STATISTICS FOR DATA SCIENCE Day 2

Agenda

(i) Histogram
(ii) Measure of Central Tendency
(iii) Measure of Dispersion
(iv) Percentile and quartiles
(v) 5 Number Summary (Box Plot)

(i) HISTOGRAM

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

How to create Histogram :

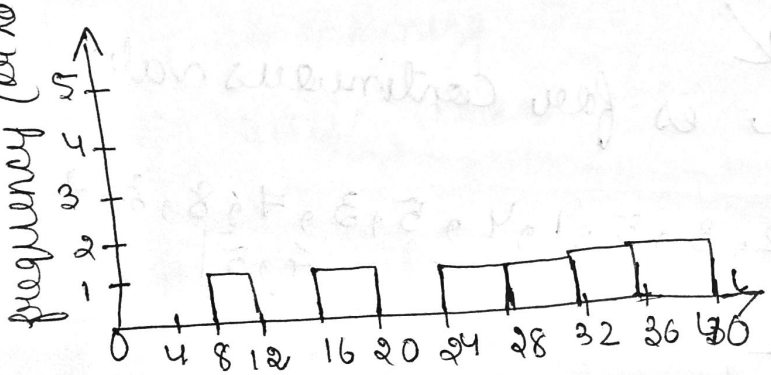
Step 1 : Sort the Numbers
Step 2 : Bins → No. of groups
Step 3 : Bins size → Size of Bins

[10, 20, 25, 30, 35, 40]
min = 10, max = 40
bins = 10
bins size = $\frac{40-10}{10}$
$= \frac{30}{10} = 3$

[10, 20, 25, 30, 35, 40]
min = 10, max = 40

0 3 6 9 12 15 18 21 24 27 30

∴ we get 10 bins.

