



STATISTICS FOR THE DATA SCIENCE

Part - 1

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

#Value_freeContent



@Krishan kumar

Statistics

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

Data facts or pieces of information → It can be major 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 information and conclusion from the providing data.

Types of Statistics:-

1. Descriptive statistic

↳ It consists 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")]

Sample data

Population data

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, 180] cm.

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

So Mean \rightarrow Average \rightarrow this is the part of descriptive stats.

$$\hookrightarrow \frac{175+180+175+180+176+160+135+180}{8} \rightarrow \text{Average height}$$

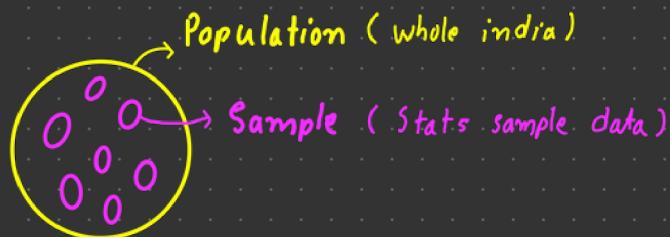
> Inferntion Question Are the height of the student in the classroom similar what you expect in the entire college.
Sample data
population data

★ Population and sample data

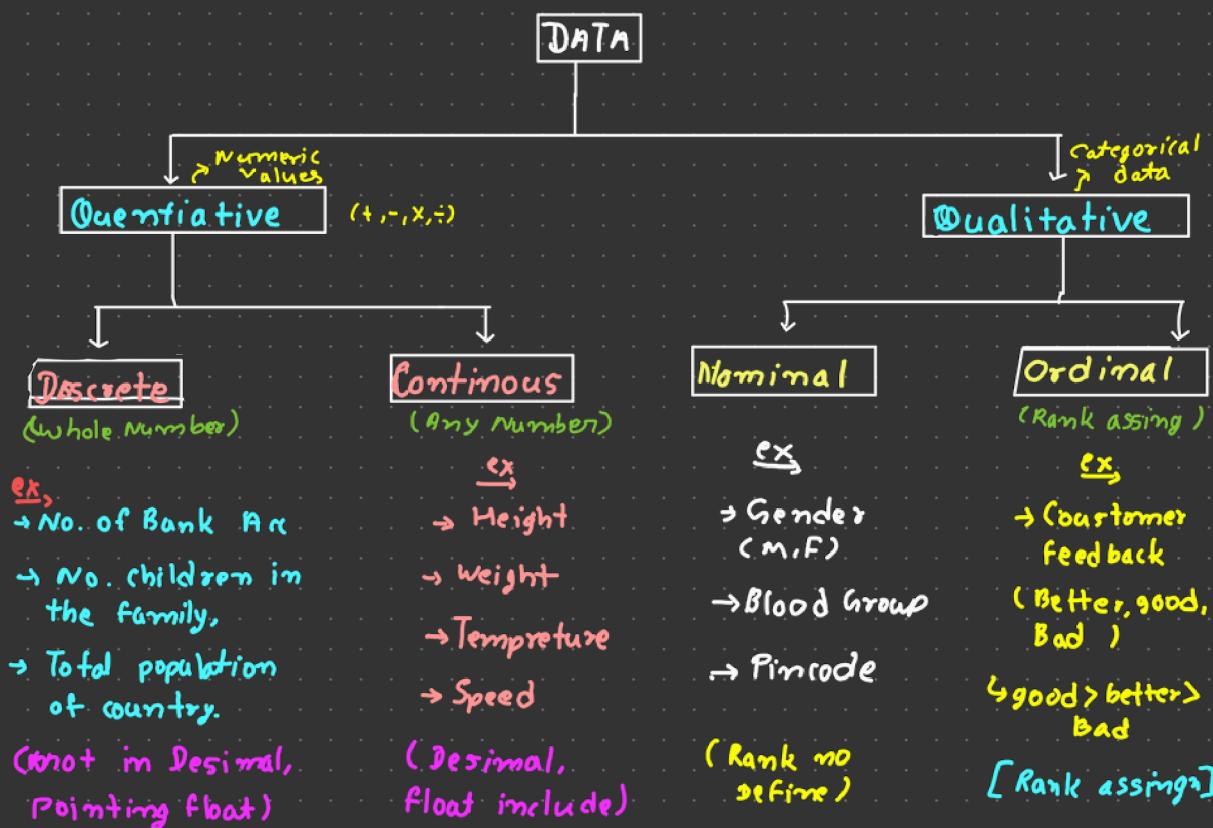
* Population Data \Rightarrow The group you are interested in studying

