Statistics

Defn:

Statistics is the science of collecting, organizing and analyzing data.

Data: "Facts or pieces of information".

Ex: Heights of students
IQs of students
Daily Activities
weights or Age of People

Types of Statistics

① Descriptive stats

Defn: It consists of organizing and summarizing data.

① Measure of central tendency
{mean, median, mode}

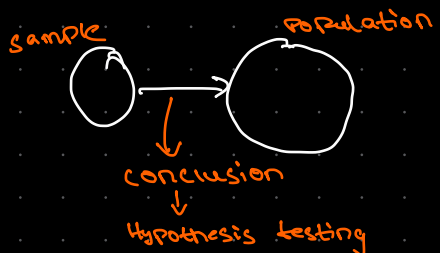
② Measure of Dispersion
{variance, standard deviation}

③ Different types of distribution of data.

Ex: Histogram,
PDF (Probability Distribution Function)
PMF (Probability Mass Function)
CDF (Cumulative Distribution Function)

② Inferential stats.

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



1) Z-test
2) t-test
3) Chi square test
4) ANOVA
5) F test

} Conclusion of sample on Population

Ex:

Lets say there are 20 classes in your college and you have collected the heights of students in the class

Heights are recorded as { 175 cm, 180 cm, 200 cm .... }

Descriptive:

"What is the average height of students in class"

(mean)

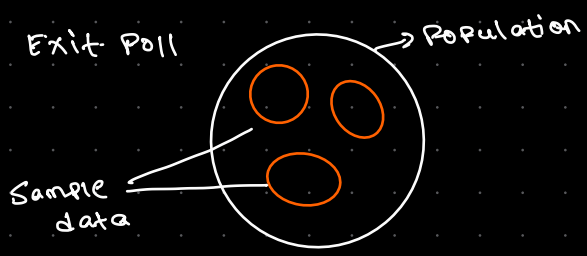
Inferential:

"Are the height of students in the classroom (sample) similar to what you expect in the college". (Population)

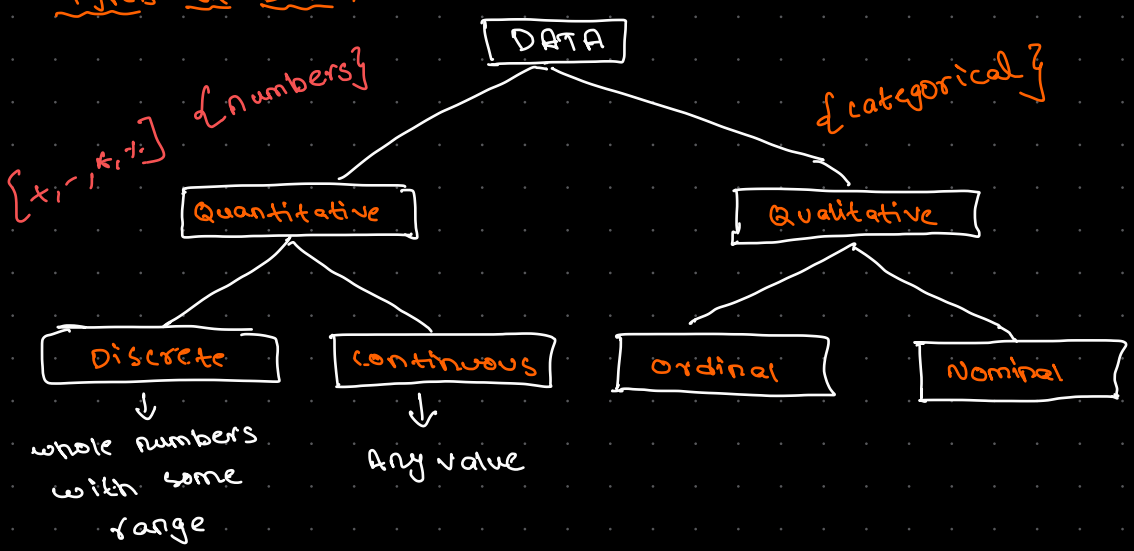
(N) (n)

Population and Sample data

Ex:



Types of Data:



Ex: * Number of bank accounts
* no. of children in a family

Ex: weight, height, Age temperature, speed etc...

Ex: Ranks [small, medium, large]
[good, better, best]

Ex: Gender (M, F)
Blood groups (O, B, A...)
colour of hair..

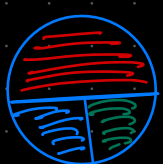
* Scales of Measurement

① Nominal scale data
② ordinal scale data
③ Interval scale data
④ Ratio scale data.

① Nominal scale data:-

→ Qualitative (Categorical data
Ex: Gender, colours, labels
→ order does not matter.

