

Statistics

Statistics is the science of collecting, organizing and analyzing data.

Data: "facts or pieces of information"

Eg: Height of students in a classroom
 $\rightarrow \{175\text{cm}, 150\text{cm}, 140\text{cm}, 130\text{cm}, 155\text{cm}\}$

Eg: Intelligence Quotient (IQ) of 5 randomly selected individuals ($109, 89, 129, 101, 105, 106$) \rightarrow Data.

Two Types

Statistics



It consists of organizing and summarizing of data.

It consists of using that you've measured to form

Conclusions

Eg: Pdf, Histogram, Box plot, Bar chart, Pie charts

Eg: Hypothesis Testing, p value Z test, t-test, Anova, Chi-square

Eg: Let's say there are 20 maths classes at your university and you've collected the ages of students in one class.

Ages $\{21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22\}$

$\min = \text{mode}$

Descriptive stats: What is the average age of student in

your maths class?

Inferrential question : Are the ages of students in this maths classroom similar to what you would expect in a normal maths class at this university?

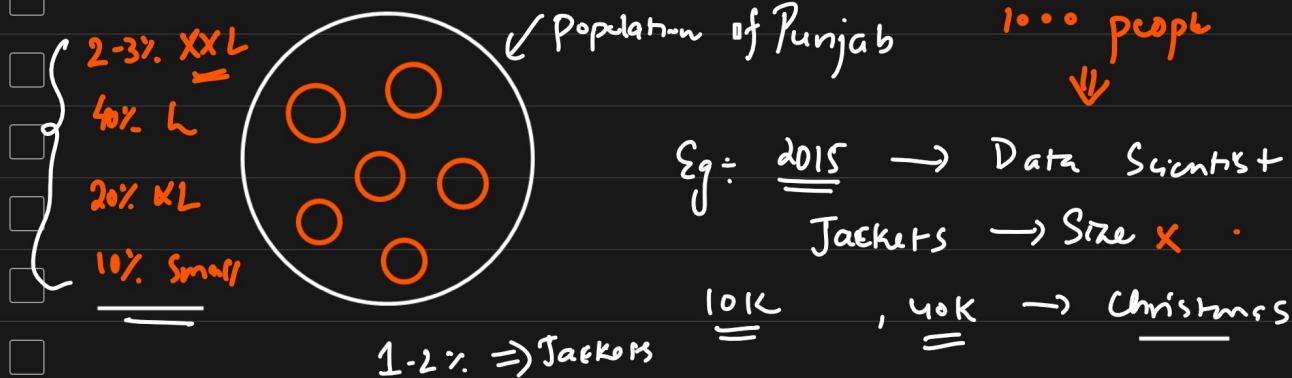


Population And Sample Data → Inferrential Statistics Results

Elections : Punjab

{ AAP, Congress }

Exit Polls ←

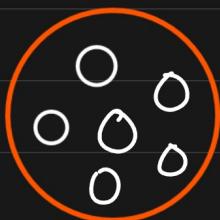


Population (N) ✓

Sample (n) ✓

Sampling Techniques

① Simple Random Sampling : Every member of the population (N) has an equal chance of being selected for your sample (n)



② Stratified Sampling

Strata → Layers
↓
Clusters
↑
Non overlapping groups

Gender → Male
Female

Blood Groups

Age groups
0-18 }
18-35 }
35-60 }

Tax Slabs
Courses

Education Qualification

Thamno

Australias

③ Systematic Sampling

Snap

Customs

(N) → Select every n^{th} individual

$\nearrow 6^{\text{th}}$
=

\downarrow
Stratified

Eg: Survey → Mail (SBI credit card)

④ Convenience Sampling : Only those people who are interested will only be participating.

