Statistics :-

1. Histogram

Let say we have data set as

Ages = {10, 12, 14, 18, 20, 21, 25, 31, 35, 36, 37, 40, 44, 50, 56, 58, 60}

To create histogram, for :-

i) Sort the numbers

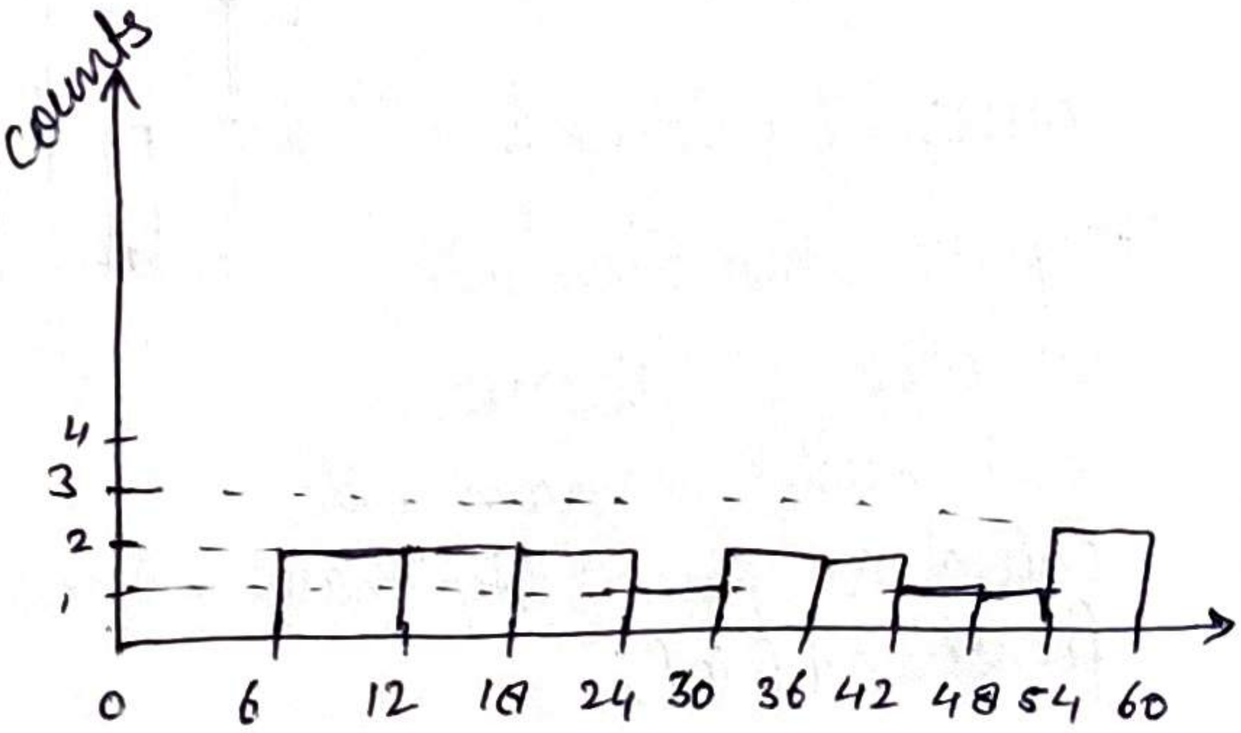
ii) Bins - no of bins to be created

iii) Size of bin - size of single bin is calculated.

If we want 10 bins then size of each bin will be $\frac{\max(Ages)}{\text{no of bins}} = \frac{60}{10} = 6$

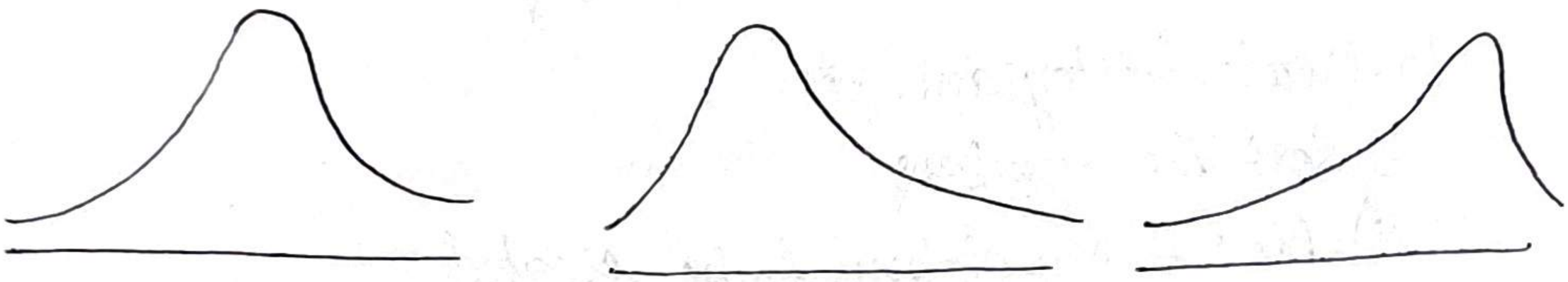
This means we need 6 as bin size.

