2nd Day of Statistics

Agenda

1. Histogram
2. Measure of Central Tendency.
3. Measure of Dispersion
4. Percentile and Quartiles
5. 5 Number Summary (Box Plot)

1. Histogram :→

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

(i) Sort the nos. (if not sorted)

(ii) Bins → No. of group.

(iii) Bins Size :- Size of Bins

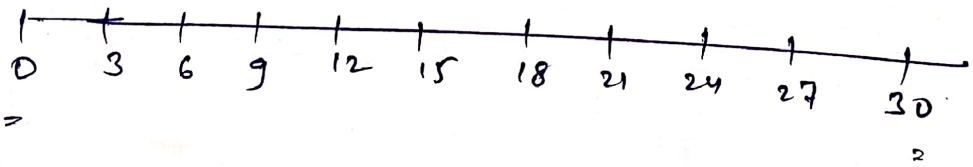
= [0, 10, 20, 25, 30, 35, 40]

min — if starts from min

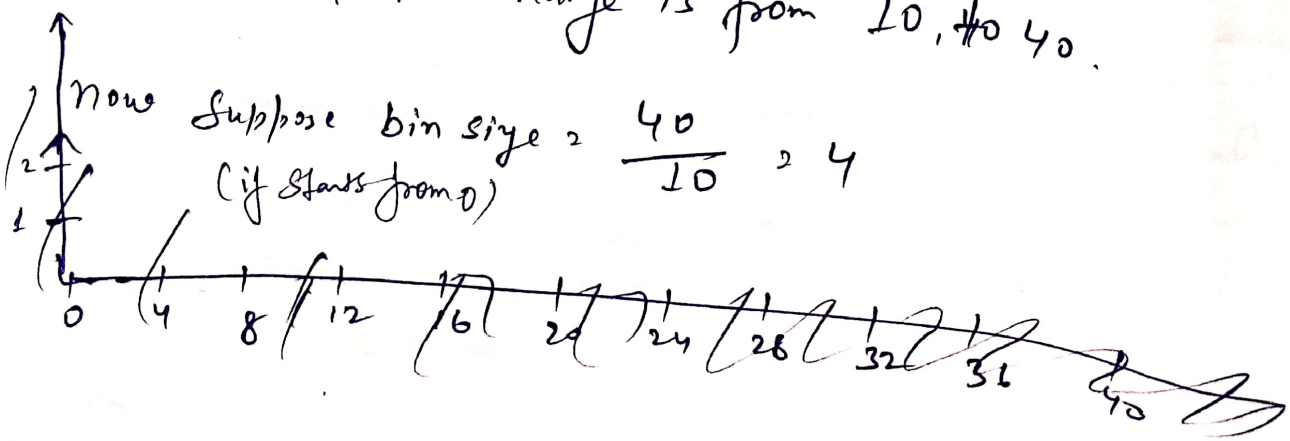
if starts from 0,

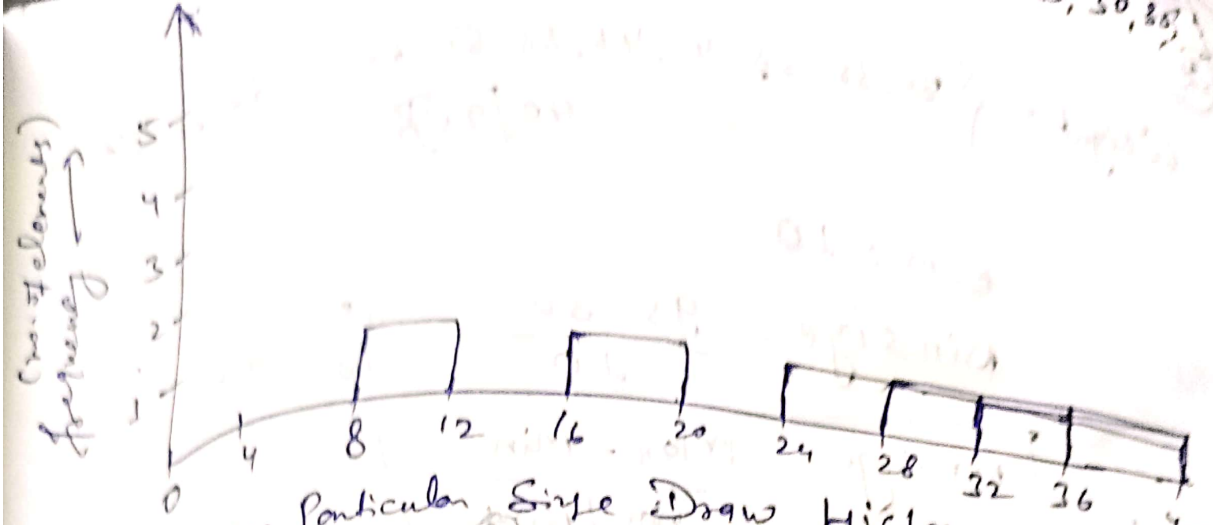
bins = 10 (divide it into 10 groups)

bin size = $\frac{40-0}{10} = \frac{40}{10} = 4$



But in our Question Range is from 10, to 40.





