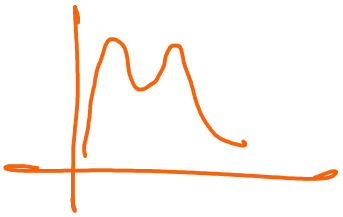
Tuesday, 9 December 2025 6:04 PM

- Continuous data

- Type of multimodal.
  - 2 peaks → Bimodal
  - 2 peak → Trimodal
  - many peak → multimodal

Multimodal → use data belonging to multiple types.
- Text
- Image
- Audio
- Video
- etc



- Occur when data has multiple clusters.
- Eg. Height of adult
  - men
  - women.

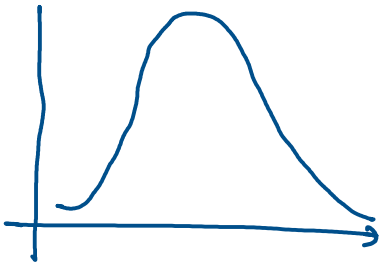
* Multinomial Distribution

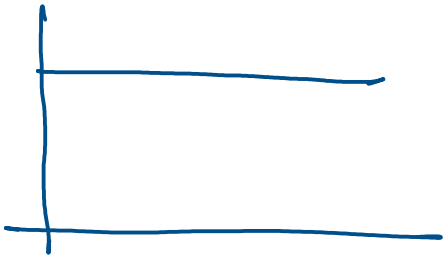
Binomial distribution → it occurs for multiple trials where each trial has 2 categories

But, if we are conducting multiple trials where each trial has more than 2 outcomes

Eg. Rolling of a dice
S = {1, 2, 3, 4, 5, 6}.

* Uniform distribution (Discrete + Continuous)
  - every outcome has an equal probability of occurring.





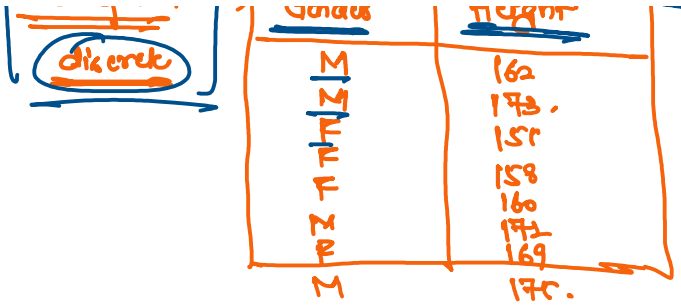
Unlike the bell shape curve → normal distribution
for uniform distribution → flat curve → because every outcome has equally likely chances to happen

- Uniform distribution (Discrete data)

Discrete data
- countable values.

eg. categories like men/female, day of week, dice outcomes (1–6).

Categorical → Numerical



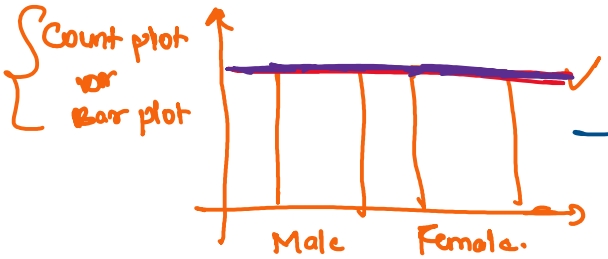

→ numerical / continuous

- Data is distributed for the gender colum.

| Categories | Count |
|---|---|
| Male | 4 |
| Female | 4 |

Since both the categories appear equally.

- uniformly distributed.
- Bernoullik distribution
  - ↳ cardinality (2)



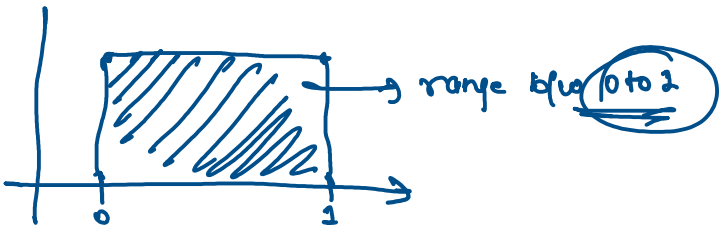

→ indicate.
- uniformly distributed
- Bernoulli distributed.

- **Continuous data (uniform distribution)**

For a continuous uniform distribution

- The probability is equally spread across a range.

Eg. Randomly generate/pick a no. b/w 0 to 1.




$P(\text{Outcome}) = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcome}}$ (∞).

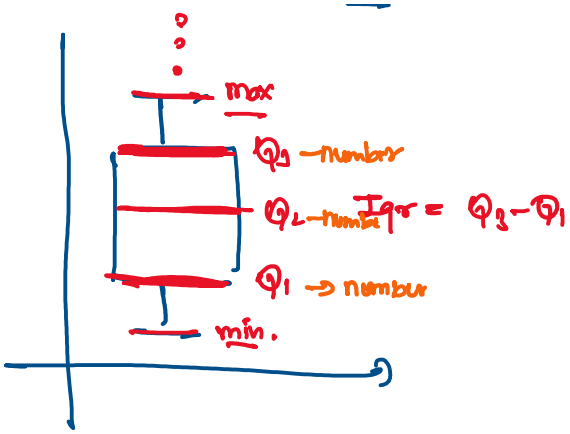
0
0.01
0.001
0.0001
0.00001
0.

