* Hare Krishna *

Statistics

1. Histograms
2. Measure of Central tendency
3. Measure of Dispersion
4. Percentiles and Quartiles
5. 5 Number Summary (Box plot).

Histogram:

Ages = {10, 12, 14, 18, 24, 30, 34, 35, 36, 37, 40, 41, 42, 43, 50, 51, 68, 78, 90, 95, 100}

1. Sort the Numbers
2. Bins → No. of groups
3. Bins Size → Size of Bin

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

bin = 10

Min = 10

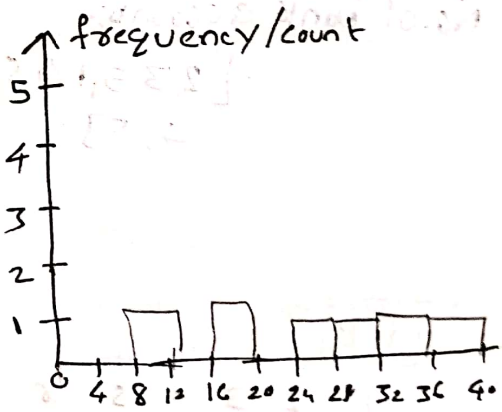
Max = 40

$\frac{40-10}{10} = \frac{30}{10} = 3$

0 3 6 9 12 15 18 21 24 27 30

$\frac{40}{10} = 4$.

0 4 8 12 16 20 24 28 32 36 40



find the bin size = 10 of Ages.

bin size = $\frac{100}{10}$ ⇒ $\frac{\text{total Ages}}{\text{bin}}$

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

bin size are desided by us // our wish //

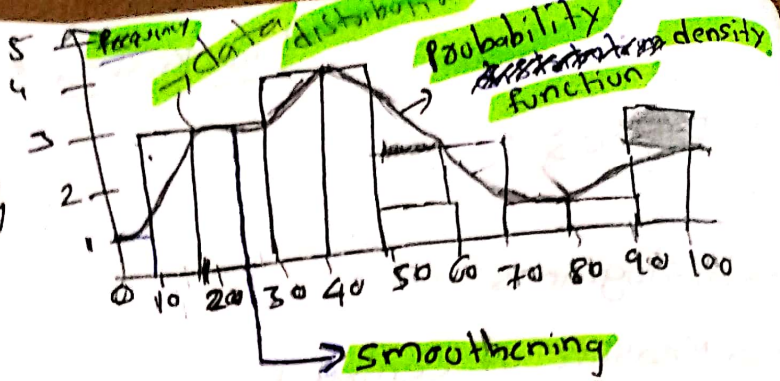
let bins = 20

bin size = $\frac{100}{20} = 5$

bin size = 5

start from 0

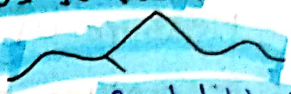
Ages = {10, 12, 14, 18, 24, 26, 30, 35
36, 37, 40, 41, 42, 43, 50
51, 65, 68, 79, 90, 95, 100}



Group = bin

Groupsize = binsize

Smoothening helps us to form distibution like this



probability distribution of data distribution

if we have negative (-ve) values our bins also get or have (-ve) values.

** {Probability density function} **

Weight = {30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 71, 80, 90, 95}

{continious values}

bins = 10

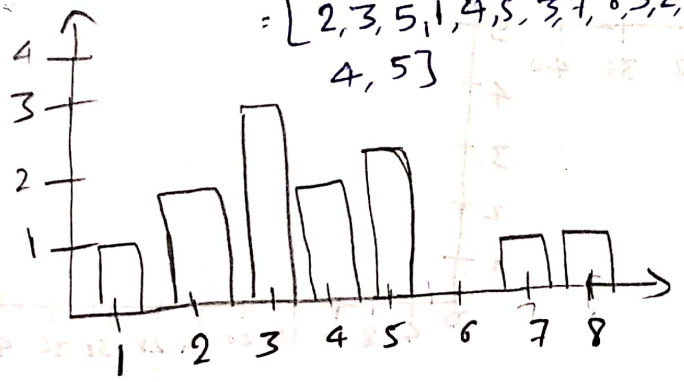
$$\text{binsize} = \frac{95-30}{10} = \frac{65}{10} = 6.5$$

~~Measures~~

Discrete continious:

No. of bank accounts:

= [2, 3, 5, 1, 4, 5, 3, 7, 8, 3, 2, 4, 5]



to Smoothen these discrete continious we use {Probability Mass function}

In discrete continious we are able to find gaps

pdf:- Probability density function → continious variable

PMF:- Probability Mass function → discrete cont^ variable

PMF

pdf: Probability density function

PMF: Probability Mass function

② → Measure of central tendency.