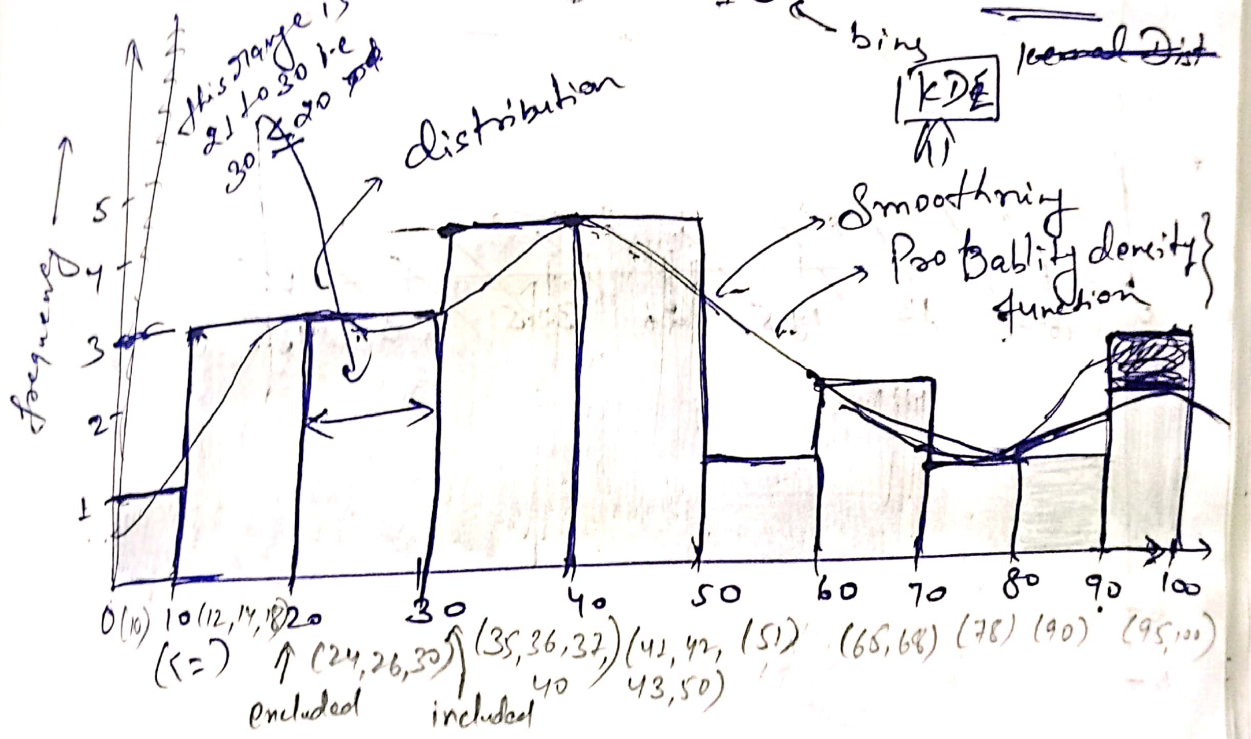
now for this particular Size Draw Histogram.

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

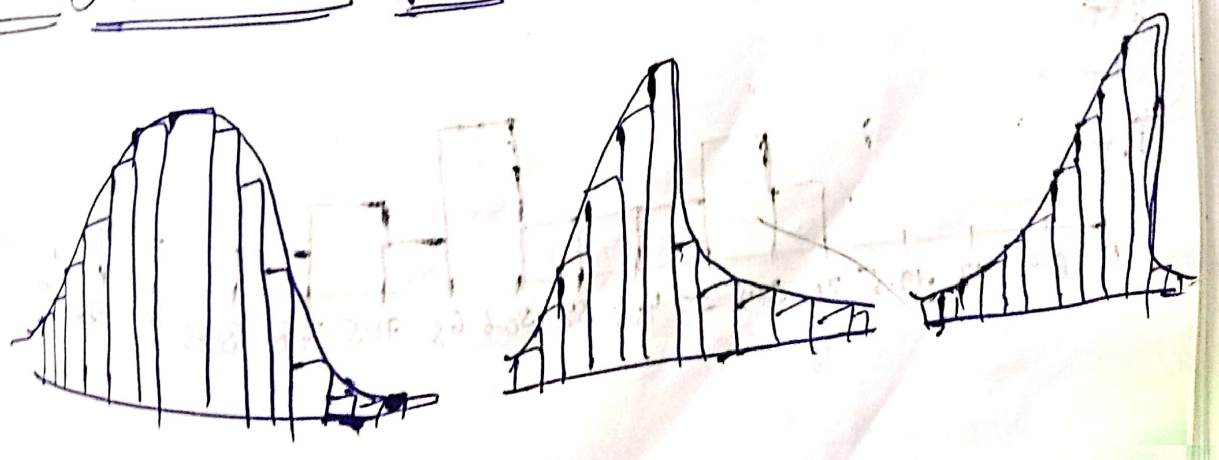
bins: 10 bin size = $\frac{100 - \text{min}}{10} = 10$.

