Statistics

Statistics → It is the science of analyzing, collecting & organizing data.

Data → Data is fact or pieace of information. collected from various sources, which may sometimes not have meaning of their own but can be processed to generate useful information.

- Data can be in the form of text, images, audio or video.

Types of statistics

- There are two main types of statistics.

1] Descriptive Statistics

- It is used to summarize and describe the main feature of a dataset.

Examples:

- Mean
- Median
- Mode
- Standard Deviation.
- Charts & Graphs.

Purpose :→ Gives a quick overview of a data.

2] Inferential statistics

→ It is used to make predictions or conclusions about a large populations using a sample of a data.

Examples:

- Hypothesis Testing
- Confidence Intervals.
- Regression Analysis
- Analysis of Variance (ANOVA) - F test.
- P value
- Z test, t test
- CHI square.

Purpose → Helps in Decision making of sample Data.

eg → Let's say there are 20 class in your college. & you have created the height of students in the class

Heights recorded are [175 cm, 180 cm, 170 cm, 135, 160, 120]

Descriptive Que?

"what is average height of the entire classroom".

Inferential Stat?

" Are height of the sample. students in classroom similar to what you expect in entire college?

Population & Sample Data.

Population Data →

Population data refers to the entire set of individuals or items that you're intresled in studying

Example:

If you're studying marks of all students in your university, the marks of every student = population data.

characteristics

1) Complete set of data.
2) Numerical value summarizing the entire population.
   1) population mean (μ)
   2) population variance (σ^2)

