Statistics: Statistics is collecting, organizing and analyzing the data.

Data is facts / pieces of information.

eg of data → height of students (cm), age, gender
- {175, 185, 190, ...}

Q The main aim of applying statistical analysis to our data is so that, it will help us in decision making process.

Types

Descriptive Stats

Defn: It consists of organizing and summarizing data

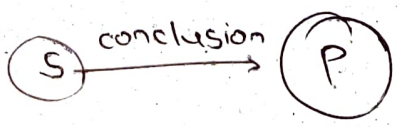
Here, we use different techniques to gain information about data.

1) measure of central tendency → mean, median, mode
2) measure of dispersion → variance, std dev
3) Histogram, Bar chart, Pie chart,

Inferential stats

Defn: It consists of using data which you have measured to form conclusions.

Sometimes, we don't have whole data. Hence, we use sample to make conclusions w.r.t the population.



1) z-test 2) t-test
Hypothesis Testing, p-value, significance

Let's say there are 50 students in a math class in the university.

[175, 180, 160, 140, 130, 140, 140, ...]

Descriptive: • Find avg height of students from the class.

$$\text{mean/avg} = \frac{\text{sum}}{n}$$

• most common height → mode.

Inferential: Are the heights of the students in the classroom similar to what you expect in the entire college?

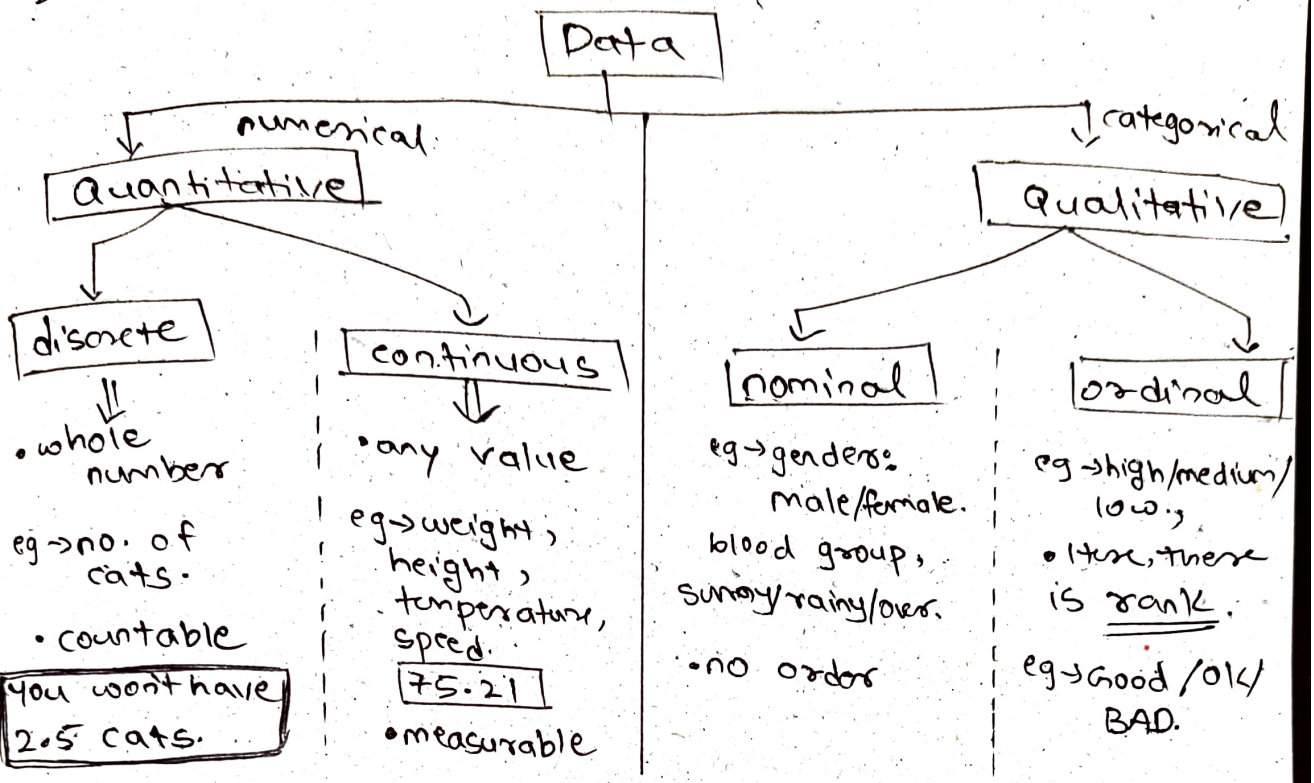
Note: Here, we have heights of one class only. (Sample) Hence, based on this sample, we can't try to make conclusion on population (college).