* Sample Data \Rightarrow A subset of population data

Ex Exit pole of election



Types of Data



Scale of Measurement

- Nominal scale data.
- Ordinal scale data.
- Interval scale data.
- Ratio scale data.

↳ Nominal scale Data :-

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

ex favorite color.
Red → 5 → 50%
Black → 3 → 30%
Blue → 2 → 20%
10 by nominal
data we find
called → cumulative freq.

there are
no any
specific rank

↳ Ordinal scale data

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

ex, 1 → Best
2 → Good
3 → Bad

ex Working Profession
Productivity

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

ex Race

1st
2nd
3rd

In ordinal difference
can't be measured.

Best work on → Friday



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 →

$$\begin{array}{l} 60^\circ F \\ 90^\circ F \\ 120^\circ F \\ 180^\circ F \end{array} \left. \begin{array}{l} \text{Ratio} \rightarrow 60:90 \rightarrow 2:3 \\ \text{Diff. can be measured} \\ 90-30 \rightarrow 60^\circ F \\ \frac{120}{180} \rightarrow \frac{60}{90} + \frac{3}{2} \end{array} \right.$$

Ratio does not mean
that Temp. increase
OR ↓ $\frac{3}{2}$ %.



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

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

$$\begin{aligned} \text{Diff.} \rightarrow 40-30 &= 10 \\ 80-50 &= 30 \end{aligned}$$

Ratio measured in
Grading.

question not in Temp. Type question

Ratio → $\frac{30}{90} = \frac{1}{3}$ This person
gains 3x marks from
that one (30marks)

Measure of centre tendency

► Mean OR Average

► Median

► Mode

► Mean or Average :-

Population (N)

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

$$\text{population mean } (N) = \sum_{j=1}^n \frac{x_j}{N}$$

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

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

$$N = 3.2$$

Sample mean = \bar{x}

$$\bar{x} \Rightarrow \sum_{j=1}^n \frac{x_j}{n}$$

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

$$\boxed{\bar{x} = 3.2}$$

► Median :-

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



Steps

→ Sort the random variable $\rightarrow [1, 2, 2, 3, 4, 5]$

→ No. of element , count $\rightarrow 6$

→ if count == even

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

$$\therefore \text{Median} \rightarrow \frac{2+3}{2} \cancel{, 2.5}$$

→ if count == odd

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

$$\boxed{\text{Median} = 3}$$

Why Median?

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

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

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

$$\bar{x} = 3$$

Dif. is mean

$$x = [1, 2, 3, 4, 5, 100] \quad \text{→ outlier}$$

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

$$\bar{x} = 19.16$$

In Median

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

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

$$\text{Median} = 3.5$$

So the result is, Mean with outlier → 3 \rightsquigarrow 19.16

Median with outlier

3 \rightsquigarrow 3.5
(every close)

Note

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

Mode :-

↳ Frequency maximum

↳ Maximum repeating Number

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

$$\text{Mode} = 2$$

Where and why we use mean, median, mode,

↳ We use these all in EDA and Feature engineering.

for ex →

Age	Numeric fe.		Gender	Mode	Mode
	weight	Salary			Degree
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