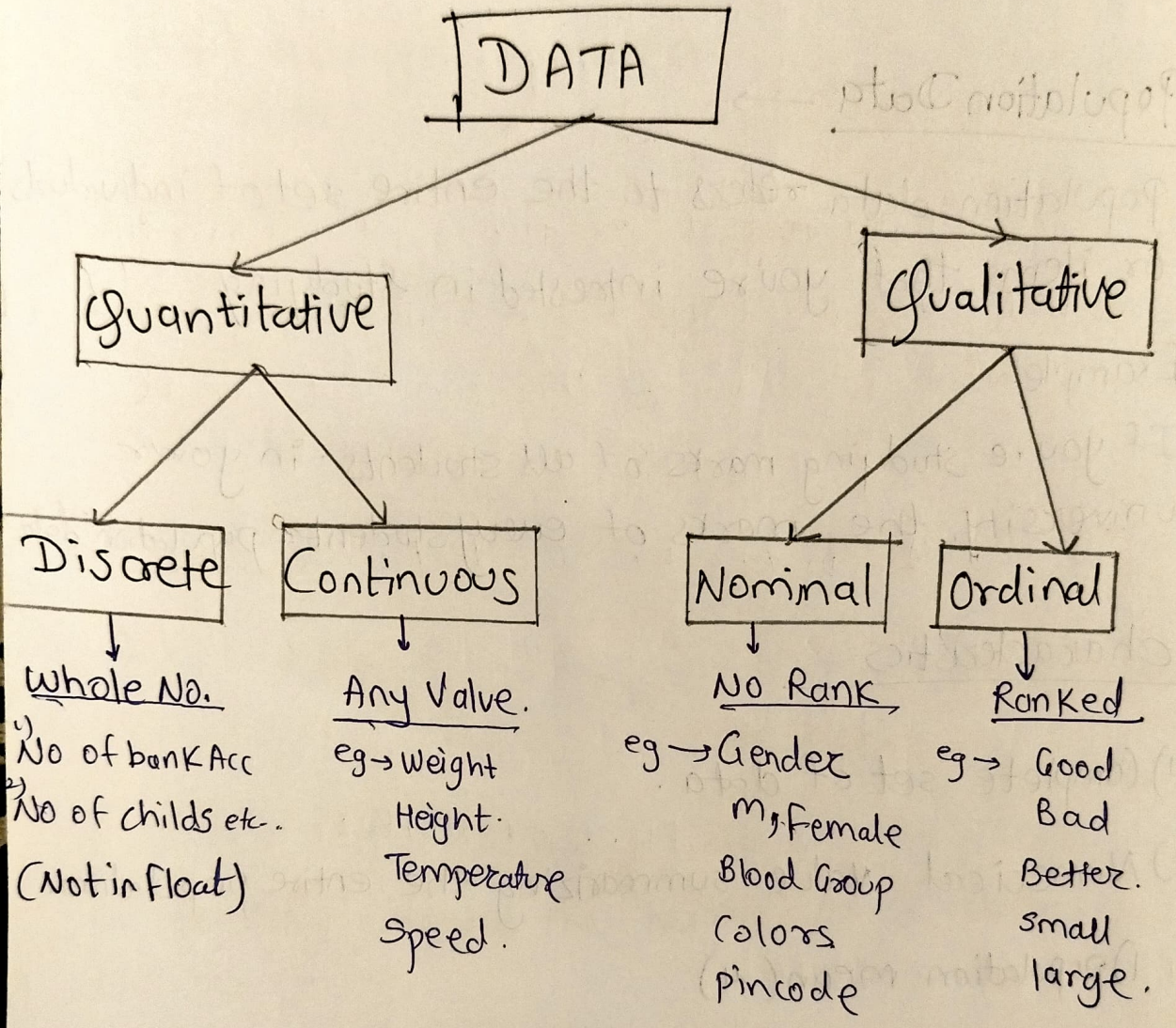
Sample data →

Sample data is a smaller part taken from the population, which is used to draw conclusions about the whole population.

Example:

If you take marks of 100 students out of 10000 that 100 students = Sample Data

Types of Data



Measure of central Tendancy

1) Mean or Average

Mean is the sum of all values divided by the number of values.

$$\mu = \frac{\sum x_i}{N}$$

x_i = numbers of element in set

N = Total Numbers count.

eg $X = \{1, 2, 3, 10, 20\}$

$$\mu = \frac{1+2+3+10+20}{5} = \frac{35}{5} = \underline{\underline{7}}$$

* Affected by extreme outliers, used for interval & ratio data

2] Median

Median is the middle value in the dataset. when the values are arranged in ascending or descending order.

If even

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

$M = \frac{3+4}{2} = \underline{\underline{3.5}}$

If odd

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

$M = 3$ (middle element)

3] Mode

Mode is the value that appears most frequently in a dataset.

Dataset X = {1, 2, 3, 3, 2, 1, 4, 5, 3}

mode = 3 (most frequent value)

Characteristics

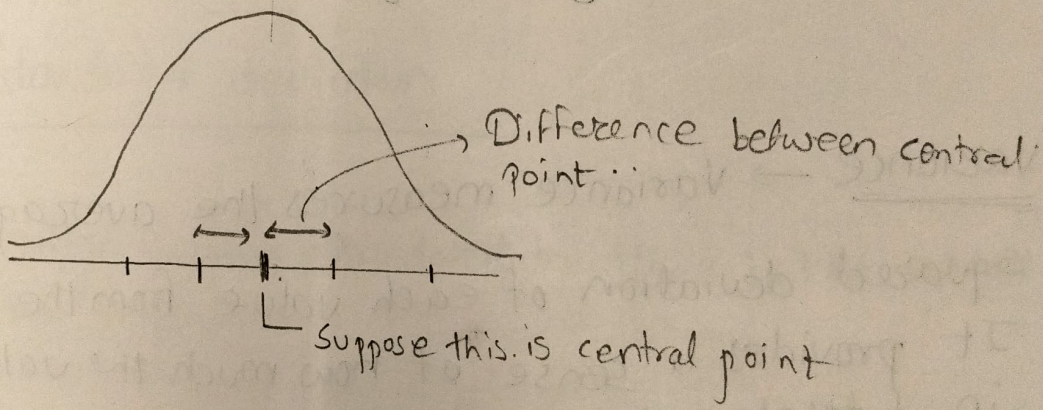
1) Not affected by extreme values.
2) Used for Nominal, ordinal, interval and ratio date.

Measure of Dispersion

- Measure of Dispersion describe the spread or variability of a dataset. They indicate how much the values in a dataset differ from the central tendancy.

Common Measure of Dispersion

1) Range
2) Variance
3) Standard Deviation
4) Interquartilic Range (IQR) → percentile.



1) **Range** :→ Range is differentce between maximum & minimum value in a dataset

Range = Maximum value − Minimum value.

Eg. Age = {14, 13, 10, 20, 25, 75, 15}

Range = ~~10 - 75~~ 75 − 10 = 65.

Characteristics :

1] Simple to calculate

2] Sensitive to outliers

3] Rough measure of dispersion.

weight = {35, 40, 45, 39, 30, 70} → outlier

Range = {45 - 30} = 15.

Range = {70 - 30} = 40

} See the diffrence.

2] Variance → Variance measures the average squared deviation of each value from the mean. It provides a sense of how much the values in dataset vary.

for,

Population Variance

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

for,

Sample Variance

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

Example → Size of flower petals.

