Statistics

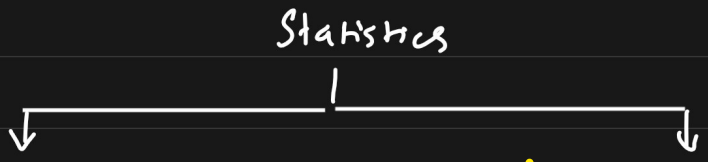
Statistics is the science of collecting, organizing and analyzing data.

Data : "facts or pieces of information"

Eg : Height of students in a classroom
→ {175cm, 152cm, 140cm, 130cm, 155cm}

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

Two Types



① Descriptive Stats

- It consists of organizing and summarizing of data.
- Eg : Pdf, Histogram, Box plot, Bar chart, Pie charts

② Inferential Stats

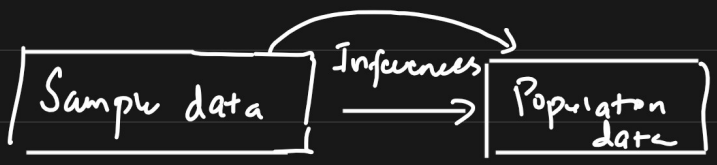
- It consists of using that you've measured to form conclusions
- Eg : Hypothesis Testing, p value, Z test, t-test, Anova, Chisquare

Eg : Lets 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}

Descriptive stats : What is the average age of student in min = mode

your maths class?

Inferential 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. → Inferential Statistics → Results

Exit Polls ←

Elections : Punjab {AAP, Congress}

↙ Population of Punjab

1000 people ⇓

Eg: 2015 → Data Scientist

Jackets → Size ✗

10K, 40K → Christmas

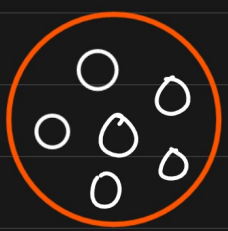
1-2% ⇒ Jackets

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 ⇓ Non overlapping groups

Clusters ⇑ groups

Gender → Male, Female

Age groups → 0-18, 18-35, 35-60

Blood Groups

Tax Slabs

Courses

Education qualification

Thanos

Australia

↳ Snap

Customs

③ Systematic Sampling

(N) → Select every nth individual

6th

3-4th

Stratified

Eg: Survey → Mail

(SBI credit card)

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

Eg: Data Science → AI

Youtube survey →

Healthcare Disease

{ Blind people }

{ Economics } → DATA Science

→ RBI → House hold Survey → Female ← UPSC

Exit Poll : Stratified + Random Sampling

Variable

A variable is a property that can take on any value

Eg: Height = 182
150
145
160

{ 182, 150, 145, 160 }
⇓
No

Two kinds of variable

① Quantitative Variable → Measured Numerically { Add, Substract, x, ÷ }

② Qualitative Variable.

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

{ Quantitative ⟶ Qualitative Variable. }

Eg: IQ

0-10 ⇓ Less IQ

10-50 ⇓ Medium IQ

50-100 ⇓ Good IQ

Quantitative

- Discrete variable
  - 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
- Continuous Variable.
  - Eg: Height = 172.5, 162.5cm, 163.5cm.
  - Rainfall: 1.35, 1.25, 1.75, 2.25cm
  - Weight
  - Temp

Stock price.

Eg: 2, 3, 4, 5

2.5, 2.75

Eg: Total no. of Employees in a Company Eg: 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? " "

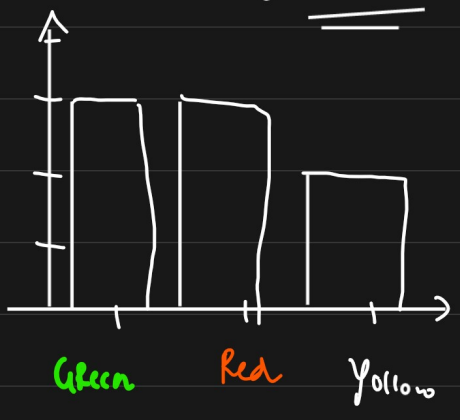
Frequency Distribution

Sample Dataset : Green, Red, Yellow, Green, Red, Yellow, Green, Red

Colors	Frequency
Green	3
Red	3
Yellow	2

① BAR GRAPH Frequency

Bar chart



⑧ Variable Measurement Scales

4 types of Measured Variable.

