Day-2

Agenda

1. Histograms
2. measure of central tendency
3. measure of dispersion
4. Percentiles and quartiles
5. 5 numbers summary (Box plot)

Histogram

→ can have continous variables as input

set 1

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 55, 58, 78, 90, 95, 100}

Steps to create

→ sort the numbers
→ Define no. of Bins (no. of groups) (It depends on user's choice, they can take any value)
→ Size of Bins

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

min = 10
max = 40
bins = 10
Bin size = $\frac{40-10}{10} = \frac{30}{10} = 3$

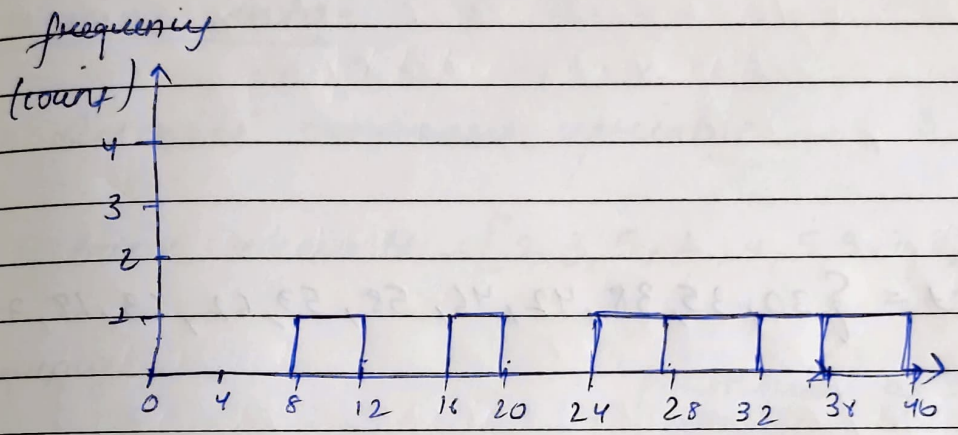
0 3 6 9 12 15 18 21 24 27 30 (*)

But this doesn't include the maxm. value

10 13 16 19 22 25 28 31 34 37 40

So we calculate it like

$$\frac{40}{10} = 4 \qquad \left(\text{Bin size} = \frac{\text{maxm. value}}{\text{no. of bins}}\right)$$



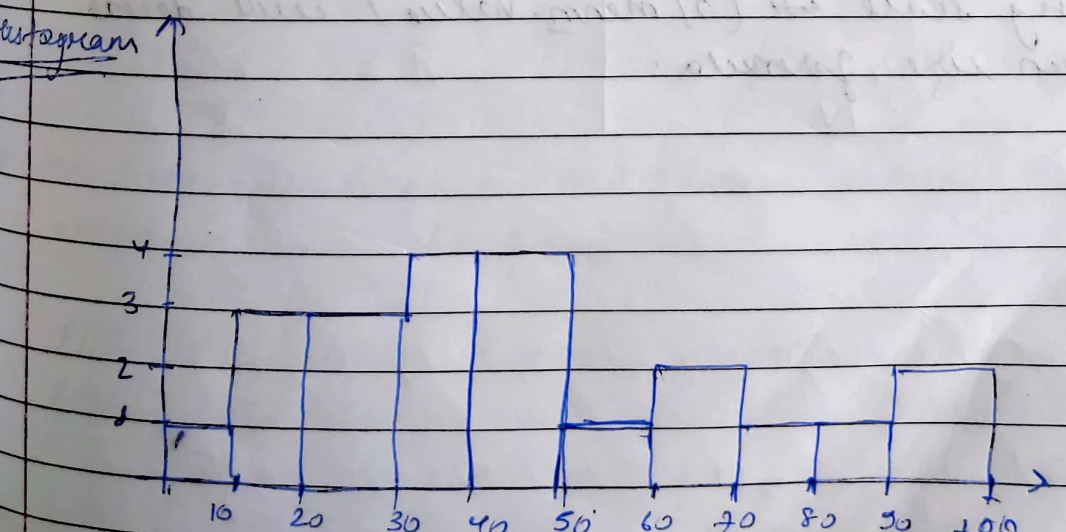
for set (A)

