Date
Page

DAY - 2

Histogram

Ages = { 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 43, 43, 50, 51, 65, 68, 78, 90, 95, 100 }

→ Descending order

1. Sort the Numbers
2. Bins :- No. of groups
3. Bin size

eg :- [10, 20, 25, 30, 35, 40]

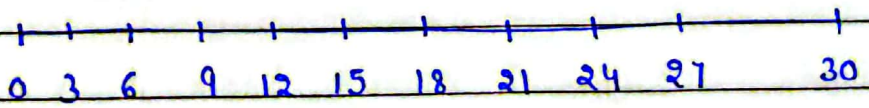
min = 10

max = 40

Now, I want to divide data into 10 groups.

⇒ bins = 10

→ bin size = $\frac{40 - 10}{10} = 3$ $\left(\frac{max - min}{bins}\right)$



Here, bins = 10 but the value is ranging b/w 0 to 30

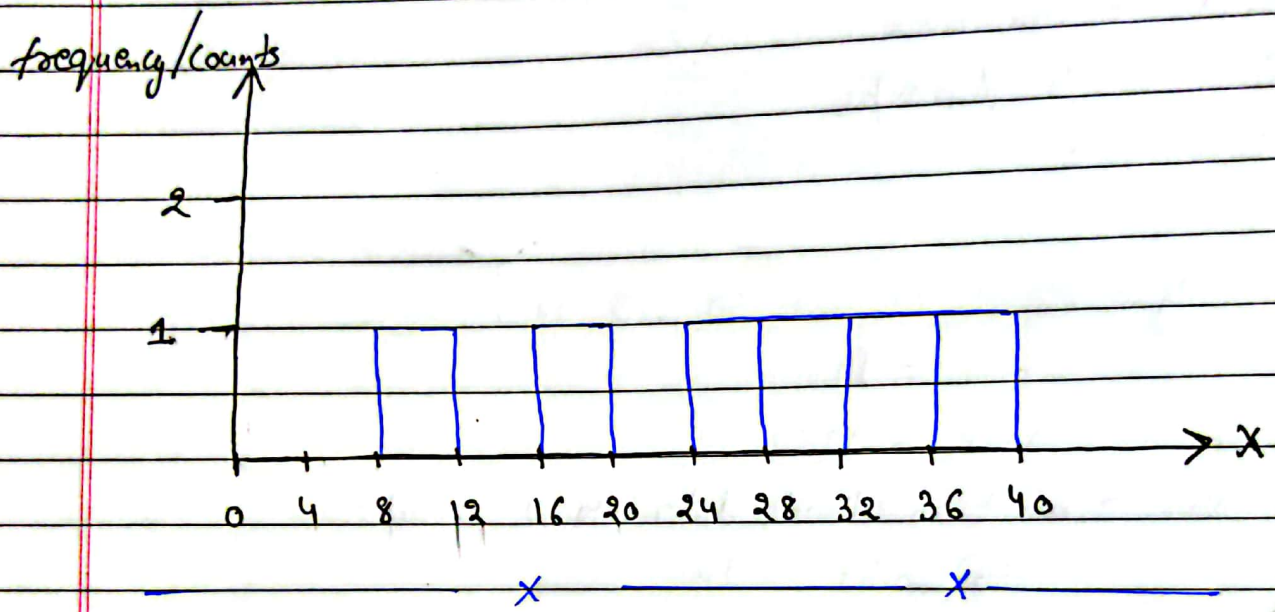
It should be ranging b/w 0 to 40.

[10, 20, 25, 30, 35, 40]