Now histogram is a graph in which on the x-axis we have bins and on y-axis we have count / frequency



- Once we smoothen the histogram we see ~~some~~ something unique that is called as "probability density function". It give insight of data distribution.

- Few example of data distribution.



This smoothning is done using Kernal ~~Dens~~ Density ~~Facade~~ estimation

- If we have discrete data set as

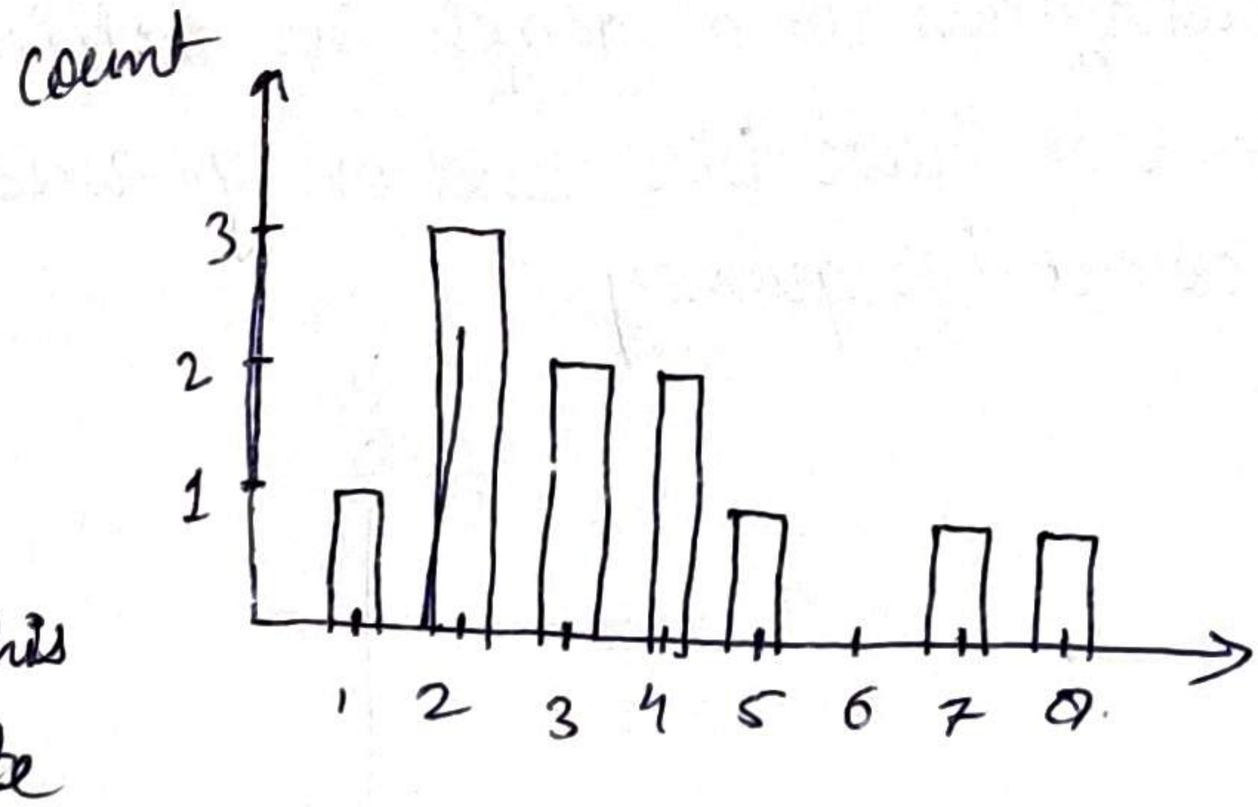
no of bank account = [1, 2, 4, 7, 5, 2, 3, 2, 3, 4, 8,]

~~Histogram~~

{discrete / continous}

interval should be 1

- If we smoothen this we will get discrete continous histogram it is called as probability mass function. (pmf)



2. Measure of central tendency

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

To understand this we have three terms-

i) Mean : ~~ii) Median~~ ~~iii) Mode~~

Eg if we have $x = \{1, 2, 3, 4, 5\}$

then mean is $= \frac{1+2+3+4+5}{5} = 3$

Population (N)	Sample (n)
mean, $\mu = \sum_{i=1}^{N} \frac{x_i}{N}$	mean $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$

- always $N > n$,
  but the same cannot be concluded for $\bar{x}$ & μ

- Application in feature Engineering :-

Age	Salary	Family size
—	—	—
—	null	—
null	—	—
—	—	—
—	—	null
—	null	—

- We cannot remove the rows with missing value as it will cause loss of information.