bins

KDE



Data Distribution Example 6 →



(2)

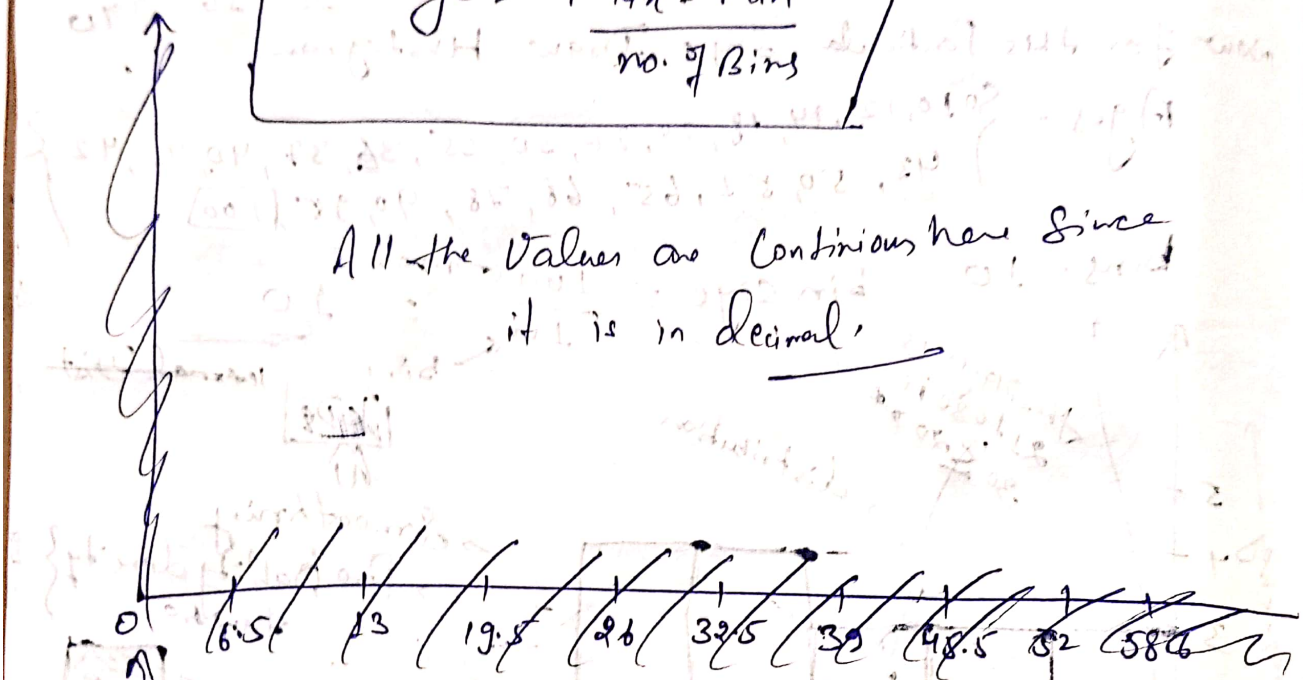
Weight = {30, 35, 38, 42, 46, 56, 59, 62, 63, 68, 75, 77, 80, 90, 95}

