## Agenda:

1. Histograms
2. Measure of central Tendency.
3. Measure of Dispersion
4. Percentiles and Quartiles
5. 5 Number Summary (Box plot)

## Histogram:

Ages = {0, 10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100} (Continuous Values)

Steps to follow:

{Bins are nothing but groups
Binsize = groupsize}

1. Sort the numbers
2. Bins → No. of groups
3. Bin size → Size of Bins

→ the numbers are already sorted.

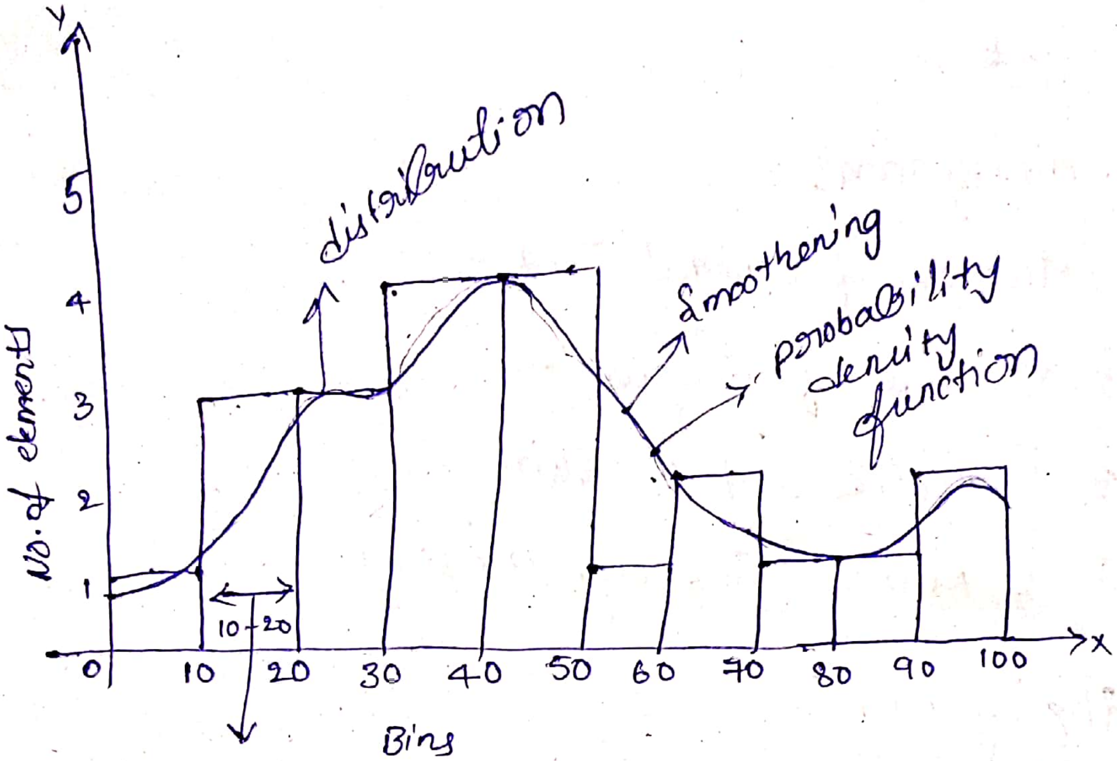
→ Bins = 10 (Here bins can taken by us that how many no of groups(bins) we want, we can take)

(Here I'm taking 10 bins)

$$\rightarrow \text{Binsize} = \frac{max - min}{bins}$$

$$= \frac{100 - 0}{10} = \frac{100}{10} = 10$$

So, Binsize = 10




between 10-20 we have 3 elements in the list, same for all.

