Data Science

- Statistic
  - ↳ sample
    - → mean
    - → std deviation
    - ↳ variance
- parameter
  - ↳ population
    - → mean
    - → std
    - ↳ variance

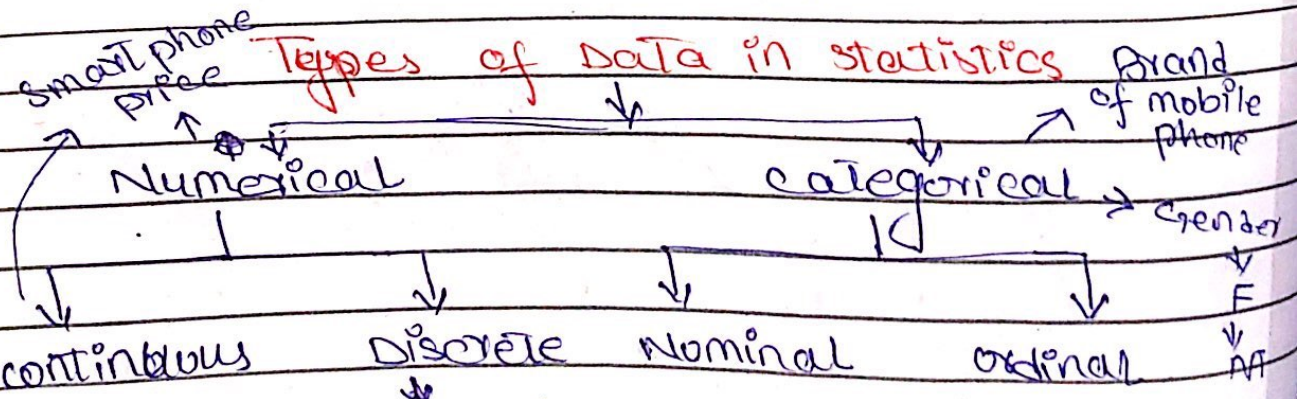
- **Descriptive statistics:**
  - → mean
  - → median
  - → mode
  - → std deviation
  - → variance
  - → MAD

  ↳ sampling from population. Draw conclusion from sample

  Bivariant (two variables)
  ↳ scatterplot

- **Inferential statistics:**
  - → confidence Interval
  - → Hypothesis Testing

Types of Data in statistics



→ order or relationship between values is ordinal.

• categorical
(nominal) → pie chart
→ Bar chart
(ordinal) → pie, Bar

Numerical
→ Histogram

→ No order or no relationship between values is nominal.

2. Visualization Technique:

Numerical → Histogram
Bivariate (two variables) → scatterplot
numerical → continuous → Univariate
→ Boxplot
→ KDEplot

~~3. Univariate~~

• Categorical (nominal, ordinal)
→ pie chart
→ Bar chart

Bivariate
• Boxplot
→ x-axis → categorical
→ y-axis → Numerical

→ one categorical, one numerical
→ Boxplot

• Both numerical values → scatter plot

• Both categorical → Bar chart
→ heatmap

2. Measure of central Tendency:

→ where our values from data occurs the most
→ mean (when data have outlier, mean don't be used)
Boxplot is used to find outlier.

→ When data have outliers, use median.

→ When range of data is high, median don't effect which is another good about median.

→ We need to find three of them (mean, median, mode), but use them properly and of to get fair and proper idea of centrality.

3. Measure of spreadi

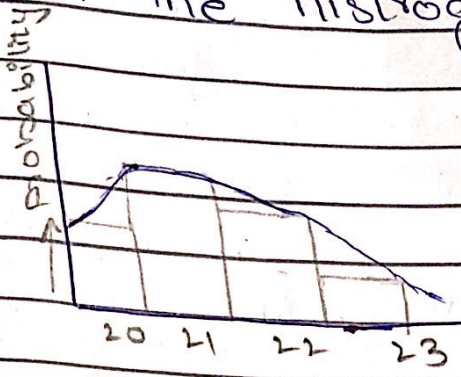
→ How much our data is spreaded

1. Range
2. Interquartile range
3. Std deviation
4. variance.

- Range → difference between smallest and largest values in data.
- percentiles → indicating the below ~~value~~ percentage of ~~given~~ observations in a group of observations falls.
- Quartiles → divide data into quarters.

