Day. 2

11/1/22

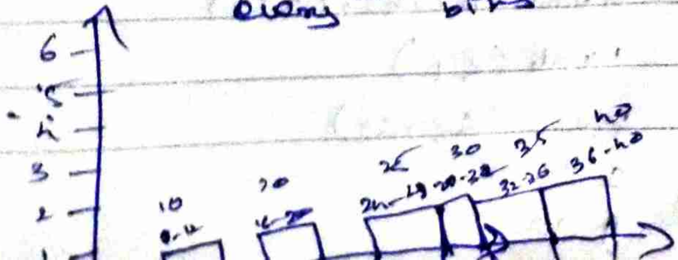
Agenda

1.) Histogram
2.) Measure g central thendency
3.) Measure of Dispersion
4.) percentile & 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, 60, 62, 72,
90, 95, 100}

i.) Sort the numbers
Ages are already sorted
ii.) Bins → No. g. groups
iii.) Bin size → size g bins.
if suppose we have in below size
[10, 20, 25, 30, 35, 40] min=10
we want 10 bins max=40
bins=10
∴ So we have to ge the 10 group g bins in means take the max (40) and divided it by bins(10)
$\frac{40}{10} = 4$ ∴ So it is size g every bins.



Ex:-1:

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

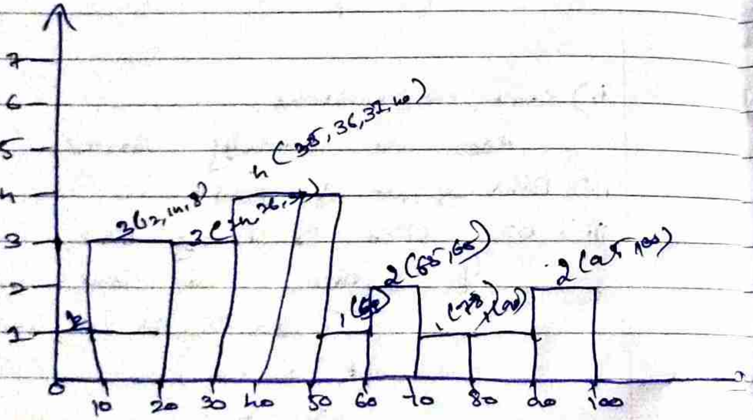
Histogram

i.) Sort the values (Asc)
  * Alrad sorted
ii.) get the number of bins
  bins = 10
iii.) get the bin size based on max value & bins

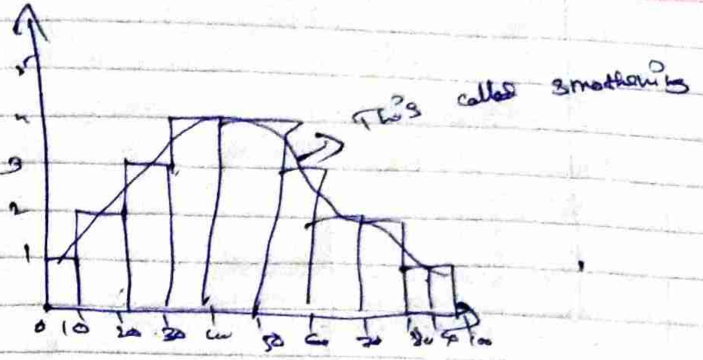
$\frac{100}{10} = 10$

binsize = 10

So,



0-10 = 1 value (10)
10-20 = 3 value (12, 14, 18)
20-30 = 3 value (24, 26, 30)
30-40 = 4 value (35, 36, 37, 40)
40-50 = 4 value (41, 42, 43, 50)
50-60 = 1 value (51)
60-70 = 2 value (68, 68)
70-80 = 1 (78)
80-90 = 1 (90)



* Smoothening concept gives the probability density function.
* Kernel ~~dens~~ distribution estimation

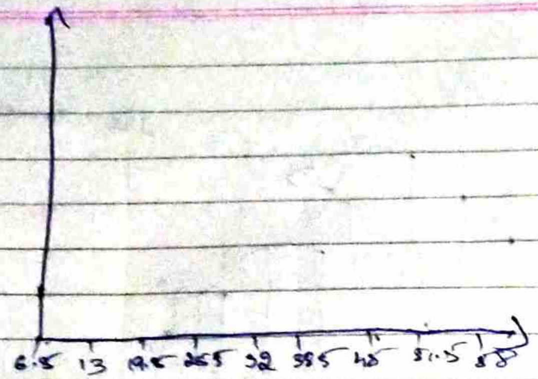
Ex2:
weight = {30, 35, 38, 42, 46, 58, 59, 62, 63
68, 75, ~~74~~ 80, 90, 95

i.) Sort the weight (Asc)

ii.) get the bin
  bin = 10

iii.) If our values start from '0' then we can take max number, and divide it by bin. we get binsize
∴ So here we have values start from 30 so we have to substract the ~~max~~ max - min