ⓐ Mean

ⓑ Median

ⓒ Mode

defination

A Measure of CT is a Single value that attempts to describe a set of data identifying the central Position

Central Position

Mean

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

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

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

Population (N)

Sample (n)

Population mean (μ)

$$\mu = \sum_{i=1}^{N} \frac{x_i}{N}$$

$N >> n$

Sample mean ($\bar{x}$)

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

never Happens

$n \geq N$ ✗

Population

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

Population (N) = 6

Population mean (μ) $= \frac{24+23+2+1+28+27}{6}$

$\mu = 17.5$

here

$N > n$

$\bar{x} \neq \mu$

$\mu \geq \bar{x}$

$\bar{x} \geq \mu$

Sample Age $= \{24, 2, 1, 27\}$

Sample (n) = 4

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

$\bar{x} = 13.5$

Pratical application { Feature Engineering }

used in

NAN = NULL Values

Not A Number

I have features like

⇓

Age	Salary	Family size
-	-	-
-	-	-
NAN	-	-

if we drop this table

we loss of info

NAN → replace with mean of Salary

NAN, NAN → replace with mean of family size.

NAN → replace with mean of age.

if we drop the table we will loss the info so that, it effects the total out come. ~~dr the~~ we have to

instead of this we have to find Mean of Age and replace the NAN value with the Mean of Age so that you cannot ~~loss~~ loss the info

Mean is talking about the Central information of data.

Ex:

Age	Salary
24	45
28	50
29	NAN
NAN	60
31	25
36	80
NAN	NAN
80	200

⇒ outliers

outliers

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

$= 29.6$

Mean of salary $= \frac{45+50+60+75+80}{5}$

$= 62$

Mean of Age after adding outliers $= \frac{24+28+29+31+36+80}{6}$

$= 38$

mean of salary after adding outliers $= \frac{45+50+60+75+80+200}{6} = 85$

before outliers mean of Age = 29.6

after outliers mean of age = 38

before outliers mean of salary = 62

after outliers mean of salary = 85

there a huge difference by adding outliers to the mean of Age, Salary,

So to avoid these difference we use

Median

Median

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

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

outlier

$\{1, 2, 3, 4, 5, \boxed{100}\}$

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

$= \frac{115}{6} = 19.16$

$\bar{x} = 3, \quad \bar{x} = 19.16$

Just by adding one outlier the mean is chaning a huge.

in order to prevent this median is used.

⇒ Steps to find out Median:-

① ⇒ Sort the Numbers

② ⇒ find the central numbers

if the no of elements are even we find the average of central elements

if the n.o of elements are odd we find the central elements

Sorted no of element are even.

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

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

if we are adding "0"

$\{0, 1, 2, 3, 4, \boxed{5}, 6, 7, 8, 9, 10\}$

now no of elements are odd

then we take the central

Median = 5 Element = 5 //

Note:- When ever we have outlier we have to use median.

{ No outliers → Mean
  with outliers → Median }

(3) Mode:- { The Most foeguently occuring elements }

Ex: { 1, 2, 2, 3, 3, 3, 4, 5 }
Mode

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

2, 3 → Mode

in recent version of python it is giving you Errors, but in previous version of python it use to give you 2, 3 Mode

Dataset:-

Types of flowers:

Lily
Sunflower
Rose
NAN ⇌ Rose
Rose
Rose
Sunflower
Rose
NAN ⇌ Rose

Mode most of time we use with categorical value variable

③ ⇒ Measure of Dispersion

ⓐ variance (σ^2) ← Spread of data

ⓑ standard deviation (σ)

