five number Summary

1) minimum

2) first Quartile (Q_1) 25%

3) median (Q_2) 50%

4) Third Quartile (Q_3) 75%

5) maximum.

Note: Choose these 5 numbers after removing the outlier from the data by finding boundary Values.

[Lower fence upper fence].

$LF = Q_1 - 1.5 \, (IQR)$

$UF = Q_3 + 1.5 \, (IQR)$

IQR = (inter Quartile Range)

$IQR = Q_3 - Q_1$

$\{1, 1, 2, 3, 4, 4, 4, 5, 5, 6, 7, 7, 8, 8, 9\}, \underbrace{\{28, 36\}}_{\text{outlier}}$ - 17th

~~$LF = Q_1$~~

$Q_1 = \frac{25}{100} \times 18$

$= 4.5$ index

$Q_1 = 3$

$Q_3 = \frac{75}{100} \times 16$

$= 13.5$ index

$Q_3 = 8$

$IQR = Q_3 - Q_1$

$= 8 - 3 = 5$

$LF = Q_1 - 1.5 (IQR)$

$= 3 - 1.5 (5)$

$= -4.5$

$UF = 8 + 1.5 (5)$

$= 15.5$

y < 4.5 and > 15.5 is outlier

minimum = 1

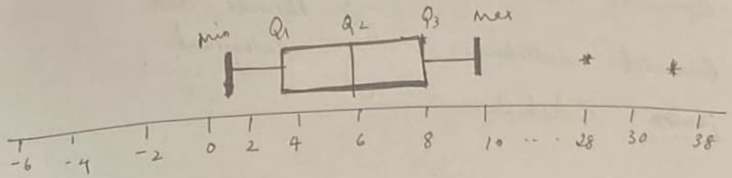
$Q_1 = 3$

$Q_3 = 8$

median = 5

max = 9

Boxplot



This graph is used to find the outlier

Different types Distribution

- To understand data patterns
- To summarise the data easily
- To calculate probability
- To make prediction and Decision
- To choose right statistical test

There are 2 category of

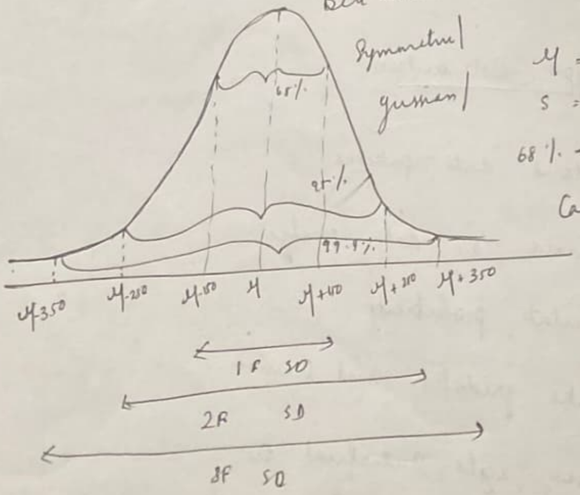
1) Continuous distribution (Number)
↳ any value including decimals (measuring)
eg: weight = 52.8 kg, time = 4.36 sec

2) Discrete distribution (Categorical distribution)
↳ specific values - you count them
eg: No. of students in class = 20, 21, 22 (not 20.5)

1) Normal distribution } continuous dist numerical
2) Standard normal distribution

3) Bernoulli distribution } Discrete dist Categorical.
4) Binomial distribution
5) poison distribution

1) Normal distribution



Bell curve / Symmetrical / gussian /

$\mu = 1.5$

$s = 0.5$

68% - 95% - 99

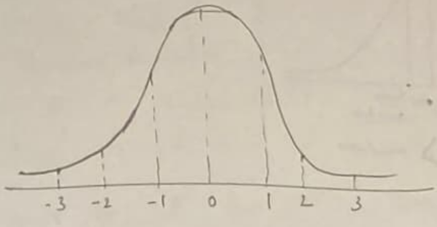
Called as confidence interval

Empirical rules

- 68% of data will be present in 1SD
- 95% of data will be present in 2SD
- 99.7% of data will be present in 3SD

2) Standard Normal distribution

$\mu = 0$ } (always)
$SD = 1$



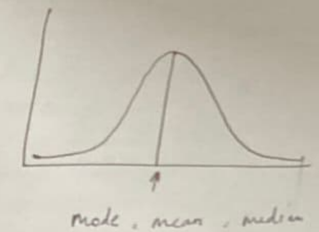
$$Z \text{ score} = \frac{x_i - \mu}{\sigma} = \frac{2 - 3.86}{2}$$