① Nominal data {Categorical data} → Ranking is not that important

Eg: Colors, Gender, Types of flowers

② Ordinal data

Percentiles

Student (Marks)	Rank
→ 100	1
96	2
57	4
85	3
44	5

Ordinal Data.

PHd → ⇓ ↙ {NLP}

Degree		Salary
PHd	1	✓
B.G	3	✓
Master	2	✓
BCA	4	✓
12	5	✓

③ Interval data ✓⇓

Assignment ⇓

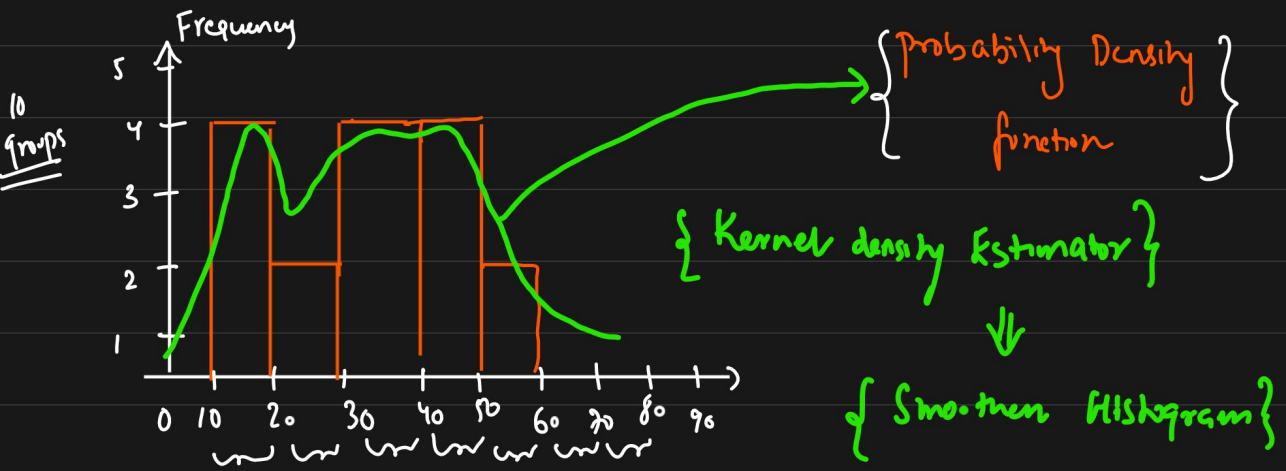
④ Ratio data ✓

⑧ Histograms ÷ Continuous

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

Histogram → Bins = 10

Mean, Median, Mode.



0 - 80 → 0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-35

Assignment

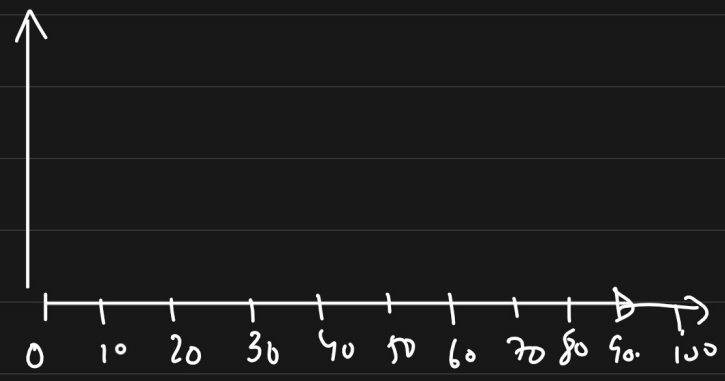
Eg: 10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