min = 10, max = 10

bins = 10

bin size = $\frac{40}{10} = 4$



now,

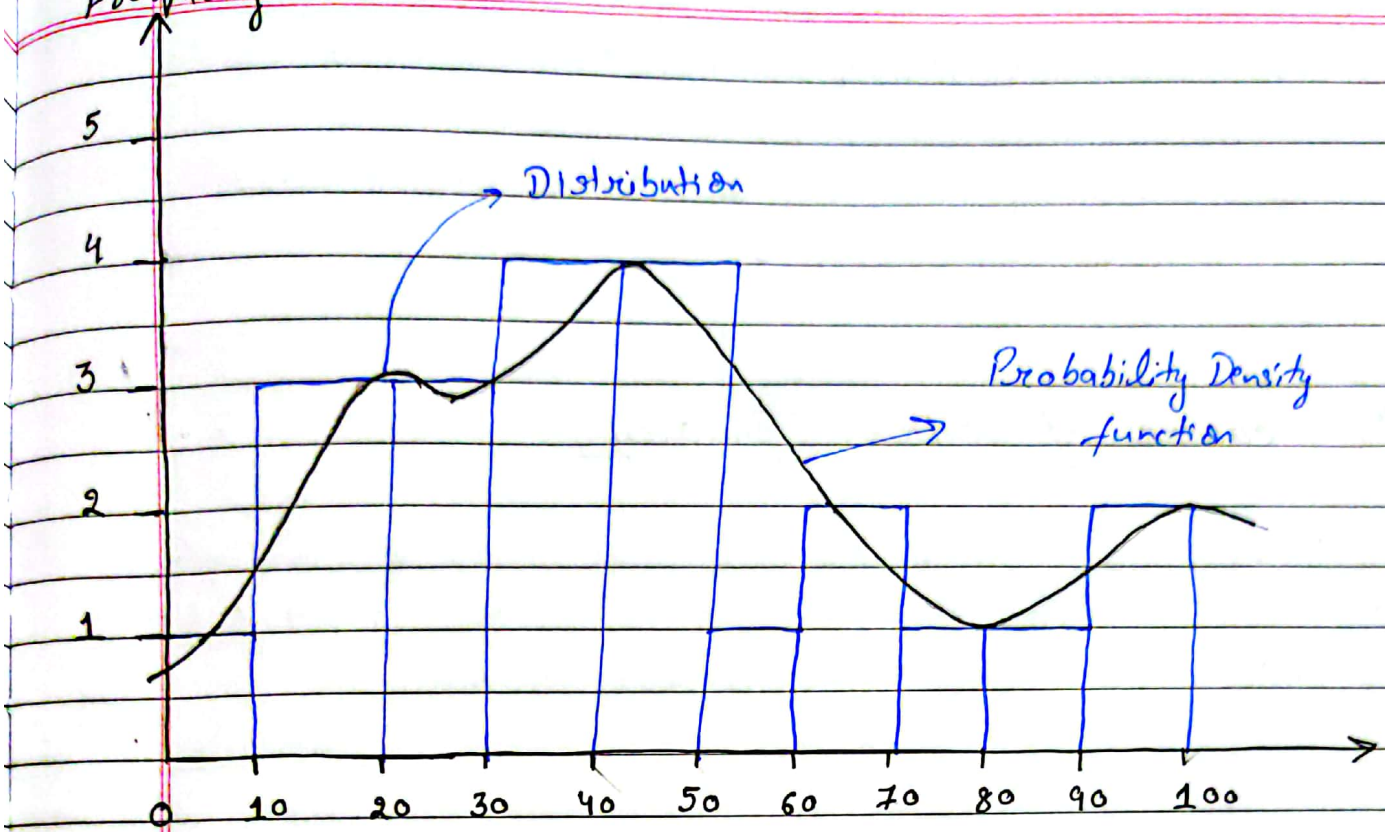
Ages = [10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100]

min = 10

max = 100

bins = 10

$\Rightarrow$ bin size = $\frac{100}{10} = 10$



→ Smoothing of Histogram creates Probability Density Function

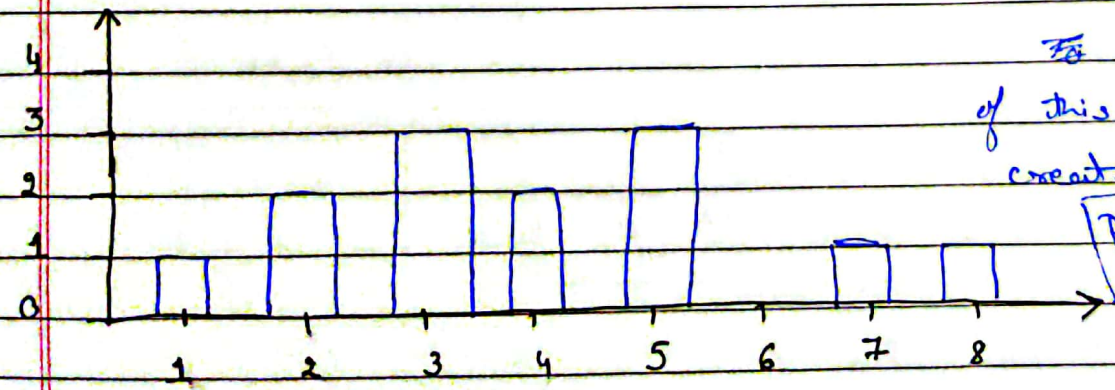
→ Through the smoothing of histogram we get distribution of data.

x x

~~Measure of Central~~

② Discrete ~~data~~ Variable

Num of bank accounts :- [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



~~For~~ smoothing of this histogram creates Probability Mass Function

Pdf : probability density function → for continuous variable

Pmf : probability mass function → for discrete variable

Measure of Central Tendency

① Mean
② Median
③ Mode

A measure of central Tendency is a single value that attempts to describe a set of data identifying the central position.