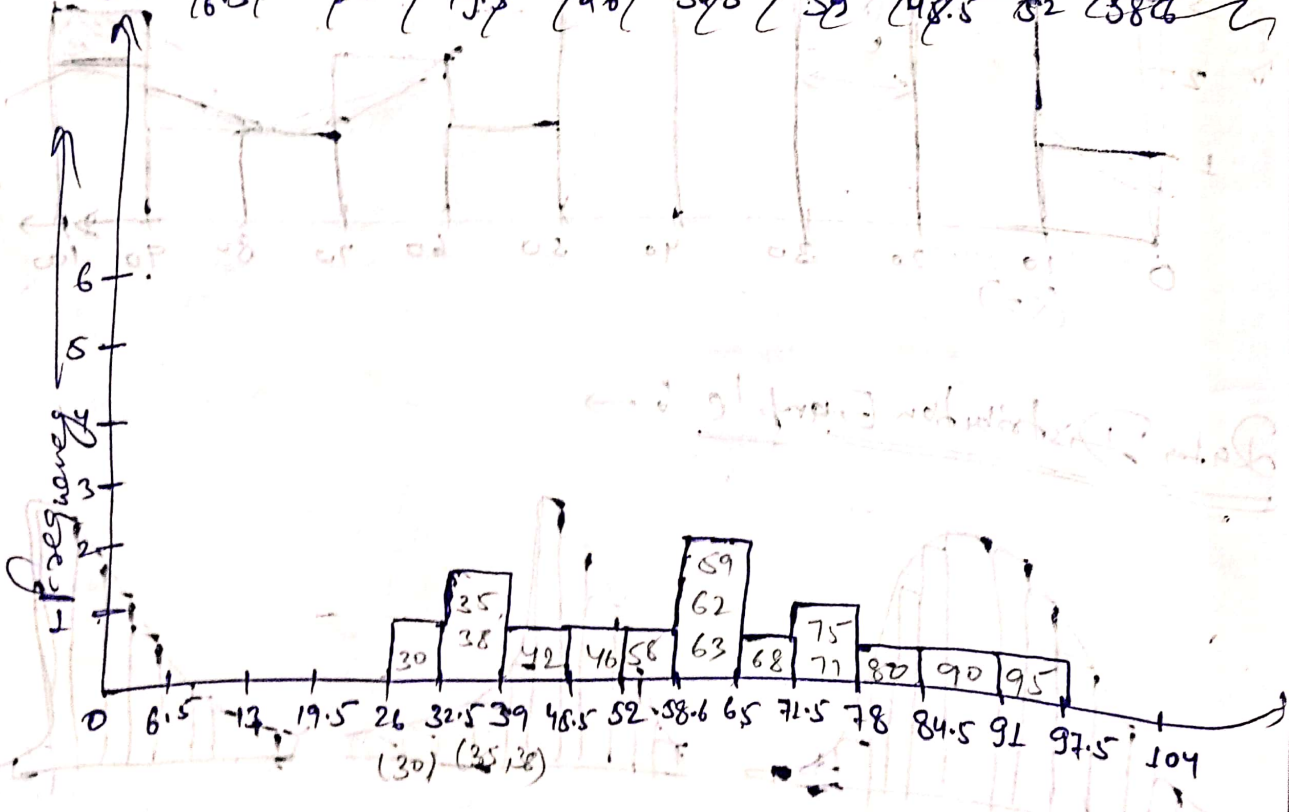
Bins = 10

Bin size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$

Bin size = $\frac{\text{Max} - \text{Min}}{\text{no. of Bins}}$

All the values are continious here, since it is in decimal.

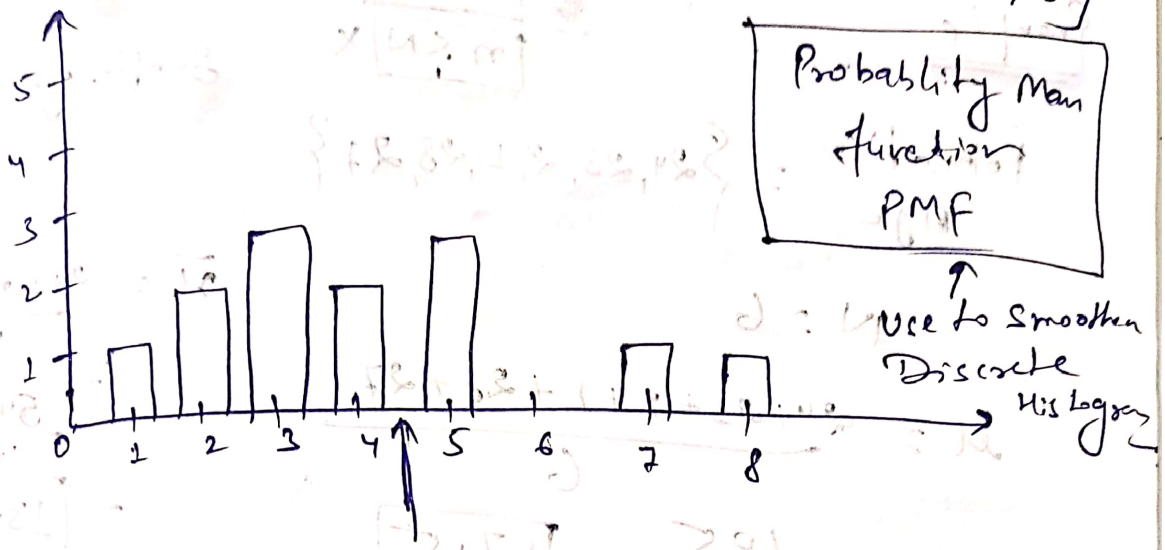




N.V.R

Discrete Continious :- →

No. of Bank a/c's = [2, 3, 5, 1, 8, 3, 7, 8, 3, 2, 4, 5]



PDF (Probability Density function) :→ Use this method to Smoothen the continious Histogram

↳ for continious variable

(38)

PMF (Probability Mass function) :→ Use this method to Smoothen Discrete Histogram.

