STATS – FSDS 2.0

Defn: Statistics is the science of collecting, organizing and analyzing data

Data = "facts or pieces of information"

Eg: Heights of students in classroom
IQ of students
Daily Activities
Weight of people, Age.

Types of Statistics

① Descriptive Stats

Defn: It consists of organizing and summarizing data

① Measure of Central Tendency
{mean, Median, Mode}

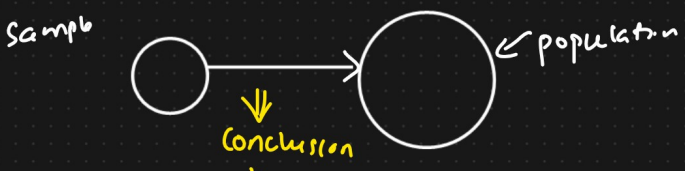
② Measure of Dispersion
{Variance, Standard deviation}

③ Different types of Distribution of data
Eg: Histogram, pdf, pmf, cdf

CLT

② Influential Stats

Defn: It consists of data you have measured to form conclusion



C.I, K, pvalue Hypothesis Testing

① Z-test
② t-test
③ Chi Square Test
④ ANOVA
⑤ F Test

⇒ Conclusion of sample on population.

Eg: Let say there are 20 classes in your college and you have collected the heights of student in the class.

Heights are recorded [175cm, 180cm, 140cm, 135cm, 160cm, 170cm]

Descriptive

"What is the average heights of the students in the classroom"

Mean = 160cms

Influential

Sample

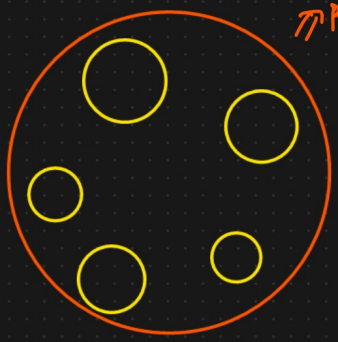
"Are the height of the students in the classroom similar to what you expect in the college"

→ population

(N) (n)

Population And Sample data

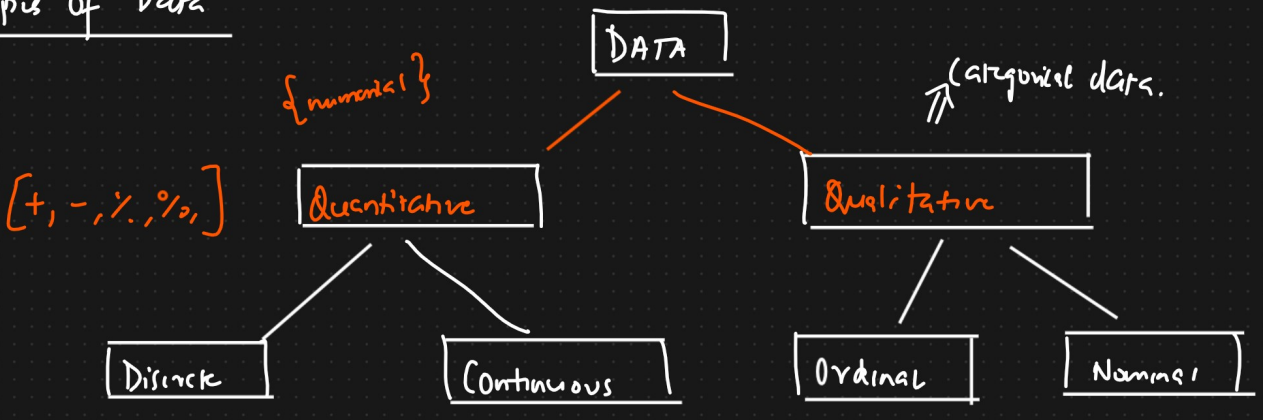
Exit Poll



Sample size = 500

Each small circle representing sample size of 500

⑥ Types of Data



Whole numbers with some range

Eg: No of bank accounts of people
No. of children in a family

Any value

Eg: Weight, Height
Age, Temperature,
Speed, Salary

Eg: Ranks

[3 Good, 2 Better, 1 Best]

Eg: Gender

M, F

Blood group

Color of Hair

Pincode

(4) Scales Of Measurement

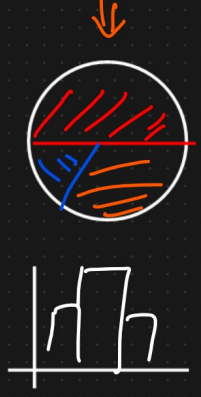
(1) Nominal Scale Data
(2) Ordinal Scale Data
(3) Interval Scale Data
(4) Ratio Scale Data.