Normal Dist data	Standard Normal dist data
2	-0.93
7	1.57
5	0.57
4	0.07
1	-1.43
3	-0.43
5	0.57

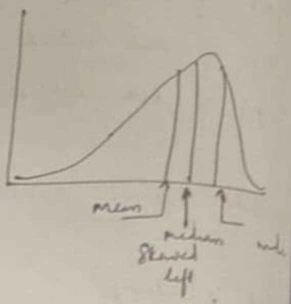
$\mu = 0$

$\sigma = 0.94$

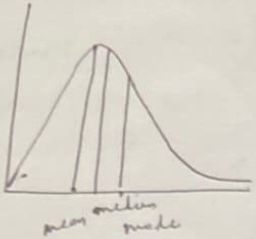
Symmetric



Symmetric



mean < median



mean > median

Positive Skew (Right Skew).

Tail on the right side is longer most data are the left

Negative Skew (Left Skew)

Tail on the left side is longer most data are on the right

Zero Skew (Symmetric)

The data is evenly distributed around the mean (like a normal distribution)

$$\boxed{\text{Skewness} = \frac{3(\text{mean} - \text{Median})}{\text{Standard deviation}}}$$

If Value near to -1 then it is -Ve Skew

If Value near to +1 then it is +Ve Skew

If Value lies in -0.5 to 0.5 Then it is Zero Ske

Kurtosis

$$K = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)$$

It measures the tailedness or peakness of d

distribution

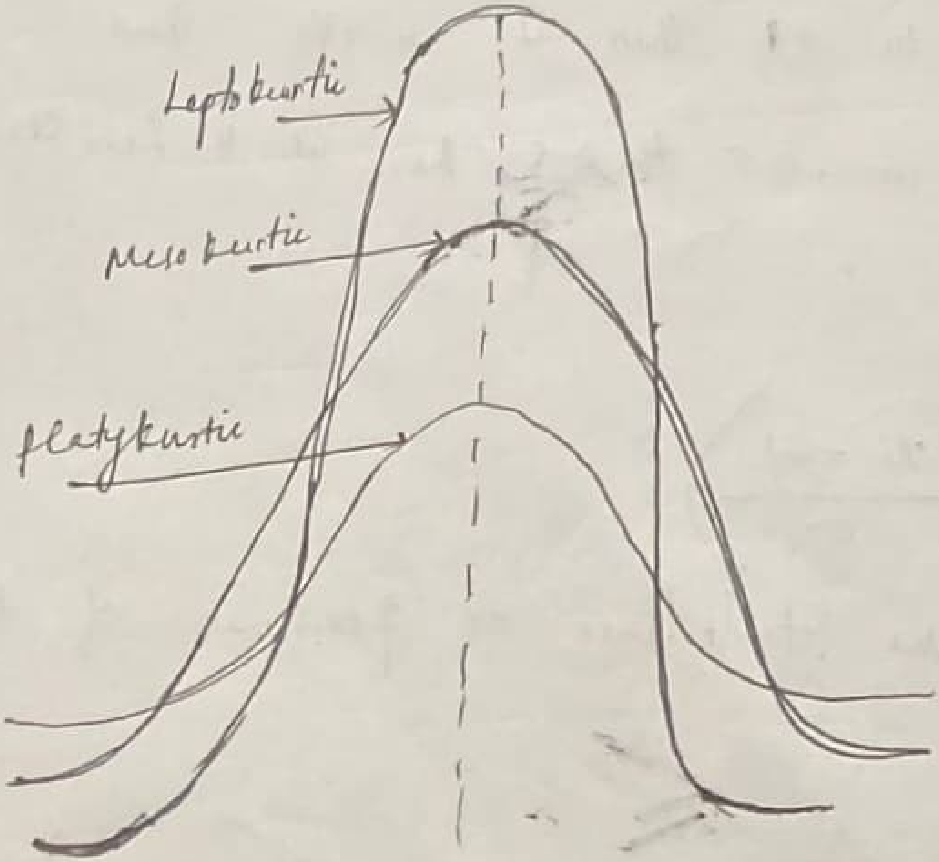
Types of kurtosis

1) Mesokurtic ($K=3$)
- Normal distribution
- No outliers
- Moderate tail and peak.

2) Leptokurtic ($K>3$)
- Heavy tails and sharp peak
- more outliers

$P(X=k) = C(n,k)\, p^k (1-p)^{n-k}$

3) platy kurtic ($K<3$)
- light tail and flat peak
- fewer outliers



5) poison distribution

It is used to model the number of event tha
occur in a fixed time interval or space and
occur independently the parameter λ represents th
avg number of event in the interval

$$\boxed{P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}}$$

λ = is the avg no of ev
k = no of argument

how many times something
happens in a fixed time or area