Descriptive Statistics

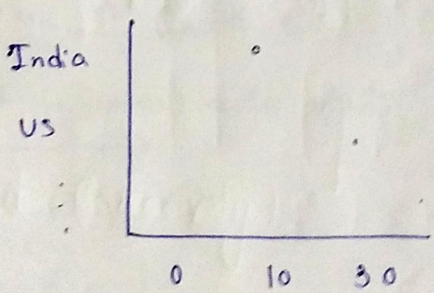
The two most important functions of descriptive statistics are:

1. Communicate information
2. Support reasoning about data.

When exploring data of large size, it becomes essensial to use summaries.

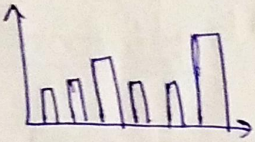
→ Graphical Summaries of Data:

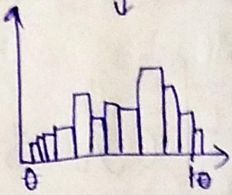
1. For (qualitative) (e.g. doctors, cartypes.) use a pie chart or a dot plot



② When the data are quantitative (i.e. number) then, they should be put on a number line.

Eg. in Bar graph ↓   Histogram ↓





- Histograms:-

Key points:- The areas of the blocks are proportional to frequency

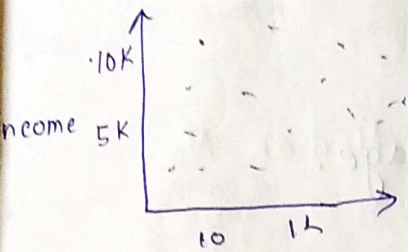
Information ~~thro~~ through histograms:-

① Density (crowding): The height of the bar tells how many subjects there are for one unit on the horizontal scale.

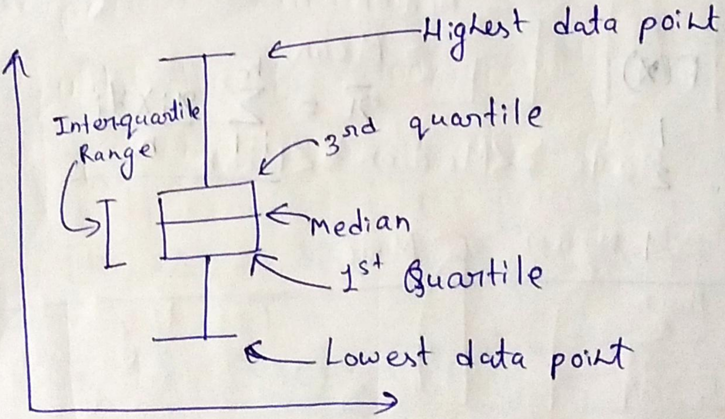
② Pecentages (Relative frequencies):
↓ Those are given by
area = height × width

③ ~~Boxplot~~ Scatterplot

The scatterplot is used to depict data that come as pairs.



④ Box and Wishker plot:-



* Measure of Central tendency:-

Central tendency refers to the measure used to determine the "center" of the distribution of data

~~① Mean:-~~

① Mean:-

It is basically the average.

(Notations)

Population Data (N)   Sample Data (n)

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

⇓

population mean

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Weighted Mean:-

X	F(X)
10	1
28	2
33	1

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i \times f_i}{n}$$

~~Types~~

Types of mean	Formula	eg:
Arithmatic mean	$\frac{\sum x}{n}$	Avg. Height
Geometric mean	$\sqrt[n]{\prod x}$	Calculation for Rate of interest
Harmonic mean	$\frac{n}{\sum(\frac{1}{x})}$	swim 1 lap at 3 Km/h and 1 lap at 2 Km/h. Calculate avg. speed.

2) Median:-

The median is the middle number of the series when ordered.

when, n = odd

[10, 28, 28, 33, 54] , $Med = X_{(n+1)/2}$

when, n = Even

[10, 28, 28, 32, 33, 54] , $Med = \frac{X_{(n/2)} + X_{(n/2+1)}}{2}$

Take avg (28, 32) = 30

③ Mode :-

It is the observation with the highest frequency

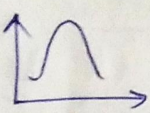
Eg:- [10, 28, 28, 33, 54]

Mod = 28

⟹ Relationship among Mean, Median, Mode.