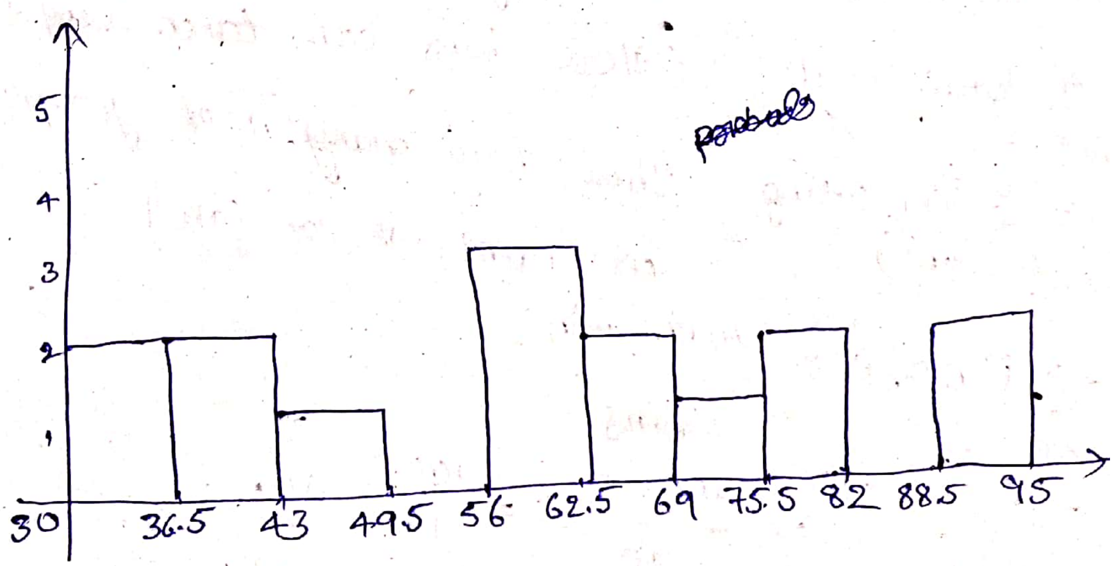
Ex-2

weights = { 30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95 }

(continuous values)

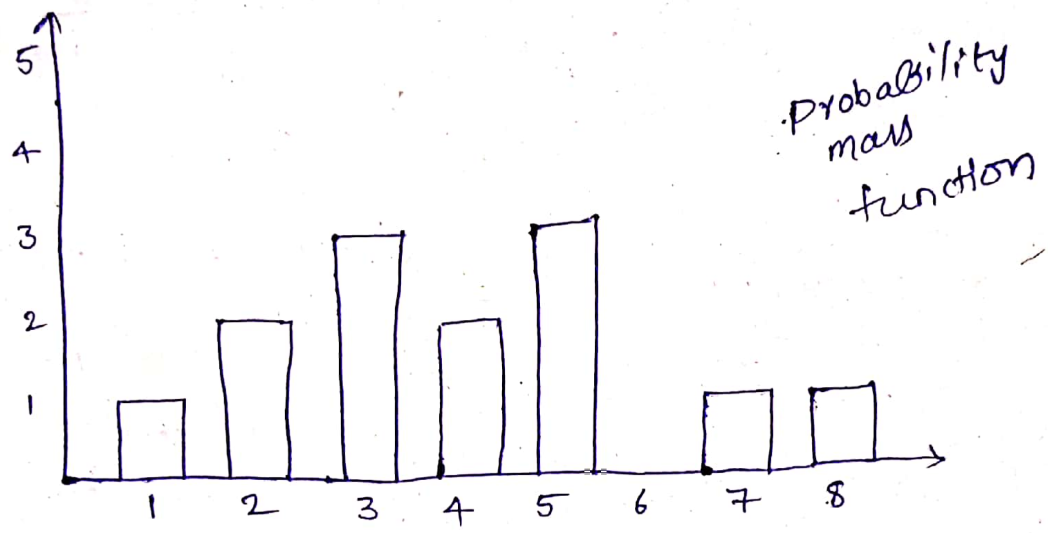
bins = 10

$$\text{binsize} = \frac{95-30}{10} = \frac{65}{10} = 6.5$$



for Discrete value.

No. of Bank accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]




pdf = probability density function } → continuous
pmf = probability mass function } → discrete

## * Measure of central Tendency:

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

→ there are three methods to identify the central position

1). Mean
2) Median
3). Mode

**Mean:**

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

$$\text{Average / mean} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

Population (N) $\boxed{N \geq n}$ Sample (n)

population mean ($\mu$) = $\boxed{\sum_{i=1}^{N} \frac{x_i}{N}}$ sample mean($\bar{x}$) = $\boxed{\sum_{i=1}^{n} \frac{x_i}{n}}$

population age = {24, 23, 2, 1, 28, 27}

$\boxed{N = 6}$

sample age = {24, 2, 1, 27}
↳ Here I picked 4 values randomly from population age.

$\boxed{n = 4}$

population mean ($\mu$) = $\frac{24+23+2+1+28+27}{6}$

$\boxed{\mu = 17.5}$

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

$\boxed{\bar{x} = 13.5}$

$\boxed{\mu \geq \bar{x}}$

$\cancel{\bar{x} \geq \mu}$

## Practical Application (Feature Engineering)

| Age | salary |
|---|---|
| 24 | 45 |
| 28 | 50 |
| 29 | NAN |
| NAN | 60 |
| 31 | 75 |
| 36 | 80 |
| NAN | NAN |

($\mu$) Age = 29.6
⇓
38 we for NAN value

Salary = 62
⇓
85 for NAN appr

If we remove total record due to NAN value then there is some loss of information.

So, by finding average (mean) we can give some values to NAN values.

## * Median:

In mean we have one problem, i.e, there may be a chance of occuring outliers, due to outliers the average value is changed.

for ex:

$\{1, 2, 3, 4, 5\}$ $\Downarrow$ for the mean $\bar{x} = 3$

$\{1, 2, 3, 4, 5, \overset{\substack{\text{outlier} \\ \uparrow}}{100}\}$ $\Downarrow$ for this $\bar{x} = 19.16$

so, we gone to median.