Healthcare Disease

Eg: Data Science → AI }
YouTube Survey → }

{ Blind people }

→ RBI → Household Survey → Female ← $\frac{\downarrow \downarrow \downarrow \downarrow \downarrow}{\text{Economics}}$ → DATA Science }

Exit Poll : Stratified + Random Sampling

Variable

A variable is a property that can take on any value

Eg: Height = 182
150
145
160

{ 182, 170, 145, 160 }
↓
No

Two kinds of Variable

① Quantitative Variable → Measured Numerically { Add, Subtract, \times , \div }

② Qualitative Variable.

↳ Eg: Gender [Male { Based on some { characteristics } we can derive categorical variables }
Female]

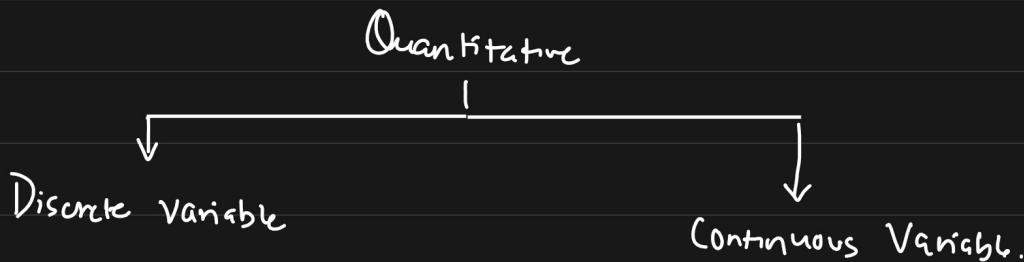
{ Quantitative → Qualitative Variable. }

Eg: IQ

0-10 10-50 50-100

↓ ↓ ↓

Low IQ Medium IQ Good IQ



Eg: whole number

Eg: No. of Bank accounts

{ 2, 3, 4, 5, 6, 7 } 2.5, X
 2.75 X

Eg: Total No. of children in a family

Eg: Height = 172.5, 162.5 cm,

163.5 cm.

Rainfall: 1.35, 1.25, 1.75, 2.25 cm

Weight

Temp

Eg: 2, 3, 4, 5

Stock price.

25, 2.75

Eg: Total no. of Employees in a Company {e.g.: 10k,

Ass:

- ① What kind of variable Marital Status is? Categorical
- ② What kind of variable Nile River length is? Continuous Quantitative
- ③ What kind of " Movie duration is? " "
- ④ What kind of Variable IQ is? " "

Frequency Distribution

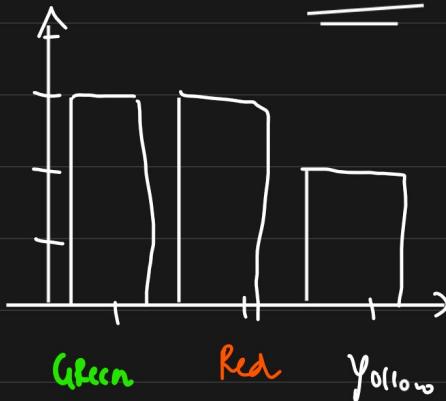
Sample Data: Green, Red, Yellow, Green, Red, Yellow, Green, Red

↓

Colors	Frequency
Green	3
Red	3
Yellow	2

① BAR GRAPH frequency

Bar Chart



① Variable Measurement Scales

4 types of Measured Variable.

① Nominal data { Categorical data }

Eg: Colors, Gender, Types of flowers

Ranking is not that important

② Ordinal data

Student (Marks)

→ 100

96

57

85

44

Rank

1

2

4

3

5

Percentiles

Ordinal Data.