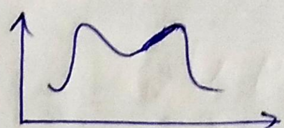
① Symmetric Distribution:-

ⓐ



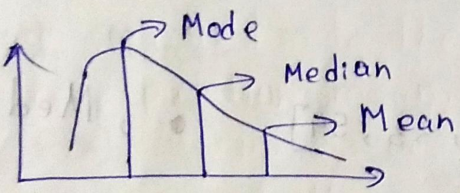
Mean = Median = Mode

ⓑ



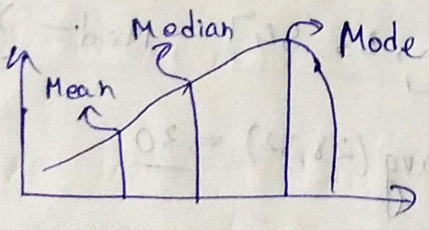
Mean = median, but bimod

② Asymmetric Distribution ~~(positi~~



(Positively sKewed)

Mode < Median < Mean



(~~Mo~~ Negetively sKewed)

~~Mod~~ ' Mean > Median > Mode

* Measure of Dispersions :-

1. Variance
2. Standard deviation.

These helps us to understand the spread of the data.

1. Variance :-
It gives us an actual value to how much the numbers in a dataset vary from the mean.

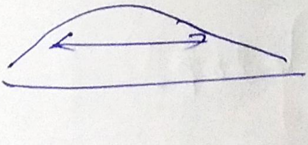
2. standard deviation :-
It measures how far apart the numbers are in a dataset.

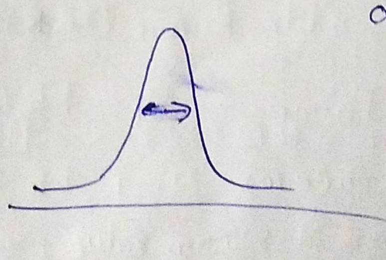
	Population (N)	Sample (n)
	$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$
Variance:	$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\boxed{n-1}}$ → Bessel's correction; Degree of freedom
S.d:	S.d = $\sqrt{\text{variance}}$	

Var → ~~Small~~ Big
S.d → "

Var → ~~Big~~ Small
S.d → "

Variance or S.d $\frac{1}{\alpha}$ height of graph
and α spread





Qus: Why Sample Variance is divided by n-1?

Our motive is too always estimate for the population (N) variance by sample variance.

But,

When we divide by n, the sample value around the sample mean... ...is always less than the value around the population mean...

$$\sum \frac{(x-\bar{x})^2}{n} < \frac{\sum (x-\mu)^2}{n}$$

(unless $\bar{x} \approx \mu$, which almost never happens)

So, if we underestimate the variation in the data around the population mean.

This is because the differences between the data and the sample mean. tend to be smaller than the differences between the data and the population mean.

To solve this problem, we divie by n-1 in sample variance ($\frac{\Sigma(x-\bar{x})^2}{n-1}$) to have a better estimation.

(Ques) Why we don't use $\Sigma|x-\bar{x}|$ instead of $\Sigma(x-\bar{x})^2$?

Since, $|x-\bar{x}|$ is not a differentiable function.

It is much harder to find the minimum value than ~~with~~ ~~w~~ in case of $(x-\bar{x})^2$

---

* Percentiles and Quartiles

→ Percentiles :- A percentile is a value below which a certain percentage of observations lie.

Eg:- 95 percentile means that the person has got better marks than 95% of the entire student population who appeared in the exam.

Percentile Ranking of x in dataset $= \dfrac{\text{No. of values below } x}{n} \times 100$

↑ No. of total values in the dataset.

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1)$$

if value = 5.25 ← (Index)

↓

$$\frac{\text{Data}[5] + \text{Data}[6]}{2}$$

→ Quartile :- Quartiles are 3 values that split sorted data into four parts, each with equal number of observations.

1^{st} Quartile (Q1) → Lower Quartile

Q1 = 25 percentile

2^{nd} Quartile (Q2) ⇒ Median

Q2 = 50 percentile

3^{rd} Quartile (Q3) ⇒ Upper Quartile

Q3 = 75 percentile.

