Discriptive statistics

1) → Measure of central tendency

- → Mean
- → Median
- → Mode

Population mean (μ)

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

Sample Mean ($\bar{x}$)

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

Median

* Sort the values either asc or dsc order
* create the mid value
* If you get mid 2 values take avg of those two values

Q. [1,2,2,3,4,5] [1,2,2,3,4,5,100] (outlier (far away from 5))

mean = 2.8 Mean = 16.7

Median = 2.5 Median = 3

* Mean will be affected by outlier where as Median won't affect by outlier
* These are calculated for null values imputation

Mode

Most repetative value.

Measure of Dispersion

Eg:-

	Person 1	Person 2
Monday	7:30 AM	8 am
Tuesday	7:45 AM	11 am
wednesday	8 AM	9 am
Thursday	7:15 AM	7 am
friday	7 AM	10 am
Prediction	7-8	9-10
	15 min	2 hr
	30 min	1 hr

Variance high
Prediction accuracy low

Variance less
Prediction accuracy high

- → Variance
- → Standard deviation
- → Range

Population Variance (σ^2)

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

Sample Variance (s^2)

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

Imp. (interview question)

[n-1 = degree of freedom.]

Q. calculate σ^2 & s^2

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

$$\sigma^2 = \frac{(1-2.8)^2 + (2-2.8)^2 + (2-2.8)^2 + (3-2.8)^2 + (4-2.8)^2 + (5-2.8)^2}{6}$$

$$= \frac{3.36 + 0.64 + 0.64 + 0.04 + 1.44 + 4.84}{6} = \frac{10.8}{6} = 1.8 \text{ km}^2$$

variance

$\sqrt{\sigma^2} = \sqrt{1.8\ km^2}$

$\sigma = 1.34$ km (Standard deviation)

$\mu = 2.8$

Standard deviation is square root of σ^2 ($\sqrt{\text{variance}}$)

Population SD (σ) $= \sqrt{\sigma^2}$

$\sigma = \sqrt{1.8}$

$\sigma = 1.34$

Sample variance (s^2)

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{10.83}{5} = 2.16$$

Sample SD (s)

$s = \sqrt{s^2}$

$s = \sqrt{2.16}$

$s = 1.46$

Range (maximum − minimum)

$= 5 - 1 = 4$

Percentile and Quantile

Percentile is the value below which a certain percent of observation be will come

* {1, 1, 2, 3, 4, 5, 5, 6, 7, 7, 8}

How much % of data will come below 6?

Percentile rank of x = $\frac{\text{No of value below } x}{N} \times 100$

$= \frac{7}{11} \times 100$

$= 63\%$ observation data value is < 6

Quantile

Quantile helps to find the value which is present at the given percentile rank.

* Which value is present at 25%?

value = $\frac{\text{Percentile}}{100} \times n+1$

$= \frac{25}{100} \times 12 = 3 \rightarrow$ index

value = 2.

90%?

$= \frac{90}{100} \times 12$

$= 10.8 \rightarrow$ Index

$10 \rightarrow$ 8 have be value before dei

value = 7

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 outliers from the data by find 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, [28, 36]} ← outliers

IQR = $\frac{75}{100} \times 17 - \frac{25}{100} \times 17$

= 13.5 − 4.5 [Index]

= 8 − 3

= 5

LF = $3 - 1.5 \times 5 = 4.5$

UF = $8 + 1.5 \times 5 = 15.5$

~~4.5~~ Anything < 4.5 & > 15.5 is a outlier. (Remove them).

Minimum = 1

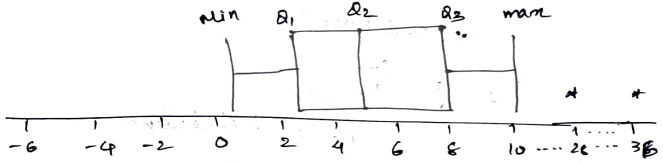
(Q_2) Median = 5

Maximum = 9

$Q_1 = 3$

$Q_3 = 8$

Box plot



This graph is used to the outliers.

Different types of Distributions

1.) To understand data patterns
2.) To summarize the data easily
3.) To calculate Probability
4.) To make prediction and Decision.
5.) To choose right statistical test.

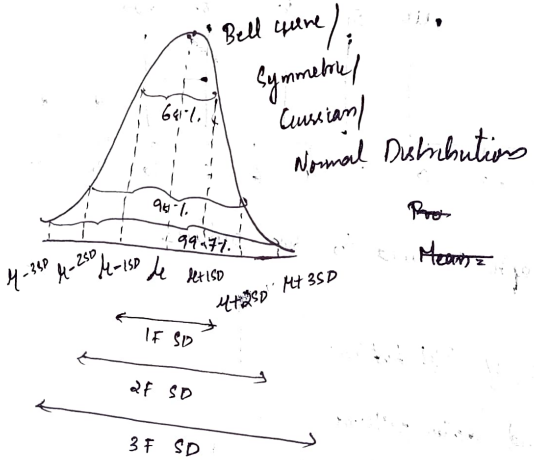
There are 2 category of distribution :-

→ Continous distribution (Number, Numerical Distribution)

→ Discrete distribution (categorical Distribution)

1) Normal Distribution } CD numerical
2) standard normal Distribution
3) Bernoulli Distribution } categorical DD.
4) Binomial Distribution
5) Poisson Distribution

1) Normal Distribution



Bell curve / Symmetric / Guassian / Normal Distribution

~~Properties~~

~~Mean =~~

Property

Mean = Median = Mode

68% - 95% - 99.7% is called as confidence Interval

Emphirical Rules :-

68% of data will present in 1SD.

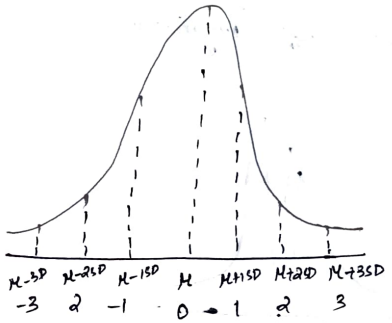
95% of data will present in 2SD.

99.7% of data will present in 3SD.

2) Standard Normal Distribution

$\mu = 0$

$SD = 1$



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

Normal Dist Data	Std Normal Dist Data $\left[\frac{x_i - \mu}{\sigma}\right]$
2	-0.93
7	1.57
5	0.57
4	0.07
1	-1.43
3	-0.43
5	0.57
$\mu = 3.86$	$\mu = 0$
$\sigma = 2.7$	$\sigma = 0.94 = 1$