(1) Nominal Scale Data

1) Qualitative / categorical Data.
   Eg: Gender, Colors, Labels
2) Order does not matter

Eg: Favorite Color

Red → 5 → 50%
Blue → 3 → 30%.
Orange → 2 → 20%.
10



(2) Ordinal Scale Data

(1) Categorical Data
(2) Ranking and order matters
(3) Difference cannot be measured

Eg: Best → 1
Good → 2
Bad → 3

Race

1st 3min
2nd 4min
3rd 5min

1st Rank → 90
2nd Rank → 70
3rd Rank → 40

③ Interval Scale Data

① The order matters

② Difference can be measured

③ Ratio cannot be measured

④ No "0" starting points

mtlb ye necessary nii hai ki temperature hmesa zero se hi start ho. -ve bhi ho skta hai

Eg: Temperature

2:1 ka mtlb ye nii hai ki 2X garmi lgega that is proving point 3

Cold or Warm

30F ← 60F ← 90F ← 120F

60:30 = 2:1

Room 1 — 30F

Room 2 — 60F

0

-15

-30

④ Ratio Scale Data

Eg: Student marks in class

① The order matter {sort this numbers} . 0, 30, 45, 60, 90, 95, 99

② Differences are measurable including ratios

③ Contain a 0 starting point

Example

① Marital Status [Nominal Scale Data]

② Favourite food based on Gender? [Nominal]

③ IQ measurements [Ratio Scale].

⇓

Ordinal

Descriptive Stats

① Measure of Central Tendency

① Mean ② Median ③ Mode.

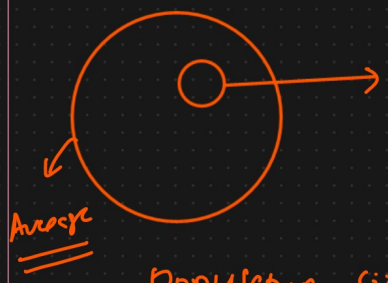
① Mean :

Population (N) | **Sample (n)**

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

$$\text{Population Mean } (\mu) = \sum_{i=1}^{n} \frac{x_i}{N} \qquad \text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$$

$$\frac{1+1+2+2+3+3+4+5+5+6}{10} = 3.2$$



Population size (N)

Sample size (n)

② Median

$X = \{4, 5, 9, 3, 2, 1\}$

Steps

① Sort the Random variable $\{1, 2, 2, 3, 4, 5\}$

② No. of elements

③ if Count == even

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

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

if count == odd

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

3 median

Why Median ?

Mean are affected by outliers

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

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

Median = 3

X = {1,2,3,4,5,100}

$\bar{x} = \frac{1+2+3+4+5+100}{1} = \frac{115}{6} \approx 19$

X = {1,2,3,4,5,100}

median = $\frac{3+4}{2} = 3.5$

Conclusion :

Median is used to find the central Tendency

When outlier is present.

③ Mode : Maximum Frequency occurring element

{2,1,1,1,4,5,7,8,9,9,10}

Mode : 1

EDA and Feature Engineering

Exploratory data analysis

- Missing Value

we can't remove the row of missing value, bcz it will cause loss of data. So for numerical features we replace the missing value with mean or median and for categorical data we use mode to fill the missing values

Age	Weight	Salary	Gender ↓ Mode	Degree ↓ Mode
24	70	40K	M	B.E
25	80	70K	F	- B.E
27 ← Outliers	95	45K	F	- B.E
24	-	50K	M	PHD
32	-	60K	- M	B.E
- Median or Mean	60	-	- M	Master
-	65	55K	- M	BSc
40	22	-	M	B.E

② Measure of Dispersion (Spread) [Spread of the data]

① Variance (σ^2)

② Standard deviation (σ)

① Variance

Population Variance

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

x_i ⇒ Data points

μ ⇒ population mean

N ⇒ populate size

Sample Variance

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

x_i ⇒ data points

$\bar{x}$ ⇒ Sample mean

n ⇒ sample size

Assignment : Why we divide sample variance by n-1?