$$\text{maxm value} = \frac{100}{10} = 10$$

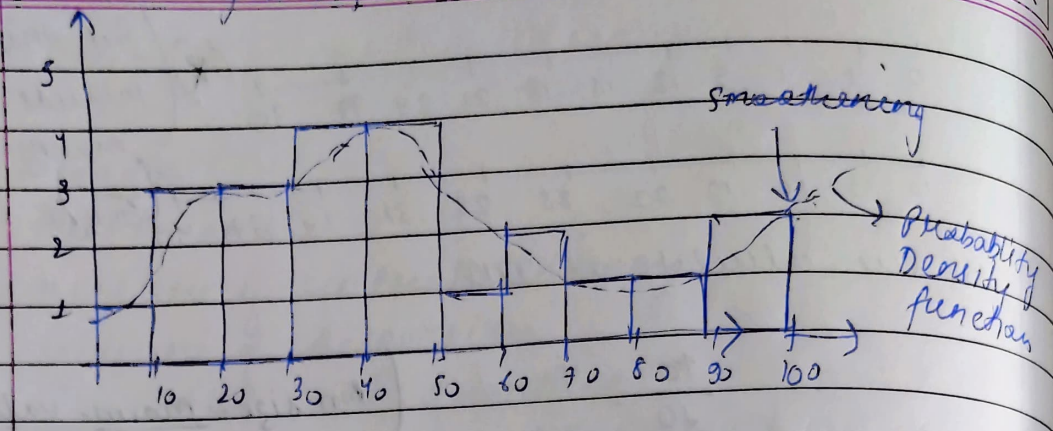
10 ← Bin size

If Bin size we want as 20, then no. of Bins will be

$$\frac{100}{20} = 5.$$



Histogram for continuous variable



Set (B) Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 80, 97}

we want Bins = 10

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

Here scale will start from 30

→ Starting scale (o/ mean value) will decide the bin size formula.

Page No.:
Date:

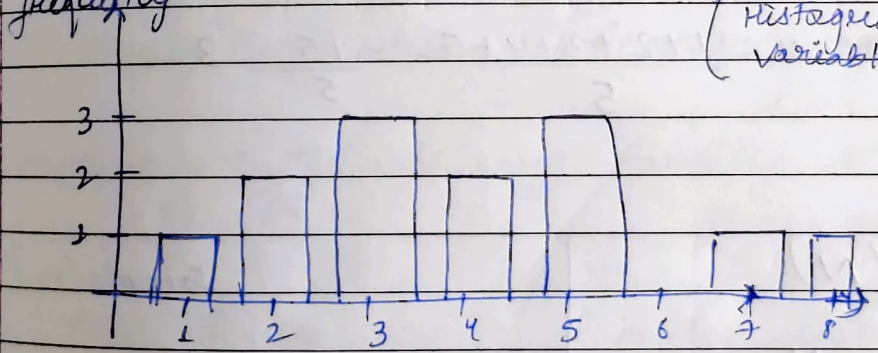
Measure of central Tendency

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

Histogram contd.

→ For Discrete ~~continuous~~ variable

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



(Histogram for discrete variables)

To smoothen we will use → Probability Mass function (PMF)

* In interviews, questions about variable will be asked on basis of graph

PDF : Probability Density Function → Continuous Variable
PMF : Probability Mass Function → Discrete variable

Measure of central tendency

1. Mean
2. Median
3. Mode

Defn: A measure of central tendency (CT) is a single value that attempts to describe a set of data identifying the central position.