$\therefore \frac{95-30}{10} = \frac{85}{10} = 6.5$



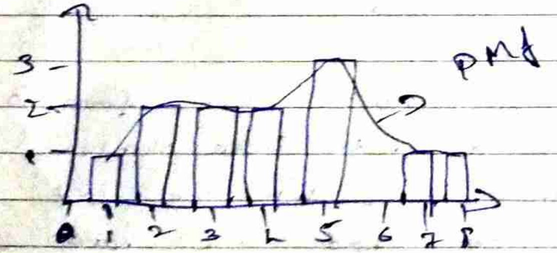
2) Measure of central Thendency

1) Mean
2) Median
3) Mode

Discrete continuous

no of Bank accounts = [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5

i) sort the values



pdf = probability density function [continuous var]
pmf = probability Mass function [Discrete var]

2) Measure of central Thendency

i) Measure of central Thendency is a single value that attempts to describe a set of data identifying the central Position.

i) Mean (Average)

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

$\bar{x} = \frac{\Sigma x}{n} = \frac{1+2+3+4+5}{5} = \frac{\overset{3}{\cancel{15}}}{\cancel{5}} = 3$

$\boxed{\bar{x} = 3}$

Population (N) Mean $N \geq n$	Sample (n) Mean
Mean → $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ $\mu \geq \bar{x}$ or $\bar{x} \geq \mu$	Mean $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

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

$N = 6$

$\text{p. Mean } (\mu) = \frac{24+23+2+1+28+27}{6}$

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

$\boxed{\mu = 17.5}$

Ex: Sample Age = {24, 2, 1, 27}

$\bar{x} = \frac{24+2+1+27}{4} = \frac{\overset{13}{\cancel{54}}}{\cancel{4}} = 13.5$

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

Practical Application [Feature Engineering]

(mean)

Age	Salary	Family size
–	–	–
–	–	–
		–
Nan	–	–
–	NAN	–
–	–	NAN
–	NAN	–

⇒ loss of Info

* So here we have to handle. this (NAN) values my put the Average of the particular column.

Age	Salary
24	45
28	50
29	NAN
NAN	60
31	75
36	80
NAN	NAN.
Avg 29.6	Avg 62

NAN ⇒ Not, A number

↓

Age	Salary
24	45
28	50
29	62
29.6	60
31	75
36	80
29.6.	62

2.) Median

$\{1, 2, 3, 4, 5\} = \{1, 2, 3, 4, 5, 100\}$ ← One layer

$\bar{x} = 3$ $\quad\quad \bar{x} = \frac{1+2+3+4+5+100}{6} = \frac{115}{6} = \boxed{19.16}$

* by adding one outlier we get more difference in mean so here median will come

Steps to find out the median

i.) Sort the values

ii.) Find the center values

i.) if values are even no we find the values average

ii.) if values are odd no we find the central elements.

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

1.) Sort = {1, 3, 4, 5 $\boxed{6, 7}$ 8, 20, 10, 120}

2.) center values $= \frac{6+7}{2} = \frac{13}{2} = 6.5$

$\boxed{\text{Median is } 6.5}$

* If there is no outliers we have to use Mean

* If there is outliers we have to use Median.

2) Mode

- Most Repeated values

Age = $\{1, 2, 2, \boxed{3, 3} \boxed{3}, 4, 5\}$

Mode = 3

Age = $\{1, \boxed{2, 2} \boxed{3, 3}, 4, 5\}$

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

Practical example

Types of flower

Rose
Lilly
Sunflower
Rose
NAN
Rose

$\boxed{\text{Mode} \Rightarrow \text{Rose}}$

→ How to handle NAN in this situation categorical Data with we have to replace Mode value instead of NAN value

↓

Rose
Lilly
Sunflower
Rose
Rose
Rose

3) Measure of Dispersion

1.) variance (σ^2)
2.) Standard (σ)

variance → spread of Data

Population variance (σ^2) - $\sigma^2 \neq s^2$	sample variance (s^2)
$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
	n-1 ⇒

∴ $x_i - \mu$ or $x_i - \bar{x}$

basically it gives the difference of x value to mean value

x → values

$\mu, \bar{x}$ ⇒ mean (N, n)

Ex: 1) $\{1, 2, 3, 4, 5\}$ & 2) $\{1, 2, 3, 4, 5, 6, 30\}$

$\mu = 3$

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

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

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

$\mu = 14.4$

$$\sigma^2 = \frac{\left[(1-14.4)^2 + (2-14\ldots + (5-14.4)^2 + (\ldots + (8-14.4)^2 + (\ldots + (30-14.4)^2\right]}{7}$$

$\sigma^2 = [719.4]$

2.) Standard deviation $(\sqrt{\sigma^2})$

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

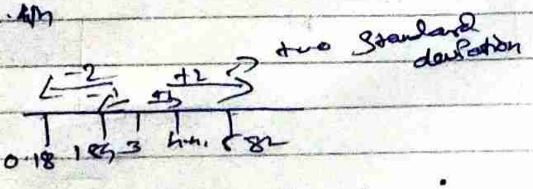
$\mu = \frac{15}{5} = 3$

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

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

$\sigma^2 = 2$

$\sigma = \sqrt{2} = 1.414$



* How many standard deviation away from the mean.