**Steps to find out median:**

1. Sort the numbers
2. Find the central number.

$\Downarrow$

case ①: if the no. of elements are even we find the average of central elements

case ②; if the no. of elements are odd we find the central elements.

Ex:

$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 100, 120\}$

median $= 5$

if $\{1, 2, 3, 4, \boxed{5, 6}\ 7, 8, 100, 120\}$, then

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

* **mode:** most frequent occuring element

Ex-1

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

$\boxed{\text{mode} = 3}$

Ex-2

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

$\boxed{\text{mode} = 2, 3}$

practical application:

Dataset

Types of flowers {categorical variable}

Lily
Sunflower
$\boxed{\text{Rose}}$
NAN → rose
$\boxed{\text{Rose}}$
Sunflower
$\boxed{\text{Rose}}$
NAN → Rose

Here rose is most frequently occuring, So, we replace NAN values with rose

* **Measure of Dispersion**

① Variance ($\sigma^2$) ← Spread of data

② standard deviation ($\sigma$)

$$x = \{1, 2, 3, 4, 5\} \quad \mu = 3$$

Variance

population variance ($\sigma^2$)

$$\boxed{\sigma^2 = \sum_{i=0}^{N} \frac{(x_i - \mu)^2}{N}}$$

sample variance $s^2$

$$\boxed{s^2 = \sum_{i=0}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Ex:

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

$\mu = 3$

$$\sigma^2 = \frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}$$

$$= \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2$$

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

Ex:2

$x = \{1, 2, 3, 4, 5, 6, 80\}$

$\mu = 14.4$

$$\sigma^2 = \frac{(1-14.4)^2+(2-14.4)^2+(3-14.4)^2+(4-14.4)^2+(5-14.4)^2+(6-14.4)^2+(80-14.4)^2}{7}$$

$$\boxed{\sigma^2 = 719.10}$$

* Here we can observe that if variance increases the spread of data is also increases

for example

when $\sigma^2 = 2$     when $\sigma^2 = 14.2$

$\sigma^2 < \sigma^2$

Scanned with CamScanner

(2) standard deviation $(\sqrt{\sigma^2}) \Rightarrow \boxed{\sigma}$

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

$\mu = 3$

$\sigma^2 = 2$ (previously we find out variance)

$\sigma = \sqrt{2} = 1.41$

Here 1.41 is the standard deviation.

* Percentiles and Quartiles:

percentage $= \{1, 2, 3, 4, 5, 6, 7, 8\}$

$$\text{percentage of Even number} = \frac{\text{no. of even numbers}}{\text{Total no. of numbers}}$$

$$= \frac{4}{8} = 0.5 \Rightarrow 50\%$$

percentiles:

Def: A percentile is a value below which a certain percentage of observation lie.

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

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

* what is the percentile rank of 10 ?

$$\text{percentile rank of } x = \frac{\text{no. of values below } x}{n}$$

$$= \frac{16}{20} = 80 \text{ percentile}$$

→ Here 16 values are present before 10 and the total values are 20.

* what is the value that exists at 25 percentile?

$$\text{value} = \frac{\text{percentile}}{100} * n+1$$

$$= \frac{\cancel{25}^{5}}{\cancel{100}_{20}} \times \cancel{20} = 5^{th} \text{ Index}$$

The value present at 5th index is 5.

* 5 number Summary:

1. Minimum
2. First Quartile (25 percentile) (Q1)
3. median
4. third quartile (75 percentile) (Q3)
5. maximum

⇒ To Remove the outliers ⇓ and to plot boxplot

Scanned with CamScanner

$\{1,2,2,2,3,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27\}$

~~first~~ to find the outliers, first we need to find lower fence and higher fence.

[Lower Fence ⟷ Higher Fence]

Lower Fence $= Q_1 - 1.5(IQR)$

Higher Fence $= Q_3 + 1.5(IQR)$

$IQR = \overset{75}{Q_3} - \overset{25}{Q_1}$

⇓

Inter Quartile Range (IQR)

$Q_1 = \frac{25}{100} \times 21 = 5.25$ Index $= 3$

$Q_3 = \frac{75}{100} \times 21 = 15.75$ Index $= \frac{8+7}{2} = 7.5$

Lower Fence $= 3 - (1.5)(4.5)$

$= -3.65$

Higher Fence $= 7.5 + (1.5)(4.5)$

$= 14.25$

* In the given list there is no values before $-3.65$, so ~~no need~~ there is no outliers at lower fence.

But in higher fence we have outlier which is greater than 14.25.

The outlier is 27. So we have to remove 27 from the data set.

$\{1,2,2,2,3,3,3,4,5,5,5,6,6,6,6,7,8,8,9,\boxed{\cancel{27}}\}$

5 number summary used to plot the box plot here

1. minimum = 1
2. $Q_1 = 3$
3. median = 5
4. $Q_3 = 7.5$
5. maximum = 9

plot all these values