13/12/23 Statistics for data science & Business Analysis

Populat^n

↓

collect^n of items of interest

-> "N"

->Parameter

(The number we'have obtained when using a populat^n are called)

->populat^n is hard to define and hard to observe in real life

(Ex: NYU campus)

-> Hard to observe

-> Hard to contact

Sample

↳ Subset of populat^n

-> denoted by "n"

-> statistics

-> less time consuming

-> less costly

->

[statistical data work with incompleteness of data

∴ work with sample data & make data-driven decis^n]

-> easy to observe

-> easy to contact

-> statistics usually based on sample data - to accurate statistical insights

2 characteristics

i) Randomness

ii) Representativeness

Randomness:-

It is collected when each member of the sample is chosen from the population strictly by chance.

Representativeness:-

It is a subset of the population that accurately reflects the members of the entire population

[Sample is not random But it is representativeness]

↓

Ex- Canteen

We can classify data in two main ways:-

Types of Data | measurement level

Categorical ↓ Categories of groups ↓ ex:- car brands (yes (or) no)

numerical ↓ represented numbers

Numerical

discrete

↓

finite man

Ex:- Grad

→ univer

→ no of b

levels of

Q

nomin

↓

Categori

like BM

merce

not n

Cannot b

Numerical data

discrete

↓

finite manner

Ex:- Grade of university

→ no of bottles

continuous

↓

infinite & impossible to count

Ex:- height, area, time. etc.

72.12345

levels of measurements

Qualitative

Quantitative

- interval → Ex:- Temperature & no true zero ↓ celsius & F
- Ratio

both represented by numbers & but one major difference

Ratio has true zero

Ex 2 apples & 8 ap

arab of 6 & 2 is 3

Kelvin ↓ true zero

nominal

↓

Categories like BMW merceded not numbers Cannot be ordered

ordinal

↓

strict order

Ex tasty, delicious

-ve to +ve (grades)

18/12/23

Types of data

↓ ↘

categorical Numerical

↓

representatⁿ — frequency distributⁿ tables,

→ bar charts

→ pie charts (% calculate chayalli)

pareto diagram →

↓

It is a special type of bar chart where categories are shown in descending order of frequency

↓

It is the no. of occurences of each item.