Ex: Favourite colour
Red → 5 (50%)
Blue → 3 (30%)
Green → 2 (20%)
10



② Ordinal scale data:-

① categorical data
② Ranking and order matters
③ Difference cannot be measured

Ex:
small → 1
medium → 2
large → 3

③ Interval scale data:-

- The order matters
- Difference can be measured
- Ratio cannot be measured.
- No "0" starting Points.

Temperature
30 F
60 F
90 F
120 F

(4) Ratio scale data

- The order matters
- Differences are measurable including ratios.
- Contains a '0' starting point.

Ex:- Students mark in class

0, 30, 45, 60, 90, 99

Examples:-

1. Marital status (Nominal scale data)
2. Favourite food based on gender (Nominal)
3. IQ measurement (Ratio scale data)
   (can also be converted to ordinal)

Descriptive statistics

① Measure of central tendency

(i) mean (ii) median (iii) mode.

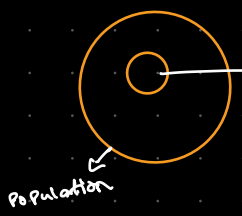
① Mean:-

Population (N) sample (n)

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

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

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



Population size (N)
sample size (n)

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

② Median

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

Steps:

① sort the numbers $\{1, 2, 2, 3, 4, 5\}$

② No. of elements

③ if count == even

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

$\frac{2+3}{2}$

$= 2.5$ (median)

if count == odd

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

median

why median? * means are affected by outlier *

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

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

$= 3$ (mean)

median = 3 (odd)

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

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

$= \frac{115}{6} \approx 19$ (mean)

median (even) $= \frac{3+4}{2}$

$= 3.5$

Conclusion:

median is used to find central tendency when outlier is present.

③ Mode:

maximum frequency occuring elements

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

mode: 1

* $\{1, 1, 2, 2, 3, 4\}$

$\{1, 2\}$

(depends) *

EDA and Feature Engineering

Age	weight	salary	Genders	Degree
24	70	40K	m	BE
25	80	70K	f	—
27	95	45K	F	—
24	—	50K	m	PhD
32	-	60K	—	BE
—	60	—	-	MS
—	65	65K	—	BSc
40	22	—	m	BE.

If outliers are present, we can use median or mean to fill the missing values (depends on usecase)

For the categorical data, we can use mode to replace the missing values.

---

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 → Population 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 $n-1$?

(unbiased estimator of population variance).

Example:

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

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

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

$\bar{x} = 3$

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

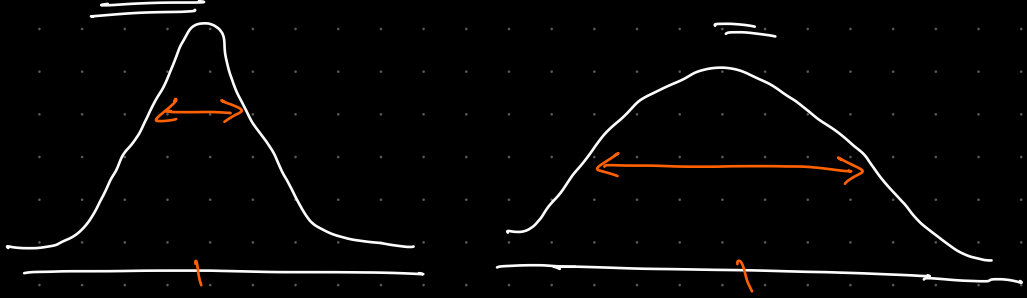
$s^2 = \frac{10}{4} = \frac{5}{2} = 2.5$

Consider X distribution and Y distribution

lets say,

$X = \{ \dots \}$ $\quad$ $Y = \{ \dots \}$

$s^2 = 2.5$ $\quad$ $s^2 = 7.5$



② Standard deviation

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

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

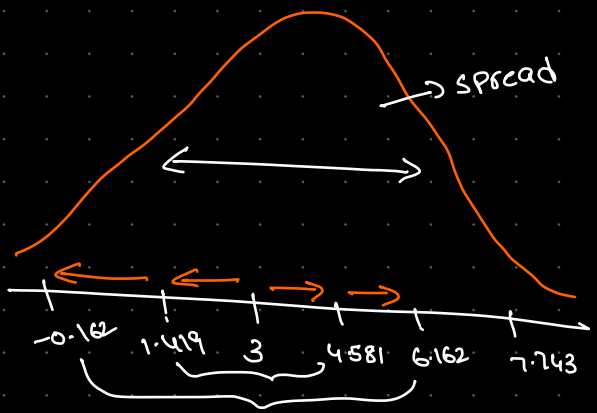
$x = \{1, 2, 3, 4, 5\}$

$\bar{x} = 3$

$s = \sqrt{\text{variance}}$

$= \sqrt{2.5}$

$s = 1.581$



3.00
1.581
4.581
1.581
6.162
1.581
7.743

→ (Standard deviation talks about how much far element is away from mean)

→ (variance is std variation square)

Assignment:

① Variance Error

---

* Random variable:-

Linear Algebra $\begin{cases} x+5=7 \\ 8=y+x \end{cases} \Rightarrow x=2$ } variables

$\boxed{y=6}$ }

* Random variable is a process of mapping the output of a random process or experiment to a number *

Ex: Tossing a coin {head, tail} → process

$$X = \begin{cases} 0 & \text{if head} \\ 1 & \text{if tail} \end{cases}$$

Rolling a Dice {1,2,3,4,5,6}

y = {sum of rolling of dice 7 times}

This helps in finding

$P_r(y > 15)$ = ____ (probability of x to be greater than 15).

$P_r(y < 10)$: ____ (probability of x to be less than 10).

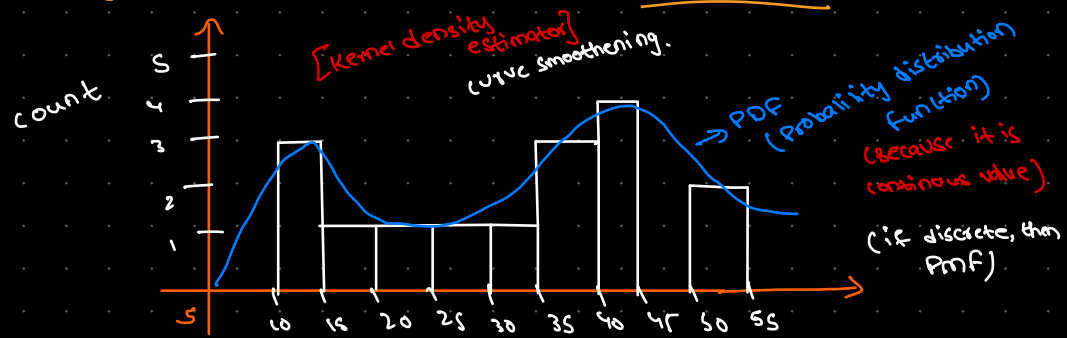
---

* Histograms and skewness → [Frequency]

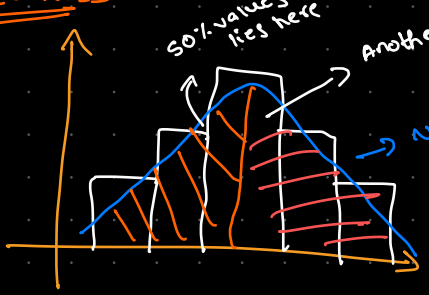
Ages: {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51} =

$\frac{50}{10} = 5$ → (bin size)

No. of bins = 10



Skewness



Also called as Symmetrical distribution

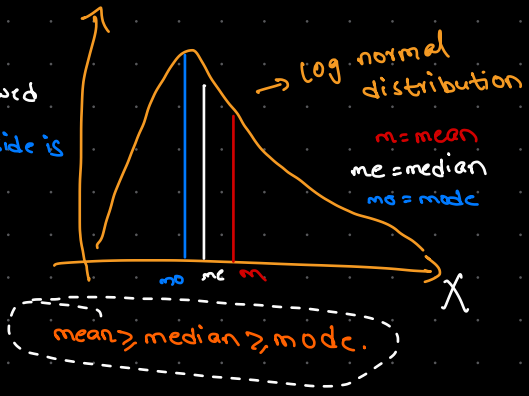
↓

No skewness

median = mean = mode

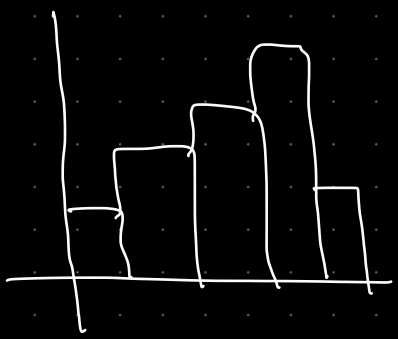
* Right Skewed

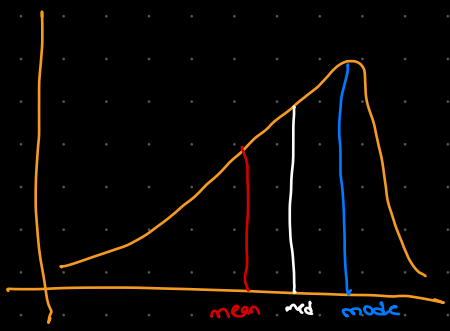
⇒ positive skewed
(right hand side is elongated)



mean ≥ median ≥ mode.

* Left Skewed





mode > median > mean