$N = 5$ $\{5, 8, 12, 15, 20\}$ ⇒ variance of this distribution

$$\mu = \frac{5+8+12+15+20}{5} = \frac{60}{5} = 12.$$

$$\sigma^2 = \frac{(5-12)^2 + (8-12)^2 + (12-12)^2 + (15-12)^2 + (20-12)^2}{5}$$

$$\boxed{\sigma^2 = 27.6}$$

(3) Standard deviation

Defination ⇒ The standard deviation is the square root of the variance.

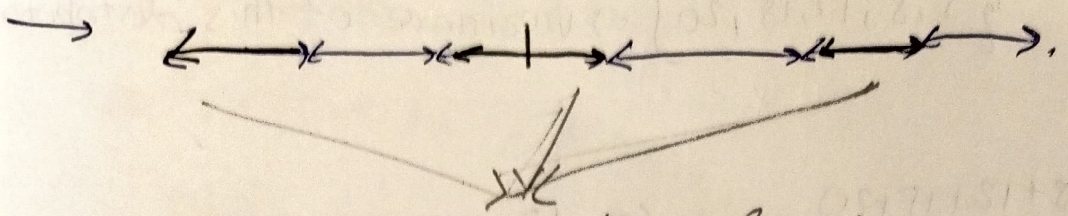
$$\sigma = \sqrt{\sigma^2}$$

$$\sigma = \sqrt{27.6}$$

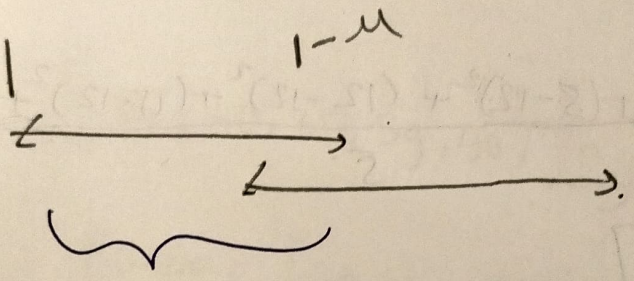
$$\boxed{\sigma = 5.25}$$

***Imp Interview question.

Q Why in Sample Variance $(n-1)$ is used ?



3 random Samples are Selected.



This distance is too large

So to reduce this the term "Bias correction" is uses so that we can reduce this huge difference by using $n-1$.

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

Random Variable

X ⟶ function ⟶ Values ⟶ Process or Experiments

$$X = \begin{cases} 0 & H \\ 1 & T \end{cases}$$

↑ assigned some values to it

Random Variable

- Discrete Random
  - eg → Tossing a Coin, Rolling a Dice.
- Continuous Random Variable
  - eg: Tommorow how much inches it is going to rain [0, 1.1, 5.5, 10.5, 10.75]

Percentage, Percentile & Quartile

1) Percentage = $\frac{\text{To count}}{\text{Total count}} \times 100$

eg = {1, 2, 3, 4, 5, 6}

odd no = $\frac{3}{6} \times 100 = 50\ \%$

2) Percentile = percentile of val x = $\frac{\#\text{ of values below } x}{n} \times 100$

X = {2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8, 9, 9, 10}

x = 9 = $\frac{11}{14} \times 100 = 78.57$ th of 9

(th is percentile symbol.)

To find Value from percentile.

Value = $\frac{\text{percentile.}}{100} \times (n+1)$ = $\frac{25}{100_4} \times (15) = 3.75 \cong 3$.

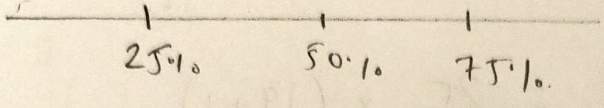
$\frac{3+4}{2} = 3.5$

3) Quartile

25% = 1st Quartile

50% = 2nd Quartile

75% = 3rd Quartile



25th percentile

Quartile is similar to the quadrant i.e $(\frac{1}{4}^{th})$ part of each

---

5 Number Summary & Boxplot

[To find outlier]

1] Minimum

2] Q1

3] Median

4] Q3

5] Maximum.

Removing the Outlier

$X = \{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 29\}$

[Lower fade ⟷ Higher fade]

IQR = Inter Quartile Range.

