Introduction to Statistics

Defn : Stats is the science of collecting, organizing and analyzing data.

Data : "facts or pieces of Information"

Eg : Height of students in classroom

No. of sales in term of revenue of a company

Salary of people in Society.

IQ of students in Classroom.

Types of Statistics

① Descriptive Stats

Defn : It consists of organizing, summarizing and visualizing data.

① Measure of Central Tendency [Mean, Median, Mode]

② Measure of Dispersion [Variance, Standard deviation, Z-score].

③ Different types of Distribution of data

Eg: Histogram, pdf, pmf, Gaussian Distr, Log Normal Dist, Exponential, Binomial, Bernoulli Distn, Poisson } EDA.

Boxplot, Violin plot, Histplot, Distribution plot

② Inferential Stats

Defn : It consists of using data you have measured to form Conclusion.

1) z-test
2) t-test
3) Chi square Test
4) Anova Test

Hypothesis Testing, Pvalue, Significance Value.

And [Feature Engineering]

EDA ⟶ Exploratory Data Analysis.

Eg: Let say there are 10 cricket camps In Bangalore. And you have Collected the height of cricketers from one of the camp.

Ans) Heights are recorded [175cm, 180cm, 140cm, 140cm, 135cm, 160, 135].
↓ Sample

Descriptive Question : [What is the average height of the entire classroom]?

[Distribution of a DATA]

→ [140cm how many Std it is away from mean] → Z-score.

Inferential Question

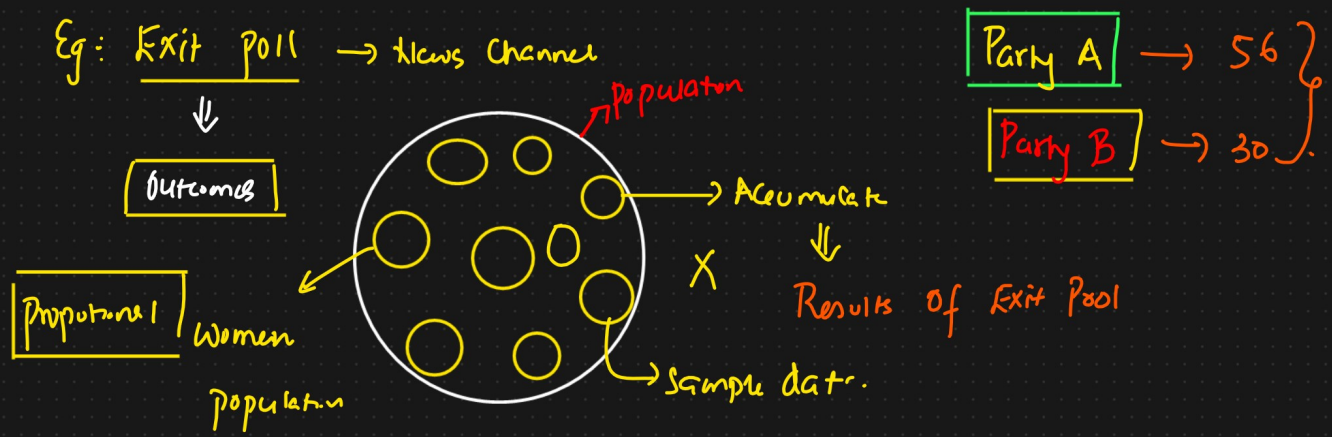
Are the average height of plAyers of camp 1 similar to that of camp2

① Population And Sample Data

Population (N) : Population is a group or a superset of data that you are interested in studying.

Sample (n) : A Sample is a subset of population data.

Eg: Exit Poll → News Channel



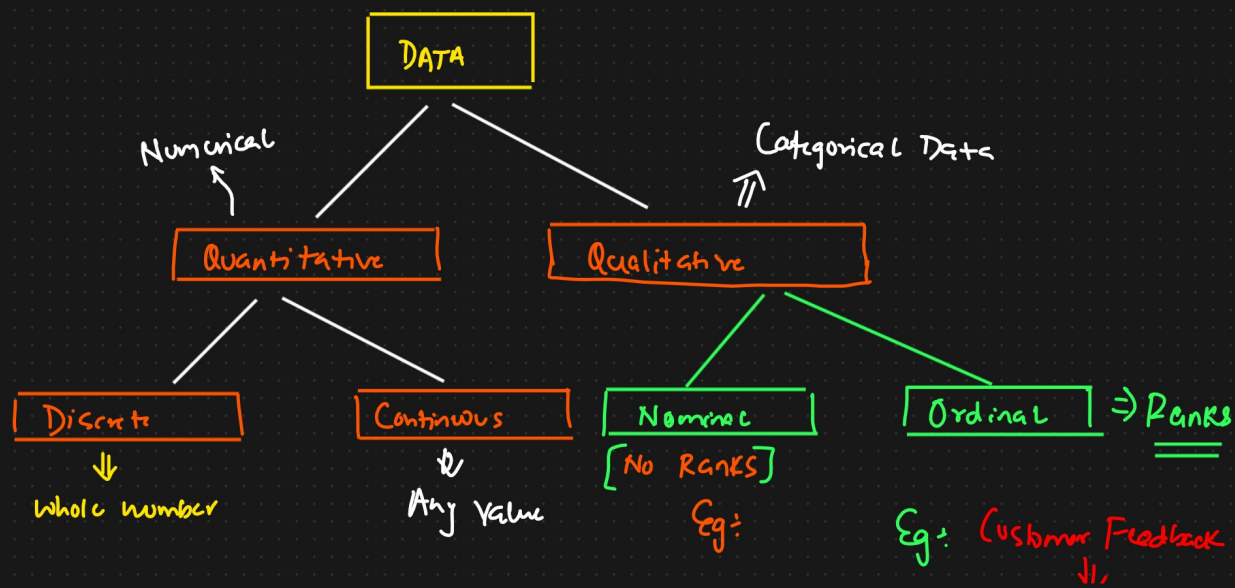
Survey ÷ DATA SCIENCE ⇒ Tech

↳ Farmers

Eg: Types of Sampling Techniques → Assignment

③ Types of Data [House Price Data].

Size	Area	No. of Room	Location	Price
1500	Posh	3	New York	—
—	—	—	Bangalore	—
			New Delhi	



Eg: No. of children in a family.

Eg: Average Rainfall

Eg: Gender, M, F

{1, 2, 3, 4, 5}.

Blood group

Weekdays

No. of accounts in banks

House price In Bangalore

Colors

No. of bikes

Length of River

Countries

Location

Cities.

DAYS

Pincode.

No. of people working in a company

Example Assignment

1. Length of Diffcient Rivers In the World?
2. Favorite Food Based On Gender?
3. Marital Status?
4. Iq measurement?
5. Weekdays For IT Employees?
6. Weekdays For KRISH?

Statistics

Agenda

1. Scale of Measurement ✓
2. Measure of Central Tendency ✓
3. Measure of Dispersion ✓
4. Random Variables ✓
5. Sets ✓
6. Histogram And Skewness.

1. Scale of Measurement

1. Nominal Scale Data ✓
2. Ordinal Scale Data ✓
3. Interval Scale Data ✓
4. Ratio Scale Data

1. Nominal Scale Data

1) Qualitative / Categorical Data
2) Eg: Gender, Color, Labels
3) Order or Rank does not matter

Eg: Favorite Color

Red → 5 → 50%.

Yellow → 3 → 30%.

Orange → 2 → 20%.

10

Gender

2. Ordinal Scale Data

i) Rank is important
2) Order matters
3) Difference cannot be measured

Eg: Race [100 mts]

→ 1st
→ 2nd
→ 3rd

{ 1st, 2nd, 3rd. }

IQ

③ Interval Scale Data

Eg: Temperature Variable

① The order matters ✓ [Rank].