Variance

- Population variance (σ^2)
- Sample variance (s^2)

Population variance (σ^2)

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

Sample variance (s^2)

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

First one ⇓ Variance ↑↑ spread ↑↑ second one ⇓

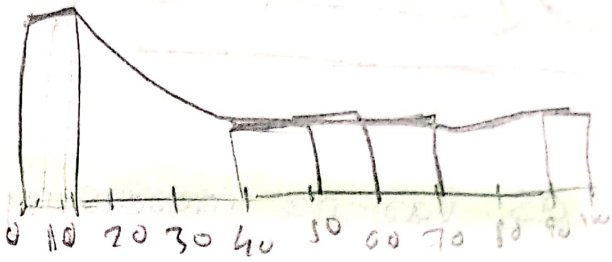
Ex:

{1, 2, 3, 4, 5, 6, 7, 8, 9, 10} , {1, 2, 3, 4, 50, 60, 70, 100}

⇓ Variance

⇓ Variance

So Variance ↑ spread ↑ as



Variance ↑↑ spread ↑↑

Ex: 2

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

$$\mu = \frac{1+2+3+4+5}{5} = \frac{15}{5}$$

$$\mu = 3$$

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

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

(b) {1, 2, 3, 4, 5, 6, 80}

$$\mu = \frac{1+2+3+4+5+6+80}{7} = \frac{101}{7} = 14.4$$

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

$$\sigma^2 = 749.10$$

by graph:

σ^2 (variance)

as variance increasing spread also increasing

So variance ↑↑ spread ↑↑

∴ variable is ~~variance~~ nothing but which indicating the spread of the data.

(*) b

Standard deviation:

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

Standard deviation = σ

$\{1, 2, 3, 4, 5\}$ $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

$\mu = 3$ $\mu = 3$

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

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

$$\sigma^2 = 2//$$

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

$\mu = 3$

When we are moving one step backword (-1) which mean's Substracting σ (1.41)

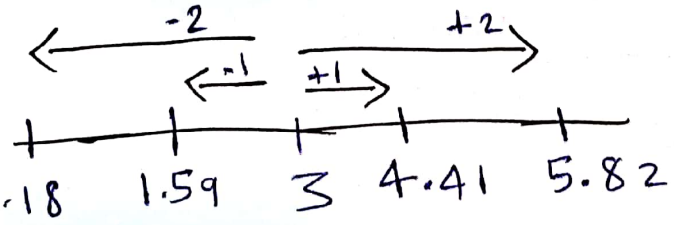
$3 - 1.41 = 1.59$

$1.59 - 1.41 = 0.81$

When we are moving one step forward (+1) which means adding σ (1.41)

$3 + 1.41 = 4.41$

$4.41 + 1.41 = 5.82$



Standard deviation tells that:-

Suppose we are taking value "4" how many standard deviation will fall away from mean.

like : ⇒ how many standard deviation "4" is falling away from mean is with in (+1) range

⇒ how many standard deviation "3" is falling away from mean is "0"

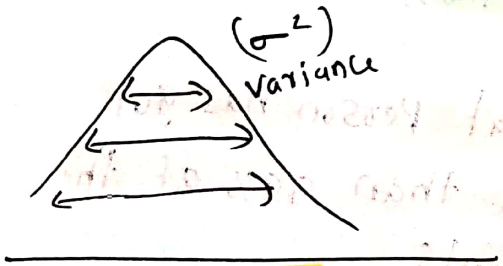
Variance talk about Spread

How the data is spread.

Variance ↑↑

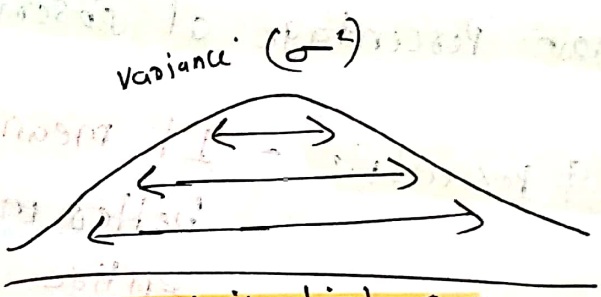
spread ↑↑

spread of data is calculated by variance.