PHD
↓
{ NLP }
↓

Degree

PHD

B.G

Master

BCA

12

Salary

✓

✓

✓

✓

✓

Assignment

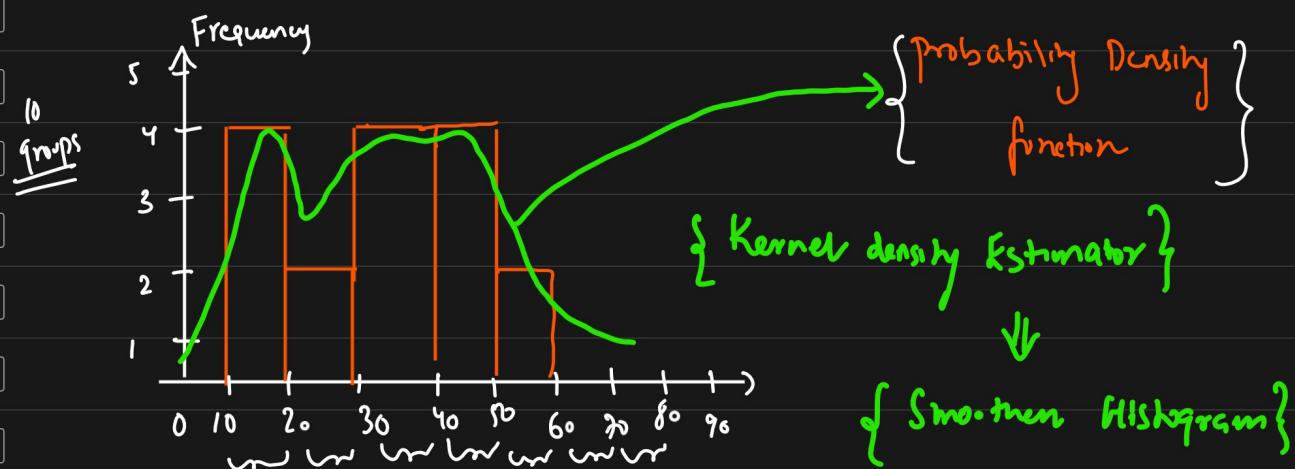
④ Ratio data ✓

③ Interval data ✓

⑥ Histograms ÷ Continuous

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

Histogram $\rightarrow \text{Bins} = 10$ \equiv Mean, Median, Mode.



0 - 10 \rightarrow 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35

Assignment

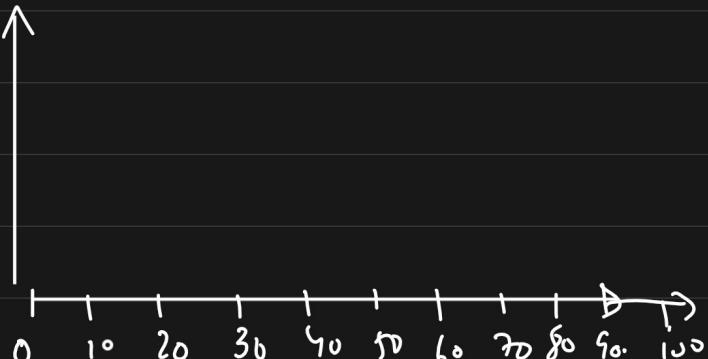
Eg: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

bins \downarrow
10

0-10 10-20 20-30 30-40

40-50 50-60 60-70

70-80 80-90 90-100

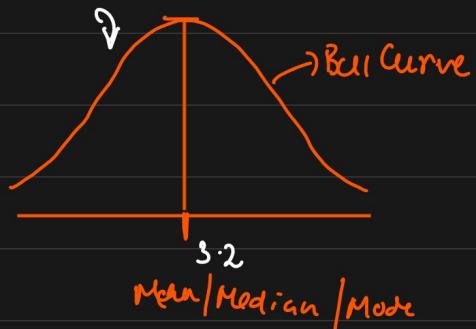


Intermediate Stats

- ① Measure of Central Tendency
- ② Measure of Dispersion
- ③ Gaussian Distribution
- ④ Z - Score
- ⑤ Standard Normal Distribution
- ⑥ Central Limit Theorem
-

- ① Measure of Central Tendency → Central position of the dataset
-

- ① Mean ✓
- ② Median ✓ { EDA & Feature Eng. }
- ③ Mode ✓
-



Population (N)