② Difference can be measured ✓

③ The Ratio cannot be measured ✓

④ No 0 starting point

30F → 60F

60 : 30 = 2 : 1

2 : 1

0F ✗

0° C

9°F

120F

④ Ratio Scale Data

Eg: Students marks in a class

① The order matter

② Differences are measurable (Ratio)

③ Contain a "0" starting point.

0, 90, 60, 30, 75, 40, 50

3 : 2

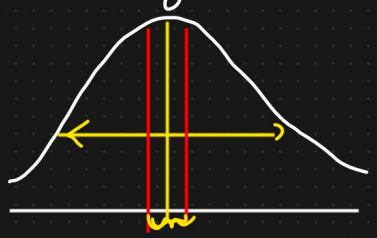
3 : 1

② Measure of Central Tendency [EDA] ⇒ Exploratory Data Analysis

① Mean or Average

② Median

③ Mode.



① Mean

Population (N)

$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$

Population mean (μ) $= \frac{\sum_{i=1}^{N} x_i}{N}$

$= \frac{[1+1+2+2+3+3+4+5+5+6]}{10}$

$= \frac{32}{10} = 3.2$

Sample (n)

$X = \{2, 3, 1, 4, 8, 9, 5\}$

(Sample mean $\bar{x}$) $= \frac{\sum_{i=1}^{n} x_i}{n}$

$= \frac{32}{7} = 4.7$

② Median

$X = \{4, 5, 2, 3, 2, 1\}$.

① Sort the Random Variable

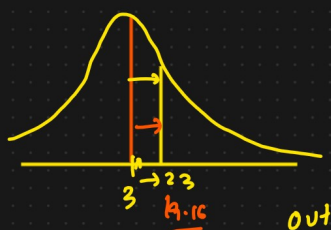
② No. of Elements
- If n = even ⇒ Middle 2 element ⇒ Average
- If n = odd ⇒ Central Element

$X = \{1, 2, \boxed{2, 3}, 4, 5\}$

$\frac{2+3}{2} = 2.5$

$\{1, 2, 3, 4, 5\}$

Mean = 3 | Median = 3



$\{1, 2, \boxed{3, 4}, 5, 100\}$ — outlier

Mean = 19.16 | Median = 3.5

$\frac{1+2+3+4+5+100}{6} = \frac{115}{6} = 19.6$

TRANSACTION FRAUDLENT {Outliers}.

$X = \{2, 8, 4, 5, 1, 7, 9, 120, 130\}$.

Mean And Median

$\{1, 2, 4, 5, \boxed{7}, 8, 9, 120, 130\}$ — outlier

Median

$\text{Mean} = \frac{1+2+4+5+7+8+9+120+130}{9} = 31.77$

$\frac{37}{7} = 5.1$

Median = 7

Outliers { Median → Calculate Central Tendency }.

$x = \{-5, -10, 1, 2, 3, 4, 5\}$. $\Rightarrow \{-10, -5, 1, 2, 3, 4, 5\}$

Mean : 0 Mean = 3 Median

③ Mode : Frequency of maximum occuring elements.

$x = \{2, 1, 1, 1, 4, 5, 7, 8, 9, 9, 10\}$.

Mode : 1.

EDA And FE

→ Missing Value or NULL Values

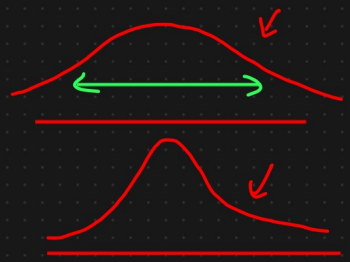
Age	Weight	Salary	Gender	Degree
24	70	40K	M	BE
25	80	70K	F	—
27	95	45K	F	—
24	—	50K	M	PHD
32	—	60K	— {Mode}	B.E
— (Mean)	60	—	—	Masters
— (Mean)	65	55K	—	BSC
40	72	—	M	B.E
Median	150			

③ Measure of Dispersion [Spread of the data].

① Variance

② Standard Deviation

① Variance



Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

{Spread}

$x_i \rightarrow$ DATA Points

$\mu \rightarrow$ population mean

$N \rightarrow$ Population Size

Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$x_i \rightarrow$ DATA POINTS

$\bar{x} \rightarrow$ Sample Mean

$n \rightarrow$ Sample Size

Assg

Why we divide sample variance by n-1? {Bessel's Correction}.

Ans) The sample variance is divide by n-1 so that we can create an unbiased estimator of the population variance

Population Std

$$\sigma = \sqrt{\text{Variance}}$$

Sample Std

$$s = \sqrt{\text{Variance}^{\text{sample}}}$$

How many std x_i is away from Mean

Eg: {1, 2, 3, 4, 5}

$\mu = 3$

Population Variance = $\sigma^2 = \frac{[(1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2]}{5}$

$= \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2$

$\sigma = \sqrt{2} = 1.41$

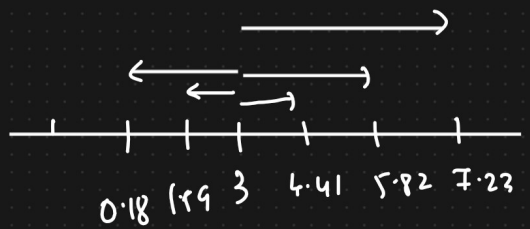
$3 - 1.41 = 1.59$

$1.59 - 1.41 = 0.18$

$3 + 1.41 = 4.41$

$4.41 + 1.41 = 5.82$

$5.82 + 1.41 = 7.23$



④ Random Variables

Variables (x, y)

$x + 5 = 7$

$x = 2$

$x = 2$

$y = 6$

$8 = y + x$

Expenment ⇑

Random Variable is a process of mapping the output of a random process or experiment to a number.

Eg: Tossing a Coin

$X = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$

Rolling a dice 6 times

$Y = \{2, 1, 4, 2, 3, 5\}$

$Y = \{$ Sum of Rolling a dice 7 times $\}$.

$Pr(y \geq 15)$ $\quad$ $Pr(y < 10)$

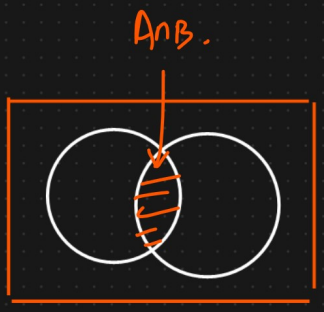
⑤ Sets

$A = \{1,2,3,4,5,6,7,8\}$

$B = \{3,4,5,6,7\}$

① Intersection

$A \cap B = \{3,4,5,6,7\}$



② Union

$A \cup B = \{1,2,3,4,5,6,7,8\}$

③ Difference

$A - B = \{1,2,8\}$

④ Subset

$A \rightarrow B \Rightarrow$ False

$B \rightarrow A \Rightarrow$ TRUE

⑤ Supreet

$A \rightarrow B \Rightarrow$ TRUE

$B \rightarrow A \Rightarrow$ Fact.

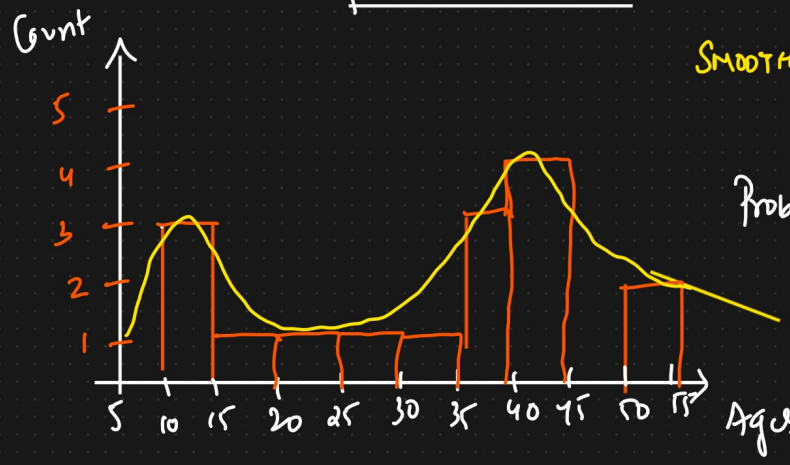
Histograms And Skewness → [Frequency]. => Distribution

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}.