Ex: what is the standard dev of 4

it is +1

because standard deviation value we should add or sub to the mean and see how much time to take to reach the given number(4)

3σ

4.) percentile And Quartiles

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

Per of even number in percentage } = $\frac{\text{No of even numbers}}{\text{Total No of Numbers}}$

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

Percentile

Def

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

99 percentile -> It means the person has got better ~~marks~~ than 99% of entire students.

Dataset

$x = \{2,2,3,4,5,5,5,6,7,8,8,8,8,9,9,9,10,11,11,12\}$

i.) What is the percentile rank of 10

i.) Sort the data first

ii.) Percentile Rank of x = $\frac{\text{No. of values below}(x)}{n}$

$= \frac{16}{20} \frac{4}{5} \frac{4}{8} = 0.8 = 80\%$

ii) What is the value that exists at 25 percentile

value $= \frac{\text{percentile th}}{100}$ → for odd numbers
for even $(n+1)$

$= \frac{25}{\cancel{100}_{5}} \times \cancel{20}$

$= 5^{th}$ index value of the given values.

iii) What is the value that exists at 20 percentile.

value $= \frac{\cancel{20}}{\cancel{100}_{5}} \times \cancel{20}+1$

$= \frac{\cancel{20}}{\cancel{100}_{5}} \times 21$

$= \frac{1}{5} \times 21 = 21/5 = 4.2 = 4^{th}$ index of given

5) 5 number Summary

1.) Minimum
2.) first quartile (25 percentile)(Q1)
3.) Median
4.) Third quartile (75 percentile)(Q3)
5.) maximum

Box plot ↑ Remove the outliers

outlier ?

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

1) To find outliers
↓ [Lower Fence ⟷ Higher Fence]

Lower fence → Q1 − 1.5 (IQR) ∴ IQR = Q3 − Q1 (75 − 25)
Inter quartile Range (IQR)

Higher Fence = Q3 + 1.5 (IQR)

$Q_1 = \frac{25}{100} \times (n+1)$

$= \frac{25}{100} \times (19+1)$

$= \frac{25}{\cancel{100}_{4}} \times 20$ 5.2

$= 5.2$

$\boxed{Q_1 = 5.2}$ Index ⇒ $\boxed{\frac{3+3}{2}} = \frac{6}{2} = 3$

$Q3 = \frac{75}{100} \times (20)$

$= 15.75^{th}$ index value so 15 & 16th value

$= \frac{7+8}{2} = \frac{15}{2} = 7.5$

$\boxed{Q3 = 7.5}$

Lower fence ⇒ 3 − (1.5)(4.5) = −3.65

Higher fence ⇒ 7.5 + 1.5 (4.5) → 14.25

(−3.65 ⟷ 14.25)

In between this values only we should have data other data is are outliers.

TO remove out lier

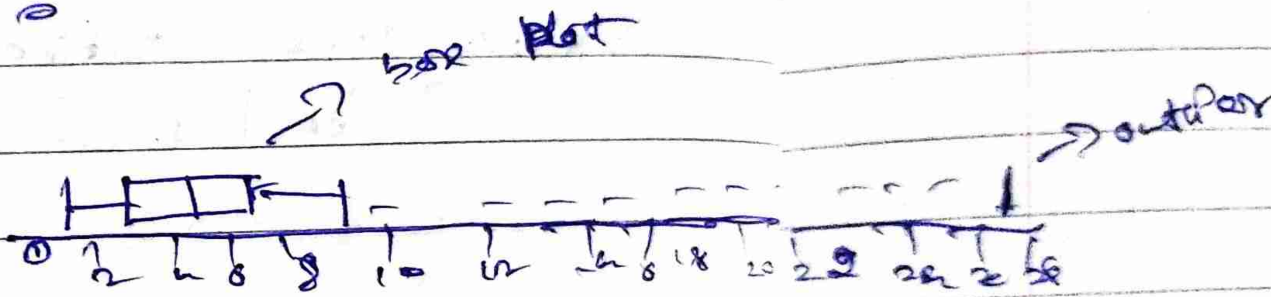
1) minimum - 1

2) Q1 - 3

3) media - 5

4) Q3 - 7.5

5) maximum - 9


box plot
outlier
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28