Lowerfade = $Q_1 - 1.5(IQR)$

Higher fade = $Q_3 + 1.5(IQR)$

$$\frac{n^{th}}{100} \times (n+1)$$

Q_1 = 25 percentile = $\frac{25}{100} \times (19+1)$

$= 5^{th}$ value. ≈ 3

Q_3 = 75 percentile = $\frac{75}{100} \times 20 = 15^{th}$ Value ≈ 7

$IQR = Q_3 - Q_1 = 7 - 3 = 4.$

Lower fade = $Q_1 - 1.5(IQR)$ Higher fade = $Q_3 + 1.5(IQR)$

$= 3 - 1.5(4)$ $= 7 + 1.5(4)$

$= -3$ $= 13.$

outlier = 19

#5 Number Summary

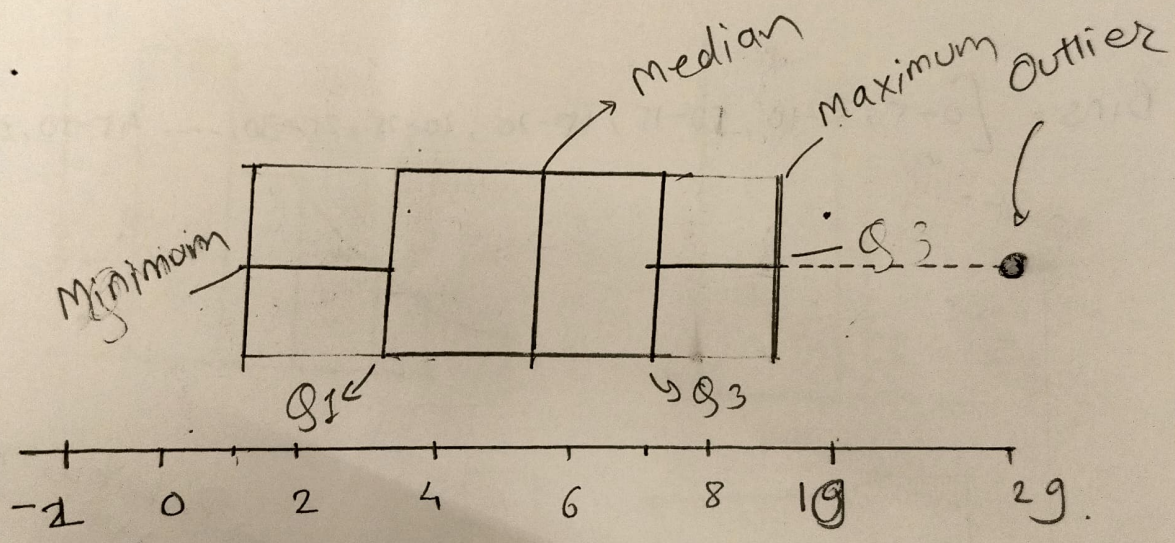
Minimum = 1. (exclude outlier)

$Q_1 = 3$

Median = 5.

$Q_3 = 7$

max = 9. (exclude outlier.)



Histogram & Skewness

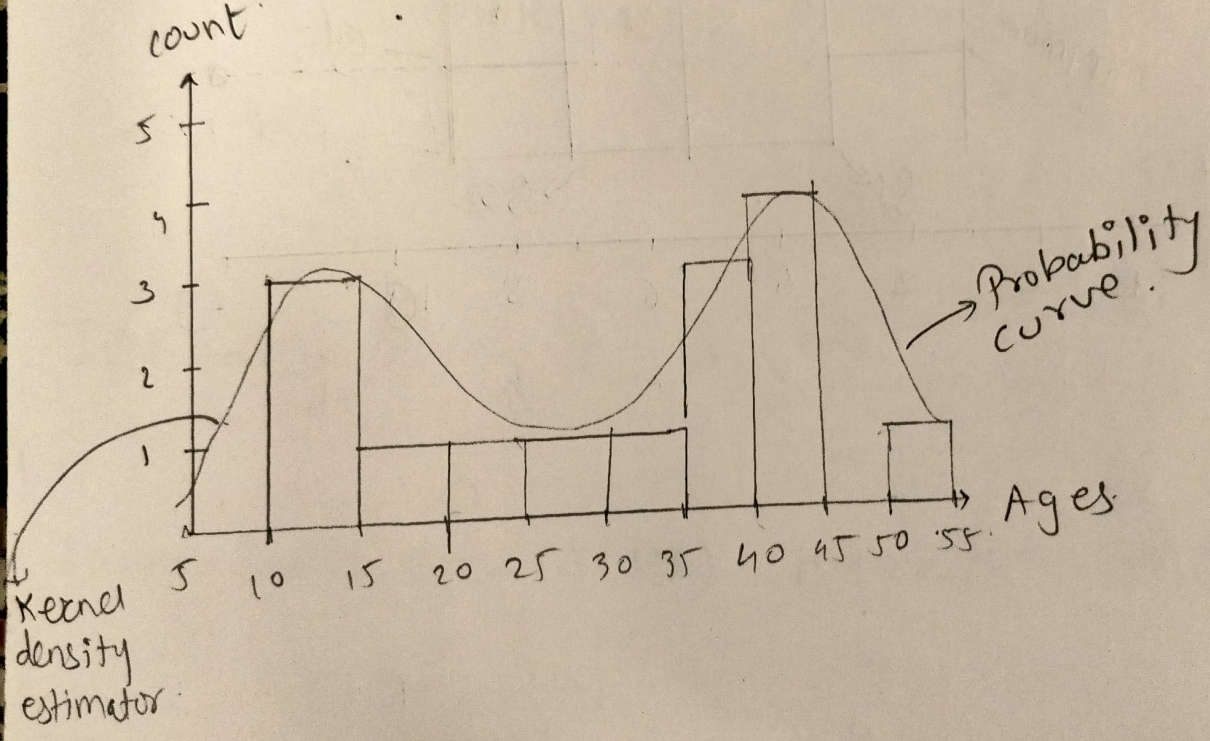
• Histogram is graphical representation of the distribution of numeric data.