↳ for Discrete Variable

(3) Measure of Central Tendency (CT) :→

(i) Mean
(ii) Median
(iii) Mode

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

(i) Mean $x = \{1, 2, 3, 4, 5\}$

$$\text{Avg / Mean} = \frac{1+2+3+4+5}{5} = 15/5 = 3$$

Population (N) $N \geq n$ Sample (n)

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

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

Example

$N \geq n$ ✓

$n \leq N$ ✗

$n = 4$

Sample Age = {24, 2, 1, 28}

Population Age = {24, 23, 2, 1, 28, 27}

$N = 6$

$$\mu = \frac{24+23+2+1+28+27}{6}$$

$$= \frac{105}{6} = 17.5$$

$$\bar{x} = \frac{24+2+1+28}{4}$$

$$= \frac{55}{4} = 13.5$$

$\mu \geq \bar{x}$

$\mu \leq \bar{x}$

Both is Possible

in Numpy:

np.nan ⇐ NULL Values

⇒ Practical Application :→

Age	Salary	Family size
—	—	—
—	—	—
NAN	—	— delete
—	—	—
—	NAN	—
—	—	NAN
—	NAN	—
NAN	—	—

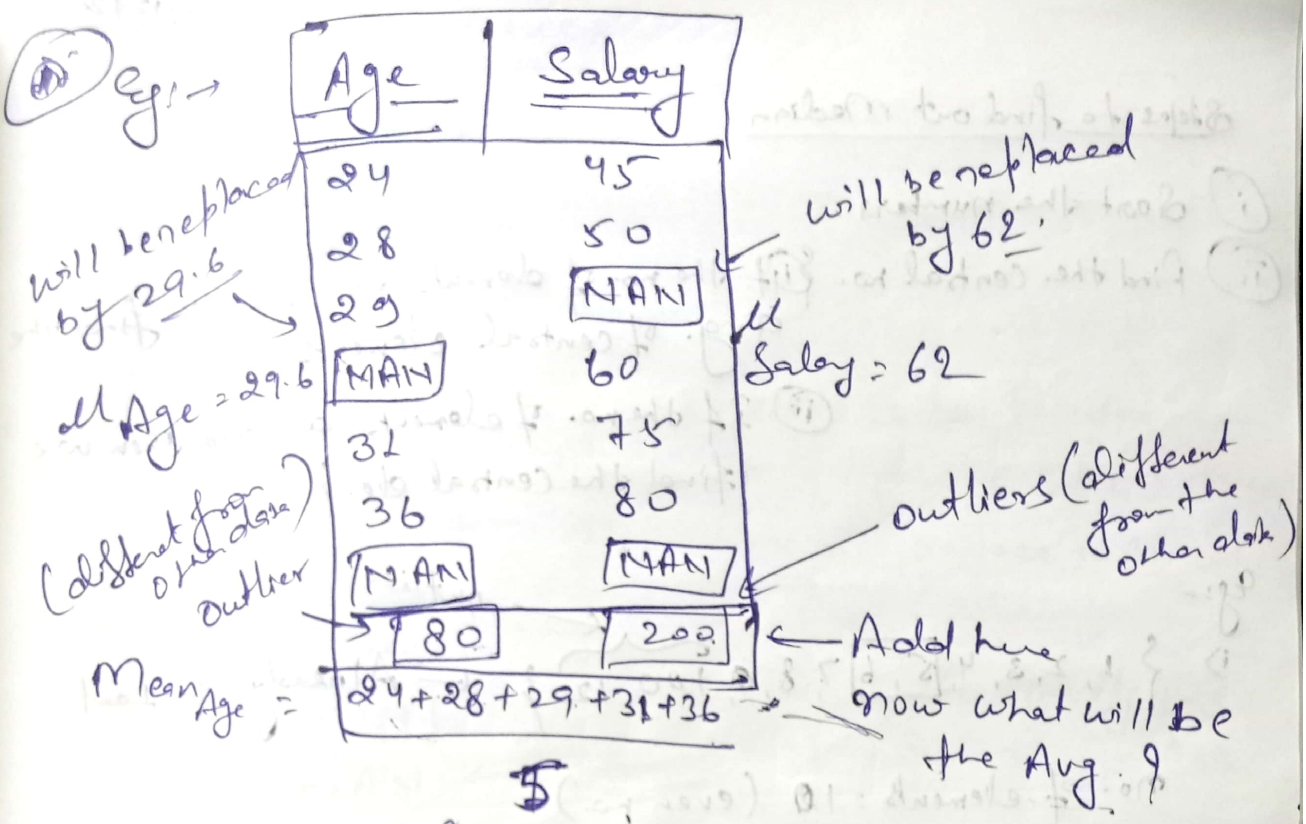
loss of info.