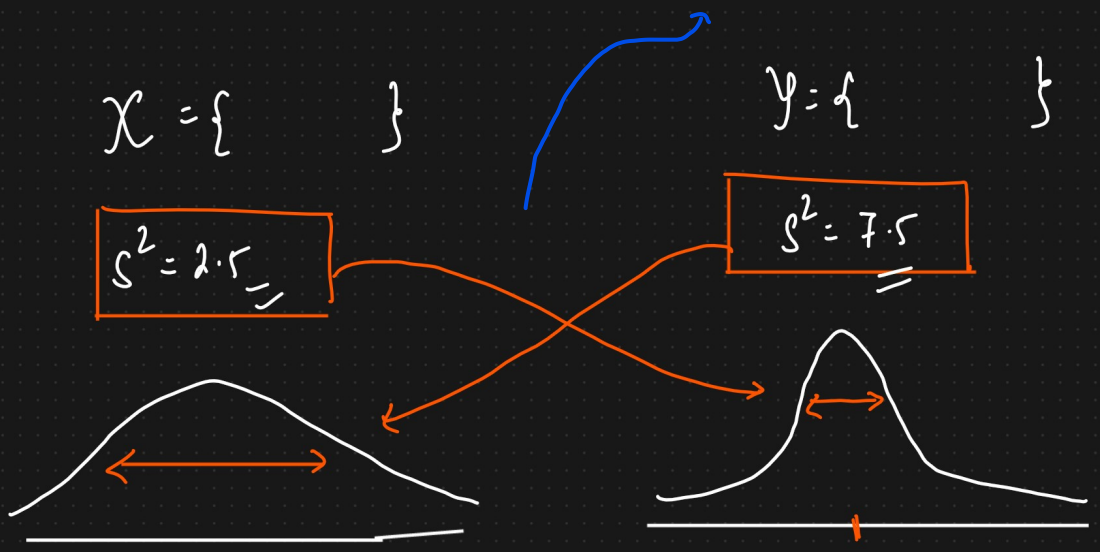
Eg: {1, 2, 3, 4, 5}.

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

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
$\bar{x} = 3$		$\sum (x_i - \bar{x})^2 = 10$

$$s^2 = \frac{10}{4} = 2.5$$

when the variance is less, spread is also less



④ Standard deviation

Population std σ = $\sqrt{\text{variance}}$

Sample std $s = \sqrt{s^2}$

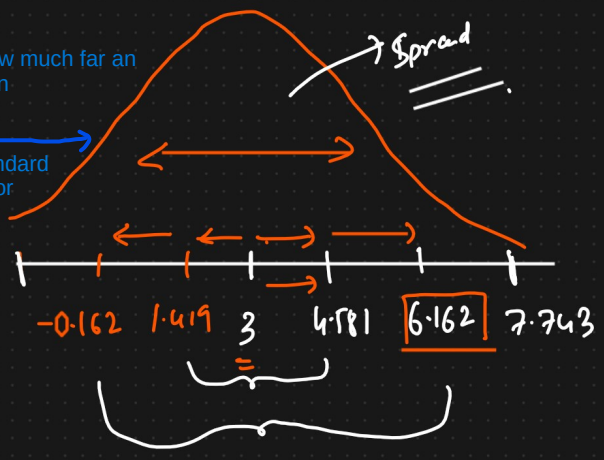
$X = \{1, 2, 3, 4, 5\}$

$\bar{x} = 3$

$s = 1.581$

Standard deviation talks about how much far an element is far away from the mean

variance is the square of the standard deviation i.e. spread of the data or we can say how well the data is spread



3.00
1.581
4.581
1.581
6.162
1.581
7.743

ⓕ Random Variable

Linear Algebra $\left\{ \begin{array}{ll} x+5=7 & \Rightarrow x=2 \\ 8=y+x & \boxed{y=6} \end{array} \right\}$ Variables

Random Variable is a process of mapping the output of a random process or experiment to a number.