Bins, Bin Size

$\frac{50}{5} = \boxed{10} \Rightarrow$ No. of bins

Bins size = 5



SMOOTHEN THE HISTOGRAM [Kde]

⇓

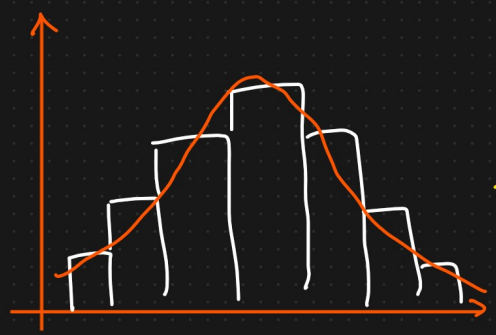
Probability distribution fn

⇓

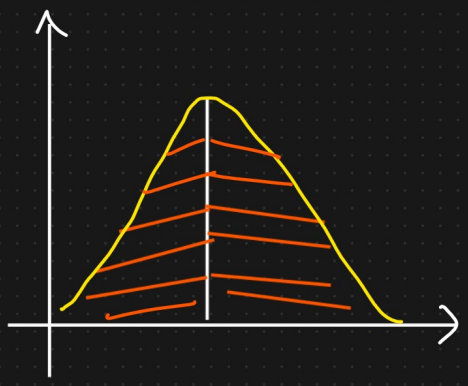
Probability density fn

Skewness

Normal/Gaussian Distribution



⇒ Symmetrical Distribution

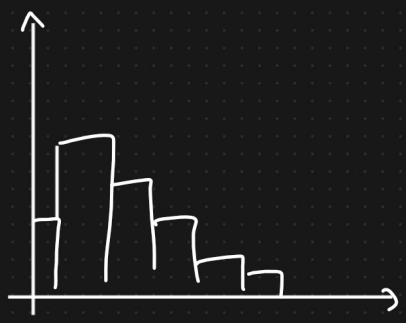


① No Skewness

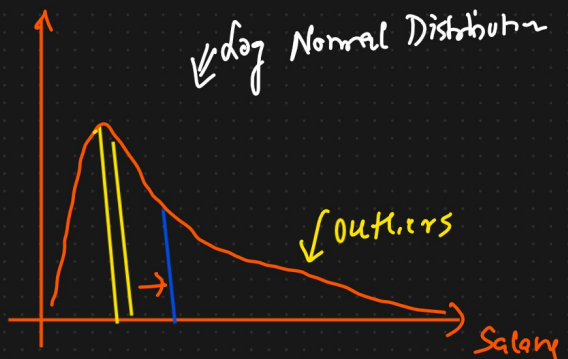
Mean, Median, Mode

Mean = Median = Mode

② Right skewed



⇒ Positive Skewed ⇒



100, 200, 300, 400, 100000

mean > median > mod

③ Left skewed

⇒ Negative Skewed

Relationship : mean ≤ median ≤ mode.

Statistics

Agenda

1. Sampling Techniques ✓
2. Covariance And Correlation
3. Probability distribution function
4. Probability density function
5. Probability Mass function
6. Cummulative Density Function
7. Types of Distribution

1. Sampling Techniques.

1. Random Sampling : Simple Random Sampling gives each member of the population an equal chance of being chosen for the sample.

Eg: Vaccination Test → 100 → Randomly Select Couple

Accidental Test → 1000 → A → 2 vehicle

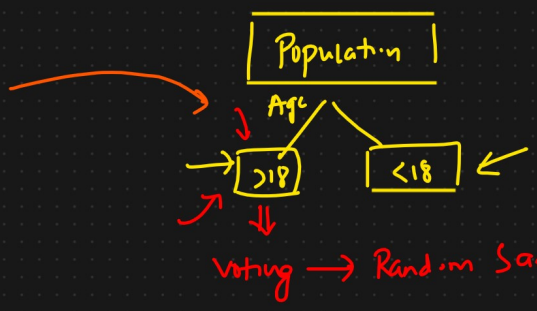
Exit Poll

Average IQ of the school → Select 10 people ⇒

2. Stratified Sampling [Stratified → layers]

Stratified Sampling Involves dividing the population into sub population that may differ in important ways

Eg: Exit Pool

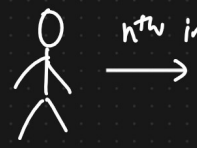


2. Salary

③ Systematic Sampling : Systematic Sampling is a statistical method involving the Selection of elements from an ordered sampling frame.

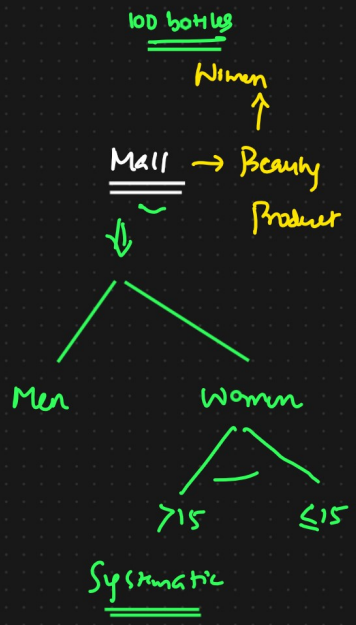
Eg: Airport

A Credit Card Company



nth individual → 7th

Mall → Fill up a form for a cause.

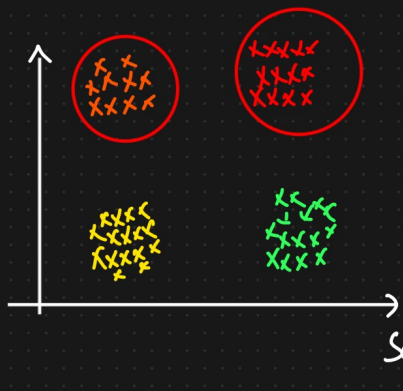


④ Convenience Sampling Assignment

⑤ Purposive Sampling → Judgemental Sampling is a method where researchers decides which members of the target population will be sampled.

⑥ Cluster Sampling

Mall iphone 28 → 100 → 20%.

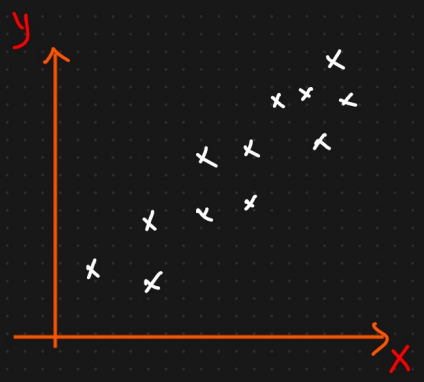


② Covariance And Correlation

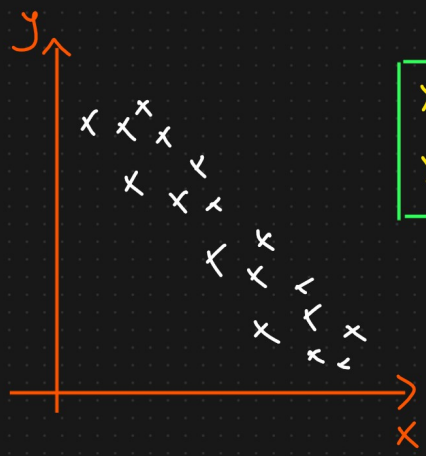
Quantify
[Relationship between X And y].
⇓
Covariance And Correlation.

X	y
2	3
4	5
6	7
8	9

X↑	y↑
X↓	y↓



X↑	y↑
X↓	y↓



X↑	y↓
X↓	y↑

Covariance

$$Cov(x,y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$Var(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

x_i → Data point of x
$\bar{x}$ → Sample mean of x
y_i → Datapoints of y
$\bar{y}$ → Sample mean of y.