Rather than deleting an entire Row we can find the mean so that we won't lose any info.

At last mean is Center of Data.

So wherever there is NAN value just

(i) Eg:→



$\mu Age = 29.6$

$\mu Salary = 62$

$\text{Mean}_{Age} = \frac{24+28+29+31+36}{5}$

$= \frac{148}{5} = 29.6$

$\mu Age = \frac{148+80}{6}$

$= 38$

$\mu Salary = \frac{45+50+60+75+80}{5}$

$= \frac{310}{5} = 62$

$\mu Salary = \frac{310+200}{6}$

$= 85$

⇒ Because of outliers there is change is mean of both Age and Salary. Hence to prevent this we use sth is called as Median.

(ii) Median :→

By adding one outlier ⇒ there is huge diff in mean

Eg:- {1, 2, 3, 4, 5} {1, 2, 3, 4, 5, 200}

outlier

$\bar{x} = 3$ —Big change→ $= \frac{115}{6} = 19.15$

Hence when there is big change like this mean it is bez of outlier

Steps to find out Median

(i) Sort the numbers.

(ii) Find the central no. {(i) If the no. of elements are even we find the avg. of central elements

(ii) If the no. of elements are odd then we find the central ele.}

Eg:-

i) {1, 2, 3, 4, 5, 6, 7, 8, 9, 100, 120} ← Already sorted (outlier)

no. of elements : 10 (even no.)

Hence find the Avg. of central ele. which is

i.e $\frac{5+6}{2} = 11/2 = 5.5$ = median

Hence median = 5

mean = $\frac{1+2+3+4+5+6+7+8+100+120}{10} = \frac{256}{10}$

mean = 25.6

Hence whenever there is an outlier we use median.

and if there is no outlier we use mean

No outliers = mean

Outliers = median

(3) Mode :→ {Most frequent occuring elements}

Eg:- (i) {1, 2, 2, 3, 3, 3, 4, 5}

Mode : 3

(ii) {1, 2, 2, 2, 3, 3, 3, 4, 5}

Mode = 2, 3

Practical appl of Mode:

Q:- Dataset { Categorical Variable }

Type of flower

Lily
Sunflower
Rose
NAN ← Rose
Rose
Sunflower
Rose
NAN ← Rose

Since Rose has Repeated max. no. of times i.e 3 times Hence we will replace NAN with Rose.

Hence Mode mostly we use with Categorical Var.

Q: If there are two values in Mode then from which value shall I replace NAN with.

Sol:
Q:- Mode = { Sunflower, Rose }

You can replace with anyone of them it's your choice.

Q: If NAN is occured more than mode what to do?

Ans: Delete that particular column. If 90% of Data is having NAN value you have to delete that particular column.

(3) Measure of Dispersion :→
(i) Variance (σ^2) → Spread of Data.
(ii) Standard Deviation (σ)

(i) Variance (σ^2): → Talks about Spread of Data.

Population Variance (σ^2)

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

Sample Variance (s^2)

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

Distance from the mean from every Data Points

Why it is (n-1) Assignment

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

$\mu = 3$

Q. Suppose we have two Dist^r

(i) $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ (ii) $\{1, 2, 3, 4, 50, 60, 70, 80, 90, 100\}$

↓ Find Variance ↓ Find Variance

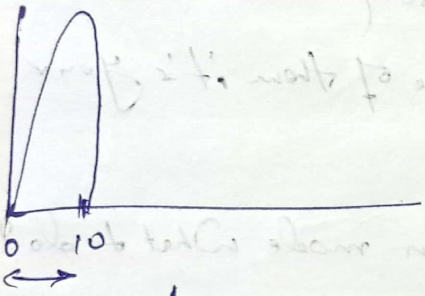
Now which Variance will be higher (i) or (ii) ?

Sol^n

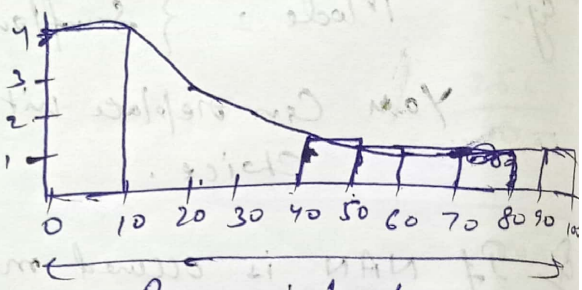
~~Second (ii) → Variance of (ii)~~

~~Variance of (i)~~

Variance of (i) < Variance of (ii)



Range is less



Range is high

Hence Variance of (i) < Variance of (ii)

Let's Calculate.