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

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

Population

mean

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

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

Sample (n)

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

Sample

Mean

Median

1, 2, 2, 3, 4, 5

↓
1, 2, 2, 3, 4, 5, 100

$$\bar{x} = \frac{1+2+2+3+4+5}{6} = \frac{17}{6} = 2.83$$

$$\bar{x} = \frac{1+2+2+3+4+5+100}{7} = \frac{117}{7} = 16.71$$

Median ✓

1, 2, 2 3 4, 5, 100

$$\bar{x} = 16.71 //$$

$$\text{Median} = 3 \\ =$$

1, 2, 2, 3, 4, 5 → odd or even
↓ $\frac{2+3}{2} = 2.5$

2.5 $\approx 2.83 //$

Mode ÷ Highest frequency: Median

1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7
↓
3

1, 2, 2, 3, 3, 4, 4, 5, 5
↓
{mode}

[2, 3, 4]

EDA

Feature Engineering

↪ NAN values ⇒ Continuous Values + outlier
= = Mean ↓
= Median

⇒ Categorical Variable.

↓
Mode

Agnl

Lidley, Sunflower, Rock, - - - , Min, Max

Measure of Dispersion → {Dispersion}

① Variance

② Standard deviation

{ }
↓

Spread ⇒ How the data is spread



① Variance

Population Variance

{
Basis (Correction)
Degree of freedom}

Sample Variance

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

Population mean

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

Sample mean
 $n-1$

Eg:

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

<u>X</u>	<u>\bar{x}</u>	<u>$x - \bar{x}$</u>	<u>$(x - \bar{x})^2$</u>
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.03
4	2.83	1.17	1.37
5	2.83	2.17	4.71

$$\left[\frac{10.84}{5} \right] = 2.168$$

↑

$n=6$

$n-1$

$$\mu = 2.83$$

{let consider

$$10.84$$

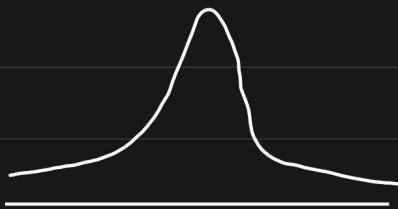
$$\frac{\sigma^2}{\text{Variance}} = \frac{6.42}{\text{as an example}}$$

Spread ↑↑

$$\frac{\sigma^2}{\text{Variance}} = 2.168$$

Variance ↑↑

Spread ↑↑



Standard deviation

$$\sigma = \sqrt{\text{Variance}} = \sqrt{2.168}$$

$$= \sqrt{1.472}$$

Variance

\downarrow

Spreadness

1, 2, 2, 3, 4, 5

2.83

-1.472

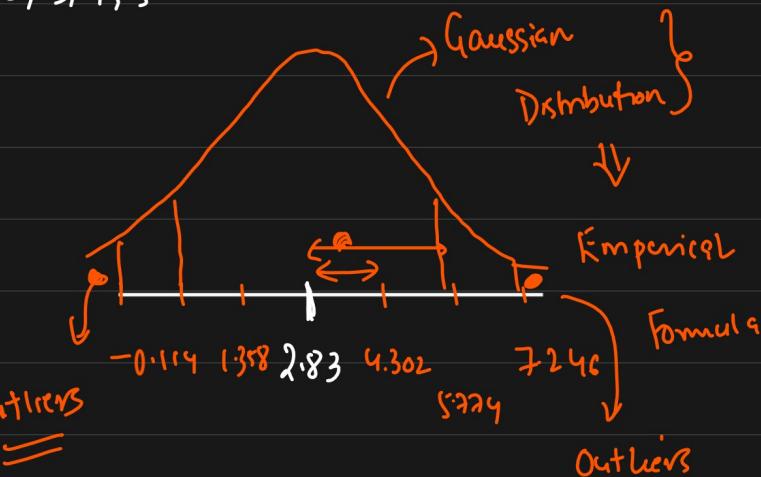
$\frac{1.358}{1.358}$

1.358

1.472

$\frac{1.358}{0.114}$

0.114



$$\begin{array}{r} 2.83 \\ 1.472 \\ \hline 4.302 \end{array}$$

$$\begin{array}{r} 1.472 \\ 5.246 \\ \hline 7.246 \end{array}$$

(*) Percentiles And Quartiles



Percentiles : 1, 2, 3, 4, 5

% of the numbers that are odd?

