what is statistics :- It is science, of collecting, organising and analyzing the data.

Data :- Facts (or) pieces of information that can be measured.

Types of Statistics :-

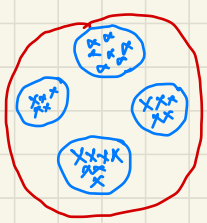
(i) Descriptive stats :- it consists of organizing and summarizing data.

(ii) inferential stats :- it is a technique where we use the data that we have measured to form conclusions.

Population and sample :- Population denoted by N
Sample denoted by n

example :- Elections :- Goa.

For exit poll's the news reporters cannot ask each and everyone whom you have voted so what they do is take random people from Random Regions & which ever party gets more votes they conclude.



The red circle is population of goa & the blue circle are the samples

Sampling Techniques :-

① Simple Random Technique :- just pick Random data.
In simple Random sampling every member of the population equal probability of getting selected

③ Stratified Sampling :- where the population (N) is split into non-overlapping groups. (strata)

③ Systematic sampling :- From the population (N) we pick every nth individual.

④ convienience sampling :- we choose only those people who are experts (or) interested in doing this.

Variables :- it is a property that can take values
ex :- height, weight.

Two kinds of variable :-
(1) Quantitative variable :- measured numerically
ex :- age, marks etc
(2) Qualitative variable / categorical variable :- Here we will have categories for the variable & we cannot perform mathematical operations
Ex :- male, female.

Quantitative variables

Discrete variable
we will have the whole numbers only
ex :- Age, no. of houses etc

Continious variable
we will have real numbers
ex :- height, weight etc.

Variable measurement Scales :- 4 types of measured variable

(i) Nominal :- The data is categorical we cannot perform any mathematical operations ex :- Colours, gender, Subjects.

(ii) ordinal :- The order of the data matters but value doesnot. ex :- 5 Student marks

Student marks	Rank
100	1
96	2
57	4
85	3
44	5

Here we are more worried about the Rank rather than the number of marks.

(iii) interval :- The value & order of the data matters, natural zero is not present.

ex :- temparature

70-80 80-90 90-100

(iv) Ratio data :-

* Frequency distribution :-

Sample dataset :- Rose, Lilly, Sunflower, Rose, Lilly, Sunflower, Rose, Lilly, Lilly.

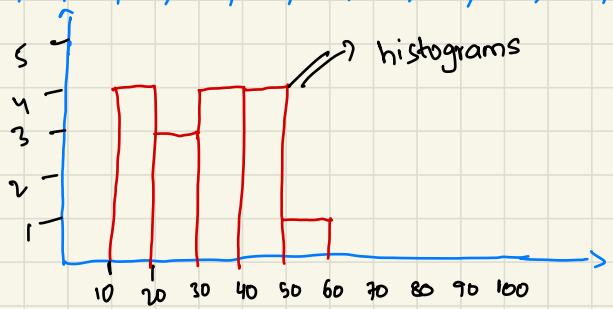
Frequency distribution table :-

Flower	frequency	Cummulative-freq
Rose	3	3
Lilly	4	7
sunflower	2	9

Note :- * if the variable is discrete then we use bar graph if the variable is continious the we will use histogram

② Histograms :- we will divide in bins, the size is 10

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



* Measure of central tendency :- It is used to measure (or) determine the center of the data.

(i) Arithematic mean for population and sample :-

Population :- (N)

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

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

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

Sample (n)

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

$$= 3.2 //$$

* Median :-

x : { 1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100 } => This is an outlier & it is also effecting the data adversly.

mean = 3.2 => This was before we added 100

$$\text{mean} = \frac{32+100}{11} = 12$$

Now we can see that there is huge difference in mean from before adding 100 (3.2) to after adding 100 (12)

Median :- {1,1,2,2,3,3,4,5,5,6,100}

1) while doing median we have to sort the numbers
2) For odd numbers :- The middle element is called median
   {1,1,2,2,3,3,4,5,5,6,100}
   For the above data 3 is the median.
3) For even numbers :- The middle 2 elements Average is the median {1,1,2,2,3,3,4,5,5,6,100,112}
   For the above data $\frac{3+4}{2}$ is the median.

*) mode :- {1,2,2,3,4,5,6,6,6,7,8,100,200}
most frequent element in the data is called the mode
∴ In the above data my mode is 6

mode is also used to fill the missing values in the data, But works good for categorical variables.

* Measure of Dispersion :- How well spread your data is.

(1) variance :-

Population variance

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

Sample variance

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

example :-

x	μ	$x-\mu$	$(x-\mu)^2$
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	1.87
5	2.83	2.17	4.71
			10.84

population variance = $\frac{10.84}{6} = 1.81$,,

Standard deviation :- it is the square root of variance.

$\sqrt{1.81} = 1.34$

* with the help of variance we can say how the data is spread.

* with the help of S.D we can give range in which a particular value is present.

Percentiles and Quartiles :-

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

ex :- {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

what is percentile Ranking of 10?

$$= \frac{16}{20} \times 100 = 80\%$$

[Here we can say that 80 percentage of the values are less than 10]

what is percentile Ranking of 11 ?

$$= \frac{17}{20} \times \frac{5}{100} = 85\%.$$

[Hence we can say that 85% of the values are less than 11]

Five number Summary :-

1) minimum
2) First Quartile (Q_1)
3) median
4) Third Quartile (Q_3)
5) maximum

Removing the outliers :-

whenever we want to remove an outlier we need to have a Lower fence & higher fence. that means.