* a curve on the same graph showing the cumulative frequency

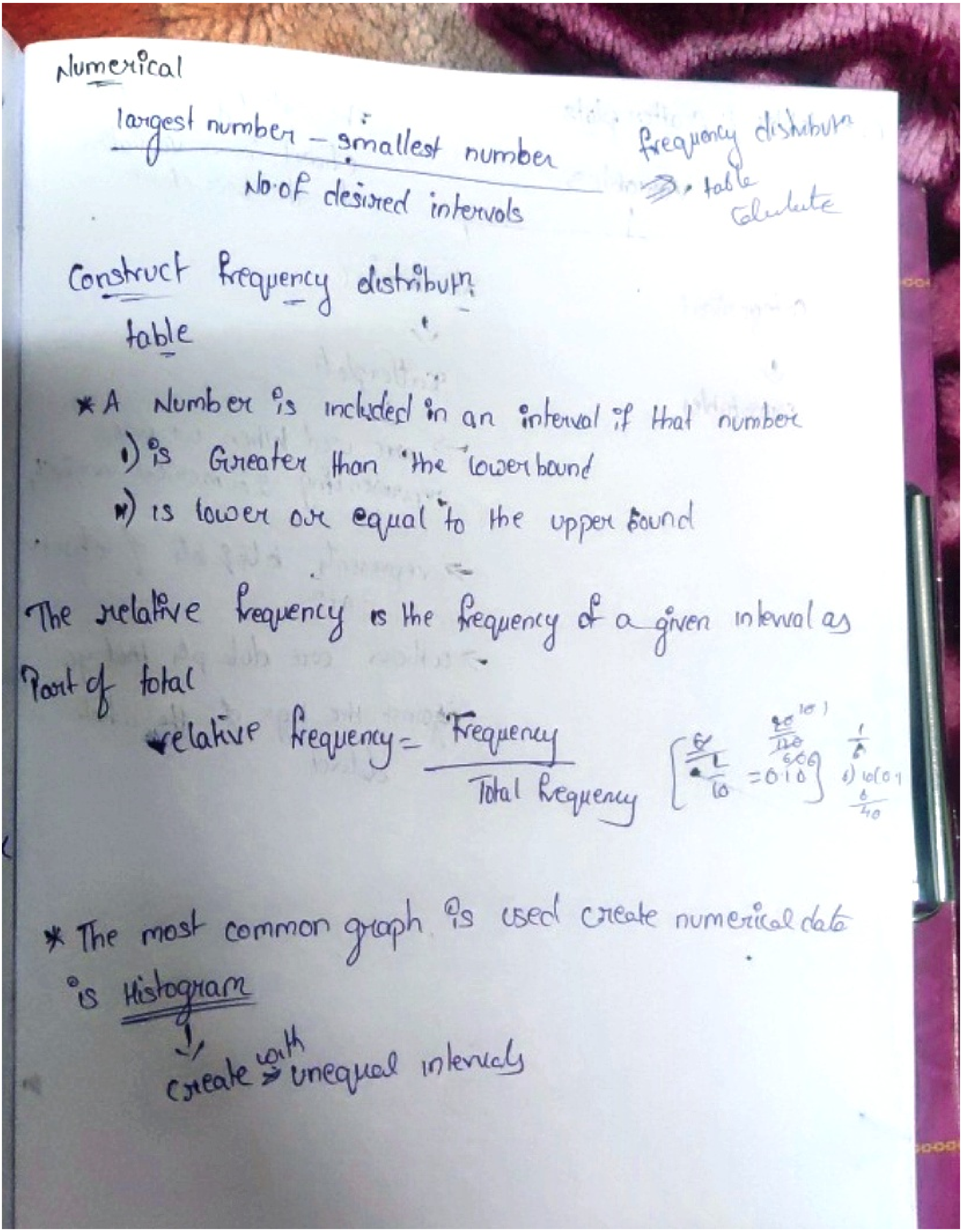
↓

it is the sum of the relative frequencies

* pareto principle :

→ 80 - 20 Rule

→ 80% of effect come from 20% of causes



Numerical

$$\frac{\text{largest number} - \text{smallest number}}{\text{No.of desired intervals}}$$

frequency distributn → table calculate

Construct frequency distributn table

* A Number is included in an interval if that number
  i) is Greater than the lower bound
  ii) is lower or equal to the upper bound

The relative frequency is the frequency of a given interval as Part of total

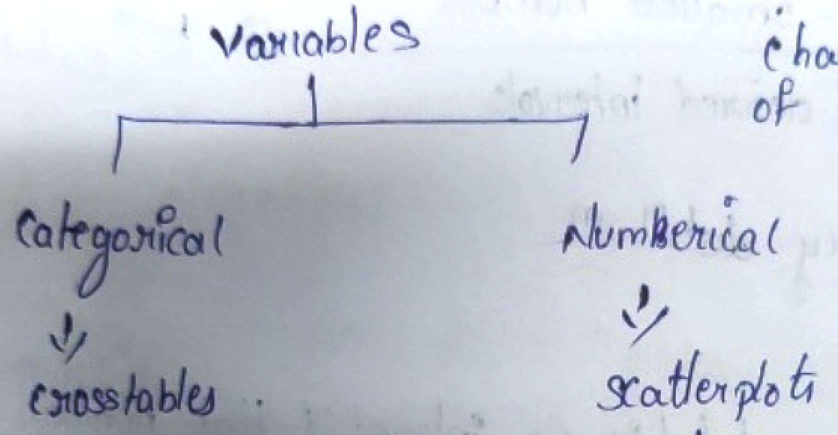
$$\text{relative frequency} = \frac{\text{Frequency}}{\text{Total frequency}} \quad \left[\frac{6/2}{10} = 0.10\right]$$

$\frac{20}{120}$ 10) 6/06 $\frac{1}{6}$ i) 10(0.1 $\frac{6}{40}$

* The most common graph is used create numerical data is **Histogram**
  ↓
  create with unequal intervals

Crosstables & Scatter plots

The side by side bar chart is a variation of the bar chart



→ are used when we are representing 2 numerical variables

→ represents lots & lots of observ-atn

→ outliers are data pts that go against the logic of the whole dataset