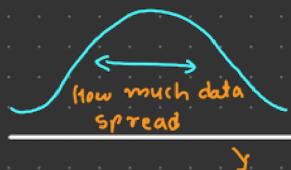
Missing data filling

- * if data is → Categorical → fill by Mode (best result)
- * if data is → Numerical → fill by Mean (best result)
- * Data with Outlier → fill by Median (best result)

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

Types

- (i) Variance
- (ii) Standard deviation



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

(i) Variance :-

(A) Population variance

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

x_j → Data point

μ → Population mean

N → Population size

(B) Sample variance

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

x_i → Data point

\bar{x} → Sample mean

n → Sample size

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

Sq The sample variance devided by $n-1$ So that we can create an unbiased estimator of the population variance

↳ This scenario → called → Bencle correction

Ex [1, 2, 3, 4, 5], $s^2 = ?$

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

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

x	\bar{x}	$(x_i - \bar{x})$
1	3	4
2	3	1
3	3	0
4	3	-1
5	3	2
Mean is 3		10

Sample variance → 2

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

$$\hookrightarrow x = 2.5$$

$$y = 7.5$$

$$s^2 = 2.5$$

$$s^2 = 7.5$$

Ans s^2 denote \Rightarrow Dispersion or spread

$$s^2 = 2.5$$



\rightarrow When spreadness $= \uparrow\uparrow$ (increase)

Height becomes $= \downarrow\downarrow$ (decrease)

► Standard deviation :- (std) (σ)

(A) Population Standard deviation

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

(B) Sample std.

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

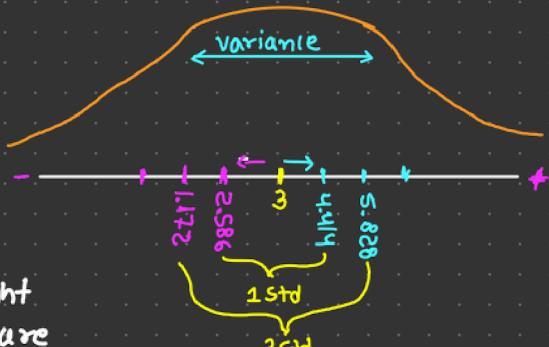
$s^2 \rightarrow$ sample variance

Ex

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

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

$$\sigma = \sqrt{3} = 1.732$$



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

Q Where the 2 element fold in the variance.

Ans \rightarrow One step left to the mean.

$$\begin{array}{r} 3.000 \\ + 1.414 \\ \hline 4.414 \\ - 2.894 \\ \hline 1.514 \\ + 1.414 \\ \hline 5.928 \end{array}$$

Random variable

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

Ex,

Tossing a coin

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

Rolling a dice

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

$$\hookrightarrow P(Y \geq 15) \quad \underline{\text{OR}} \quad P(Y < 10)$$

Probability of

Sets

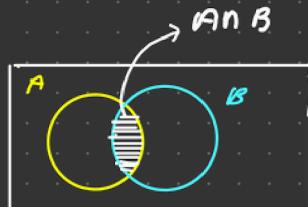
$$A = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

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

Intersection ($A \cap B$)

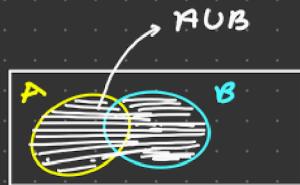
(common values)

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



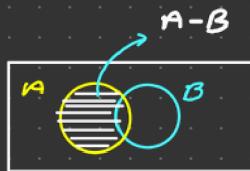
Union

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



Difference

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



Subset (son)

$$A \rightarrow B = \text{False} \quad (\because \text{all element of } B \text{ Present in } A)$$

$$B \rightarrow A = \text{True} \quad (\because \text{All element of } A \text{ Present in } A)$$

Superset

$$B \rightarrow A = \text{False}$$

$$A \rightarrow B = \text{True}$$