Ages = {11, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50}

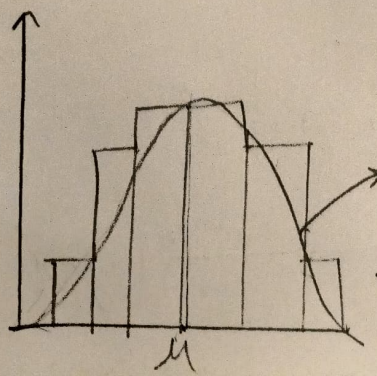
0 - 50

1) No of bins = 10 = $\frac{50}{10}$ = 5 bin unitsize

Bins = [0-5, 5-10, 10-15, 15-20, 20-25, 25-30, ---- 45-50, 50-55]



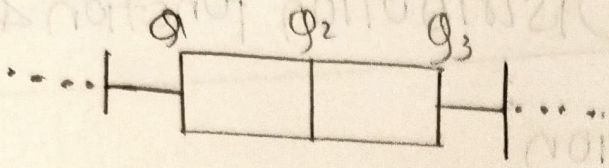
★ Skewness



Normal / Gaussian Distribution

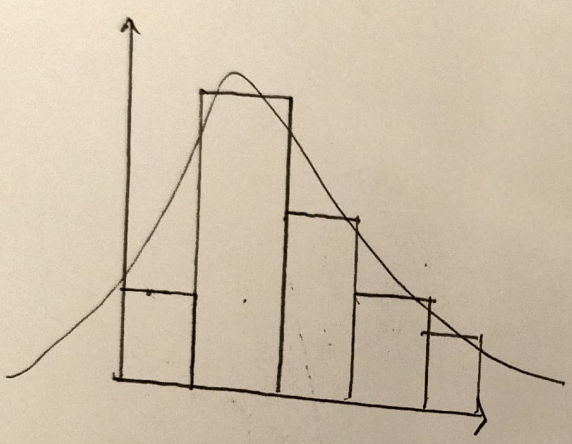
→ Symmetric distribution.

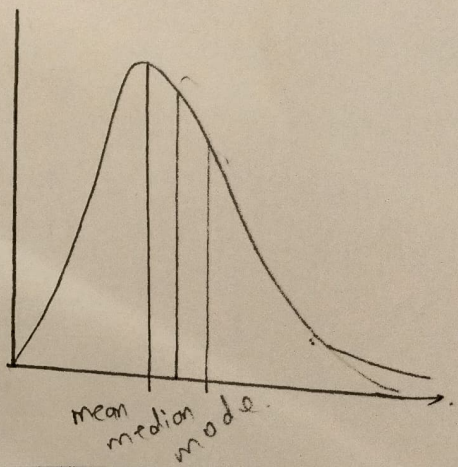
⇒ No skewness



The mean, median & mode are all perfectly at center

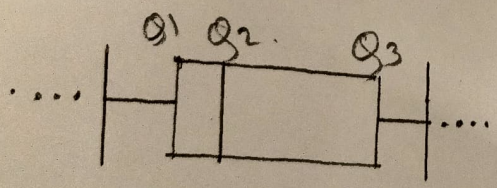
② Right skewed.





mean ≥ median ≥ mode

Box plot.



$Q_3 - Q_2 \geq Q_2 - Q_1$

③ left skewed → Vice versa