- So the missing values can be replaced with mean of entire column.

If we have outliers present in our dataset then it will significantly change the mean value. So it is not advisable to remove the missing value with mean. Instead we use something called as median.

ii) Median :-

Median is middle of dataset in arranged in an order.

To find the median follow following steps.

a) Sort the number

b) If no of element is even then we find avg of central element of is median

If no of element is odd then median is central element

Eg. {0, 1, 2, 3, 4, 5, 6, 7, 8, 100, 120}

mean = 23.27 } with outliers
median = 5

mean = 4 } without outliers
median = 4

Conclusion: When we have outliers our mean is significantly effected but the median has very small change.

So, in case of outliers median is better metrics to rely on.

4. Measure of dispersion :-

i) Variance (talks about spread of data)

Age = { 1, 1, 5, 5 } ← distribution 1

$\mu = \frac{1+1+5+5}{4} = 3$

spread of data is more

Age 2 = { 2, 2, 4, 4 } ← distribution 2.

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

spread of data is less

This spread is calculated via variance & Std. deviation will tell how much distance a point is from center in terms of standard deviation

For population data.

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

For sample data.

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

→ not n rather n-1 is to underestimate the true population variance.

This is also called Basel correction

Eg If in a population mean is 2050 but not known. Random sample is choosen to be 2051, 2053, 2050, 2055, 2051

$\bar{x} = 2052$

Actual variance - ie of population = $\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n}$

$= \frac{(2051-2050)^2 + (2053-2050)^2 + \cdots + (2051-2050)^2}{5} = 7.2.$

Since the actual mean is not known so calculating variance from sample data. ie

$s^2 = \frac{(2051-2052)^2 + (2053-2052)^2 + \cdots + (2051-2052)^2}{5} = 3.2.$

If we use (n-1) then $s^2 = 4$ which is closer to the actual variance 7.2.

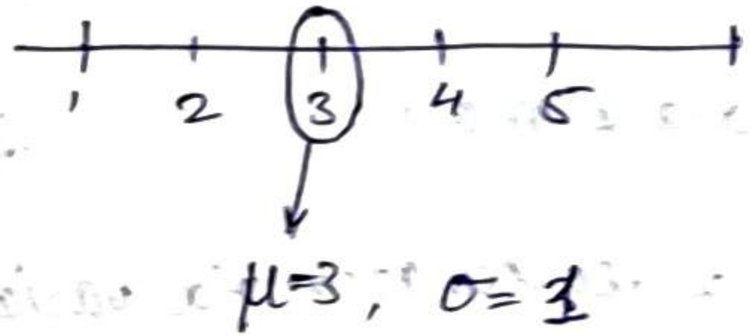
Variance will give how spread the data is from mean.

ii) Standard Deviation.

This shows how far is data from the mean.

Eg.

Then we will say 4 is one std. deviation away from the mean.