$Var(x) = Cov(x,x)$ ⇒ Spread

$Cov(x,y)$

X↑	y↑
X↓	y↓

+ve Covariance

X↑	y↓
X↓	y↑

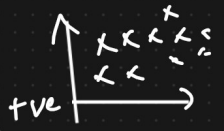
-ve Covariance

X	y
→ 2	3
→ 4	5
6	7
$\bar{x}=4$	$\bar{y}=5$

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}_i)}{n-1}$$

$$= \frac{[(2-4)(3-5) + (4-4)(5-5) + (6-4)(7-5)]}{2}$$

$$= \frac{4+0+4}{2} = 4 //.$$



X	y
8	6
7	5
6	4
3	1
2	0
$\bar{x}=5.2$	$\bar{y}=3.2$

+ve Covariance

X	y
8	6
9	5
10	4
12	2
11	1
10	3.6

−ve Covariance

−3

6.7

Advantages

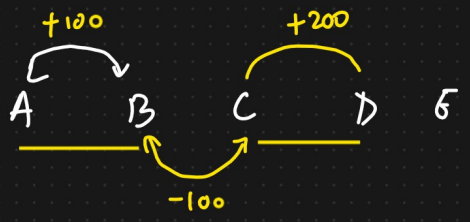
① Relationship between X and y

+ve or −ve value

A ⟷ B + infinity, − infinity.

Disadvantages

① Covariance does not have a specific limit value



② Pearson Correlation Coefficient [-1 to 1]

$$\rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y} \Rightarrow -1 \text{ to } 1$$

① The more the value towards +1, the more +ve correlated X, Y is.

② The more the value towards -1, the more -ve " X, Y is

X	Y
8	6
9	5
10	4
12	2
11	1
$\bar{X} = 10$	$\bar{Y} = 3.6$

$\sigma_x = 1.58$ $\sigma_y = 2.073$

$$\text{Variance} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$\sigma = \sqrt{\text{Variance}}$

$Cov(x,y) = -3$

$$\rho_{(x,y)} = \frac{-3}{1.58 * 2.073} = \frac{-3}{3.27} = -0.917 \Rightarrow \text{Negative Correlated} \Rightarrow 91.7\%.$$

$$\frac{[4+1+0+4+1]}{4} = \frac{10}{4} = 2.5$$

$\sigma = \sqrt{2.5}$

X↑ Y↓ => 91.7%

1 unit X↑ Y↓ 0.917.

③ Spearman Rank Correlation [-1 to 1].

$$r_s = \frac{Cov(R(x), R(y))}{\sigma(R(x)) * \sigma(R(y))}$$

X	y	R(x)	R(y)
1	2	2	2
3	4	3	3
5	6	4	4
7	8	5	6
0	7	1	5
8	1	6	1

Ascending Order.

{ Non linear Relationship }

Spearman Rank Correlation

⇓

Non linear Relationship

DATA SCIENCE

Feature Selection

Housing Price DATASET

Size of House ↑ — No. of Rooms ↑

95%

-95%

+ve or -ve Correlation

-ve correlation

Haunted

≈0

No. of people staying

Price ↑

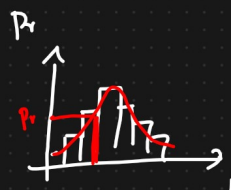
O/P

Statistics

1. Probability Distribution function
2. Probability Density function (pdf)
3. Probability mass function (pmf)
4. Different types of distribution
5. CDF {Commulative distribution fn}.

X = {Continuous} or Discrete Values

{Distribution of Data}.



Probability Distribution function

Discrete Value → No. of Bank Account

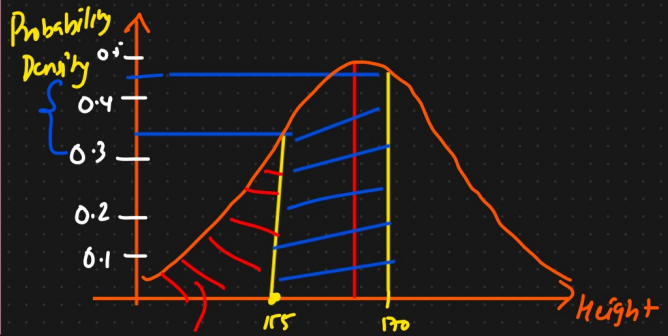
Probability density function

Probability mass function

1. Probability Density function

1. Continuous Random Variable

Eg: Height of Students in classroom



Formula.

$Pr(H \leq 155)$ ⇒ Formula

$Pr(H \geq 155 \text{ and } H \leq 170) =$

↓ 165

Area under this curve

② Probability Mass Function (pmf)

① Discrete Random Variable

Eg: Rolling a Dice $\{1,2,3,4,5,6\}$

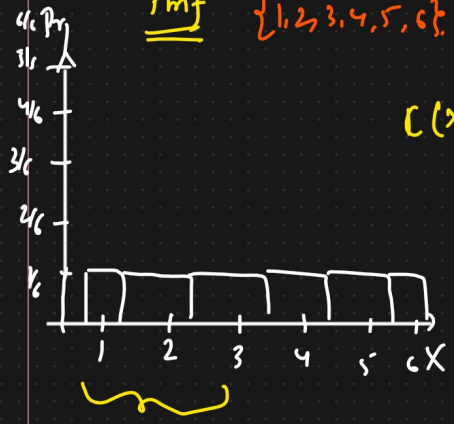
$Pr(1) = 1/6$ $Pr(3) = 1/6$

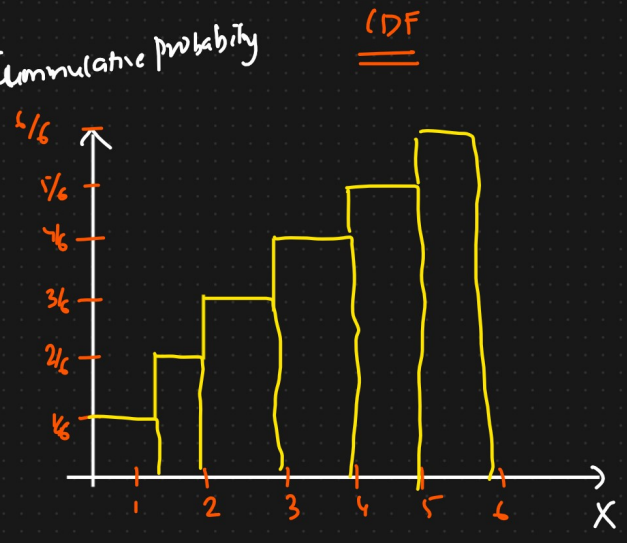
$Pr(2) = 1/6$ $Pr(4) = 1/6$.

$$Pr(X \leq 4) = Pr(X=1) + Pr(X=2) + Pr(X=3) + Pr(X=4)$$
$$= 1/6 + 1/6 + 1/6 + 1/6 = 4/6 = 2/3.$$