→ <u>Five number system and Box plot :-</u>

1. Minimum
2. ~~First Q~~. Q_1
3. Median (Q_2)
4. Q_3
5. Maximum

⇒ Box plot → <u>Outlier detector</u>

To detect ~~outlie~~ outliers we ~~generall~~ generally define [Lower fence ↔ Higher fence]

→ Range = Maximum − Minimum

→ Inter Quartile Range (IQR) = $Q_3 - Q_1$

→ Lower Fence (LF) = $Q_1 - 1.5$ (IQR)

→ Higher Fence (HF) = $Q_3 + 1.5$ (IQR)

<u>Imp</u>

→ Maximum = Maximum value present in data < HF

→ Minimum = Minimum value present in data > LF

Probabilities

<u>Probability</u> :- It is the numerical description of how likely an event is to occur.

$$P(A) = \frac{\text{no. of favourable outcomes}}{\text{Total no. of outcomes}}$$

$$0 \leq P(A) \leq 1$$

- <u>Sample space</u> :- Set of all the possible outcomes
- Complement of event (A', $\bar{A}$ or A^c) :- ($\bar{A}$ = not A)

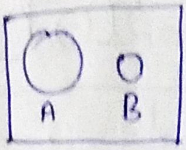
$$P(A) + P(\bar{A}) = 1$$

- Odds in favour of an event = $\frac{\text{No. of favourable cases}}{\text{No. of unfavourable cases}}$
- Odd's against an event = $\frac{\text{No. of unfavourable cases}}{\text{No. of favourable cases}}$
- Equally likely events : $P(A) = P(B)$

- **Mutually exclusive or disjoint events:-**

(Intersection)

$n(A \cap B) = 0$



- **Exhaustive set of event:-**

$\{A\} + \{B\} =$ Sample space

- **Independent events:-**

Events not depending on other events.

* Addition principle of probability:-

(Non-Mutually Exclusive)

① $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

② $P(A-B) = P(A) - P(A \cap B)$

For A, B, C three events

③ $P(A \cup B \cup C) = P(A) + P(B) + P(C)$
$- P(A \cap B) - P(A \cap C) - P(B \cap C)$
$+ P(A \cap B \cap C)$

④ P(Exactly two of the events).
$= P(A \cap B) + P(B \cap C) + P(C \cap A) - 3P(A \cap B \cap C)$

⑤ P(Exactly one of the events)
$= P(A) + P(B) + P(C) - 2P(A \cap B) - 2P(B \cap C) - 2P(A \cap C)$
~~+3(PA~~ $+3P(A \cap B \cap C)$

→ If A, B and C are mutually Exclusive events:-

⑥ $P(A \cup B) = P(A) + P(B)$

⑦ $P(A \cup B \cup C) = P(A) + P(B) + P(C)$

* Conditional Probability :-

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

A happening given B already happend

→ Multiplication principle of probability :-

① $P(A \cap B) = P(A) \cdot P(B/A)$

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

Addition principle → Union of sets
Multiplication principle → Intersection of sets

If A, B and C are independently events:-

$\boxed{P(A/B) = P(A)}$

③ $P(A \cap B) = P(A) \cdot P(B)$

④ $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

- Total Probability theorem:-

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)$$

- Baye's Theorem:-

$$P(B_i/A) = \frac{P(B_i) \cdot P(A/B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A/B_i)}$$

Note

Prior	Event	Posterior
$P(A)$	E	$P(A/E)$

* Random Variables:-

→ Random Variables:-

Random variable is a real valued function which assigns a real number to each sample point in the sample space.

→ Types of Random Variables:-

- Categorical R.V. (Gender of a person)
- Numerical R.V.
  - Discrete R.V. (No. of people in family)
  - Continuous R. (Salary, loan et

This function of random variable maps the output of a random process or experiment to a number.

---

* Probability Distribution Function:-

Probability Distribution Function gives the probability of ~~the~~ all the possible outcomes of any random variable.

Two functions are used to describe the probability Distribution of the function.

Probability Distribution Function

Probability Density Function (PDF) ⇑ Continuous R.V.

