

STATS - FSDS 2.0

Dfn: 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

Dfn: It consists of organizing and summarizing data

① Measure of Central Tendency

{mean, Median, Mode}

② Measure of Dispersion

{Variance, Standard deviation}

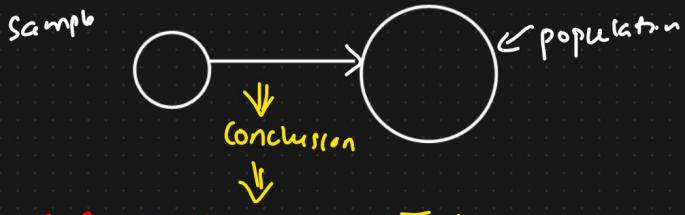
③ Different types of Distribution of data

Eg: Histogram, pdf, pmf, cdf

CLT

② Inferential Stats

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



C.I, P-value Hypothesis Testing

① Z-test

② t-test

③ Chi-Square Test

④ ANOVA

⑤ F-test

} \Rightarrow 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 height of the students in the classroom"

$$\underline{\text{Mean}} = 160 \text{ cms}$$

Inferential

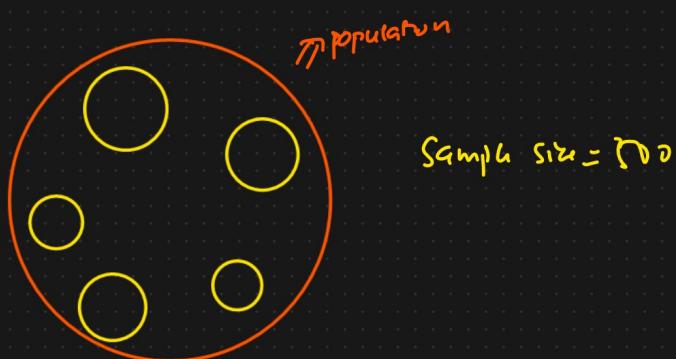
"Are the height of the students in the classroom similar to what you expect in the college"

Sample
π

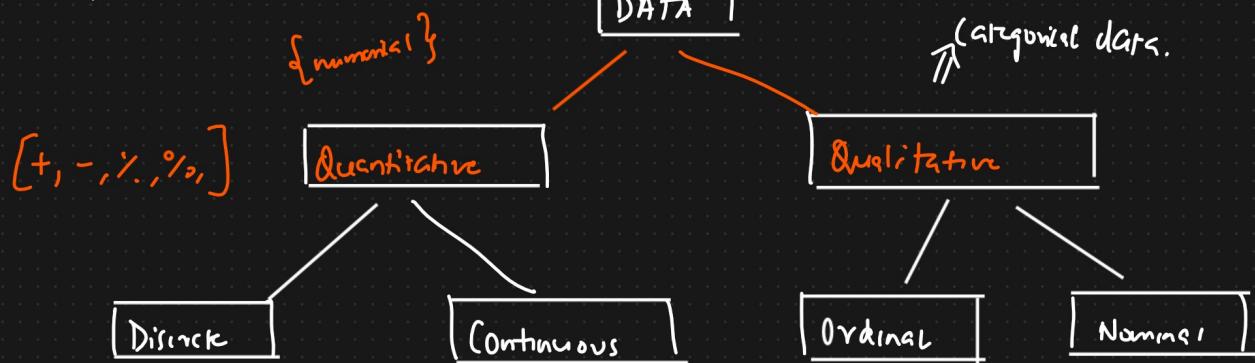


(N)
Population And Sample data

Exit Poll



(f) Types of Data



\Downarrow Whole numbers with some range	\Downarrow Any value	\Downarrow <u>Eg: Ranks</u> $\begin{bmatrix} 3 & 2 \\ \text{Good, Better,} & \text{Best} \end{bmatrix}$ Age, Temperature, Speed, Salary	\Downarrow <u>Eg: Gender</u> M, F
<u>Eg: No. of bank accounts of people</u> <u>No. of children in a family</u>	<u>Eg: Weight, height</u> <u>Age, Temperature,</u> <u>Speed, Salary</u>	<u>Blood group</u> <u>Color of hair</u> <u>Pancake</u>	

④ Scales Of Measurement

- ① Nominal Scale Data
- ② Ordinal Scale Data
- ③ Interval Scale Data
- ④ Ratio Scale Data.

① Nominal Scale Data

i) Qualitative / Categorical Data.

Eg: Gender, Colors, labels

ii) Order does not matter

\Downarrow
Eg: Favorite color

Red \rightarrow 5 $\rightarrow 50\%$

Blue \rightarrow 3 $\rightarrow 30\%$

Orange \rightarrow 2 $\rightarrow 20\%$



Race

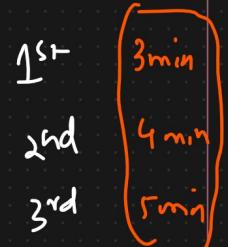
② Ordinal Scale Data

① Categorical Data

② Ranking and order matters

③ Difference cannot be measured

\Downarrow
Eg: $\begin{cases} \text{Best} \rightarrow 1 \\ \text{Good} \rightarrow 2 \\ \text{Bad} \rightarrow 3 \end{cases}$



1st Rank $\rightarrow 90\%$
 2nd Rank $\rightarrow 70\%$
 3rd Rank $\rightarrow 40\%$

③ Interval Scale Data

- ① The order matters
- ② Difference can be measured
- ③ Ratio cannot be measured
- ④ No "0" starting points

Eg: Temperature

$$\begin{array}{c} \text{Cold or} \\ \text{warm} \end{array} \left[\begin{array}{l} \leftarrow 30^{\circ}\text{F} \swarrow \\ \downarrow \\ \leftarrow 60^{\circ}\text{F} \swarrow \end{array} \right] \quad 60 : 30 = \boxed{2 : 1}$$

$$\begin{array}{c} \text{Hot or cold} \\ \left[\begin{array}{l} \leftarrow 90^{\circ}\text{F} \swarrow \\ \downarrow \\ \leftarrow 120^{\circ}\text{F} \swarrow \end{array} \right] \end{array} \quad 0$$

$$\begin{array}{ccc} \downarrow & \downarrow & \\ \text{Room 1} & \text{Room 2} & \\ 30^{\circ}\text{F} & 60^{\circ}\text{F} & -15 \\ & & -30 \end{array}$$

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)

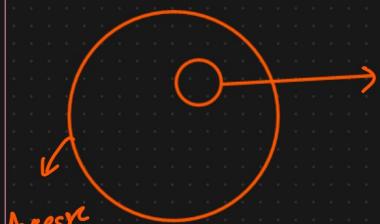
$$X = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6\}$$

Sample (n)

$$\text{Population Mean} (\mu) = \sum_{i=1}^n \frac{x_i}{N}$$

$$\text{Sample mean} (\bar{x}) = \sum_{i=1}^n \frac{x_i}{n}$$

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



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 \leq even

if count = odd

$$\{1, 2, \boxed{3}, 4, 5\}$$



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

$$\{1, 2, 2, \boxed{3}, 4, 5, 6\}$$



3 median

Why Median?

Mean are affected by outliers

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

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

$$\text{Median} = 3$$

$$X = \{1, 2, 3, 4, 5, \downarrow 100\}$$

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

$$X = \{1, 2, \boxed{3, 4}, 5, 100\}$$



$$\text{Median} = \frac{3+4}{2} = 3.5$$

Conclusion:

Median is used to find the central Tendency

When outlier is present.

③ Mode: Maximum Frequency occurring element

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

$$\text{Mode} = 1$$

EDA and Feature Engineering

- Missing Value

	Age	Weight	Salary	Gender	↓ Mode	↓ Mode
	24	70	40K	M	B.E	
	25	80	70K	F		- B.E
Outliers	27	95	45K	F		- B.E
	24	-	50K	M	PWD	
↓ Median	32	-	60K	[M]	B.E	
{	[]	60	-	[M]	Master	
Mean	[]	65	55K	[M]	BSC	
	40	72	-	M	B.E	

② Measure Of Dispersion [Spread of the data]

① Variance (σ^2)

② Standard deviation (σ)

① Variance

Population Variance

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

$x_i \Rightarrow$ Data points

$\mu \Rightarrow$ Population mean

$N \Rightarrow$ population size

Sample Variance

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

$x_i \Rightarrow$ data points

$\bar{x} \Rightarrow$ Sample mean

$n \Rightarrow$ Sample size

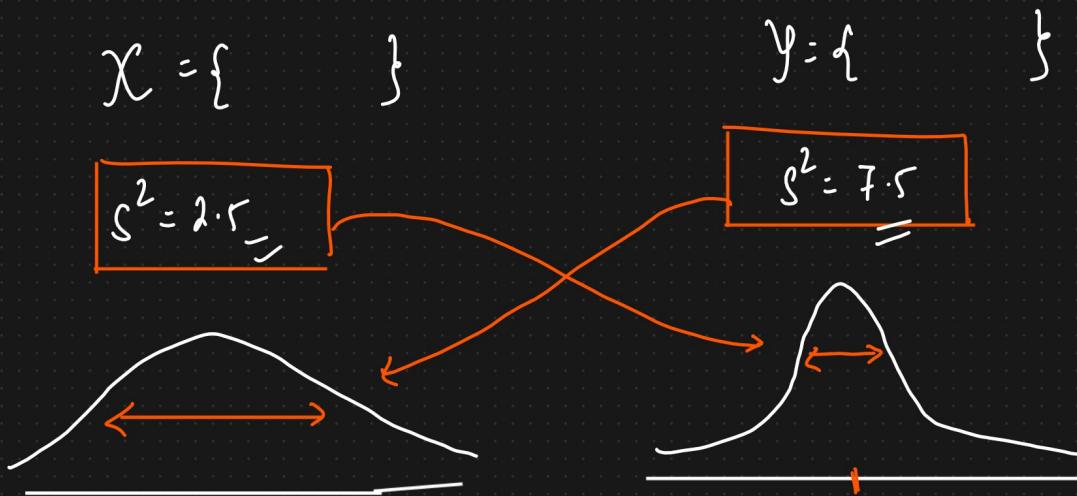
Assignment : Why we divide Sample Variance by $n-1$?

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

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



④ Standard deviation

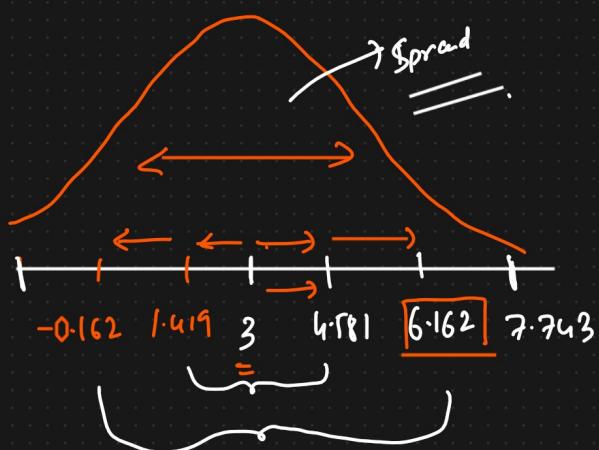
Population Std $\sigma = \sqrt{\text{Variance}}$

Sample Std $S = \sqrt{S^2}$

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

$$\bar{x} = 3$$

$$S = 1.581$$



$$\begin{array}{r}
 3.00 \\
 1.581 \\
 + \quad \quad \quad \\
 \hline
 4.581 \\
 1.581 \\
 \hline
 6.162 \\
 1.581 \\
 \hline
 7.743
 \end{array}$$

④ Random Variable

linear
 algebra

$$\left\{
 \begin{array}{l}
 n+5=7 \Rightarrow n=2 \\
 8=y+x \\
 \boxed{y=6}
 \end{array}
 \right\} \text{Variables}$$

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

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

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

$G = \{\text{Age of student in a class}\}$

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

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



$$\Pr(Y > 15) = \quad \quad \quad \Pr(Y < 10)$$

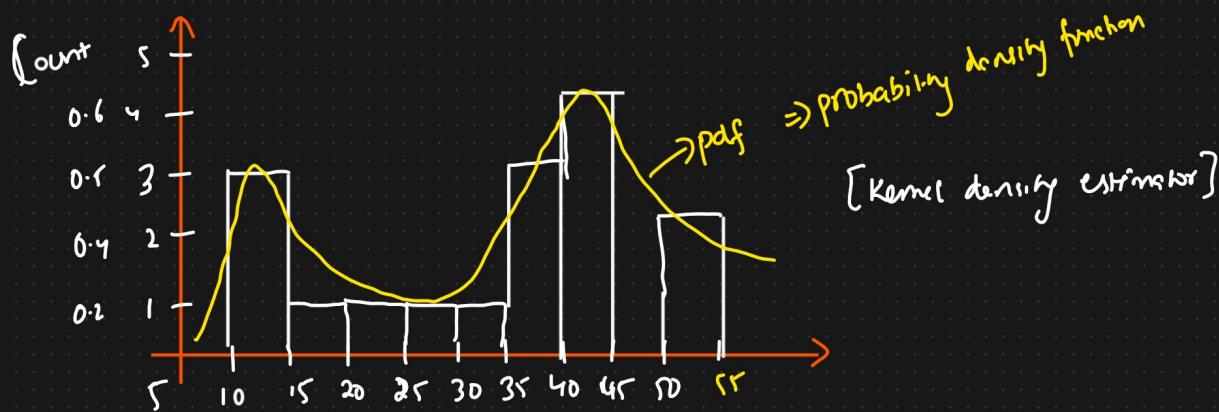
$$\Pr(40 \leq Y \leq 15) =$$

① Histograms And Skewness \rightarrow [Frequency]

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

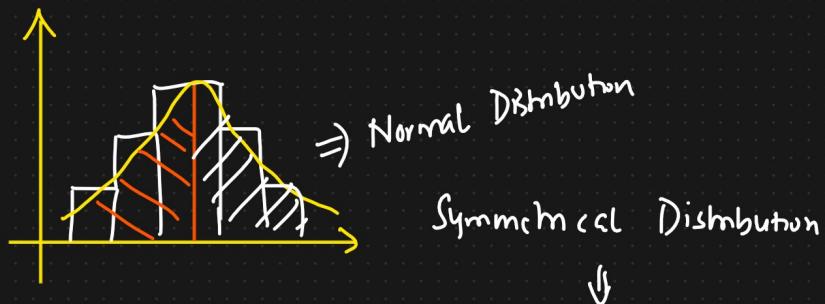
$$\frac{50}{10} = 5 \rightarrow \text{bin size}$$

No. of Bins = 10 \rightarrow buckets



Skewness

$$\chi =$$

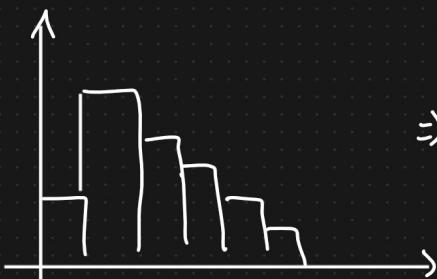


Median = Mean = Mode

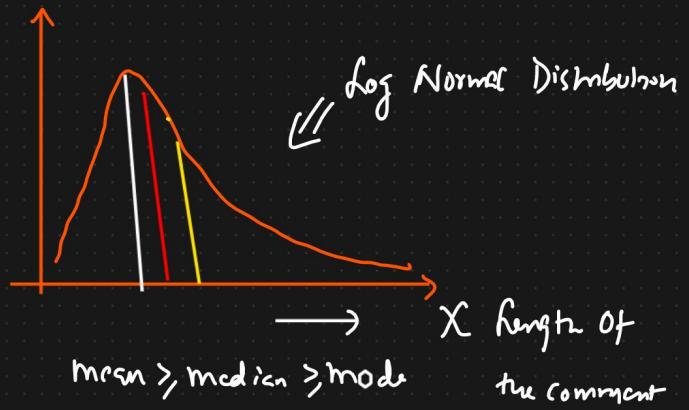


No skewness

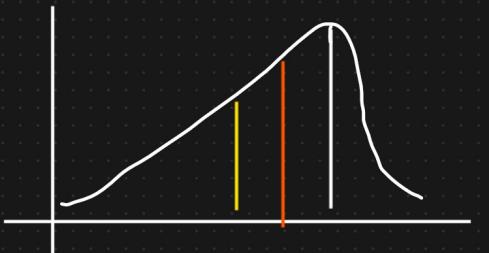
② Right skewed



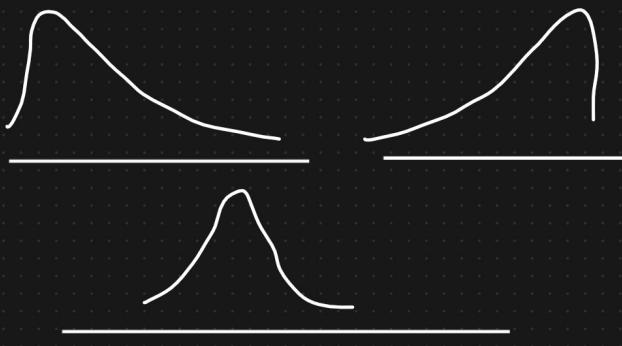
\Rightarrow Positive Skewed



③ left skewed



mode > median > mean



Knowledge sharing → Profile Building