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

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

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

Example:

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

$$\text{Population mean } (\mu) = \frac{24+23+2+1+28+27}{6} = \frac{105}{6} = 17.5$$

Page No.:
Date:

Sample ~~Age~~ Age(n) = {24, 2, 1, 27}

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

$= \frac{54}{4}$

$\bar{x} = 13.5$

$N > n$	✓
$n > N$	✗
$\mu \geq \bar{x}$	✓
$\bar{x} \geq \mu$	✓

Practical Application (Feature Engineering)

Features → Age, Salary, Family Size

Age	Salary	Family Size
—	—	—
—	—	—
NAN	—	—
—	—	—
—	NAN	—
—	—	NAN
—	NAN	—
NAN	—	—

— If we drop any row with NAN value then there will be loss of Info

we can replace NaN with mean values of respective features

	Age	salary	
	24	45	
	28	50	
	29	NAN	← 62
29.6 →	NAN	60	
	31	75	
	36	80	
29.1 →	NAN	NAN	← 62

Age mean = 29.6
Salary " = 62

Suppose we are adding a new row with values as

80 200

then mean for
Age = 38
Salary = 85

values 80, 200 are outliers since they are outside the range values

$\{1,2,3,4,5\}$ $\bar{x} = 3$
$\{1,2,3,4,5,100\}$ $\bar{x} = 19.16$

↑ outliers

Page No.:
Date:

Median

1. Sort the numbers
2. Find the central numbers

→ No. of elements → even → Find the average of central elements

→ No of elements → Find the central element

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

median = $\frac{5+6}{2}$ = 5.5 , mean = 25.1

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

median = 5 , mean = 25.6

* If no outliers then use mean
If outliers are present then use median

Mode

→ Most frequent occuring element

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

mode = 3

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

mode = 2, 3

(In older version of Python it would return 2 & 3. But in newer versions it returns an error.)

Types of flower

Lily
Sunflowers
Rose
NAN ← Rose
Rose
Sunflowers
Rose
NAN ← Rose

Replacing data with m

Page No.:
Date:

Measure of dispersion

(i) Variance (σ^2) ← spread of Data
(ii) standard deviation (σ)

Variance

(i) Population variance (σ^2)
(ii) sample variance (s^2)

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

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

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

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

Population Variance

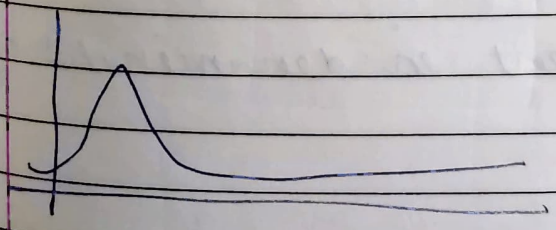
{1,2,3,4,5,6,7,8,9,10}

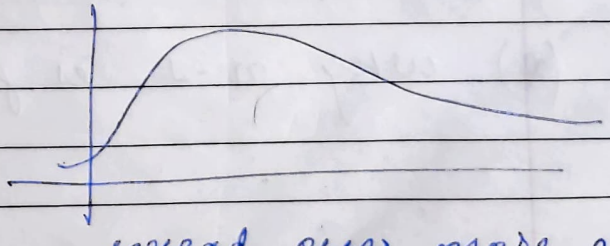
variance = less

{1,2,3,4,50,60,70,100}

variance = more

Distribution





spread over more area

Page No.:
Date:

{ 1, 2, 3, 4, 5 }

$\mu = 3$

{1, 2, 3, 4, 5, 6,

$\mu = 14.4$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

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

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

$$\sigma^2 = \frac{(1-14.4)^2 + (2-14.4)^2 + \div --}{7}$$

variance

variance ↑ spread ↑
(σ^2)

Assignment

Assignment :-

(Q) Why $n-1$ is present in denominator?

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

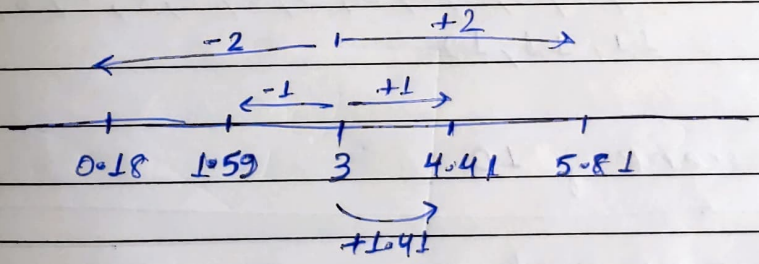
$\{1, 2, 3, 4, 5\}$

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

$$\sigma^2 = \frac{10}{5} = 2$$

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

→ To find out how much away a number is present from the mean



Q) where does 4 fall

Ans → 1 standard deviation towards right

variance ↑ ⇒ Standard deviation ↑

Percentiles and Quartiles

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

% age of even numbers = 50%

Defn: A percentile is a value below which a certain percentage of observations lie

99 per percentile = it means the person has better marks than 99% of the entire students

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

(Q) Percentile rank of 10

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

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

→ Dataset should be sorted while calculation

→

(a) what is the value that exists at 25 percentile

$$\text{value} = \frac{\text{Percentile}}{100} \times (n+1) \qquad (\text{Even } n)$$

$$= \frac{25}{100} \times (19+1) = 5^{th} \text{ index}$$

5

$n \rightarrow$ Total no. of values in the list

2, 2, 3, 4, 5, [5], 5, 6, 7, ...

index: 0 1 2 3 4 [5]

* index will start from 0

$$\text{value} = \frac{\text{Percentile}}{100} \times n \qquad (n - \text{odd})$$

value at 95 percentile

$$\frac{95}{100}$$

5 number summary

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

Remove the outliers

{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}
0 1 2 3 4 5 6

[lowies fence ⟷ highes fence]

lowes fence = Q1 − 1.5(IQR)
highes fence = Q3 + 1.5(IQR)

~~IQR~~

IQR = Q3 − Q1
↓
(Inter quartile Range)

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

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

Since 5.25 index doesn't exist so avg. of value

at index 586 will be taken,

$\therefore \frac{3+3}{2} = 3$

$\therefore Q1 = 3$

$Q3 = \frac{75}{100} \times 21 = 15.75$ index

$\frac{8+7}{2} = 7.5$

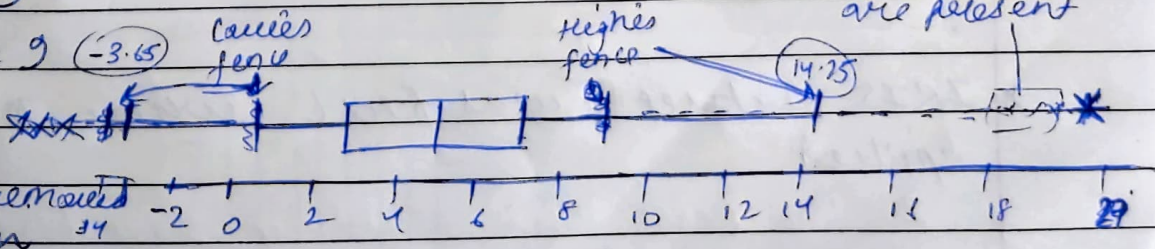
$IQR = Q3 - Q1 = 7.5 - 3 = 4.5$

Lower fence $= Q1 - 1.5(IQR)$
$= 3 - 1.5(4.5) = -3.65$

Higher fence $= Q3 + 1.5(IQR)$
$= 7.5 + 1.5(4.5) = 14.25$

Minimum = 1
Q1 = 3
Median = 5
Q3 = 7.5
Maximum = 9

Any value outside Lower & higher fence is called outliers



Here 27 is removed so it's an outliers