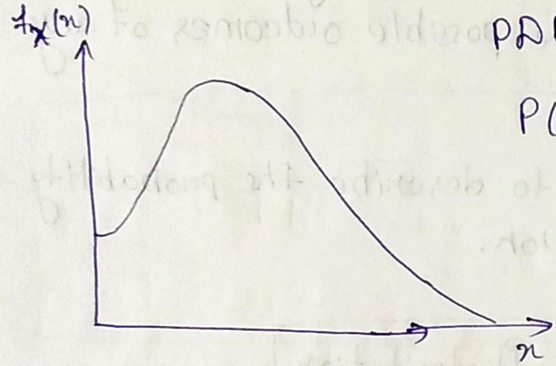
Probability Mass Function (PMF) ⇑ Dicrete R.V.

→ Probability Density Function (PDF) :-

(Continuous R.V.)

Probability Density function (PDF) is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value.

This probability is given by the integral of this variable's PDF over the range.



PDF @ $f_X(x)$

$P(a<X<b) =$ area under $f_X(x)$

Note :-

$f_X(x)$ needs to satisfy:-

- It is defined for all numbers
- $f_X(x) > 0$
- Area under $f_X(x) = 1$

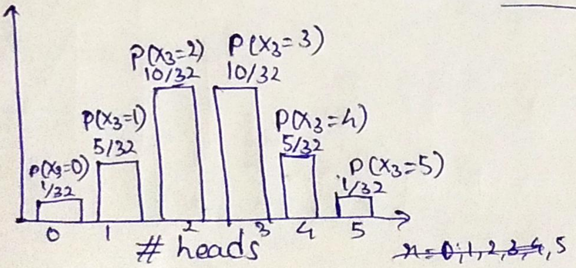
→ Probability Mass Function:- (PMF) (Discrete R.V.)

Probability Mass Function is a function that gives the probability that a Discrete Random Variable is exactly equal to some value. Sometimes it is also Known as the discrete probability density function.

The PMF is often the primary means of defining a discrete probability distribution.

$\boxed{PMF : P_X(x)} = P(X_n = x), x = 0,1,2,3,4$

Eg:-



Note :-

It must satisfy:-

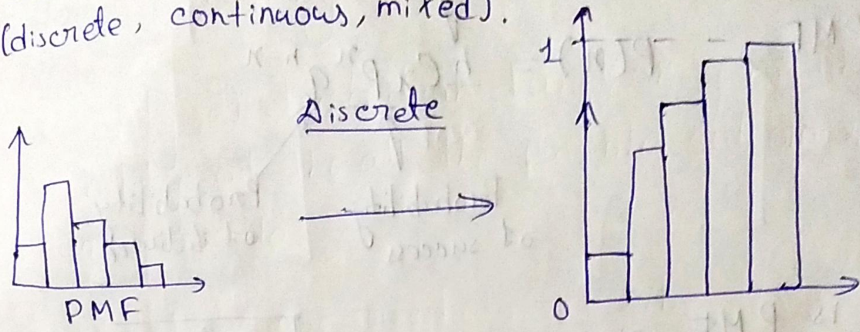
- $P_X(x) \geq 0$
- $\sum_x P_X(x) = 1$

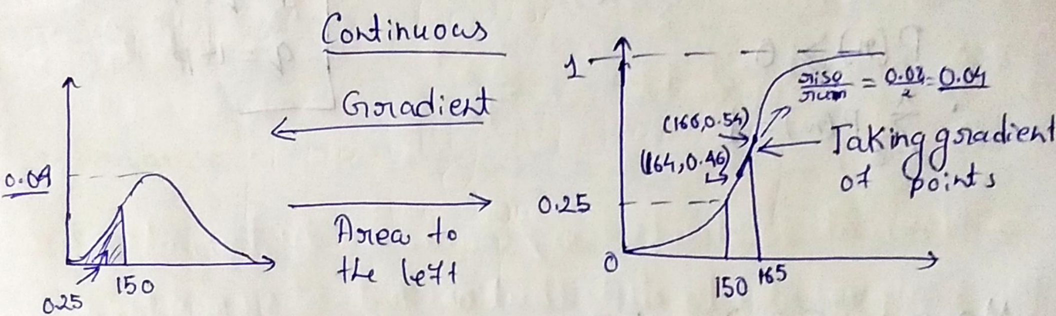
Cumulative Distribation Function: (CDF)

→ ~~Cumm~~. Cumulative

The cumulative Distribution Function (CDF) of a random variable is another method to describe the distribution of random variables.

