STATISTICS FOR THE DATA SCIENCE

Part - 1

- WHY STATISTICS IN DATA SCIENCE?
- TYPES OF STATISTICS: DESCRIPTIVE VS. INFERENTIAL
- POPULATION VS. SAMPLE DATA
- SCALE OF MEASUREMENT
- MEASURES OF CENTER TENDENCY
- MEASURES OF DISPERSION
- SETS IN STATISTICS

#Value_freeContent

@Krishan kumar

Statistics

Defination Statistics in the science of collecting, organizing and analyzing data

Data facts or pieces of imformation → It can be mejor and collect

ex → ① Height in the classroom of the student [175, 180, 160] cm

② IQ [100, 90, 80, 60]

Why statistics ?

To innovate any product, to bring value to any product, the role of data is very important. Statistic give a lots of information about data because statistic provide such tools, we can get many infraction and concussion from the providing data.

Types of Statistics :-

1. Descriptive statistic →

↳ It consits of organizing and summarizing data.

(A) Measure of centre Tendency →

[Mean, Median, Mode]

(B) Measure of Dispersion →

[Variance, Std (Standard deviation)]

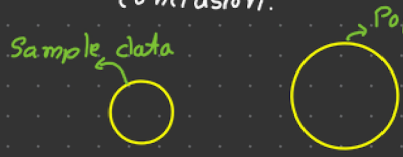
(C) Different type of Distribution of data →

tools :- [Histogram, PMF (Probability Function)

, PDF (" Density ")

2. Infernfion statistics →

↳ It consists of using data you have measured to from conclusion.



Problem lets say there are 20 Statistic class at your college and you have collected the height of the student in the class. heights are recorded [175, 180, 175, 180, 176, 160, 135, 190] cm.

* **Descriptive Question** → What are the average height of the entire classroom.

Sol ⇒ Mean → Average → this is the part of descriptive stats.

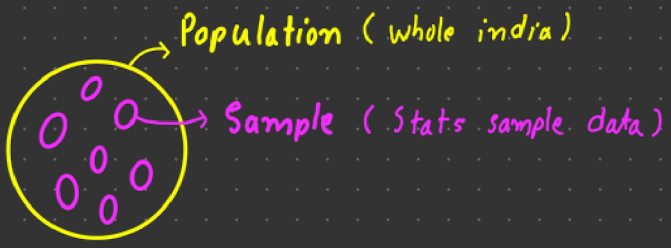
↳ $\frac{175+180+175+180+176+160+135+190}{8}$ → Average height

▶ **Infernition Question** Are the height of the student (Sample data) in the classroom smilliar what you expect in the entire college. (population data)