(i) All the numbers above the higher fence are outliers
(ii) All the numbers below the Lower fence are outliers.

Lower fence = $Q_1 - 1.5 * (IQR)$
upper fence = $Q_3 + 1.5 * (IQR)$

IQR (inter quartile range) = $Q_3 - Q_1$
(75%) − (25%)

ex :- {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27} n = 19
(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19)

Calculate the higher and the Lower fence?

$Q_1 = \frac{25}{100} \times (20) = 5 = 3$

$Q_3 = \frac{75}{100} \times (20) = 15 = 7$

Lower fence : $3 - (1.5)(4)$
$= 3 - 6$
$= -3$

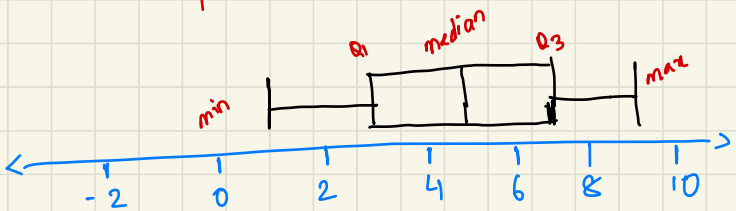
upper fence : $7 + (1.5) * 4$
$= 7 + 6$
$= 13$

Data after removing outliers :-

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

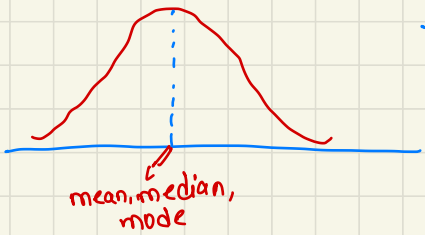
min = 1
Q1 = 3
median = 5
Q3 = 7
max = 9

Box plot :-

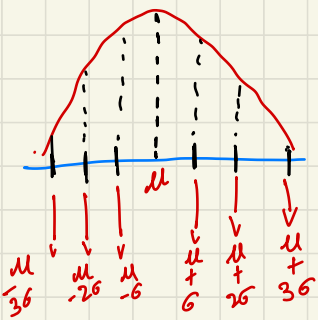


① Distribution :- it is used to get a brief idea about the dataset.

(i) Gaussian / normal distribution :-



The 2 halfs of the bell curve are symmetrical.



emperical formula :-

68-95-99.7% Rule

Between the $\mu \pm \sigma$ (1st SD) 68% of the data is present.
$\mu \pm 2\sigma$ (2nd SD) 95% of the data is present.
$\mu \pm 3\sigma$ (3rd SD) 99.5 of the data is present.

Z-Score :- it will help us in telling how much standard deviation it is away from the mean.

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

Note :- when you apply zscore for a particular distribution it will be converted to Standard normal distribution.

In Standard normal distribution mean = 1 & S.D = 0.

practical Application :-

Datasct :-

Age (=> years)	Salary (=> Rs)	weight (=> kg)
24	40K	70
25	80K	80
26	60K	55
27	70K	45

For this datasct we can see that the features have different units Hence to make them similar we do standarization, such that mean = 0 & $\sigma = 1$

* The procesn of converting your data in a way that mean = 0 and SD = 1 is called as standardization

Normalization :- Here you will have an option to convert all the values blw zero and one (0, 1)

$$\text{min, max scalar} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

practical Applications :- { Ind vs SA }

(2021)
① ODI Series :-
Scores Avg = 250
SD = 10
R.pant score = 240

(2020)
② ODI Series :-
Avg = 260
SD = 12
R.pant = 245

Q.) In which year R.pant has better score ?

2021 (Z score)

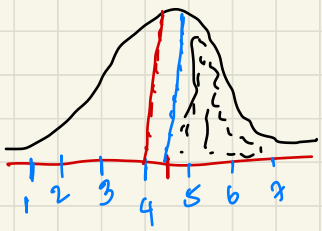
$= \frac{-10}{10} = 1$

2020 (Z score)

$= \frac{-15}{12} = -1.25$

interview problem :-

a) what percentage of scores fall above 4.25?



1) do the Zscore = $\frac{4.25 - 4}{1}$ = 0.25 [look this at z-table]

2) Since it is bell curve total consider as 1

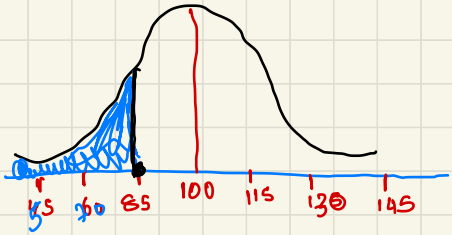
So percentage above 4.25 is = 1 − (0.25 value in z-table)
= 1 − 0.5987
= 40%.

Q) in india the average IQ is 100, with a standard deviation of 15. what percentage of the population would you expect to have an IQ lower than 85%.

$$Z\text{score} = \frac{x - \bar{x}}{\sigma}$$

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

$$= -1$$



∴ Lower IQ people are 15.86%.