$$\% \text{ of odd} = \frac{3}{5} = \underline{\underline{60\%}}$$

Percentiles : $\{CAT, GATE, SAT\} \Rightarrow \underline{\underline{99\%}}$

Defn : A percentile is a value below which a certain percentage of observations lie

99 percentiles mean the person has got better marks than 99% of the students.

Data set : 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

What is the percentile ranking of 10? $n=20$

Percentile Rank of $x = \frac{\text{# of values below } x}{n} \times 100$

$$= \frac{16 + 0.8}{20} = \underline{\underline{80}} \text{ percentile}$$

$$= \frac{17}{20} = 85$$

② What value exists at percentile ranking of 25%?

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

$$= \frac{25}{100} \times (21) = \underline{\underline{5.25}} \rightarrow \text{Index}$$

Value = 5

Quartiles (25%)

Five Number Summary

- ① Minimum
- ② First Quartile (25%) Q_1
- ③ Median
- ④ Third Quartile (75%) Q_3
- ⑤ Maximum

Removing the Outliers

Inter Quartile Range: (75% - 25%)
 $Q_3 - Q_1$

$$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, \cancel{10}\}$$

[Lower Fence \longleftrightarrow Higher Fence]

$$\text{Lower Fence} = Q_1 - 1.5(\text{IQR}) \quad (25\%) \quad Q_1 = \frac{3+5}{2} \times 20^{\text{th}} \text{ index}$$

$$\text{Higher Fence} = Q_3 + 1.5(\text{IQR})$$

$$\text{IQR} = Q_3 - Q_1 = 7 - 3 = 4 \quad (75\%) \quad Q_3 = \frac{7+8}{2} \times 20^{\text{th}} = 15^{\text{th}} \text{ index}$$

$$Q_3 = 7$$

$$\text{Lower Fence} = 3 - 1.5(4) = 3 - 6 = -3$$

$$\text{Higher Fence} = 7 + 1.5(4) = 7 + 6 = 13$$

$$[-3 \longleftrightarrow 13] \quad -\text{ve} \quad \underline{\underline{3}}$$

$$\text{Remaining} \quad \frac{5+5}{2} = 5$$

$$1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, \cancel{10}$$

5 Number Summary

Minimum = 1

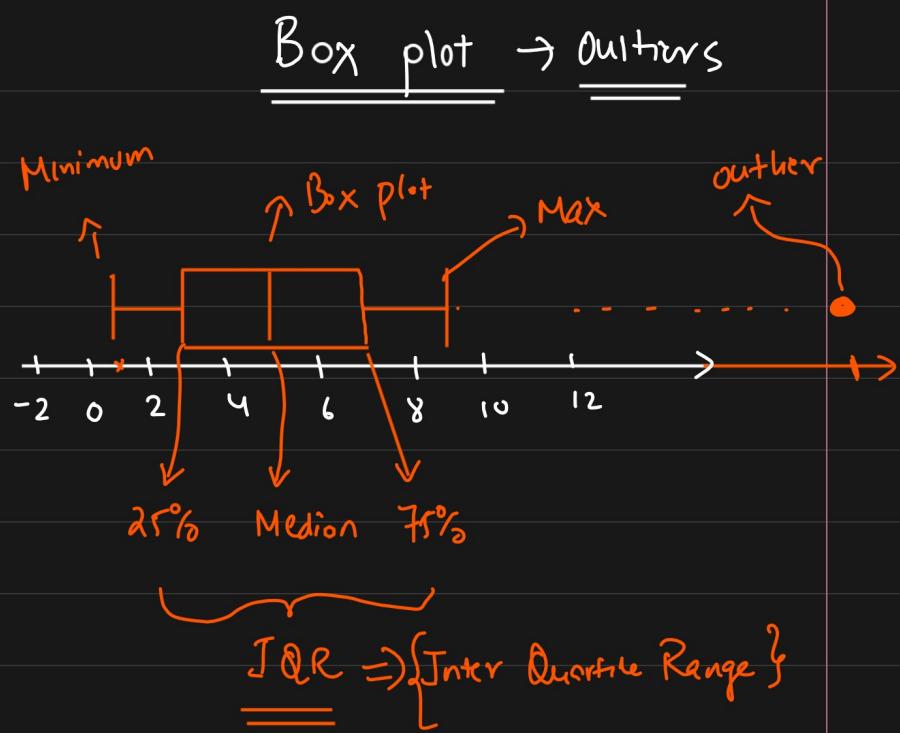
Q₁ = 3

Median = 5

Q₃ = 7

Max = 9

Use of Box plot



① Distributions

① Normal / Gaussian Distribution ✓

② Standard Normal Distribution ✓

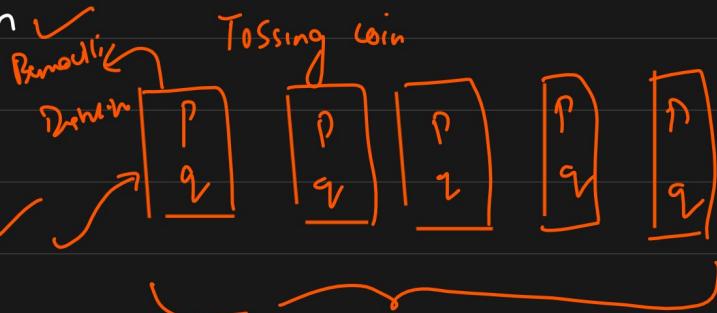
③ Z-Score ✓

④ Log Normal Distr ✓

⑤ Bernoulli's Distribution ✓

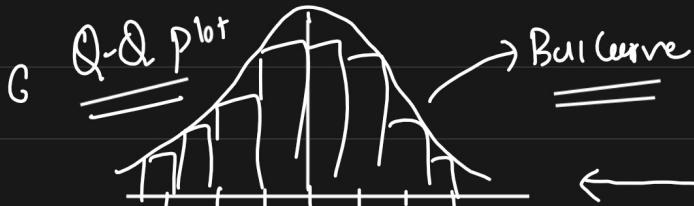
⑥ Binomial Distribution }

① Gaussian / Normal Distribution



Properties

{ Power Law }



① Empirical Rule of
Gaussian Distribution \Downarrow
80-20%

↳ DATASET → IRIS Dataset } → Petal, Sepal length
domain expansion }

② Weight of human brain

③ Height → Doctor

$$68.2, -95.4, -99.7$$

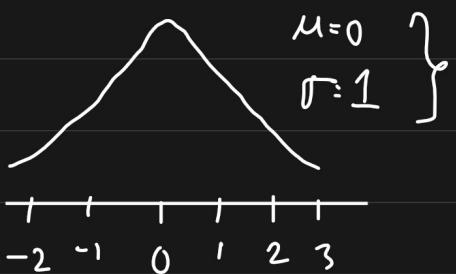
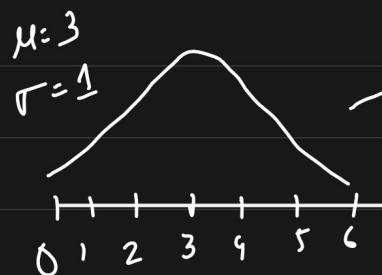
Outliers