- If the data is continuous numerical values

| Height |
|---|
| 150.01 |
| 150.12 |
| 150.34. |

$0 < no < 1$

→ plot
- Count plot

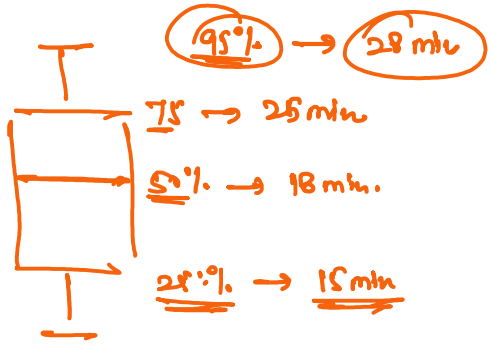



- box plot
- plots such as
  - Histogram
  - KDE plot

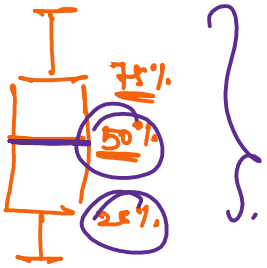
Box plot doesn't tell us about distribution

1% → 10
99th % → 30

Domain



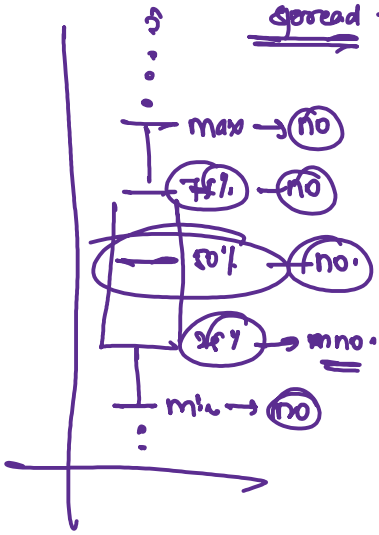

Titanic → Age

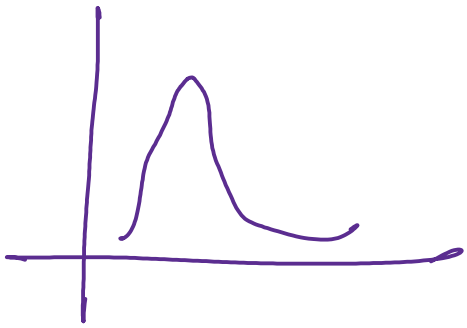


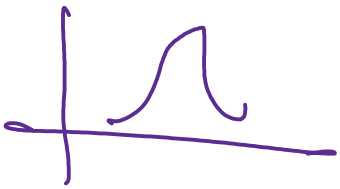

$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Box plot → Not going to tell us about distribution
↳ Instead it tell us about spread + outliers.

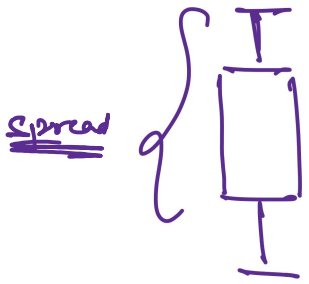
$\frac{1+2+3+4+100}{5} = \frac{110}{5} = 22$

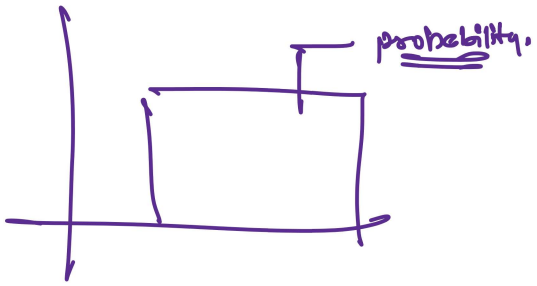


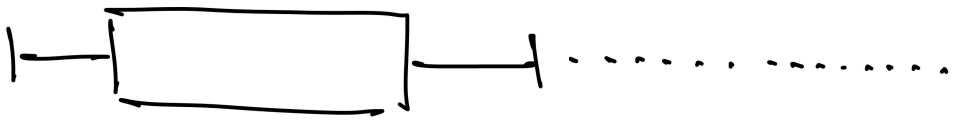





median → 16 → 50% → 16



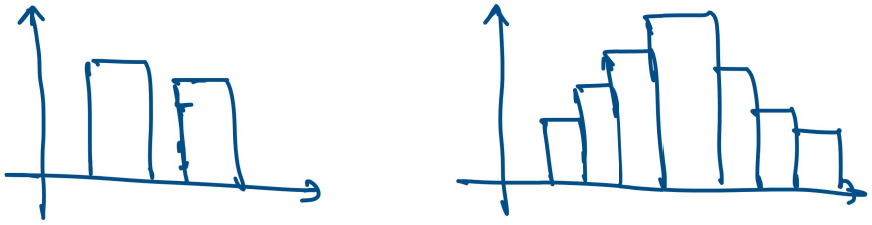



100% — curve

Distribution
- Box plot → (X)
- KDE.
- histogram.