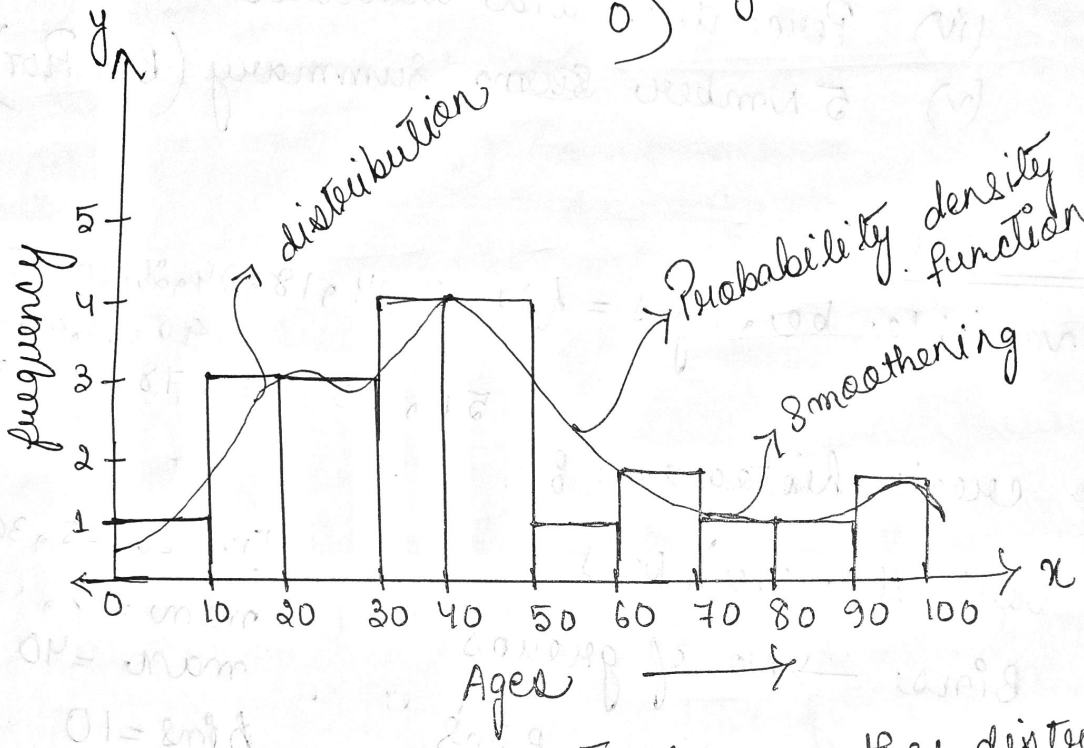


e.g. Let, the bin = 10 for this data

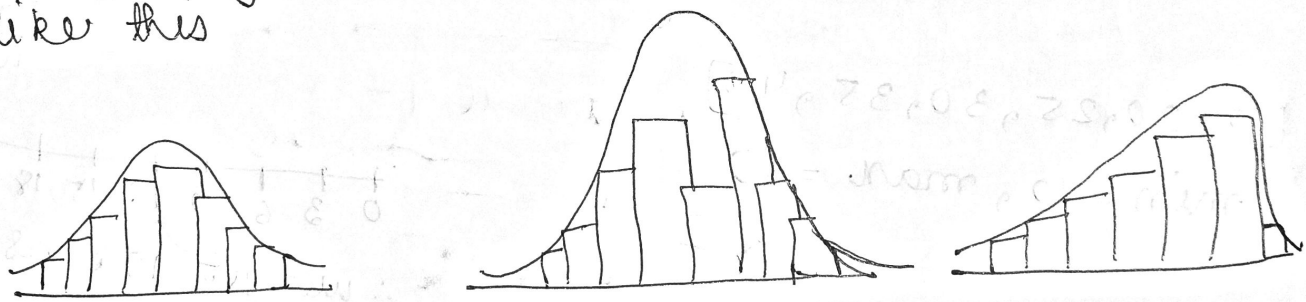
Age = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}

min = 10, max = 100

bin size = $\frac{100-0}{10} = 10$ (We are starting from 0 that's why we are subtracting by 0)



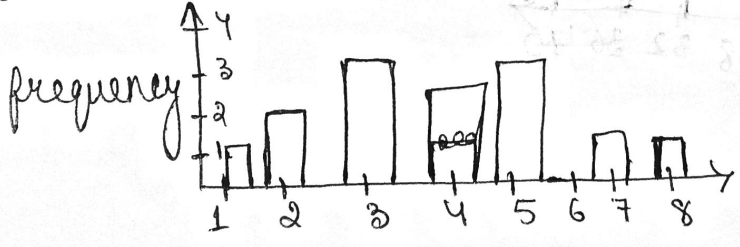
Smoothening will help to form the distribution like this