The advantage of CDF is that it can be defined for any kind of random variable (discrete, continuous, mixed).





The CDF shows how much probability the variable has accumulated until a certain value.

That means that, $F_X(x) = P(X \leq x)$ ← It is defined for every real number

Properties

- $0 \leq F_X(x) \leq 1$
- Left "endpoint" is "0"
- Right "endpoint" is "1"
- Never decreases
- (End points can be infinity)

→ Binomial Distribution :- (Many trials)

1. All the trials are independent
2. Number of n trials is finite
3. The probability p of success is same for each trial.

It's PMF, $P(x) = \overset{\text{no. of trials}}{n} C_x \underset{\text{Probability of success}}{p^x} \underset{\text{Probability of failure}}{q^{n-x}}$

P(x) is PMF, since,

$\boxed{\begin{matrix} p+q=1 \\ q=1-p \end{matrix}}$

- $P(x) \geq 0$
- $\sum_{x=0}^{n} P(x) = 1$

- Mean: $E(x) = np$
- Variance: $E(x^2) - (E(x))^2 = npq = \sigma_x^2$
- SD: $\sqrt{npq}$

→ Bernoulli Distribution :- (Only One trial)

Bernoulli distribution is a discrete probability distribution.

It describes the probability of achieving a "sucess" or "failure" from a Bernoulli trial.

A Bernoulli trial is an event that has only two possible outcomes (success or failure).

It's PMF, $P(x) = p^x(1-p)^{(1-x)}$

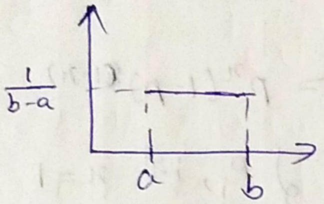
$$P(x) = \begin{cases} p, & \text{if } x=1 \\ 1-p, & \text{if } x=0 \end{cases}$$

- Mean: $E(x) = p$
- Variance : $E(x^2) - (E(x))^2 = p(1-p) = pq$
- SD : $\sqrt{pq}$

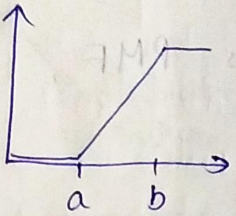
→ Rectangular or Uniform Distribution:-

A continuous random variable x is said to follow a continuous uniform or rectangular distribution over a interval (a, b), if its PDF is given by,

$$f(x) = \begin{cases} K, & a < x < b \\ 0, & \text{otherwise} \end{cases} \quad \boxed{K = \frac{1}{b-a}}$$



CDF →

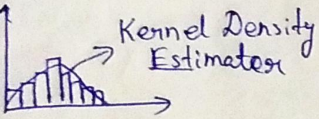


- Mean: $(E(x)) = \frac{a+b}{2}$
- Variance: $E(x^2) - (E(x))^2 = \frac{(a-b)^2}{12}$
- SD: $\sqrt{\frac{(a-b)^2}{12}}$

$$F_X(x) = \begin{cases} 0, & x < 0 \\ \frac{x-a}{b-a}, & 0 \leq x \leq 1 \\ 1, & 1 \leq x \end{cases}$$

→ Normal Distribution / Gaussian Distribution:-

It is a probability distribution that is symmetric about mean, showing the data near the mean are more frequent in occurrence than data far from the mean.



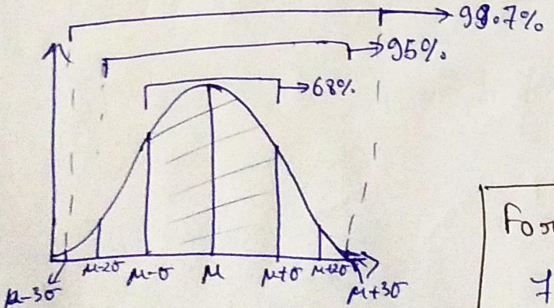
Empirical Rule

↓

{68 - 95 - 99.7}

↑

3-sigma Rule



Example:- Height, weight

Formula:-

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

→ Log-normal distribution :-

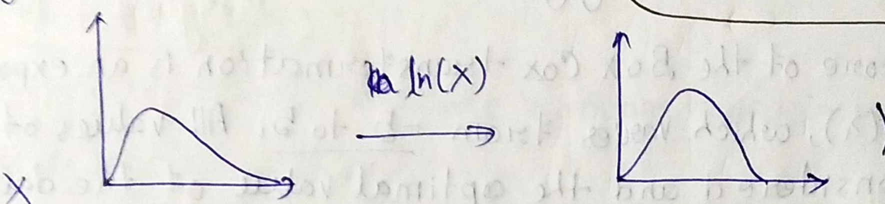
is a continuous probability distribution

A Log-normal distribution, of a random variable whose logarithm is normally distributed.