bins = 10

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

40-50 50-60 60-70

70-80 80-90 90-100

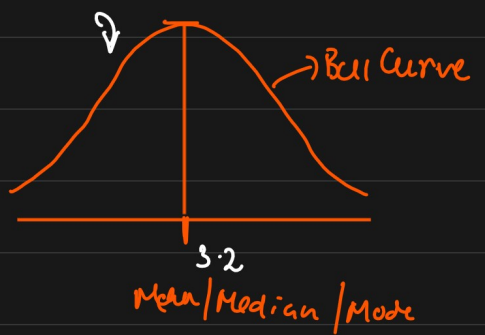


Intermediate Stats

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)

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

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

μ ↓ Population mean

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

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

Sample (n)

Sample → $\bar{x}$ ↓ Sample Mean

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

Median

1, 2, 2, 3, 4, 5

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

1, 2, 2, 3, 4, 5, 100

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

Median ✓

1, 2, 2, 3, 4, 5, 100

$\bar{x} = 16.71$ //

Median = 3

1, 2, 2, 3, 4, 5 → odd or even

1, 2, 2, 3, 4, 5

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

2.5 ≈ 2.83 //

Median

1, 2, 2, 3, 3, 4, 4, 5, 6

Error

[2, 3, 4]

Mode ÷ Highest frequency.

1, 2, 2, 3, 3, 3, 4, 5, 6, 6, 7

3

EDA

Feature Engineering

↳ NAN values ⇒

Continuous Values + Outlier

↳ Mean

Median

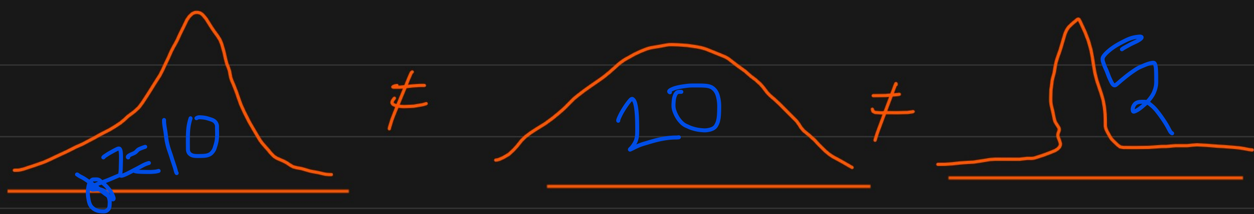
⇒ Categorical Variable.

Mode

Agri   Lilly, Sunflower, Rose, - - - -, NAN, NAN

Measure of Dispersion → {Dispersion}
⇓
Spread ⇒ How the data is spread

① Variance }
② Standard deviation }



① **Variance**   { Bessels Correction / Degree of freedom }

Population Variance   **Sample Variance**

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

μ → population mean

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

$\bar{x}$ → sample mean

$n-1$ → n-1

Eg:

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

x	$\bar{x}$	$x - \bar{x}$	$(x - \bar{x})^2$
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