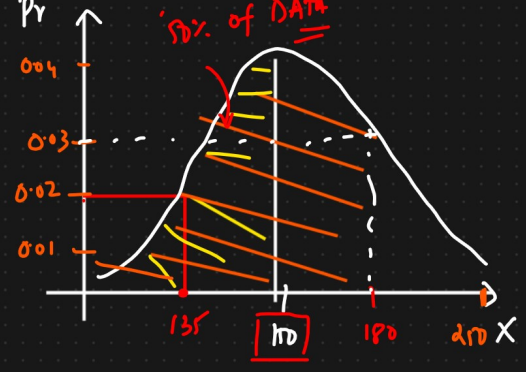
③ Cummulation Distribution Function

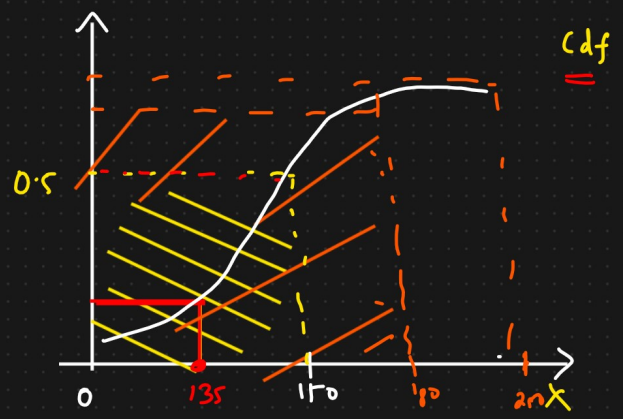
Eg: Rolling a dice

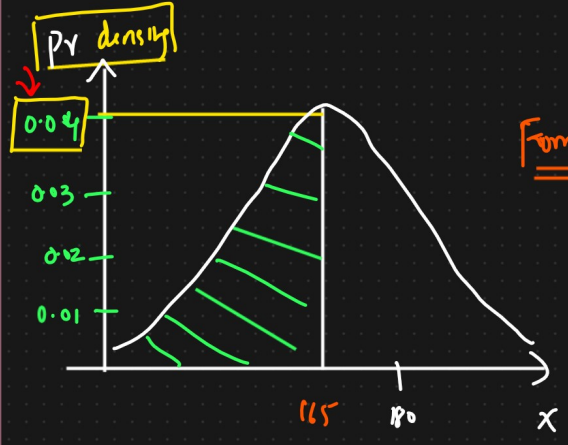




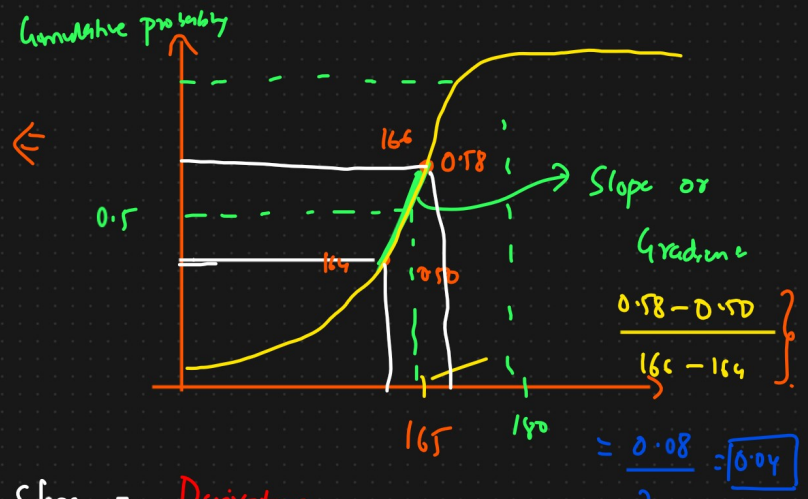
② Pdf and cdf







Formula ⇐



Slope = Derivative

Probability Density of pdf ⇒ Gradient or derivative of cdf

pdf is the derivative of cdf.

cdf is the integration of pdf.

Types of Probability Distribution

1. Normal/Gaussian Distribution (pdf) ✓
2. Bernoulli Distribution (pmf) → Outcomes are binary only. ✓
3. Uniform Distribution (pmf)
4. Poisson Distribution (pmf) ✓
5. Binomial Distribution (pmf) ✓
6. Log Normal Distribution (pdf)

① Bernoulli Distribution

① Discrete Random Variable {pmf}

② Outcomes are Binary

Eg: Tossing a Coin {H, T}.

$Pr(H) = 0.5 = p$ //

$Pr(T) = 0.5 = 1-p = q$ //.

p => probability of one outcome

q => probability of other outcome

Eg: Whether the person will pass/Fail

$Pr(Pass) = 0.7 = p$

$Pr(Fail) = 1-p = 0.3 = q$.

$P(T) = 0.8 \Rightarrow p$

$q = 1-0.8 = 0.2$

Parameters

$0 \leq p \leq 1$

$q = 1-p$

$K = \{0,1\}$ => Outcomes

Heads and Tail

$K = 0, 1$

p q

Pr(H) Pr(T)

PMF

$$Pmf \begin{cases} q = 1-p & \text{if } K=0 \\ p & \text{if } K=1 \end{cases}$$

Formula

$$PMF = p^{K} * (1-p)^{1-K}$$

$K \in \{0,1\}$

if K = 1

$PMF = p$ //

if K = 0

$Pmf = p^0 * (1-p)^1$

$= (1-p) = q$ //

② Mean of Bernoulli Distribution

$$E(K) = \sum_{i=0}^{K} K \cdot p(K)$$

$Pr(k=1) = 0.6 \Rightarrow p$

$Pr(k=0) = 0.4 \Rightarrow 1-p$

$$= [0 * 0.4 + 1 \cdot 0.6]$$

$$= 0.6 = p //.$$

$\{1, 2, 3, 4, 5, 6\}$

④ Binomial Distribution

Note → Refer Wikipedia

① Discrete Random Variable.

② Every Experiment outcome is Binary.

③ These experiment is performed for n trials

Eg: Tossing a coin 10 times

$Pr(H) = 0.5 = p$

$Pr(T) = 0.5 = q = 1-p.$

Notation: $B(n, p)$

Parameters: $n \in \{0, 1, 2, 3 \cdots\} \Rightarrow$ no. of trials

$p \in \{0, 1\} \rightarrow$ Success probability for each trial

$q = 1-p.$

Support = $K \in \{0, 1, 2, \cdots n\} \Rightarrow$ Number of Successes.

PMF

$$Pr(k,n,p) = {}^{n}C_{k}\, p^{k} (1-p)^{n-k}$$

for $k = 0, 1, 2, \ldots n$ where

$${}^{n}C_{k} = \frac{n!}{k!(n-k)!}$$

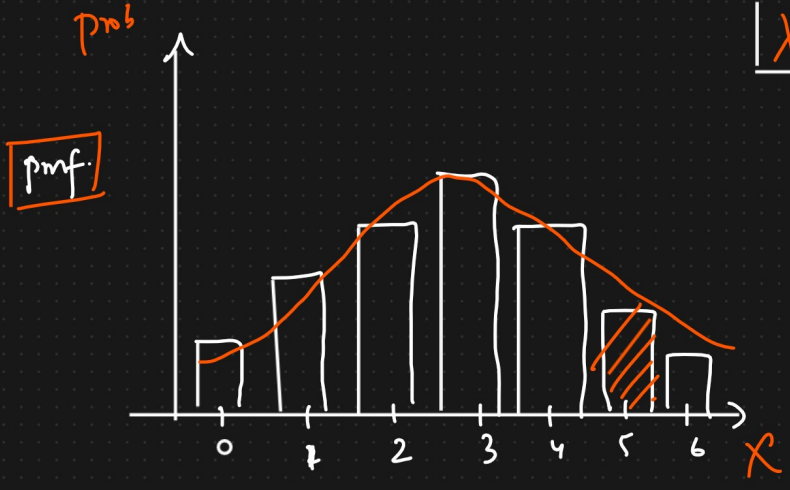
⑧ Poisson Distribution

① Discrete Random Variable (pmf)

② Describe the number of event occuring in a fixed time interval.