Thus, if the random variable X is log-normally distributed, then Y = ln (X) has a normal distribution.

Equivalently is Y has a normal distribution then the exponential function of Y, X = exp (Y), has a log normal distribution.

Eg :- Distribution of wealth in world

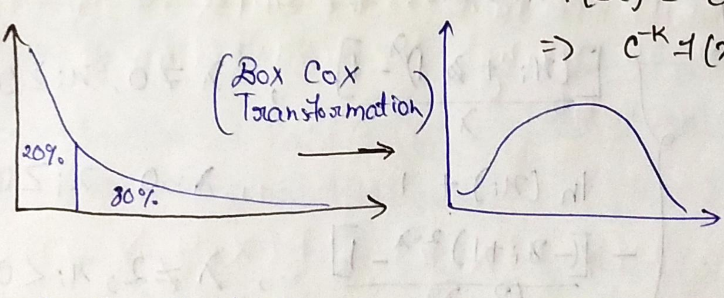


→ Power Law Distribution & Petro Distribution :-

Power Law Distribution is a functional relationship between two quantities, where a relative change in one quantity results in a relative change in the other quantity proportional to a power of the change, independent of the initial size of those quantities: One quantity varies as the power of another.

$$f(cx) = a(cx)^{-K}$$

$$\Rightarrow c^{-K} f(x) \propto f(x)$$



Eg :- 20% of team winning in 80% of matches

Pareto Distribution is a power-law probability distribution that is used in description of social, quality control, geophysical and many other types of observable phenomena.

(Survival function also tail function) $\bar{F}(x) = \Pr(X > x) = \begin{cases} \left(\frac{x_m}{x}\right)^{\alpha}, & x \geq x_m \\ 1, & x < x_m \end{cases}$

x_m = scale parameter

α = shape parameter or tail index (Pareto index in wealth distribution model)

⇒ Box Cox Transformation:-

A Box Cox transformation is a transformation of non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log(y), & \text{if } \lambda = 0. \end{cases} \quad \Big| \; y \geq 0$$

At the core of the Box Cox transformation is an exponent, lambda (λ), which varies from -5 to 5. All values of λ are considered and the optimal value of the data is selected.

The "Optimal Value" is the one which results in the best approximation of a normal distribution curve.

→ Yeo-Johnson Transformation:-

This transformation is somewhat of an adjustment to the Box-Cox transformation, by which we can apply it to negative numbers.

$$x_i^{(\lambda)} = \begin{cases} \frac{[(x_i + 1)^{\lambda} - 1]}{\lambda}, & \lambda \neq 0,\ x_i \geq 0 \\ \ln(x_i) + 1, & \lambda = 0,\ x_i \geq 0 \\ -\frac{[(-x_i+1)^{2-\lambda} - 1]}{(2-\lambda)}, & \lambda \neq 2,\ x_i < 0, \\ -\ln(-x_i+1), & \lambda = 2,\ x_i < 0 \end{cases}$$

* Some more Important probability distributions :-

→ Geometric Distribution :-

→ The geometric Distribution is a probability distribution that models the number of Bernoulli trials needed to achieve the first success.

→ It is applicable in situations where there are only two ~~people~~ possible outcomes in each trial, often referred as success and failure.

→ The geometric distribution is memoryless, meaning that the outcome of previous trials does not affect the probability of success in subsequent trials.