① Mean → $X = \{1, 2, 3, 4, 5\}$

$$\text{Average / Mean} = \frac{1+2+3+4+5}{5} = 3$$

→ Mean should be defined based on 2 factors

Population (N) | Sample (n)

$$\text{Population mean } (\mu) = \sum_{i=1}^{N} \frac{X_i}{N}$$

$N \geq n$

$$\text{Sample mean } (\bar{x}) = \sum_{i=1}^{n} \frac{X_i}{n}$$

Examples

→ Population Age = [24, 23, 2, 1, 28, 27]

↓

$N = 6$

Population mean (μ) = $\frac{24+23+2+1+28+27}{6} = 17.5$

Sample Age = [24, 2, 1, 27]

↓

$n = 4$

sample mean ($\bar{x}$) = $\frac{24+2+1+27}{4} = 13.5$

* ① $N \neq n$

② $\mu \neq \bar{x}$ or $\bar{x} \neq \mu$

→ Practical Application (Feature Engineering)

Age	Salary	Family Size
–	–	–
–	–	–
NaN	–	–
–	NaN	–
–	–	NaN

← If we drop whole row with NaN there will be loss of info

So we can find ~~mean~~ mean of each column and replace NaN values with that.

② Median :-

eg :- [1, 2, 3, 4, 5] → $\bar{x} = 3$

[1, 2, 3, 4, 5, 100] → $\bar{x} = 19.16$ (100 is an outlier)

Steps to find out median

① Sort the Numbers

② find the central number given two conditions.

(a) If num of elements are even we find average of central elements.

(b) If num of elements are odd we find ~~average~~ central element.

eg :-

[1, 2, 3, 4, 5, 6, 7, 8, 100, 200]

median = $\frac{5+6}{2} = \frac{11}{2} = 5.5$

100, 200 are outliers

[1, 2, 3, 4, 5, 6, 7, 8, 100]

median = 5

③ Mode :- Most frequent occuring element

[1, 2, 2, 3, 3, 3, 4, 4, 5] → 3

[1, 2, 3, 2, 3, 3, 3, 4, 4, 5] → 2, 3

Practical Example

Types of Flowers

Lily
Sunflower
Rose
NaN
Rose
Sunflower
Rose
NaN

Here we can use mode value

* Measure of Dispersion

① Variance (σ^2) → Spread of data

② Standard Deviation (σ)

Variance

Population Variance (σ^2)

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

Sample Variance (S^2)

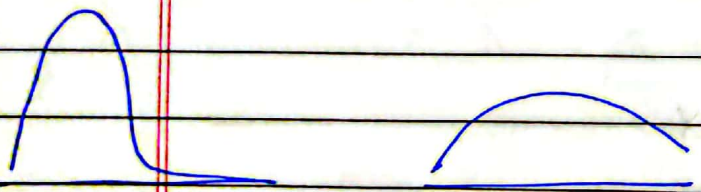
$$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$x_i - \mu \rightarrow$ Distance from Mean

① [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

② [1, 2, 3, 4, 50, 60, 70, 100]

Variance 1 < Variance 2



As variance ↑ Spread ↑

(2) Standard Deviation (σ)

Standard Deviation ($\sqrt{\sigma^2}$)

$\{1, 2, 3, 4, 5\}$

$\sigma^2 =$ ~~[illegible]~~, $\mu = 3$

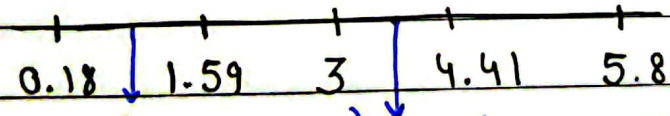
$\mu = 3$

$\sigma^2 = 2$

$\sigma = \sqrt{2} = 1.14$

2 std to left $\leftarrow -2 \quad +2 \rightarrow$ 2 std. to the right

1 std. to left $\leftarrow -1 \quad +1 \rightarrow$ 1 std. to the right



(1 falls in 2nd std dev to left) 4 → (so 4 falls in 1st std. deviation)

Std. ~~[illegible]~~ tells that how many std. away a number falls away the mean.