$n = 6$

$$\frac{10.84}{5} = 2.168$$

↑ n-1

$\mu = 2.83$ {let consider as an example}

10.84

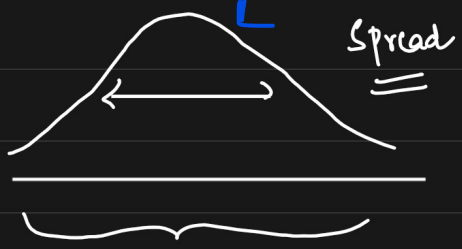
Variance ⇑⇑ Spread ⇑⇑

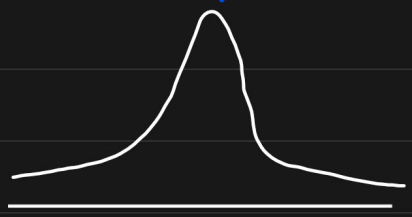
σ^2 ⇓ ✓

Variance = 6.42 Spread ⇑⇑

σ^2

Variance = 2.168



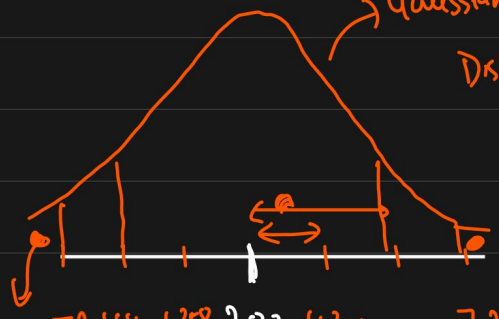


Standard deviation

$\sigma = \sqrt{\text{Variance}} = \sqrt{2.168} = \{1.472\}$

Variance ⇓ Spreadness ⇓

1, 2, 2, 3, 4, 5



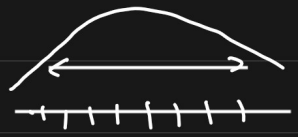
2.83 0 − 1.472 = 1.358

1.358 + 1.472 = 0.114

2.83 + 1.472 = 4.302

4.302 + 1.472 = 5.774

5.774 + 1.472 = 7.246



Percentiles And Quartiles

Percentages : 1, 2, 3, 4, 5

% of the numbers that are odd?

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

Percentiles : {CAT, GATE, SAT} ⇒ 99%

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

99 percentile mean the person has got better marks than 99% of the Students.

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

Average = 5

What is the percentile ranking of 10 ? n = 20

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

$= \frac{16}{20} = 0.8 = 80$ percentile

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

$P = \frac{\text{Rank}}{\text{Data}} \times 100$

② What value exists at percentile ranking of 25% ?

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

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

Value = 5

Quartiles (25%)

Five Number Summary

1. Minimum
2. First Quartile (25%) Q1
3. Median
4. Third Quartile (75%) Q3
5. Maximum

Removing the outliers

Inter Quartile Range = (75% − 25%)
Q3 − Q1

{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~29~~}

[Lower Fence ⟷ Higher Fence]

Lower Fence = Q1 − 1.5(IQR)

Higher Fence = Q3 + 1.5(IQR)

(25%) $Q1 = \frac{25}{100} \times (20) = 5^{th}$ index

Q1 = 3

IQR = Q3 − Q1 = 7 − 3 = 4

(75%) $Q3 = \frac{75}{100} \times 20^{15} = 15^{th}$ index

Q3 = 7

Lower Fence = 3 − 1.5(4) = 3 − 6 = −3

Higher Fence = 7 + 1.5(4) = 7 + 6 = 13

[−3 ⟷ 13]

[−3 ⟷ 13]