Eg: No. of people visiting hospital every hour

No. of people visiting banks at 11am



Threshold

$\lambda = 3$ ⇒ Expected no. of event occur at every time interval

$x = 5$

$$Pr(X=5) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

λ

$\lambda = 3$ ⇒ Let us consider $\lambda = 3$

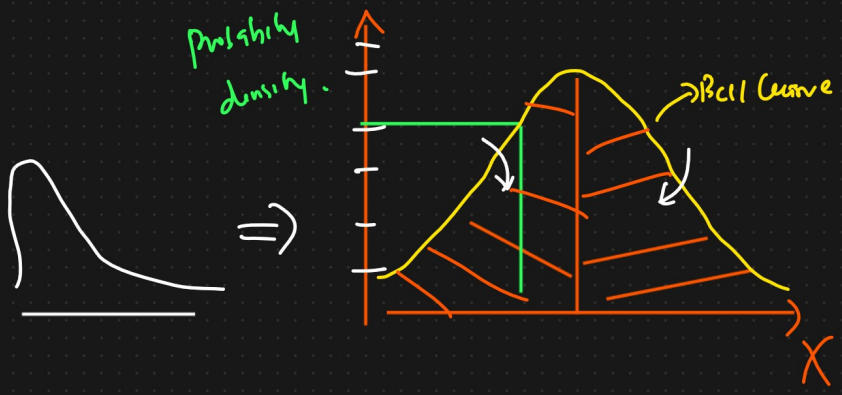
$$= \frac{e^{-3}\, 3^{5}}{5!} = 0.101 = 10.1\%$$

4pm and 5pm.

$$Pr(X=4 \text{ and } X=5) = Pr(X=4) + Pr(X=5)$$

⑧ Normal / Gaussian Distribution (pdf)

⑧ Continuous Random Variable.



Ages, Height, weight.

$x = \{ \ldots \}$

⇓

Dr.

IRIS

Notation : $N(\mu, \sigma^2)$

Parameters : $\mu \in R$

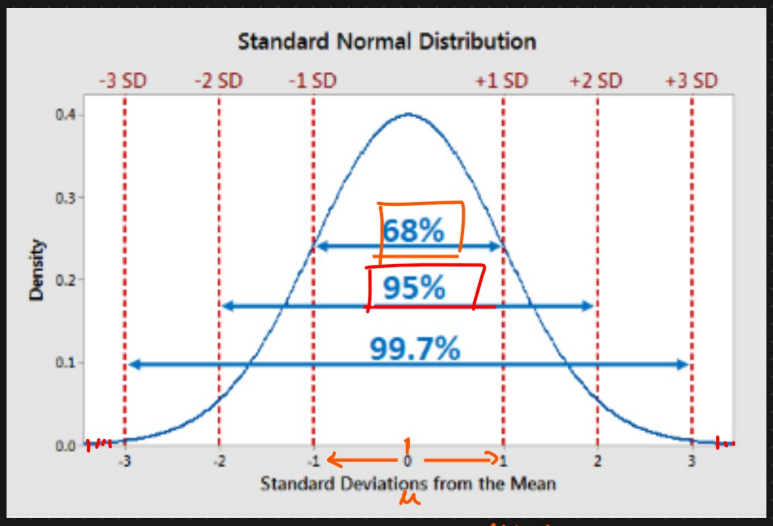
$\sigma^2 \in R > 0$ = variance

$x \in R$

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

Mean = μ Variance = σ^2 Std = $\sqrt{\text{Variance}}$.

Empirical Rule of Normal Distn



$\mu-3\sigma$ $\mu-2\sigma$ $\mu-\sigma$ $\mu+\sigma$ $\mu+2\sigma$ $\mu+3\sigma$

$X = \{ \_ \_ \ldots \}$ (100)

⇓

Normal Distribution

68% of the entire distribution will fall in 1st Std.

$$\mu-\sigma \le x \le \mu+\sigma \approx 68\%$$
$$\mu-2\sigma \le x \le \mu+2\sigma \approx 95\%$$
$$\mu-3\sigma \le x \le \mu+3\sigma \approx 99.7\%$$

$\mu = 10$ $\sigma = 5$ ⇒ Normal Distribution

25

f1 ⇒ f2 ⇒ f3 ⇒

Statistics

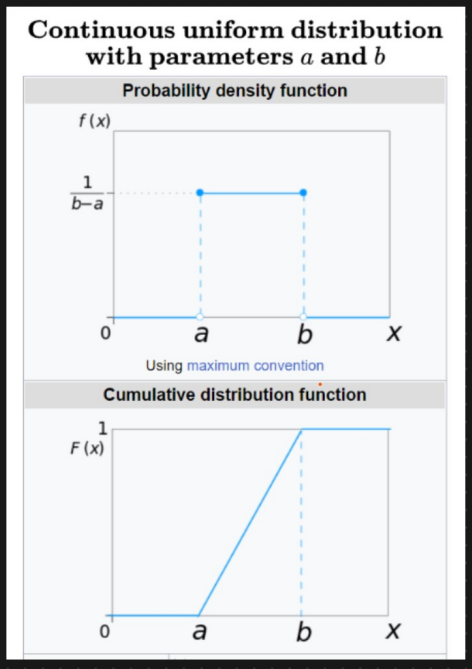
Agenda

1. Uniform Distribution
2. Standard Normal Distribution And Z-score
3. Central limit Theorem
4. Log Normal Distribution

1. Uniform Distribution

1. Continuous Uniform Distribution (pdf)
2. Discrete Uniform Distribution (pmf)

1. Continuous Uniform Distribution [Continuous Random Variable]



Notation = $U(a,b)$

Parameters = $-\infty < a < b < \infty$

$$\text{Pdf} = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a,b] \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{cdf} = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{x-b} & \text{for } x\ [a,b] \\ 1 & \text{for } x > b \end{cases}$$

$$\text{Mean} = \frac{a+b}{2} \qquad \text{Median} = \frac{a+b}{2}$$

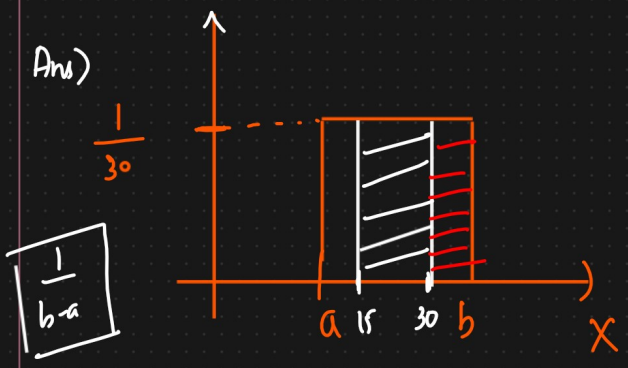
Eg: The number of candies sold daily at a shop is uniformly distributed with a maximum of 40 and a minimum of 10.

(i) Probability of daily sales to fall between 15 and 30?

$x_1 = 15$

$x_2 = 30$

Ans)



$a = 10$ $b = 40$

$$Pr(15 \le x \le 30) = (x_2 - x_1) * \frac{1}{b-a}$$

$$= 15^{1} * \frac{1}{30\ 2} = 0.5 //.$$

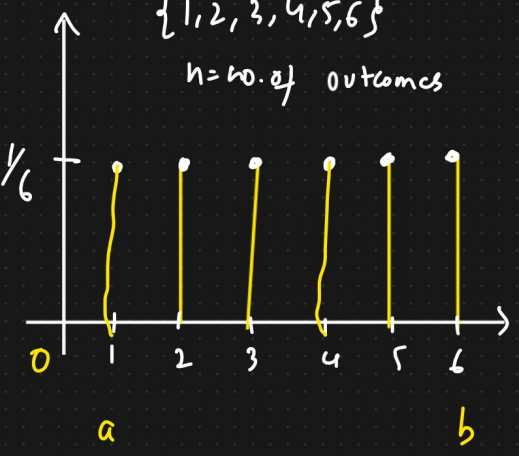
$$Pr(x \ge 30) = (40-30) \times \frac{1}{30} = 1\cancel{0} \times \frac{1}{3\cancel{0}} = \frac{1}{3} = 0.33 = 33\%.$$

② Discrete Uniform Distribution (pmf)

Eg: Rolling a dice

{1, 2, 3, 4, 5, 6}

n = no. of outcomes



$n = 6$

$$n = b - a + 1$$

$$= 6 - 1 + 1$$

$$n = 6$$

$$Pr(1) = \frac{1}{n} = 1/6$$

Notation $U(a, b)$

Parameter a, b $b > a$

$PMF = \frac{1}{n}$

$Mean = \frac{a+b}{2}$

$Median = \frac{a+b}{2}$.