* Percentiles and Quartiles

Percentage = { 1, 2, 3, 4, 5, 6, 7, 8 }

% of even nos = $\frac{\text{no. of even nums}}{\text{Total elements}} \times 100$

$= \frac{4}{8} = 50\%$

Percentiles :- A percentile is a value below which a certain percentage of observations lie.

eg:-

99 percentile → It means the person has got better marks than 99% of entire students

Ascending order

Dataset :- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

(16 values) (20 values)

Q What is the percentile rank of 10

Percentile Rank of x = $\frac{\#\text{ no. of values below x}}{\text{sample size (n)}}$

Percentile Rank of 10 = $\frac{16}{20}$ = 80 percentile

Percentile Rank of 8 = $\frac{9}{20}$ = 45 percentile

↓

45% of entire values is less than 8

Q What is the value that exist at 25 percentile?

$$\text{Value} = \frac{\text{Percentile}}{100} \times (n+1) \rightarrow \text{for odd}$$

$\longrightarrow$ Value $= \frac{25}{100} \times 20$

$= 5^{th}$ index

$\downarrow$

5 will be the output

now 95 percentile value

$$\text{value} = \frac{95}{100} \times 20 = 19.95 \text{ index}$$

$= 12 \rightarrow$ output

* 5 number Summary

1. Minimum
2. First Quartile (25 percentile) (Q1)
3. Median
4. Third Quartile (75 percentile) (Q3)
5. Maximum

⇒ Remove the outliers

⇓

Box plot

eg:- [1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 9, 27]

27 ↓ outlier

↓

Here, we will try to create a fence and value should range b/w lower fence and higher fence value.

[Lower fence ⟷ Higher fence]

→ Lower fence = Q1 − 1.5 (IQR)

IQR = Q3 − Q1

↓

Inter Quartile Range

→ Higher fence = Q3 + 1.5 (IQR)

↓

1.5 std dev. to the mean

$$Q1 = \frac{25}{100} \times (n+1)$$

$$= \frac{25}{100} \times 21 = 5.25 \text{ index}$$

↓

Take average of 5th & 6th index value

↓

Output = 3

$$Q3 = \frac{75}{100} \times 21 = 15.75 = \frac{8+7}{2} = 7.5$$

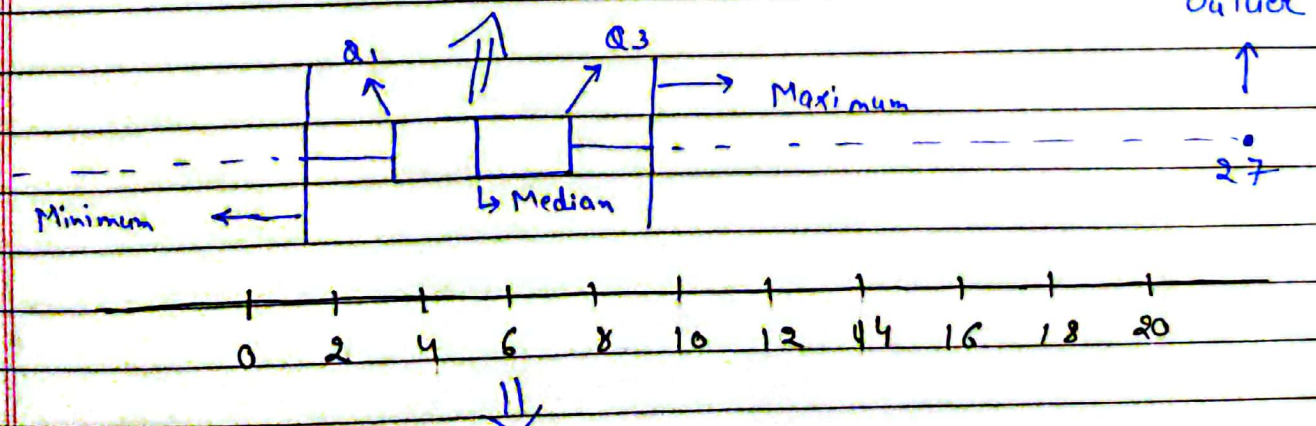
$\Rightarrow$ Lower fence $= 3 - (1.5)(4.5) = -3.65$

Higher fence $= 7.5 + (1.5)(4.5) = 14.25$

[1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27]

① minimum = 1
② Q1 = 3
③ Median = 5
④ Q3 = 7.5
⑤ Maximum = 9

Box Plot



To Treat Outliers