$\mu_{(i)} = \frac{1+2+3+4+5+6+7+8+9+10}{10}$

$= \frac{55}{10} = 5.5$

$\mu_{(ii)} = \frac{1+2+3+4+50+60+70+100}{8}$

$= \frac{290}{8} = 35$

$\sigma^2 = \frac{(1-5.5)^2 + (2-5.5)^2 + (3-5.5)^2 + (4-5.5)^2 + (5-5.5)^2 + (6-5.5)^2 + (7-5.5)^2 + (8-5.5)^2 + (9-5.5)^2 + (10-5.5)^2}{10}$

$= \frac{(4.5)^2 + (3.5)^2 + (2.5)^2 + (1.5)^2 + (0.5)^2 + (0.5)^2 + (1.5)^2 + (2.5)^2 + (3.5)^2 + (4.5)^2}{10} = 2$

$$= \frac{20.25 + 12.25 + 6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25 + 12.25 + 20.25}{10}$$

$$= \frac{82.5}{10} = 8.25$$

(ii) $\sigma^2 = \frac{(1-35)^2 + (2-35)^2 + (3-35)^2 + (4-35)^2 + (50-35)^2 + (60-35)^2 + (70-35)^2 + (100-35)^2}{8}$

$$= \frac{(-34)^2 + (-33)^2 + (-32)^2 + (-31)^2 + (15)^2 + (25)^2 + (35)^2 + (65)^2}{8}$$

$$= \frac{1156 + 1089 + 1024 + 961 + 225 + 625 + 1225 + 4225}{8}$$

$$= 10530/8 = 1316.25$$

$\sigma^2_{(i)} < \sigma^2_{(ii)}$

which is $8.25 < 1316.25$

Example (ii)

(i) $\{1, 2, 3, 4, 5\}$

$\mu = 3$

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

$= \frac{(-2)^2 + (-1)^2 + (0)^2 + (-1)^2 + (2)^2}{5}$

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

(ii) $\{1, 2, 3, 4, 5, 6, 80\}$

$\mu = 14.4$

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

$= (-13.4)^2 +$

$$= (13.4)^2 + (12.4)^2 + (-11.4)^2 + (-10.4)^2 + (-9.5)^2 + (8.5)^2 + (65.6)^2$$

$$7$$

$$= 179.56 + 153.76 + 129.96 + 108.25 + 90.25 + 72.25 + 4303.36$$

$$7$$

$$= \frac{5037.3}{7} \approx 719.61$$

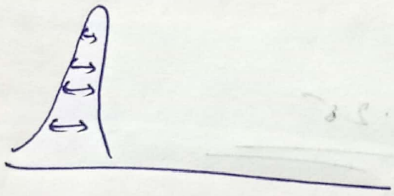
Now we can see:

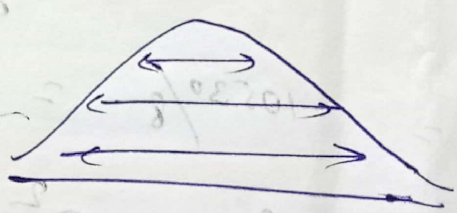
$$\sigma^2 < 719.61$$

i.e. $\sigma^2_{(i)} < \sigma^2_{(ii)}$

So can we conclude that as Variance keeps on increasing data set spread also increases.

i.e Variance ↑↑ Spread ↑↑





$\sigma^2_{(i)} <$ $\sigma^2_{(ii)}$

Lower spread ↓

Lower Variance ↓

Higher spread (width)

Higher Variance

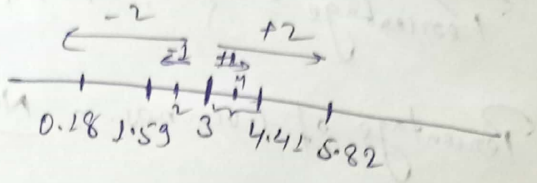
(ii) Standard Deviation ($\sqrt{\sigma^2}$) :→

Eg: {1, 2, 3, 4, 5}

$\mu = 3$

$\sigma^2 = 2$

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



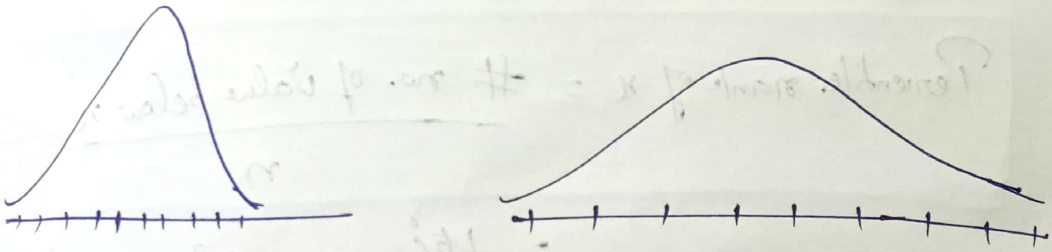
It tells that how many standard deviation away from the mean (each distr) it falls. whether +1, -1, +2 or -2.