−ve 3

Remaining

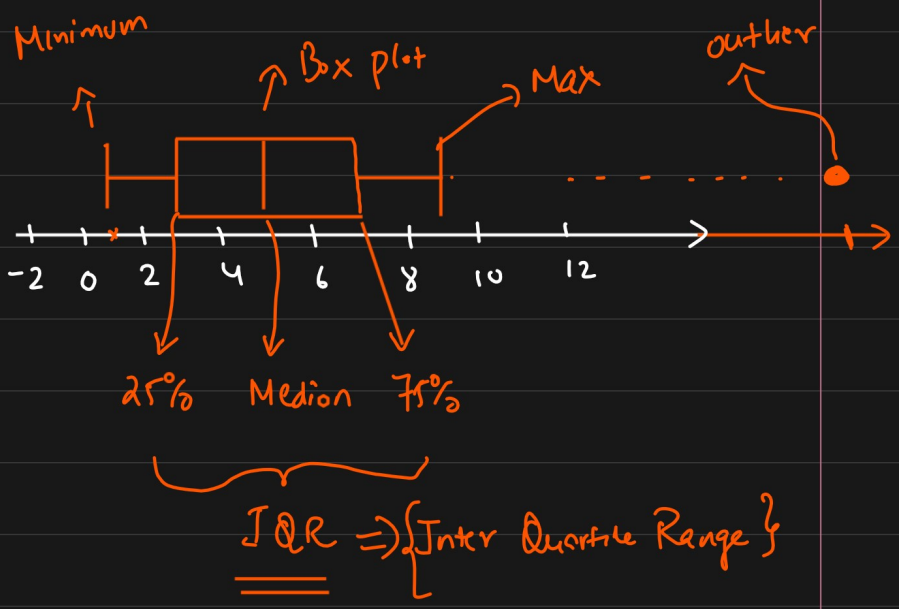
$\frac{5+5}{2} = 5$

1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~20~~

5 Number Summary

Minimum = 1
Q1 = 3
Median = 5
Q3 = 7
Max = 9

Box plot → Outliers



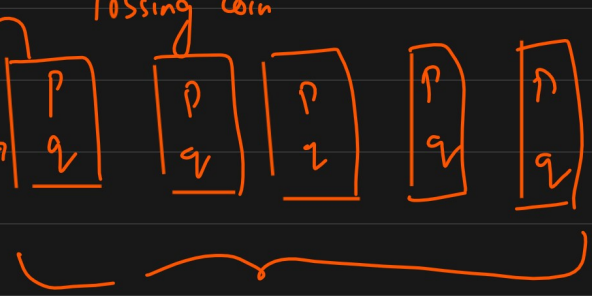
IQR ⇒ {Inter Quartile Range}

Use of Box plot

Distributions

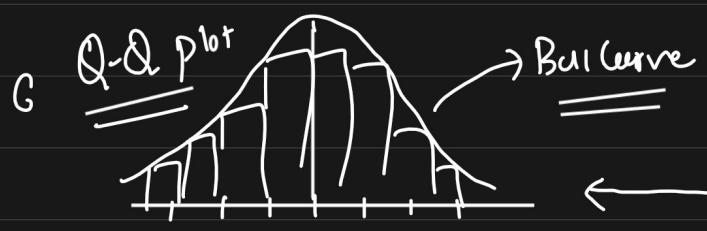
1. Normal / Gaussian Distribution ✓
2. Standard Normal Distribution ✓
3. Z - Score ✓
4. Log Normal Distr ✓
5. Bernoulli's Distribution ✓
6. Binomial Distribution }

Bernoulli Distr



1) Gaussian / Normal Distribution

Properties { Power Law }



1) Empirical Rule of Gaussian Distribution ⇓ 80-20%

↳ DATASET → IRIS Dataset } → Petal, Sepal length

⇓ Domain expertise }

2) Weight of human beign 3) Height → Doctor

68.2, 95.4 – 99.7

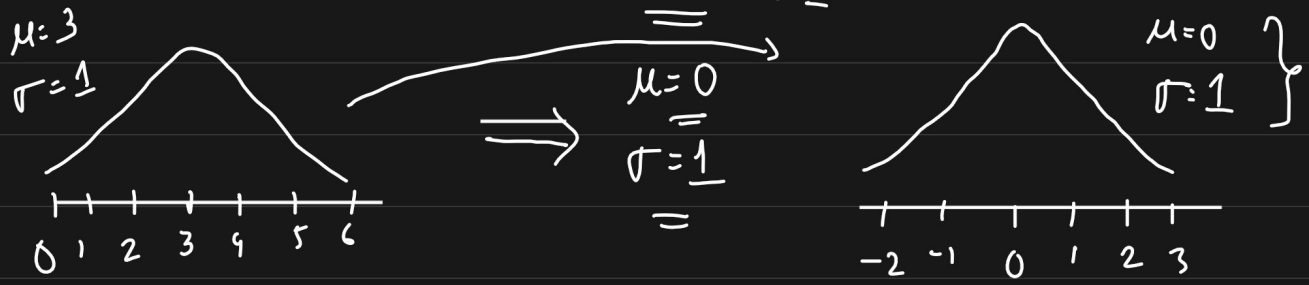
Outliers

Standard Normal Distribution

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

$\{1,2,3,4,5\}$ $\quad \mu = 3$

$\sigma = 1.414 \approx 1$



$\{1,2,3,4,5\}$ $\quad \left\{ Z\text{-score} = \frac{x-\mu}{\sigma} \right\}$

$\frac{2-3}{1}$ $\quad \frac{3-3}{1} = 0$ $\quad = \frac{1-3}{1}$

$= -2$

Why ??

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

Standardization Vs Normalization

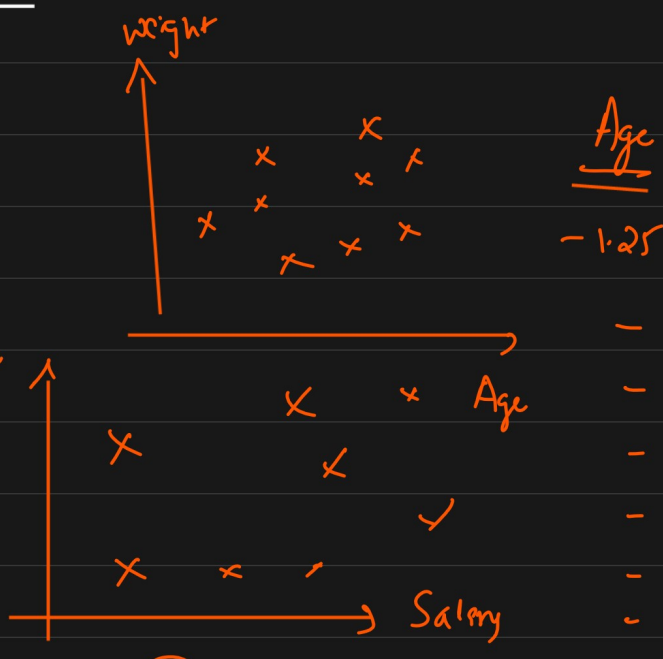
Age (years)	Weight (kg)	Salary (INR)
25	75	25K
26	80	30K
28 ✓	85	40K
30	60	80K
32	70	

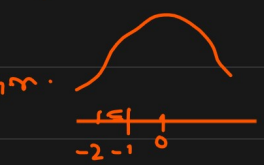
Different unit

Same unit scale ??

Maths → Scale

$\frac{25 - 28.2}{2 \text{ to } 6}$





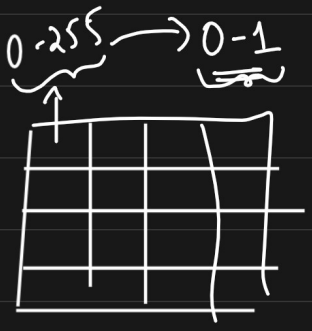
ML Library

Normalization [Min Max Scaler] → 0 to 1 / 0 to 1

{Standardization} & Normalization

{Min Max Scaler}

Normalization ⇓ CNN ← Convolutional Neural N/w.



Normalization (0-1)