It's, PMF, $P(x) = (1-p)^{x} \times p$

$= q^{x} \times p$

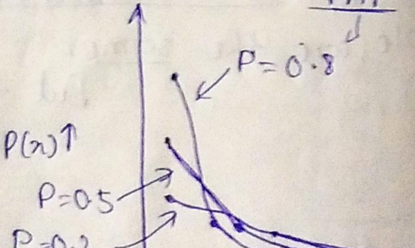
p = No. of success

$q = 1-p$ = No. of failures

x = No. of trials to get the first success

- Mean: $E(x) = \frac{1}{p}$
- Variance: $Var(x) = \frac{1-p}{p^2} = \frac{q}{p^2}$
- SD : $\sqrt{q/p^2}$

PMF



, Y) is called

✱ Hypergeometric Distribution :-

→ The hypergeometric distribution is a probability distribution that describes the number of successes in a sequence of draws without replacement from a finite population containing a specific number of success and failures.

→ This distribution is used when the outcomes of the draws are not independent, as they are in the binomial distribution, because the sample size is not fixed and decreases with each draw.

It's PMF, $P(x) = \dfrac{{}^{K}C_{x} \times {}^{N-K}C_{n-x}}{{}^{N}C_{n}}$, where,

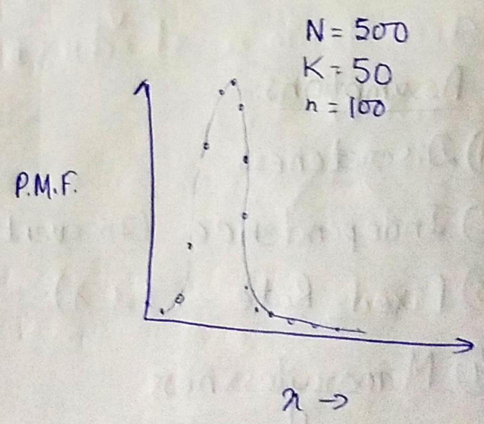
N = No. of all the items in the population

K = The total no. of items in the population that are classified as success

n = The no. of items drawn from the population without replacement.

- Mean: $E(x) = \dfrac{nK}{N}$
- Variance: $Var(x) = \dfrac{nK(N-K)(N-n)}{N^2(N-1)}$
- SD: $\sqrt{Var(x)}$

$= \sqrt{\dfrac{nK(N-K)(N-n)}{N^2(N-1)}}$



→ Poisson Distribution :-

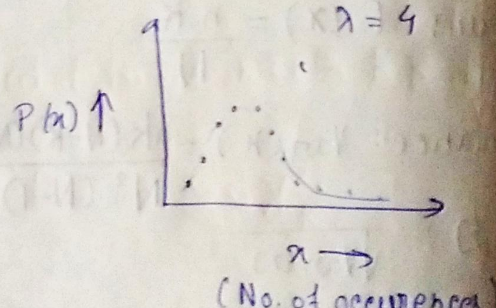
→ The poisson Distribution is a discrete probability distribution that describes the number of events that occur within a fixed interval of time or space.

→ The poisson distribution is often used to model rare events where the avg. rate of occurance is known, but the exact timing on occurence of events is random.

→ The poisson distribution has only one parameter, denoted by λ (lambda), which represents the average rate of event occurence in the given interval.

It's PMF ; $P(x) = \dfrac{e^{-\lambda} \cdot \lambda^{n}}{n!}$, where, n = No. of events

- Mean: $E(x) = \lambda$
- Variance: $Var(x) = \lambda$



Assumptions:-

① Discreteness
② Independence (In event happening)
③ Fixed Rate (of λ)
④ Memorylessness

called a

→ Exponential Distribution :-

- The exponential distribution is a continuous probability distribution that models the time between independent events occurring at a constant rate.

→ It is commonly used to describe the time until the next event in a sequence of events where the events occur continuously and indepently at a constant average rate.

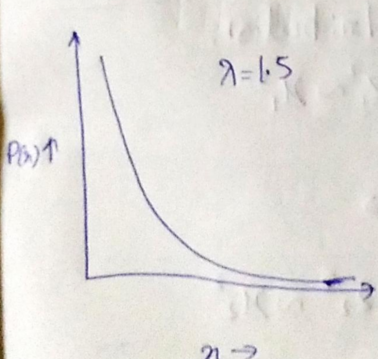
→ The exponential distribution has a single parameter, denoted by λ, which represents the rate of event occurences.