→ It a histogram the count plot for numerical data.



→ Yes, we can consider histogram a count plot for numerical data because

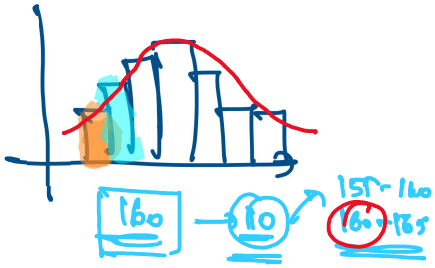
- The y-axis shows. → count (frequency).
- Bars represents → how much data falls in each group.

But in histogram, we have bins (intervals)

Height

150 – 155 cm
155 – 160
160 – 165
⋮

} each group.

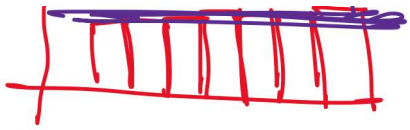



bin → bin size (5)

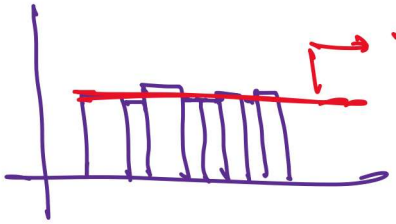
155 – 160
160 – 165 } .

155 – 165 → bin size (10)

→ look like flat or equally spread

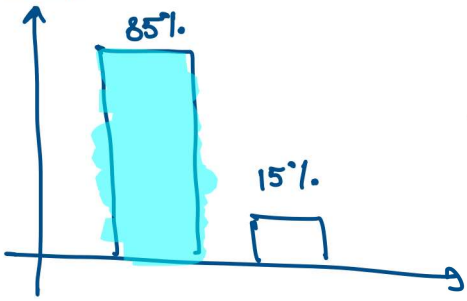


↳ approx



→ yes, the data is uniformly distributed.
means similar no. of values exist across the range.

- why.?
  - Crucial role of uniform distribution → imbalance detection.




→ Data → ML algo.

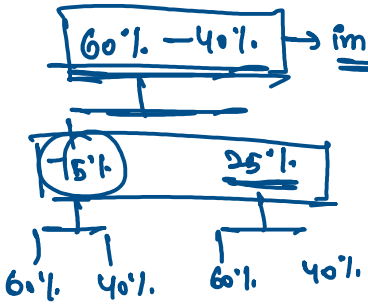
ML algo + data → model → learns from data → 85% → ML
15%

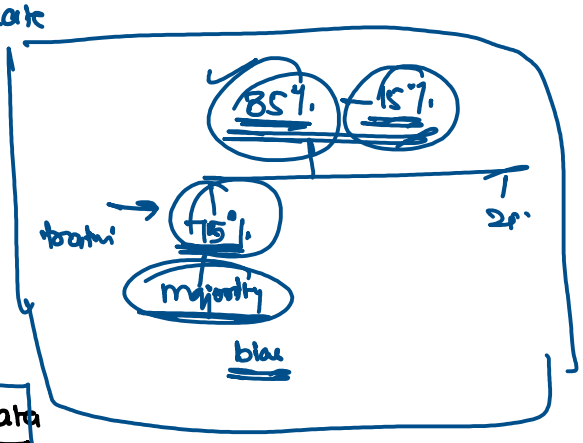
SMOTE

50%-50% → cannot exist

60%-40%.
70%-30%.
75%-25%. } allowed → capture real time instance
85%-15%.
90%-10%.
95%-5%. } not allowed.

60%-40% → imbalances in data

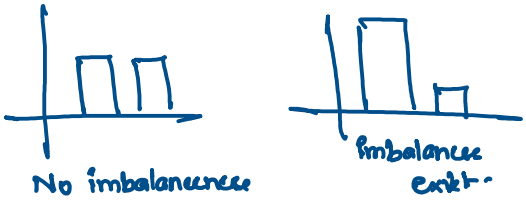



Imbalance data

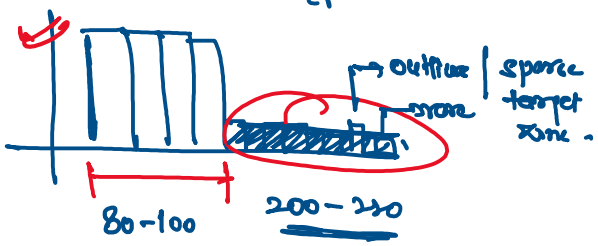
Classification imbalance.

- Detected by count plot.
- One target class occurs more frequently than the other.




Regression imbalance

- Detected b histograms.




- In the histogram
  - One region → high frequency.
  - Other region → very low frequency.

- Why imbalanceness in a problem

Classification

↳ model learns mostly from the majority class → poor prediction. (bias models)

Regression

- If certain values are rare
  - The model will not learn those pattern well
  - Prediction near those values becomes inaccurate
  - Errors MSE / RMSE increase sharply.

Imbalanceness introduces

→ high error
→ high bias.