Sample data and Population Data:

eg → Exit Poll:

Media can't ask each and every person about whom they voted. Hence, they take a sample of population from ~~certain~~ & different populations and based on result, they find maximum chances about which party would win in which region.

⑭ Types of Data:



⑮ Scale of Measurment of Data:

1) Nominal Scale Data:

- Qualitative / categorical variable (gender, color)
- order does not matter.

eg → survey (10.)

Red → 5 ---> 50%
Blue → 3 ---> 30%
Yellow → 2 ---> 20%

(Focus more on count & distribut°)

2) Ordinal Scale Data:

- Ranking and order matters.
- Difference cannot be measured.

eg → Qualification

1) PHD → Rank 1
2) mTech → Rank 2
3) BTech → Rank 3

(focus on order)

3) Interval Scale Data:

- Rank & order matters
- Difference can be measured (excluding ratio)
- Doesn't have "0" (zero) starting value.
  - ↳ You can have −20F as temp

eg → Temperature

30 F] 60 − 30 = 30
60 F
80 F
90 F 90 − 30 = 60

$\frac{60F}{30F} = \frac{2}{1} \Rightarrow 2:1$

doesn't necessarily mean temperature is double as there can be external factors like AC

4) Ratio Scale Data:

- Order & rank matters
- Differences & ratios are measurable
- It does have a "0" (zero) starting value.

eg → Grades

100, 90, 60, 35, 45, 50

100, 90, 60, 50, 45, 35

$\frac{100}{50} = \frac{2}{1} \Rightarrow 2:1$ clearly indicates a boy got double the marks of another boy.

~~eg → Length of different rivers in the world.~~

Note: Median is least affected by Outliers

Measure of central Tendency.

1) Mean :

Let population → N
& sample → n

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

- Population mean $(M) = \boxed{\sum_{i=1}^{N} \frac{X_i}{N}}$
- Sample mean $(S) = \boxed{\sum_{i=1}^{n} \frac{X_i}{n}}$

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

---

2) Median : Central point of distribution.

- $\{4, 5, 2, 3, 2, 1\}$ (even)

sort → 1, 2, 2, 3, 4, 5

Median → $\frac{2+3}{2} = \boxed{2.5}$

- $\{1, 2, 2, 3, 4, 5, 7\}$ (odd)

median → central element $= \boxed{3}$

---

3) Mode : maximum frequency.

$\{2, 1, 1, 1, 4, 5, 7, 8, 9, 10\}$ → $\boxed{\text{mode} = 1}$

Why is median important?

→ Consider sample {1, 2, 3, 4, 5}

$$\text{mean (s)} = \frac{1+2+3+4+5}{5} = 3 \qquad \text{median} = 3$$

Here, mean & median are same hence, both are equally important.

Now, suppose we add an outlier in the data

{1, 2, 3, 4, 5, 100}.

$$\text{mean (s)} = \frac{115}{6} \approx 18. \qquad \text{median } 3.5$$

Here, we can see a drastic change in mean but not in median.

Hence, if our dataset contains outliers, it is good to fill missing values with median instead of mean, for numeric columns.

eg →

Type of flower	Age
Lily	10
Rose	3
Sunflower	—
—	5
Rose	8

Rose →

In left column, you can take mode & add Rose in missing cell.

In right column, as these are no outliers we can fill the missing cell with either mean or median of the Age column.

(After this, pratical implementation)