Q. 4 is falling how many standard deviation away from the mean?

Ans: within the +1 Range.

Similarly 2 is falling -1 away from the mean.

Eg:-



σ^2 < σ^2 (Variance)

$\sqrt{\sigma^2}$ < $\sqrt{\sigma^2}$ (S.D)

Variance can lie in both right or left. Hence we want to know which side it lie right side or left side for that we use Standard deviation.

(4) Percentile & Quantiles :→

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

Percentage of Even nos. = $\frac{\text{No. of Even nos.}}{\text{Total no. of nos.}} = \frac{4}{8} = 0.5$

= 50%

⇒ Percentiles :→ GATE, CAT, ILETS, SAT, GRE, JGE, NPTT ⇒ Ra

Defⁿ :→ A percentile is a value below which certain percent of observation lie.

99 Percentile = It means that the person has got better marks than 99% of entire student.

Index :- 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

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

What is the percentile rank of 10

$$\text{Percentile rank of } x = \frac{\text{\# no. of value below } x}{n}$$

$$= \frac{16}{20} = 0.8 = 80 \text{ percentile}$$

All the elements should be in ascending order if not just sort it.

Q: Can 2 people get same percentile ?

Ans: YES.

Percentile of 8 = $\frac{9}{20} = 0.45 = 45$ Percentile.

this means that 45% of people are less than 8.

for all 8 it will be same i.e 45 Percentile.

Percentile of 6 = 7/20 = 0.35 = 35 percentile

Percentile of 9 = 14/20 = 70 Percentile

(4) What is the Value that exists at 25 percentile?

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

for even 'n' = ~~for odd n~~ multiply with (n+1) ~~multiply with n~~

not same

$$= \frac{25}{100} \times 20^{5} = 5^{th} \text{ index.}$$

Hence O/P = 5 ← 5th index Value is 5.

(4) What is Value ~~at~~ that exists at 95 percentile?

~~Value = 95/100 × 20~~

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

~~19 index~~

Hence O/P is 12. ← 19th index

(5) 5 number Summary :→

(i) Minimum
(ii) First Quartile (25 percentile) (Q1)
(iii) Median
(iv) Third Quartile (75 percentile) (Q3)
(v) Maximum

All these we use to remove the outliers.

⇓

create Box Plot

Data set : { 1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27 }

(1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19)

27 ↓ Outlier

How to know 27 is an outlier for that we are creating a fence called

[Lower fence ⟷ Higher fence]

many experiments they found out after ~~different~~

Lower fence = $Q1 - 1.5(IQR)$ where, $IQR = Q_3 - Q_1$

(It specifies 1.5 ~~means~~ deviation from the mean)

Inter Quartile Range

Higher fence = $Q_3 + 1.5(IQR)$

$Q1 = \frac{25}{100} \times (n+1) = \frac{25}{100} \times 21 = \frac{525}{100} = 5.25^{th}$ index

Since we don't have 5.25 index so we will take the avg. of 5th & 6th index i.e

$Q_1 = \frac{3+3}{2} = 3$ ⟸

$Q_3 = \frac{75}{100} \times 21 = \frac{63}{4} = 15.75^{th}$ index

Hence take the Avg. of 15th & 16th index i.e

$Q_3 = \frac{7+8}{2} = 15/2 = 7.5$ ⟸

Lower fence = $Q1 - 1.5(IQR)$

$= 3 - 1.5(7.5 - 3)$

$= 3 - 1.5 \times 4.5$

$= -3.75$

Higher fence $= 7.5 + 1.5(7.5-3)$

$= 7.5 + 1.5 \times 4.5$

$= 7.5 + 6.75$

$= 14.25$

Hence all the values be in the range of

-3.75 to 14.25

Hence 27 doesn't lie in this range Hence it is an outlier.

Hence we can remove the outlier.

Now,

Middle element

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

remove outlier

Now find

(i) Minimum = 1

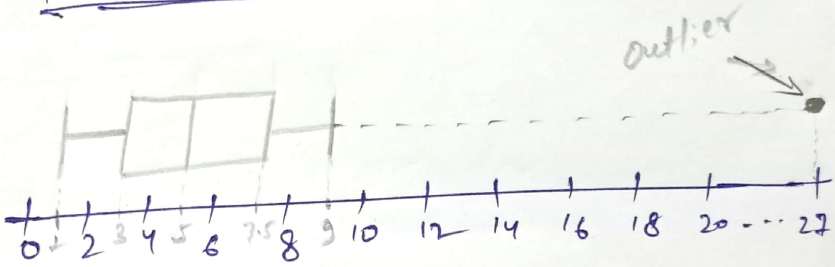
(ii) $Q_1 = 3$

(iii) Median = 5

(iv) $Q_3 = 7.5$

(v) Maximum = 9

Now create Box Plot.



We create Box Plot to Treat outliers.