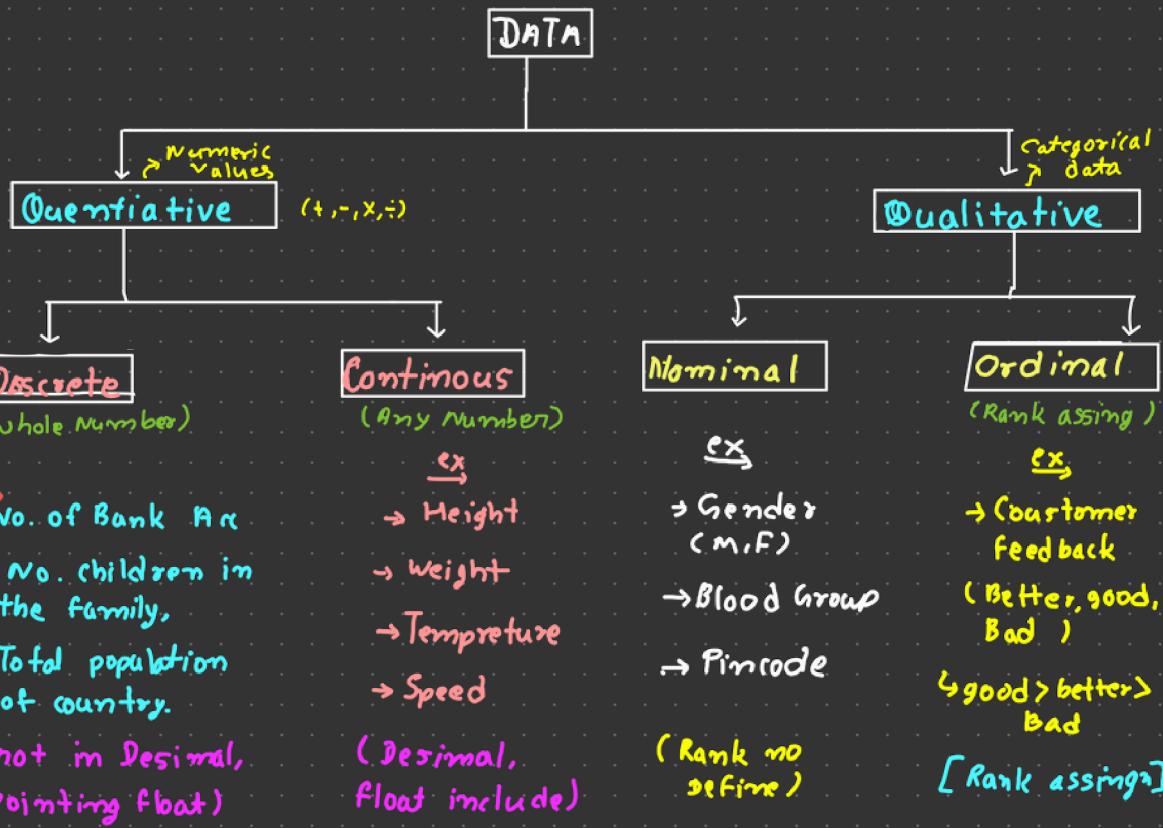
⭐ **Population and sample data**

* Population Data ⇒ The group you are intrested in studying
* Sample Data ⇒ A Subset of population data

Ex. Exit pole of election



Types of Data



Scale of Measurement

- Nominal scale data.
- Ordinal scale data
- Internal scale data
- Ratio scale data.

Nominal scale Data :-

→ Qualitative OR categorical data
→ Order does not matter
↳ ex → Gender / color / habit

ex Fevorite color.

Red → 5 → 50%
Black → 3 → 30%
Blue → 2 → 20%
10

there are No any Specific Rank

↳ by nominal data we find called → cumulative freq.

Ordinal scale data

* Ranking is imp.
* Order is matter
* Difference can't be m measured

ex → 1 → Best
2 → Good
3 → Bad

ex Working Profession Productivity

Hours		Day
1	-	Mon.
2	-	Tue.
3	-	Wed.
4	-	Th
5	-	Fri

Best work on → friday

ex → Race →

1st
2nd
3rd

→ In ordinal Difference can't be measure.

Internal scale data

* The order matter
* Difference can't be measured.
* Ratio can't measured
* No '0' starting fixed

It can be -ve OR +ve

Ex, Temp Variable →

60F, 90F → Ratio → 60:90 → 2:3

→ Diff. can be measure

120 F, 180F → 90 - 30 → 60 F

$\frac{180}{120} \rightarrow \frac{60}{40} \rightarrow 3/2$

Ratio does not mean that Temp. increase OR ↓↓ 3/2 ↑.

Ratio scale data

* The order matter
* Diff. are measurable (including ratio)
* Contain a '0' Starting point.

Ex, Student marks in a class.

0, 90, 60, 30, 75, 80, 85, 50

Assending order → 0, 30, 50, 60, 75, 80, 85, 90

Diff. → 60 - 30 = 10

80 - 50 = 30

Ratio → $\frac{30}{90}$ = 3:1 This person gain 3X marks from that one (30 mark)

Ratio measured in Grading.

Question not in Temp. Type Question

Measure of centre tendency

- Mean OR Average
- Median
- Modde

Mean OR Average :-

Population (N)

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

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

$\Rightarrow \frac{1+1+2+2+3+3+4+4+5+5}{10}$

$\Rightarrow \frac{32}{10} = 3.2 \Rightarrow$ Average

$N = 3.2$

Sample mean = $\bar{x}$

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

$\bar{x} = \frac{32}{10} = 3.2$

$\bar{x} = 3.2$

Median :-

$x = [4, 5, 2, 3, 2, 1]$

Steps

- Sort the random variable → [1, 2, 2, 3, 4, 5]
- No. of element , count → 6
- if count == even

  [1, 2, 2, 3, 4, 5]

  - Median → $\frac{2+3}{2}$ → 2.5

- if count == odd

  [1, 2, 2, 3, 3, 4, 5]

Median = 3

Why Median ?

↳ Median is used to find centre of tendency when outlier is present.

$x = [1,2,3,4,5]$

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

$\bar{x} = 3$

Diff. in mean

$x = [1,2,3,4,5,100]$ → outlier

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

$\bar{x} = 19.16$

In Median →

$x = [1,2,3,4,5,100]$

Median → $\frac{3+4}{2} \rightarrow \frac{7}{2}$

Median = 3.5

So the result is, Mean with outlier → 3 (Diff) 19.16

Median with outlier

3 → 3.5 (very close)

Note

Whenever we find out centre of tendency we should do median.