Eg: Tossing a Coin $\{Head, Tail\} \Rightarrow$ Process

$$X = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$$

$X = \{$ Ages of student in a class $\}$

Eg: Rolling a dice $\{1,2,3,4,5,6\}$

$Y = \{$ Sum of rolling of dice 7 times $\}$

$\Downarrow$

$Pr(Y \geq 15) =$ _____ $Pr(Y < 10)$

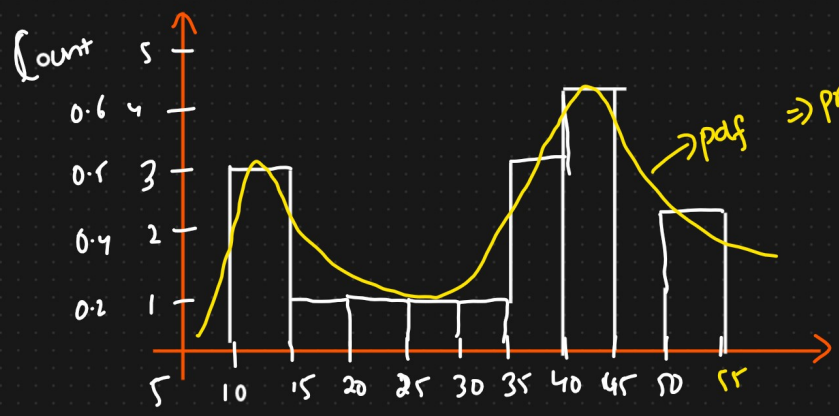
$Pr(40 \leq Y \leq 15) =$

① Histograms And Skewness → [Frequency]

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

$\frac{50}{10} = 5$ → bin size

No of Bins = 10 → buckets



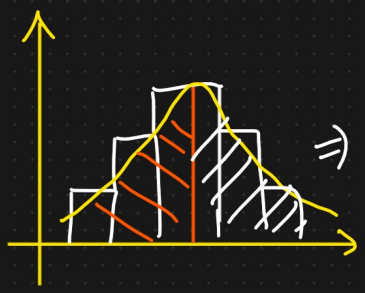
⇒ probability density function

[Kernel density estimator]

Smoothening is happening through this process, will discuss later

Skewness

$\chi =$



⇒ Normal Distribution

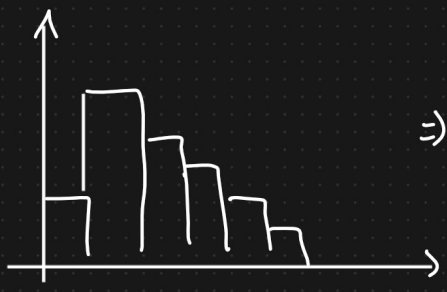
Symmetrical Distribution

⇓

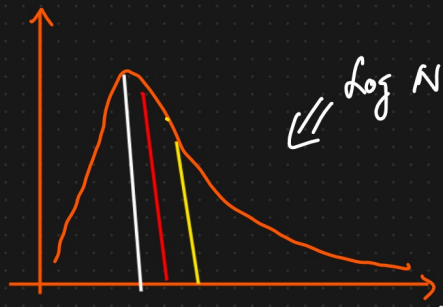
No skewness

Median = Mean = Mode

② Right skewed



⇒ Positive skewed



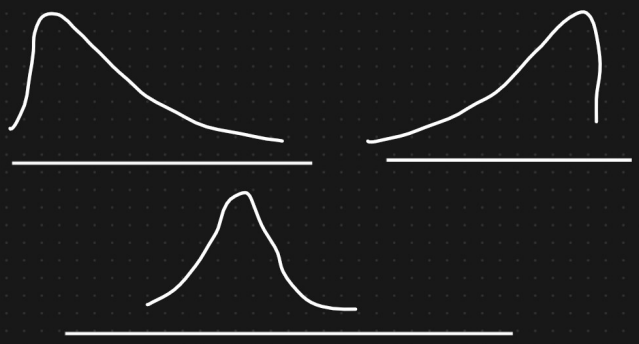
Log Normal Distribution

→ X Length of the comment

mean > median > mode

③ Left skewed

mode > median > mean



Knowledge sharing ⟶ Profile Building