This one is for continuous value

*Discrete Continuous

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

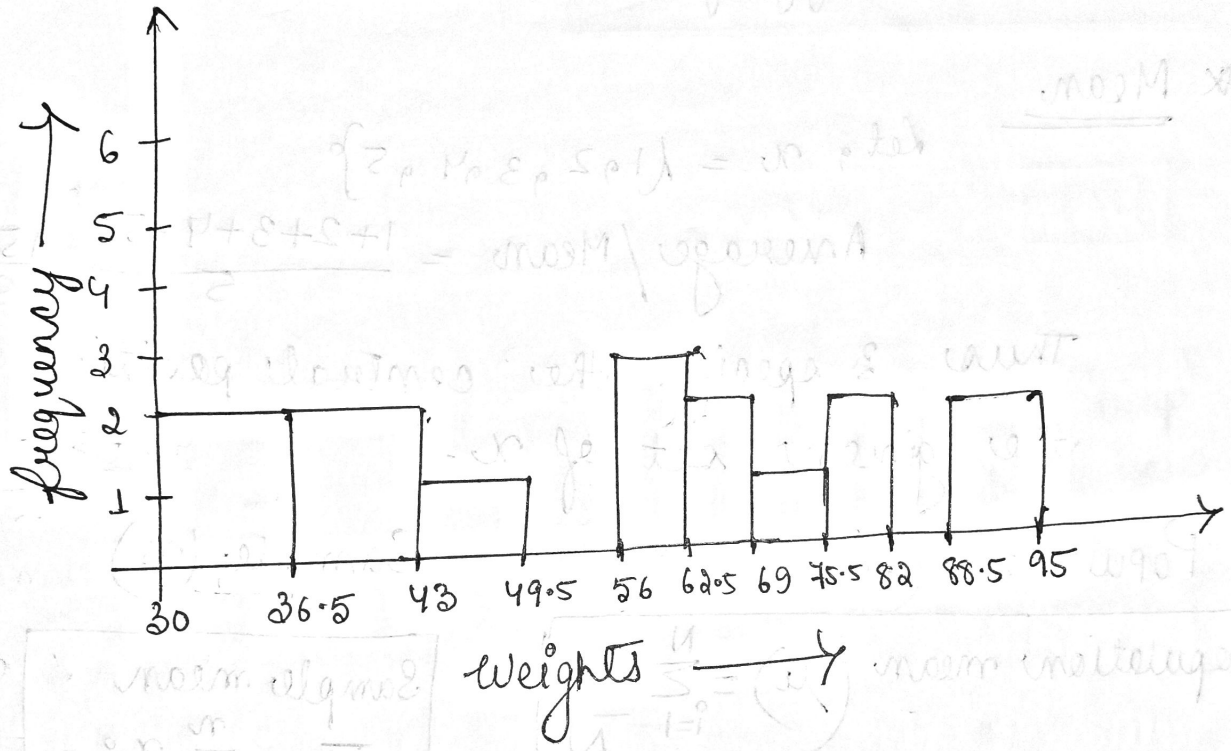


E.g. weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 80, 90, 95}

bins = 10

If we want to start from 30, then

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



* For smoothening the continuous histogram we will use Probability density function (pdf)

* For smoothening the discrete continuous histogram we will use Probability mass function (pmf)

(ii) Measure of Central Tendency

(a) Mean (b) Median (c) Mode

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

* Mean

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

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

Thus 3 specify the central position of the given set of x.

Population (N) Sample (n)

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

and off course N >>> n

Example

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

$\therefore N = 6$

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

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

Sample Age = {24, 2, 1, 27} [Let us pic randomly 4 value from Population Age]

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

$= \frac{54}{4} = 13.5$

PRACTICAL APPLICATION (Feature Engineering)

Let, the three features be

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

* If we drop the NAN row then we will loss some of our information
* We will use mean here in place of NAN value

NAN → Null or Not number

Age	Salary
24	45
28	50
29	NAN
NAN	60
31	75
36	80
NAN	NAN

Age mean $= \frac{24+28+29+31+36}{5}$

Replacing NAN value by 29.6

If we add one more value 80 to the set then the NAN value will change to 30

(here 80 is outlier)

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

Replacing NAN value by 62 on Salary set

If we add one more value 200 to the set then the NAN value will change to 85

(here 200 is outlier)

Median

{1, 2, 3, 4, 5}

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

$= 3$

{1, 2, 3, 4, 5, 100} ← outlier

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

$= 19.16$

By adding one outlier, the mean is changing $\bar{x} = 3 \longrightarrow \bar{x} = 19.16$ which is huge.

so we will use median here

Steps to find out median

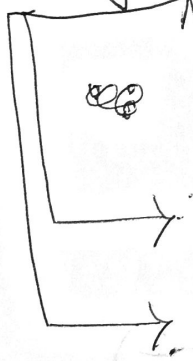
(i) Sort the numbers

(ii) Find the central number

(a) If the no. of elements are even we find the avg. of central elements

(b) If the no. of elements are odd we find the central elements.

For e.g. {1, 2, 3, 4, 5, 6, 7, 8, 100, 120}



∵ no. of elements is even [100 & 120 are outlier]

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

mean = 25.6 (since mean >>>> median)

* when there is no outlier we use mean for NAN value

* when there is outlier we use median for NAN value

MODE : Most frequent occurring element

e.g. {1, 2, 2, 3, 3, 3, 4, 5}
⇓
mode = 3

{1, 2, 2, 2, 3, 3, 3, 4}
⇓
mode = 2 & 3.

Practical Application

Dataset

Types of flower

Lily

Sunflower

Rose

NAN

Rose

Sunflower

Rose

NAN

* Mode we will use in categorical dataset

since here mode is Rose, so will replace NAN value by Rose

(iii)

Measure of Dispersion

(i) Variance (σ^2)

(ii) Standard deviation (σ)

* Variance (σ^2) ⟶ Spread of Data

Population Variation (σ^2)

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

Sample Variance (s^2)

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

$x_i - \mu$ → distance from mean with every data point

Assignment (1): Why sample variance is divided by 1?