Sample std. deviation

$S = \sqrt{\text{Sample Variance}}$

$\sigma = \sqrt{\text{population variance}}$

6. Percentiles and Quartiles:-

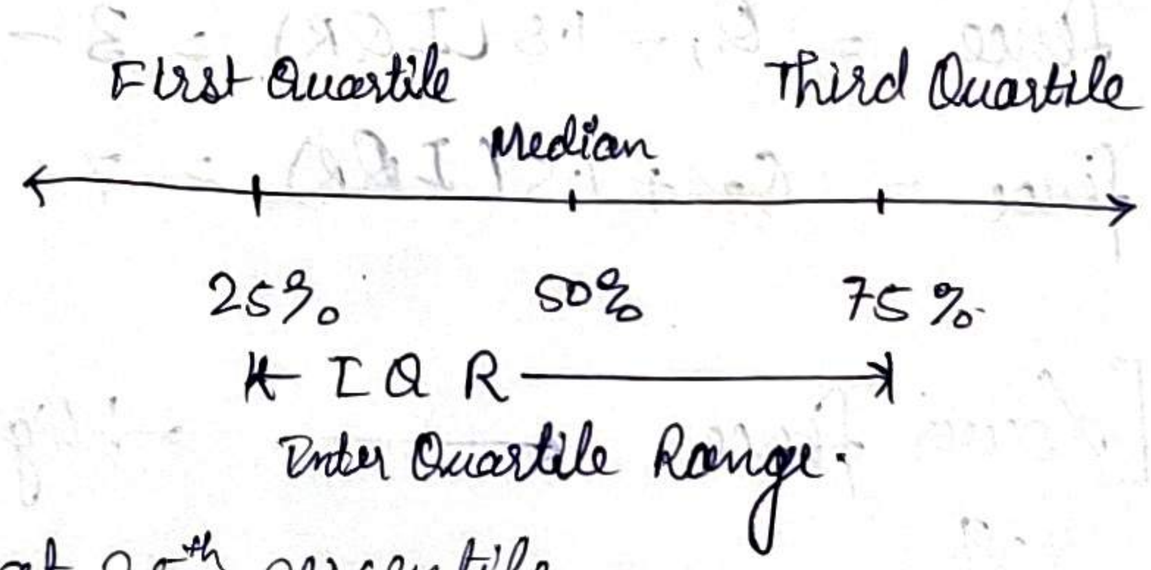
- Percentage = $\frac{\text{no of value of intrest}}{\text{no of observation}} \times 100$

- Percentile – It is a value below which certain percentage of observations lies in a data set.

$$\left.\begin{matrix}\text{Percentile of } x \\ \text{in data set}\end{matrix}\right\} = \frac{\text{no of values below } x}{\text{Total no of values}} \times 100$$

Percentile denotes how many data points are less than 'x'

- Quartile :-



Ques. What value exist at 25th percentile -

→ data = {2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12}

index value = $\frac{\text{Percentile}}{100} \times (n+1) = \frac{25}{100} \times 21 = \frac{5}{4} \approx 5$ ie 5th index

so the value is - 5

7. 5 number summary :- (used to remove outliers)

data set - {1, 2, 2, 2, [3], 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 8, 8, 9, 9, 27}

5th index

~~5 minimum value~~ :

To find outlier we need to create fence.

[Lower Fence ⟷ Higher Fence]

Lower fence = $Q_1 - 1.5 * IQR$

Higher fence = $Q_3 + 1.5 \times IQR$.

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

$= 5^{th}$ index

5^{th} index has value 3.

Q_3 (75% percentile) = $\frac{75}{100} \times (20 \cdot 1.9 + 1) = 15^{th}$ index

15^{th} index has 7

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

Lower fence = $Q_1 - 1.5(IQR) = 3 - 1.5(4) = -3$

Higher fence = $Q_3 + 1.5(IQR) = 7 + 1.5 \times 4 = 13$

[Lower fence ⟷ Higher fence].

-3 13

So 27 is outliers.

Data set after removing ~~quantile~~ outliers.

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

min value = 1,

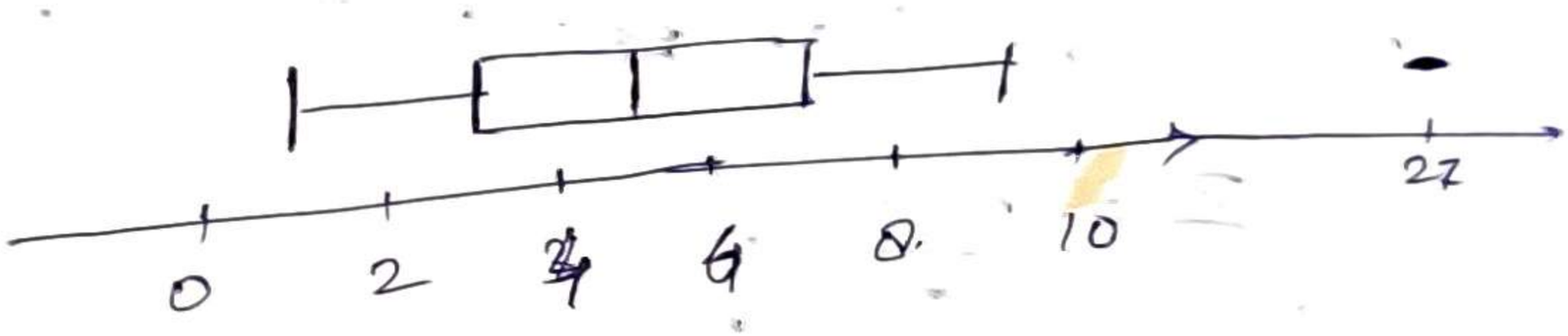
$Q_1 = 3$

Median = 5

$Q_3 = 7$

max = 9.

To create Box plot :-



2. Histogram: