value, and reduces symmetrically as we move away from it.

$X \sim N(\mu, \sigma)$
- μ → mean - location parameter
- σ → standard deviation - scale parameter

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

$E(x)$ = mean = μ

$Var(x) = \sigma^2$

CDF = $F(x) = P(X \leq x) = \int_0^x f(x)\,dx$.

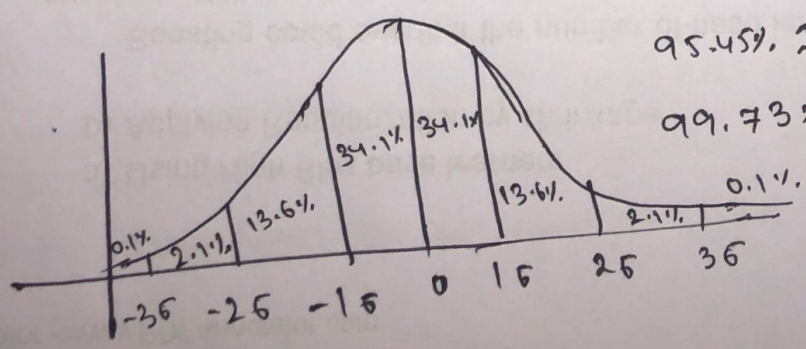
S.D = σ

68-95-99.7 rule:-

$68.27\% \approx Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma)$

$95.45\% \approx Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$

$99.73\% \approx Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$



The log normal distribution :-

→ R.v 'x' is said to be log normally distributed when log of 'x' is normally distributed.

→ log normal distribution can only contain +ve real values, and assymetric around mean and have heavy tails.

$$\ln(x) \sim N(\mu, \sigma)$$

Transformation of sampled data to gaussian :-

→ normal distribution is regarded as ideal and it is often assumed by many statistical methods. Hence it is often the practice that data is transformed to normal whenever possible.

→ commonly used transformation
(i) The Box-Cox transform
(ii) The Yeo-Jhonson transform
(iii) The Quantile transforms.

EXPLORATORY DATA ANALYSIS :-

- UNIVARIATE → single variable analysis
- BI-VARIATE → two variable analysis
- MULTIVARIATE → multiple variable analysis.

① MEAN — $\frac{1}{n}\sum_{i=1}^{n} x_i$

② Variance — $\frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2$ — Std-dev = $\sqrt{\text{Variance}}$

③ MEDIAN — ODD — $\frac{n+1}{2}$

EVEN — Mean $(n/2 \text{ & } \frac{n+1}{2})$

④ PERCENTILE — rank of x — $\frac{\#\text{values below } x}{n} \times 100$

Quantiles → 25th, 50th, 75th, 100th

↓ median

percentile value — $\frac{\text{percentile}\ (n+1)}{100}$

↳ 5.25 → 5th val + (0.25)(6th − 5th)

IQR = $Q_3 - Q_1$

$[Q_1 - (1.5 * IQR),\ Q_3 + (1.5 * IQR)]$ → points lying outside are called outliers.

⑤ MAD — MEDIAN$(|x_i - \text{median}|)_{i=1}^{n}$ (or) Mean

⑥ AXIOMS — (i) $0 \le P(E) \le 1$ (ii) $P(S) = 1$ (iii) $E_1, E_2, E_3 \ldots E_n$ are mutually exclusive $P(E_1 \cup E_2 \cup E_3 \ldots E_n) = \sum_{i=1}^{n} P(E_i)$

* $P(E^c) = 1 - P(E)$, * $P(E \cup E^c) = P(S) = 1$

⑦ PRINCIPAL OF INCLUSION & EXCLUSION:

$$|P_1 \cup P_2 \cup P_3 \ldots P_n| = \sum_{i=1}^{n} P_i - \sum_{\substack{i=1 \\ j=2}}^{n} (P_i \cap P_j) + \sum_{\substack{i=1 \\ j=2 \\ k=3}}^{n} (P_i \cap P_j \cap P_k) \ldots (-1)^{n-1} |P_1 \cap P_2 \cap P_3 \ldots P_n|$$

* if $E \subseteq F$ then $P(E) \le P(F)$

Conditional probability = $P(E|F) = \frac{P(E \cap F)}{P(F)}$ $P(F) \neq 0$ (2)

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

MULTIPLICATION THEORM:

$P(E_1 E_2 \ldots E_n) = P(E_1)\, P(E_2|E_1)\, P(E_3|E_2 \cap E_1) \ldots P(E_n|E_1 \ldots E_{n-1})$

MATCHIN PROBLEM - no one picked hat - 0.3678

Prob of exactly k from n picked - $^nC_k \sum_{i=0}^{n-k} \frac{(-1)^{n-k}}{i!} \frac{(n-k)!}{n!}$

INDEPENDENT EVENTS - $P(E|F) = P(E)$ & $P(E \cap F) = P(E)\,P(F)$

MUTUALLY EXCLUSIVE - $E \cap F = \phi$, $P(E \cap F) = 0$

E & F independent — E & F^c also independent

* infinite sequence trails, INDEPENDENT TO EACH OTHER

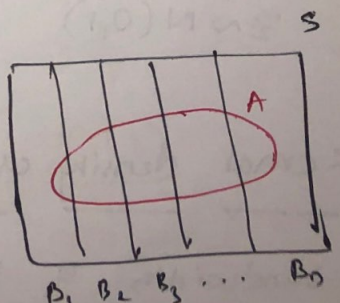
'p' SUCESS PROBABILITY.

ATLEAST ONE SUCCESS - $1-(1-p)^n$

EXACTLY K SUCCESS o/t of N - $^nC_k\, p^k (1-p)^{n-k}$

ALL N SUCCESS - p^n

* LAW OF TOTAL PROBABILITY :-



Mutually exclusive $B_i \cap B_j = \phi$

Mutually exhaustive $B_1 \cup B_2 \ldots B_n = S$

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\, P(B_i)$$

BAYES THEORM - $P(A/B) = \frac{P(B/A)\,P(A)}{P(B)} = \frac{P(B/A)\,P(A)}{P(B/A)P(A) + P(B/A^c)P(A^c)}$ ③

ODD IN FAVOUR of A - $\frac{P(A)}{P(A^c)} = \frac{P(A)}{1-P(A)}$

PROBABILITY MASS FUNCTION - $P(X=i) = C \cdot \frac{\lambda^i}{i!}$, $C = e^{-\lambda}$

CDF - $F_X(a) = \sum_{x \le a} P(X)$

(i) $0 \le P(X=a) \le 1$ (ii) $F(-\infty) = 0$ & $F(\infty) = 1$; b

$a \ge b$ then $F(a) \ge F(b)$

EXPECTATION - $E(X) = \sum_x x \cdot p(x)$ ↑PMF / $\int_x x \cdot f(x)\,dx$ ↓PDF

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

$E[X+Y] = E[X] + E[Y]$, $E[X \cdot Y] = E[X]\,E[Y]$

BERNOULI DISTRIBUTION:- $X \in \{0, 1\}$ Discrete distribution

p - success prob

$q = (1-p)$ - failure prob

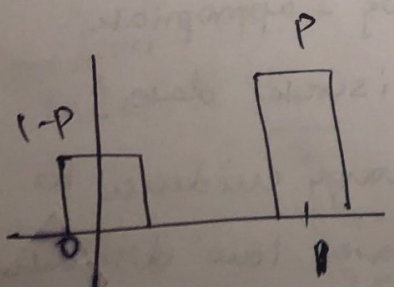
$X \sim Bernoulli(p)$

MEAN - p

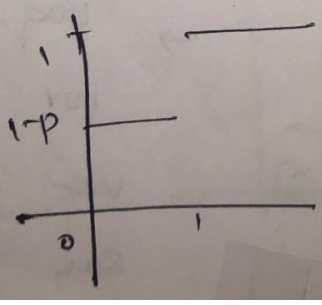
Variance - pq

Estimation - $\frac{n(1)}{n(0)+n(1)}$

PMF:-



CDF:-



BINOMIAL DISTRIBUTION:-

$X \sim (n, p)$ — probability of success ④
↳ # of trials

PMF — ${}^{n}C_{K}\, p^{K} q^{(n-K)}$

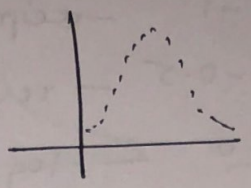
CDF — $\sum_{i=0}^{K} {}^{n}C_{i}\, p^{i} (1-p)^{(n-i)}$

MEAN — np

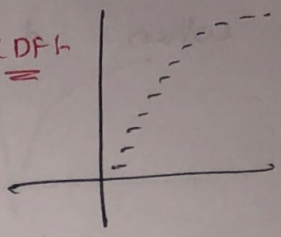
Variance — $npq = np(1-p)$

Restimation — $\frac{\sum_{i=1}^{m} n_i}{n \times m}$

PMF :-



CDF :-



Discrete curves

POISSON DISTRIBUTION:-

$X \sim \text{poisson}(\lambda)$ ← rate → Discrete

PMF — $\frac{\lambda^{K} e^{-\lambda}}{K!} = P(X = K)$

CDF — $P(X \le K) = \sum_{n=0}^{K} e^{-\lambda} \cdot \frac{\lambda^{n}}{n!}$

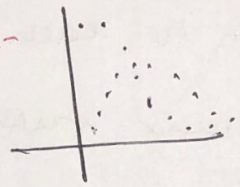
MEAN — λ

Variance — λ

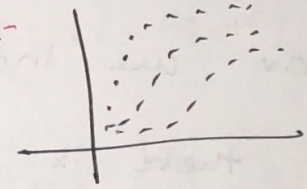
ESTIMATION OF λ :-

$\hat{\lambda} = 1/n \sum_{i=1}^{n} K_i$

PMF :-



CDF :-



UNIFORM DISTRIBUTION:-

$X \sim \text{UNIF}(a, b)$ (a → min, b → max) $\quad x \in [a, b]$

PDF — $\begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

CDF — $\begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } x \in [a, b) \\ 1 & \text{if } x \ge b \end{cases}$

median & Mean — $\frac{a+b}{2}$

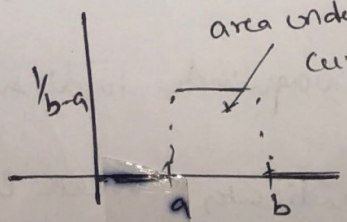
Variance — $\frac{(b-a)^2}{12}$

ESTIMATION of a & b :-

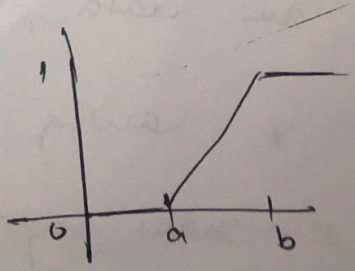
a → minimum of all the values

b → maximum of all the values

PDF



CDF

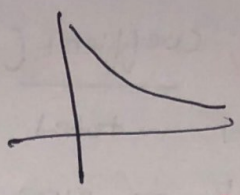


EXPONENTIAL DISTRIBUTION:- $X \sim Exp(\lambda)$ (rate)

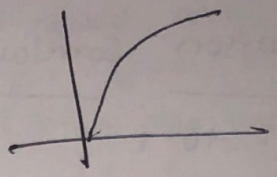
PDF – $f(x) = \lambda e^{-\lambda x}$ $x \geq 0$; 0 $x < 0$

CDF – $F(x) = 1 - e^{-\lambda x}$ $x \geq 0$; 0 $x < 0$

PDF



CDF



MEAN – $1/\lambda$

Variance – $1/\lambda^2$

ESTIMATION – $\frac{n}{\sum_{i=1}^{n} x_i}$

* MEMORYLESS DISTRIBUTION
→ exponential
→ Geometrical

Normal distribution:- $X \sim N(\mu, \sigma^2)$

PDF – $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$, $\sigma \neq 0$

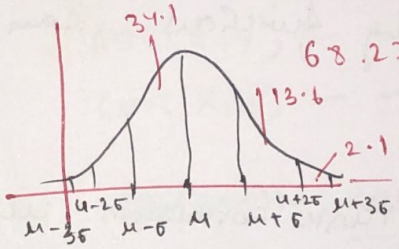
CDF – $\frac{1}{2}\left[1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]$

Mean – $\frac{\sum_{i=1}^{n} x_i}{n}$

Variance – $\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$

* Binomial is normal for large n, and for p not too close to '1' or '0'

$X \sim N(\mu, \sigma^2) \rightarrow Z = \frac{x-\mu}{\sigma} \sim N(0,1)$ (standard normal dist)



68.27 – 95.45 – 99.73

SKEWNESS:-

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^{3/2}}$$

$$= \frac{E[x^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3}$$

* normal distribution's skewness = 0.

KURTOSIS:- outliers, univariate normal distribution = 3

$$\text{Excess kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} - 3$$

LOWER EXCESS KURTOSIS : LOWER OUTLIER | Lighter tails

HIGHER EXCESS KURTOSIS : higher outliers | Leaviar tails

CLT — $X \sim N(\mu, \sigma^2) \rightarrow \bar{x}_i \sim N(\mu, \sigma^2/n)$ as $n \to \infty$

CHEBYSHEV'S INEQUALITY — $P(|X-\mu| \geq k\sigma) \leq 1/k^2$

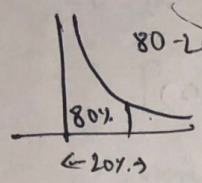
$P[\mu - k\sigma < X < \mu + k\sigma] > 1 - 1/k^2$

* valid only for finite mean & finite variance

LOG NORMAL DISTRIBUTION:- $X \sim \log\text{-normal}(\mu, \sigma^2)$

→ R.V takes only +ve & real values $Y = \log_e(x) \sim N(\mu', \sigma^{2'})$

POWER LAW (PARETO) DIST:- $X \sim \text{pareto}(x_m, \alpha)$ (x_m: scale like μ, α: shape like σ^2)



$\alpha \uparrow$ – tails become less fatter.

BOX-COX transformation:-

$$y_i = \begin{cases} \frac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x_i) & \text{if } \lambda = 0 \end{cases}$$

COVARIANCE:-

$$\text{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

cov(x,y) – +ve then x ↑↓ y ↑↓

cov(x,y) – -ve then x ↑↓ y ↓↑

PEARSON CORRELATION COEF:- $\rho_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \cdot \sigma_y}$ $-1 \leq \rho \leq 1$

SPEARMAN RANK CORR-COEF:- $r = \rho_{r_x, r_y}$

LINEAR BIASED REMOVED

K-S TEST:- $D_{n,m} = \sup |F_{1,n}(x) - F_{2,m}(x)|$

Ho rejected if $D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \times m}}$, $c(\alpha) = \sqrt{-1/2 \ln(\alpha/2)}$ (significance level)

p-value = $2 \cdot e^{-2D^2\left(\frac{nm}{n+m}\right)}$

Decision based on sample	Truth: Ho true	Truth: Ho False
Fail to reject Ho		type 2 error
Reject H_0	Type 1 error	