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

{1, 2, 3, 4, 5}

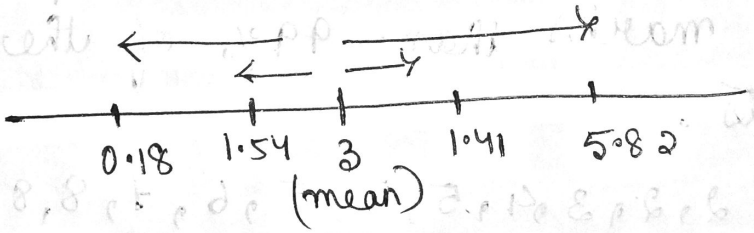
mean = 3

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

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

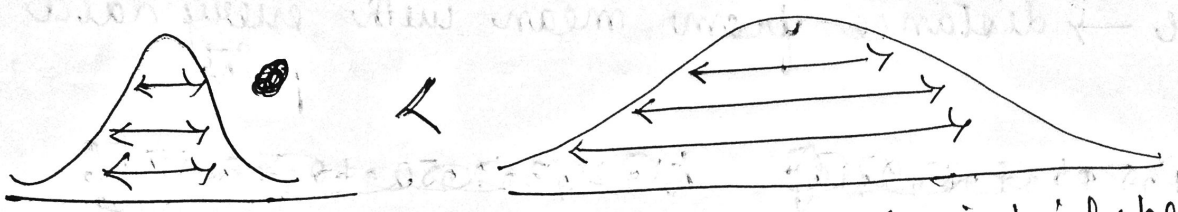
$\Rightarrow \sigma^2 = 2 \quad \Rightarrow \sigma = \sqrt{2}$

$= 1.41$



Standard deviation mean how many standard deviation away the no. from the mean

So, in the data {1, 2, 3, 4, 5}, the no. 2 is away by -1 standard deviation from 3.



spread is less here

spread is high here
⇓
variance is high
⇓
Standard deviation is high

(iv) Percentile And Quartiles

~~Percentage~~ ⇒ %

Percentiles : MATE, CAT, SAT, JEE, NEET ⇒ marks are count on basis of Percentile

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

99 percentaile = It means that the person has got 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. what is the percentile rank of 10

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

Percentile rank of 10 = $\frac{16}{20}$ = 80 percentaile

Percentile rank of 8 = $\frac{9}{20}$ = 45 percentile

Q what is the value that ~~consist~~ exist at 25 percentile

Value = $\frac{\text{Percentile}}{100} * (n$

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

Q what is the value that exist at 95 percentile

Value = $\frac{95}{100} \times 21 = 19.95$

5 numbers Summary

(i) Minimum (ii) First Quartile (25 percentile) (Q1) (iii) Median

(iv) Third Quartile (75 percentile) (Q3) (v) Maximum

These 5 numbers Summary used to remove outliers.

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

Here 27 is the outlier.

[Lower Fence ⟷ Higher Fence]

75% 25%

IQR = Q3 − Q1

↓

Inter Quartile Range

Lower Fence = Q1 − 1.5(IQR)

Higher Fence = Q3 + 1.5 (IQR)

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

$= \frac{25}{100} \times 21 = 5.25 \Rightarrow$ Index $= \frac{3+3}{2} = 3$ (Adding 5th & 6th index value)

$Q3 = \frac{75}{100} \times 21 = 15.75 \Rightarrow$ Index $= \frac{8+7}{2}$

$= \frac{15}{2} = 7.5$

Lower Fence $= Q1 - 1.5(IQR) = 3 - 1.5(4.5)$

$= -3.65$

$IQR = Q3 - Q1$

$= 7.5 - 3$

$= 4.5$

Higher Fence $= Q3 + 1.5(IQR)$

$= 7.5 + 1.5 \times 4.5$

$= 14.25$

Now,

1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

① Minimum = 1

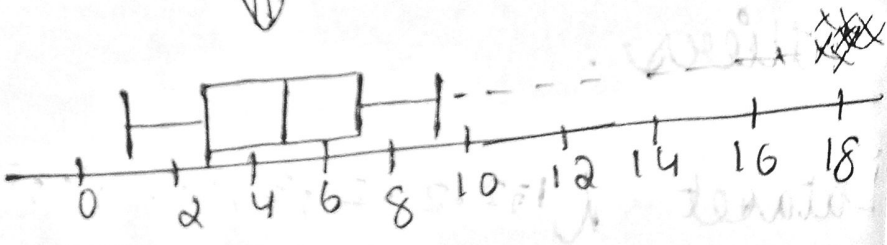
② Q1 = 3

③ Median = 5

④ Q3 = 7.5

⑤ Maximum = 9

Boxplot



To treat outliers