It's PDF : $\begin{cases} \lambda e^{-\lambda x} & , \text{for } x \geq 0 \\ 0 & , \text{for } x < 0 \end{cases}$

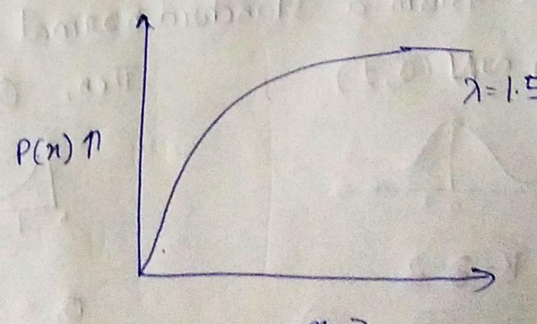
It's CDF : $\begin{cases} 1 - e^{-\lambda x} & , \text{for } x \geq 0 \\ 0 & , \text{for } x < 0 \end{cases}$

· Mean: $E(x) = \frac{1}{\lambda}$

· Variance: $Var(x) = \frac{1}{\lambda^2}$



~~PMF~~ PDF



CDF

→ Chi-squared distribution:

→ The chi-squared distribution is a continuous probability distribution that arises in statistics and probability theory.

→ It is used in hypothesis testing, confidence interval estimation etc.

→ The chi-squared distribution is denoted by as χ^2.

→ It has a ~~parameter~~ parameter called degrees of freedom (ν)

→ If $Z_1, \ldots, Z_K$ are independent, standard normal random variables, the sum of their squares,

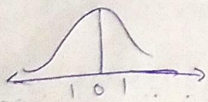
$$Q = \sum_{i=1}^{K} Z_i^2 ,$$

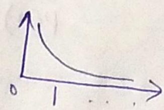
is distributed like the chi-squared distribution with K degrees of freedom.

This is denoted as, $Q \sim \chi^2(K)$ or $Q \sim \chi^2_K$.

→ Also, it can be thought of as the "square" of a selection taken from a standard normal distribution.

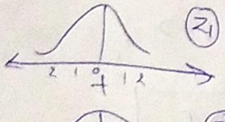
Eg:- $Z_1 \sim N(0,1)$, Then, $Q = Z_1^2 \sim \chi_1^2$

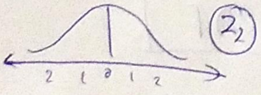




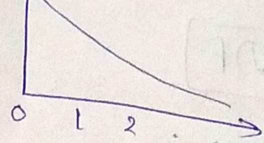
If $K = 2$

$Q = Z_1^2 + Z_2^2 \sim \chi_2^2$





=



It's PDF: $P(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} \times e^{-\frac{x}{2}}, & \text{for } x > 0 \\ 0, & \text{for } x \leq 0 \end{cases}$

where, $\nu \rightarrow$ degrees of freedom

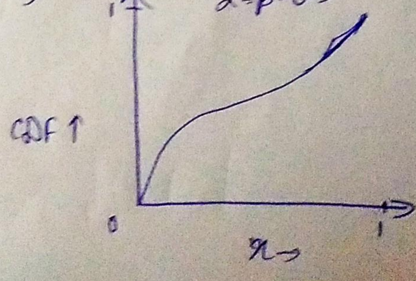
$\Gamma(.) \rightarrow$ gamma function

- Mean : $E(x) = \nu$
- Variance : $Var(x) = 2\nu$

→ Beta Distribution:-

- The Beta distribution is a continuous probability distribution defined on the interval $[0, 1]$.
- It is a versatile statistics and is often employed as a prior distribution for probabilities. [like for binomial and Bernoulli distribution]
- It's PDF : $P(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$, where $B(\alpha, \beta)$ is the beta function. α, β are (+ve) shape parameters

- Mean: $E(x) = \frac{\alpha}{\alpha+\beta}$
- Variance : $Var(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$



→ Gamma distribution :-

→ The Gamma distribution is a continuous probability distribution that generalizes the exponential distribution.

→ It is used in various fields, including reliability engineering, queuing theory, and statistical modeling of waiting times.

It's PDF, $f(x;\alpha,\beta) = \dfrac{x^{\alpha-1} \times e^{-x/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}$, where,

$x \rightarrow$ Random variable

$\alpha \rightarrow$ (+ve) shape parameter

$\beta \rightarrow$ (+ve) scale parameter

$\Gamma(\alpha) \rightarrow$ Gamma function

- Mean : $E(x) = \alpha\beta$
- Variance : $Var(x) = \alpha\beta^2$