Mode :-

↳ Frequency maximum

↳ Maximum repeating Number

$x = [1,2,3,4,5,8,9,1,2,3,2,8,9,1,2,3,4]$

Mode = 2

Where and why we use mean, median, mode,

↳ We use these all in EDA and feature engineering.

for ex →

Age	weight (Numeric fe.)	Salary	Gender (mode)	Degree (mode)
24	70	40K	M	BE
25	80	50K	F	- } → Data is missing
27	95	70K	M	-
24	—	35K	-	BE
32	-	70K	M	PHD
-	60	40K	-	Master
-	65	-	F	BSc
40	72	-	m	BE

Note

Missing data Filling

* if data in → Categorical → Fill by Mode (best result)
* if data in → Numerical → Fill by Mean (best result)
* Data with Outlier → Fill by Median (best result)

Measure of Disperson :- (How much our data spread)

Types → (i) Variance
(ii) Standard deviation

To calculate the this spreaded data we use → Measure of disperson

(i) Variance :-

(A) Population variance

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

x_i → Data point
μ → Population mean
N → Population size

(B) Sample variance

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

x_i → Data point
$\bar{x}$ → Sample mean
n → Sample size

Q) Why we devide Sample variance by n-1 ?

Sol) The sample variance devided by n-1 so that we can create an unbiased estimater of the population variance
↳ This scenerio → called → Bencle correction

Q) → [1,2,3,4,5] , s^2 = ?

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

⇒ $s^2 = \frac{10}{5}$ ⇒ 2

x	$\bar{x}$	$(x_i - \bar{x})$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	4
means 15/5 = 3		10

Sample variance → 2

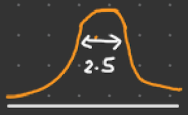
* What difference in these two sample variance ? x, y

↳ $X = 2.5$ $\quad y = 7.5$

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

Ans → s^2 = denote → Dispersion of spread

$s^2 = 2.5$



→ When spreadness = ↑↑ (increase)

Height becames = ↓↓ (decrease)

Standard deviation :- (std) (σ)

(A) Population Standard deviation

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

(B) Sample std.

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

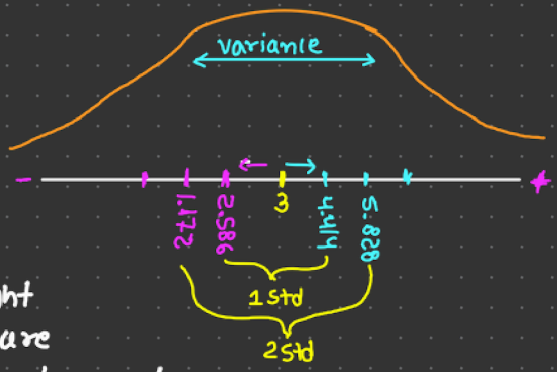
s^2 → sample variance

ex →

$X = [1, 2, 3, 4, 5]$

Sample mean = $\bar{x} = 3$

$\sigma = \sqrt{3} = 1.414$



When one step towards the right and one step towards the left are combained then we called them 'one std'

Q where the 2 element fold in the variance.

Ans → one step left to the mean.

3.000
-1.414
2.586
-1.414
1.172

3.000
+1.414
4.414
+1.414
5.828

Random variable

Random variable is a process of maping the output of a random process or experiments to a number there are not any fix value.

Ex,

Tossing a coin

$X = \begin{cases} 0, \text{ if Head} \\ 1, \text{ if Tell} \end{cases}$

Rolling a dice

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

↳ $P(Y \geq 15)$ OR $P(Y < 10)$

Probability of

Sets

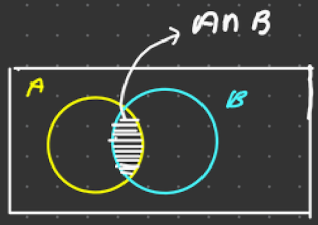
$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$

$B = \{3, 4, 5, 6, 7\}$

Intersection ($A \cap B$)

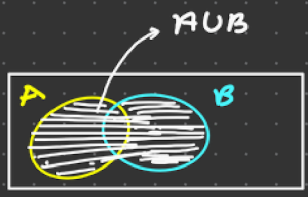
(common value)

$A \cap B = \{3, 4, 5, 6, 7\}$



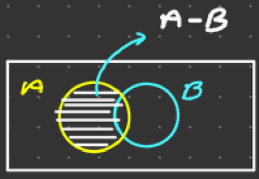
Union

$A \cup B = \{1,2,3,4,5,6,7,8\}$



Difference

$A - B = \{1,2,8\}$



Subset (Som)

$A \rightarrow B$ = False (∵ all element of B Present in A)

$B \rightarrow A$ = True (∵ All element of A Present in A)

Superset

$B \rightarrow A$ = False

$A \rightarrow B$ = True