Data Science

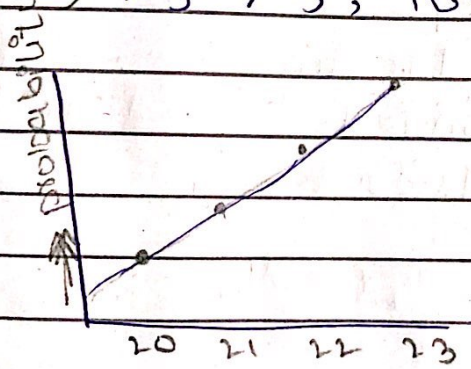
showing
→ probability's ~~funct~~ density function (PDF) is KDE
→ smoothen the histogram is simply PDF



- PDf → tell information about only the certain point

CDF :

20 → 5
21 → 7, 7+5 = 12
22 → 6, 12+6 = 18
23 → 3, 18+3 = 21



- CDF → tells about all values before the certain point.

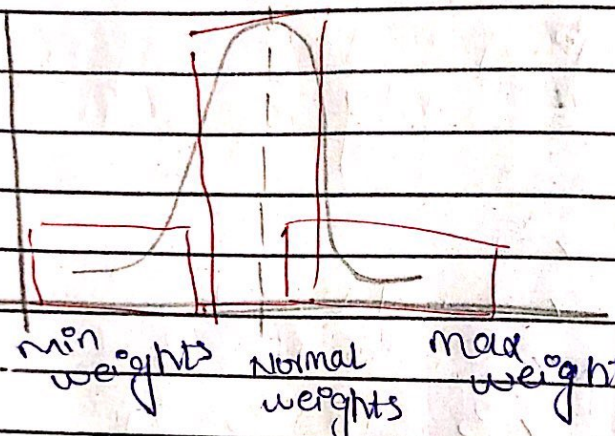
- Pdf = count / sum(count)
- CDf = cumsum(pdf)

→ CDf tells us how much probability of our result is be correct and how much ~~there~~ probability of ~~of~~ our result be wrong.

→ CDF supports PDF by additional of sum results.

- **Normal Distribution:**

→ Piosson
→ Binomial
→ Bernoulli
→ Normal Distribution



equation

→ **Central Limits Theorum:**
We can convert any distribution into normal distribution

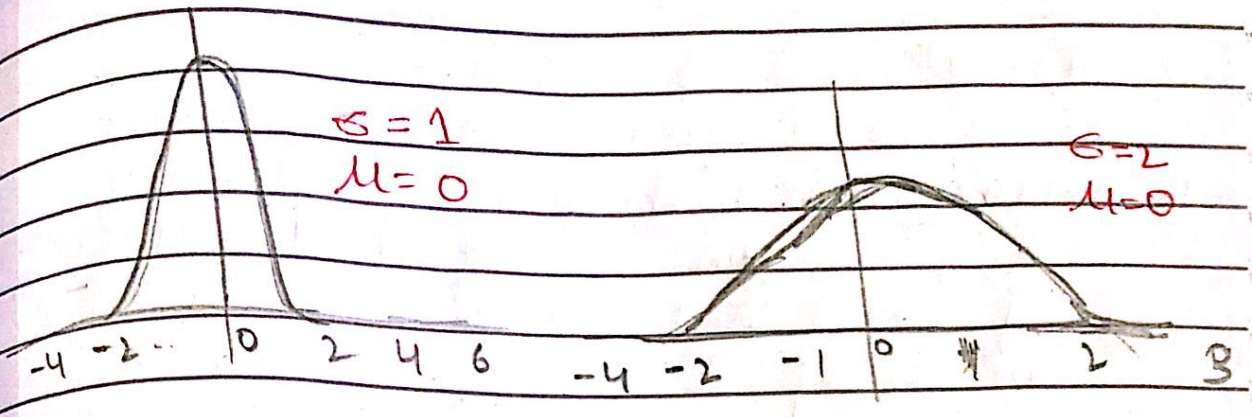
$$\rightarrow p(x=x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

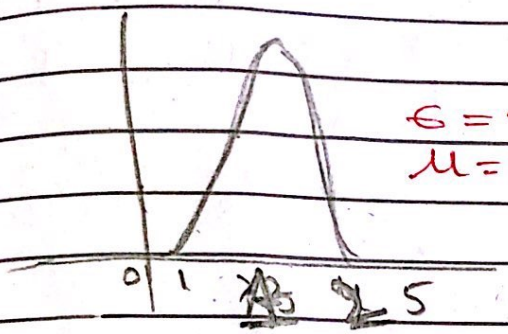
μ → mean
σ → std deviation
σ^2 → variance.

when $\mu = 0$, $\sigma = 1$ (gaussion distribution)

$$p\xi = \boxed{\frac{1}{\sqrt{2\pi}}} e^{\frac{1}{2}x^2}$$

$\boxed{\sqrt{2\pi}}$ → constant





→ calculating z-score

→ $z = \frac{\text{Total score} - \text{mean}}{\text{std}}$

→ z value in z-table

→ than multiply that z-table value with 100.

→ Mean, Mode, Median is always same for Normal distribution.