⑦ Standard Normal Distribution And Z-Score

Z-Stats

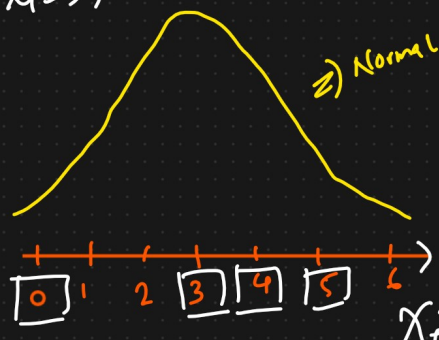
$X = \{1, 2, 3, 4, 5\}$ ⇒ Normally Distributed

$\mu = 3$

$\sigma = 1.414 \approx 1$

$\mu = 3, \sigma = 1$

⇒ Normal Distribution

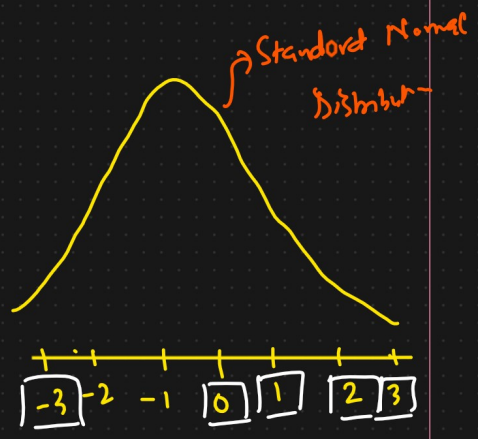


$X = \{1, 2, 3, 4, 5, 6\}$.

$$\Rightarrow \text{Z\_score} = \frac{x_i - \mu}{\sigma}$$

$\mu = 0$ & $\sigma = 1$

⇒



$$\text{Z\_score} = \frac{1-3}{1} = -2 \qquad \frac{3-3}{1} = 0 \qquad \frac{0-3}{1} = -3$$

$$= \frac{2-3}{1} = -1 \qquad \frac{4-3}{1} = 1$$

$$\frac{5-3}{1} = 2$$

$\mu = 0$ $\sigma = 1$

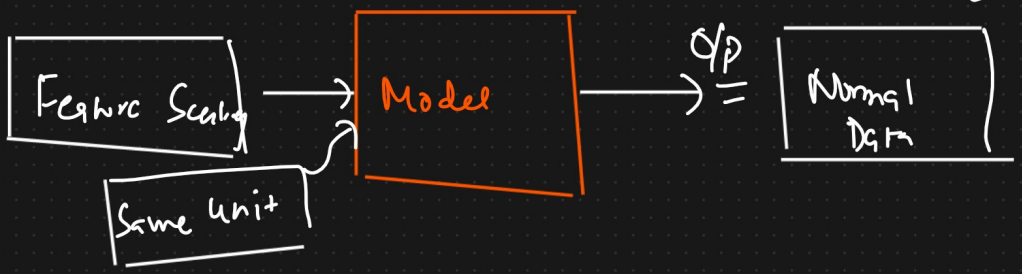
	Age (year)	Weight (kg)	Height (cm)	Salary (IN₹)
-3 to +3	24	70	–	–
	32	80	–	–
	36	75	–	–
	38	60	–	–
	20	30	–	–

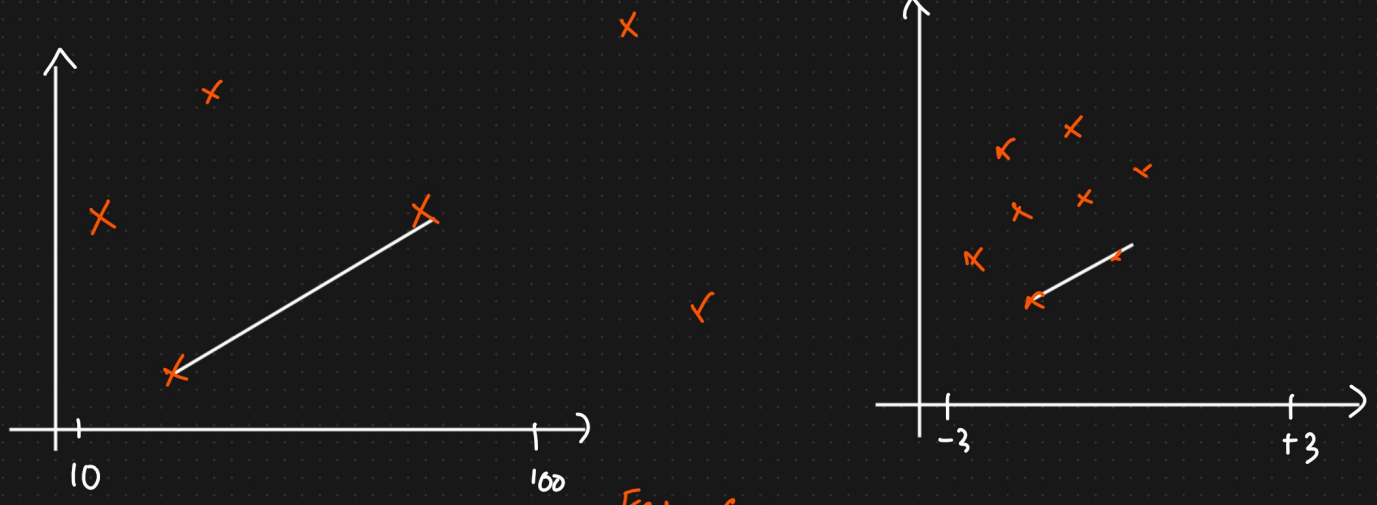
⇒ Model ≥ Mathematical Equation

↳ Feature Scaling

Z-Score :-

Normalization





(-1, 1) ← (0, 1) ⇒ Deep Learning

Normalization And Standardization [Assignment].

⇓

TRANSFORMATION

⇓

Assignment.

Feature Scaling

⇓

↑ ⇓ ↑

Z-score

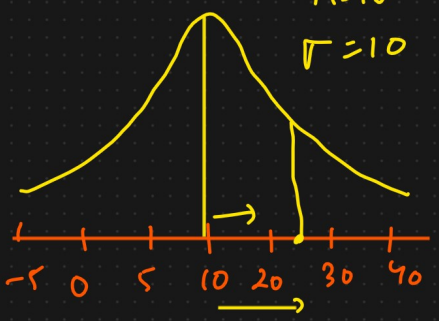
$\mu = 0, \sigma = 1$

⇓

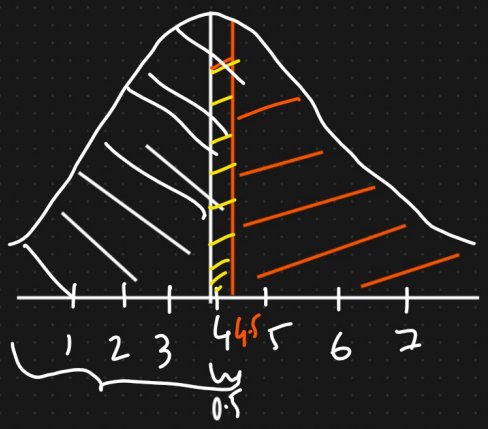
Feature TRANSFORMATION

$\mu = 10$

$\sigma = 10$



$$Z\text{-score} : \frac{25 - 10}{10} = \frac{15}{10} = \boxed{1.5}$$



① How many standard deviation 4.5 is away from mean? ⇒ 0.5

② What percentage of data if falling above 4.5?

Z table ⇒ Area Under the Curve

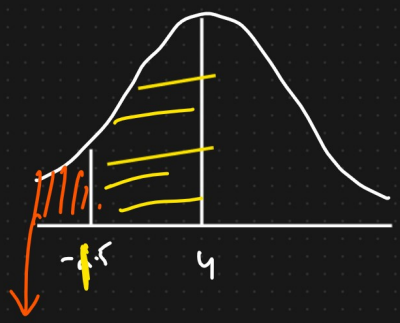
$$Z\text{-score} = \frac{4.5 - 4}{1} = 0.5$$

0.6915

Area under the curve (≥ 4.5) $= 1 - 0.6915 = 0.3085 = 30.85\%$

③ Percentage of data falling below 2.5?

$$Z.\text{score} = \frac{2.5 - 4}{1} = -1.5$$



Z table = 0.0668
= 6.6%

0.5 − 0.0668

6.6%

Prob In India the average IQ is 100, with a standard deviation of 15. What is the percentage of the population would you expect to have and IQ lower than 85?