Standard Normal Distribution

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

$$\mu = 3$$

$$\sigma = 1.414 \approx 1$$



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

$$\left\{ Z\text{-Score} = \frac{x-\mu}{\sigma} \right\}$$

$$\begin{aligned} &= \frac{3-3}{1} = 0 \\ &= \frac{2-3}{1} = -1 \\ &= \frac{1-3}{1} = -2 \end{aligned}$$

$$\text{Why } 22 \quad \boxed{\begin{array}{l} \mu=0 \\ \sigma=1 \end{array}}$$

✓

Standardization vs Normalization

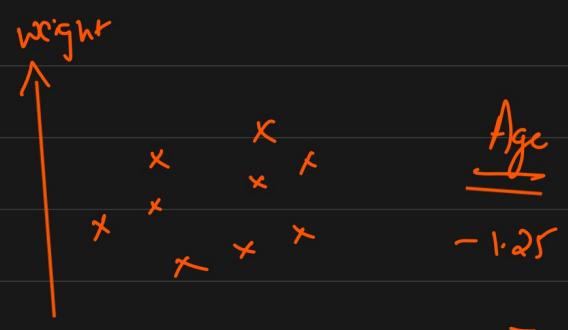
Years
Age ↑
Different unit
Weight ↗ kg

$$\begin{array}{ll} \text{Age} & \text{Weight} \\ 25 & 75 \\ 26 & 80 \\ 28 & 85 \\ 30 & 60 \\ 32 & 70 \end{array}$$

$$\begin{array}{ll} \text{Salary} & \text{INR} \\ 25K & \\ 30K & \\ 40K & \\ 60K & \\ \sim & \end{array}$$

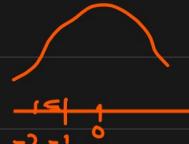
$$\frac{25 - 28.2}{25.6}$$

Same unit scale ??



Maths → Scale

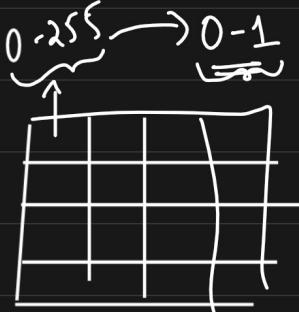
Standardization



Normalization

[Min Max Scalar]

$$\begin{matrix} \downarrow \\ 0 \text{ to } 4 \end{matrix}$$
$$\left. \begin{matrix} \downarrow \\ 0 \text{ to } 1 \end{matrix} \right\}$$



Convolutional

Neural Netw.

ML Disease
✓
Standardization

g

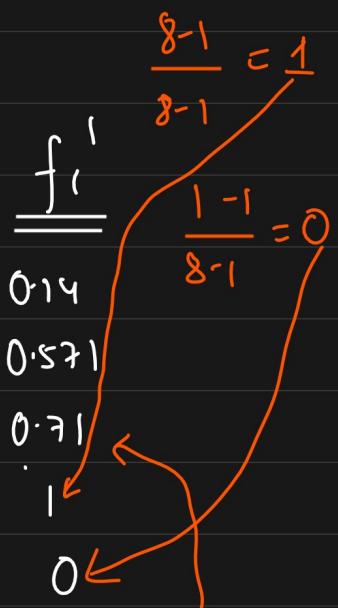
Normalization

$\leftarrow \boxed{\text{CNN}}$

f1

Normalization

(0 - 1)



$$\left\{ \begin{array}{l} x_{\text{Norm}} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \\ \parallel \end{array} \right.$$