$$\left\{ x_{Norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \right\}$$

⇓

Min Max Scaler = 0 to 1

f1
2
5
6
8 ✓
1 ✓

f1'
0.14
0.571
0.71
1
0

$\frac{8-1}{8-1} = 1$

$\frac{1-1}{8-1} = 0$

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

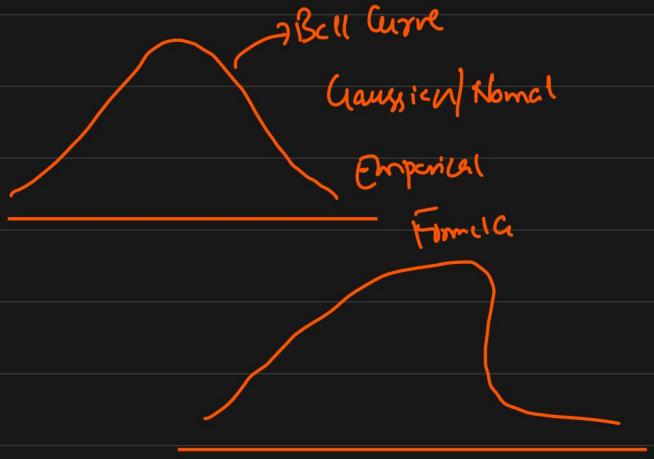
$= \frac{2-1}{8-1} = \frac{1}{7} = 0.142$

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

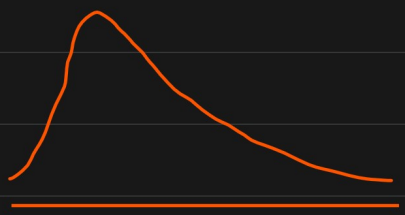
https://www.geeksforgeeks.org/normalization-vs-standardization/

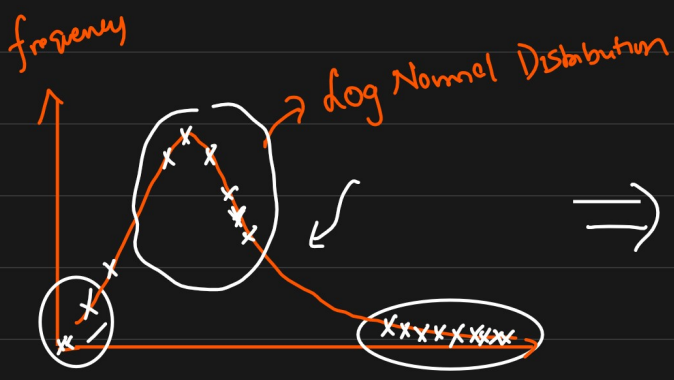
Log Normal Distribution

A lognormal distribution is a continuous probability distribution of a random variable in which logarithm is normally distributed. Thus, if the random variable has a lognormal distribution, then Y=lnhas a normal distribution. Likewise, if has a normal distribution, then X=exphas a lognormal distribution.



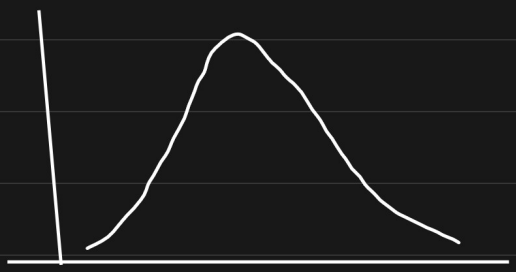
Skewed Curve

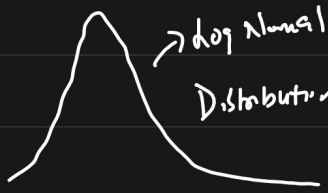




Gaussian Distribution

Normal Distribution

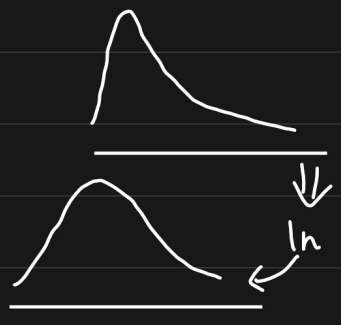




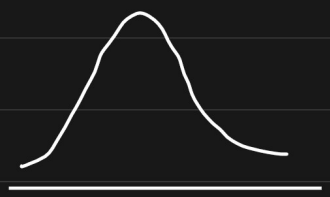
X = Log Normal Distributed

$\{Y = \ln(x)\}$ Gaussian Distribution

$\{X = \exp(y)\} \rightarrow e^{y}$



X	$Y = \ln(x)$
25	=
30	=
40	=
35	=



(*) Bernoulli's Distribution.