Ans) $\mu = 100 \quad \sigma = 15$

① $Z\text{-score} = \frac{x_i - \mu}{\sigma}$

Refer Z table

$= \frac{85 - 100}{15}$

$= -1$

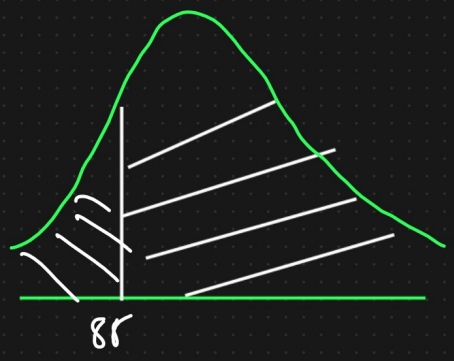
Area under the curve $= |0.15866|$

$0.5 - 0.15866$

$= 0.34134.$

Area under the curve $(> 85) = 1 - 0.15866$

$= 0.8413..$



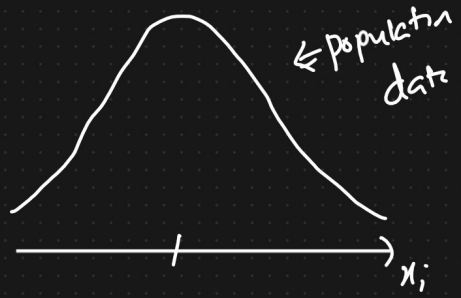
Central limit Theorem

① $X \approx N(\mu, \sigma)$ (↓↓)

n samples

$n = 30$

Sampling mean



$S_1 = \{ x_1, x_2, x_3 \dots x_n \} = \bar{x}_1$

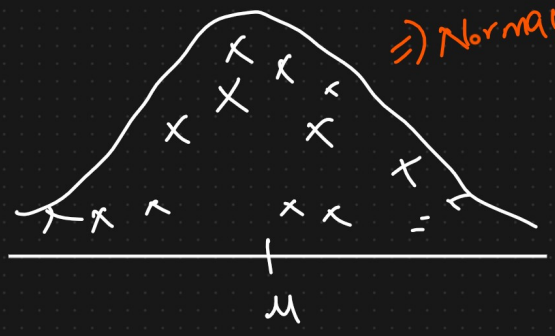
$S_2 = \{ \quad \} = \bar{x}_2$

$S_3 = \{ \quad \} = \bar{x}_3$

$\vdots$

$S_m = \{ \quad \} = \bar{x}_m$

$\bar{X} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots \bar{x}_n\}$ ⇒ Sampling distribution mean



Important for Interview

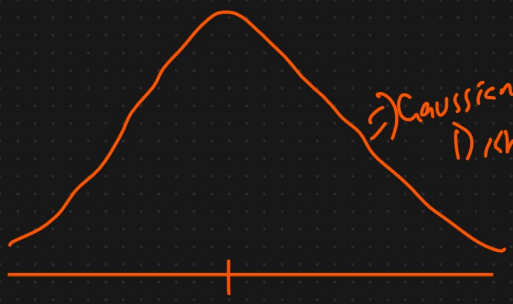
$X \sim N(\mu, \sigma)$

σ = population std

μ = population mean

n = sample size

Sampling distribution of mean

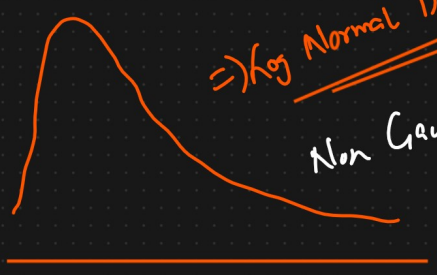


$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

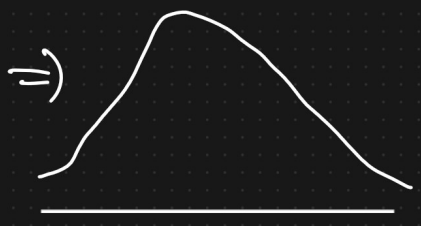
n can be any value

$\frac{\sigma}{\sqrt{n}}$ ↓ Standard Error

$n \geq 30$



Sampling Distribution mean

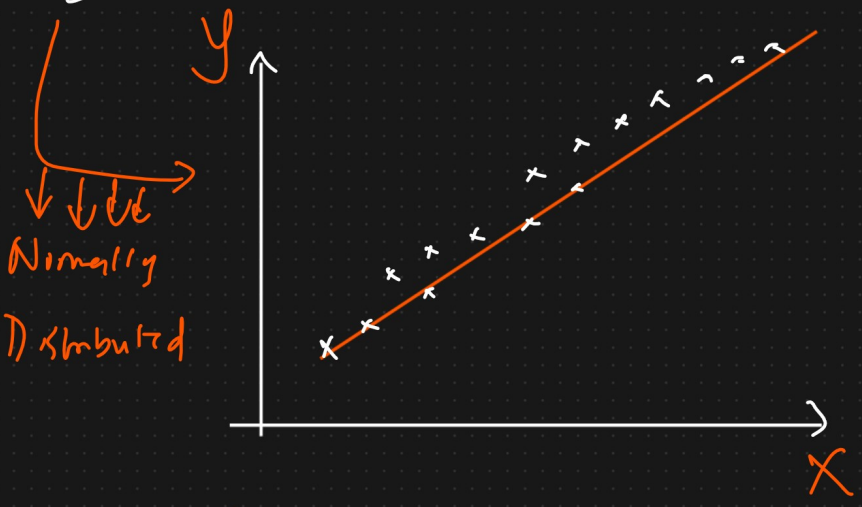


⑭ How to check whether a distribution is Normal or Not

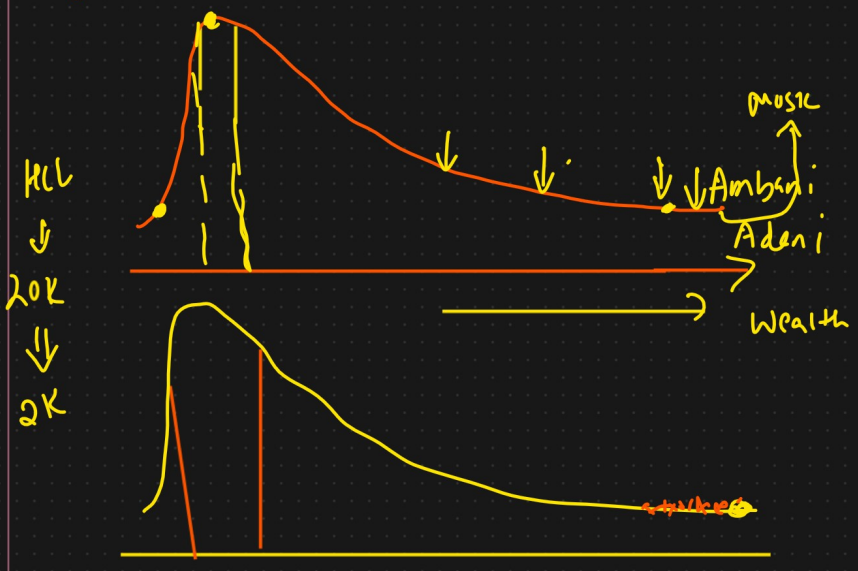
Q-Q plot

Normal Distribution

$X = \{ - \ldots \}$ $Y = \{ - \ldots \}$



Log Normal Distribution



Wealth distribution of people in world