Min Max Scalar

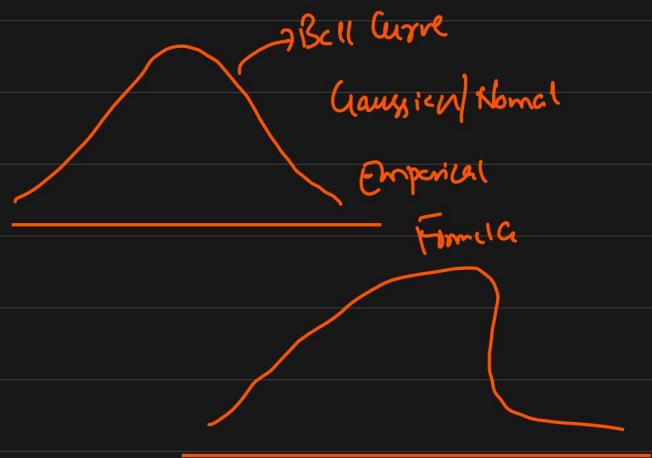
 = 0 to 1

$$= \frac{2 - 1}{8 - 1} = \frac{1}{7} = 0.142$$

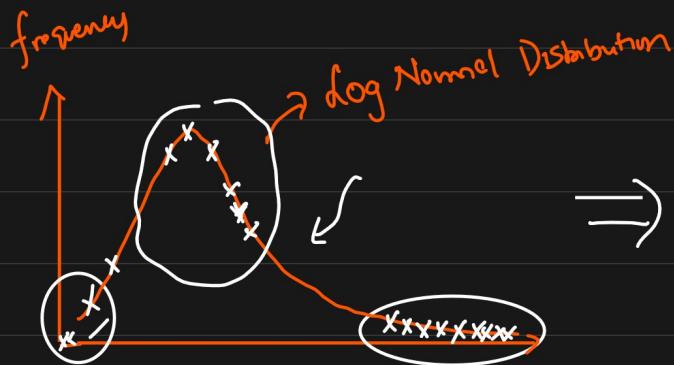
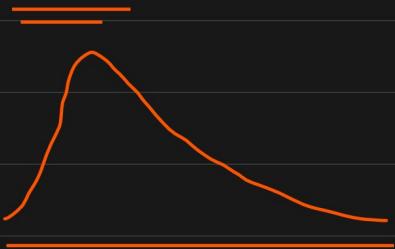
$$\frac{6 - 1}{8 - 1} = \frac{5}{7}$$

$$\frac{5 - 1}{8 - 1} = \frac{4}{7} = 0.571$$

Log Normal Distribution

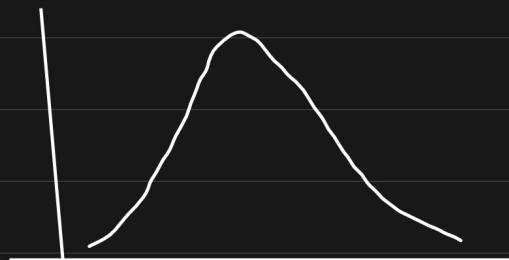


Skewed Curve



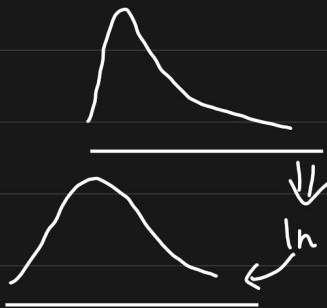
Gaussian Distribution

Normal Distn.



$X = \text{log Normal Distributed}$

$$\left\{ Y = \ln(X) \right\} \quad \begin{array}{l} \text{Gaussian} \\ \text{Distribution} \end{array}$$



$$\left\{ X = \exp(Y) \right\} \rightarrow c^y$$

X

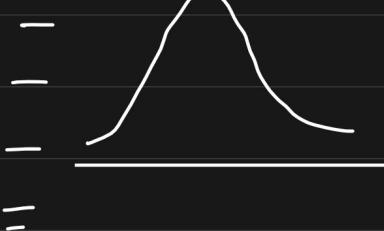
$\mathcal{Y} = \ln(x)$

25

30

40

45



① Bernoulli's Distribution