spread is low ↓↓

variance is low ↓↓



spread is high ↑↑

variance is high ↑↑

Standard deviation say's that {1, 2 .... 1000} in side these number, if i take any number i will be able to find how many standard deviation is away from the mean.
↓
group of

* Percentiles & Quartiles:-

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

Percentage of even Number

$$= \frac{\text{No of even Numbers}}{\text{total no of numbers}}$$

$$= 4/8 = 0.5 = 50\%$$

Percentiles:- Gate, LAT, IELS, SAT, GRE, JEE, NEE

↓ Percentiles

Defination:- A percentile is a value below which a certain Percentage of observations lies.

99 percentile = It means that Person has got better mark's than 99% of the entire students.

Asscending order:

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

index 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

what is the Percentile rank of 10

$$\text{Percentile Rank of } X = \frac{\#\text{ No of value below } X}{n}$$

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

what is the value that exist at 25 percentails.

for even. $\text{Value} = \frac{\text{Percentails}}{100} * n$

$= \frac{25}{100} \times 21 = 5^{th}$ index,

output = 5

for odd $\text{Value} = \frac{\text{Percentails}}{100} * n.$

$= \frac{95}{100} \times 21 = 19.95^{th}$ index

= output = 19

* 5 number Summary

① Minimum

② First Quartile (25 Percentile) (Q1)

③ Median

④ Third Quartile (75 Percentile) (Q3)

⑤ Maximum.

these 5 summary used to remove the outliers.

Ex: {1, 2, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 22}

[lower fence ⟷ higher fence]

lower Fence = $Q_1 - 1.5(IQR)$

$IQR = Q_3 - Q_1$ (75, 25)

IQR ⇒ Inter Quartile Range (IQR)

lower Fence = $Q_1 - 1.5(IQR)$

Higher Fence = $Q_3 + 1.5(IQR)$

$Q_1 = \frac{25}{100} * (n+1) = \frac{25}{100} * 21 = 5.25 \Rightarrow$ index

5.25 ⇒ index = 3,

we dont have 5.25 so we tala avarage of $\frac{3+3}{2} = \frac{6}{2} = 3$,

$Q_3 = \frac{75}{100} * 21 = 15.75$ index $= \frac{8+7}{2} = 7.5$

$Q_3 = 7.5$

lower Fence = $3 - (1.5)(4.5) = \boxed{-3.65}$

Higher Fence = $7.5 + (1.5)(4.5) = \boxed{14.25}$

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

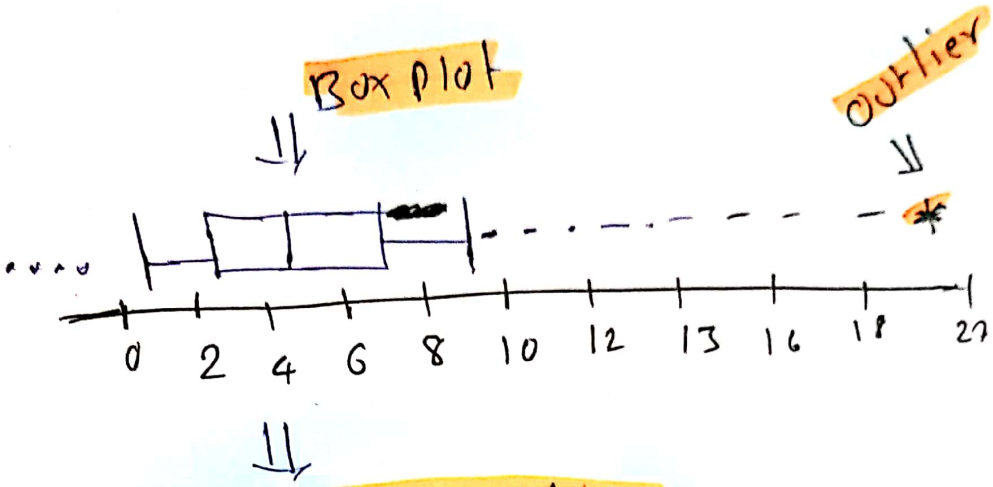
① minimum = 1

② Q_1 = 3

③ median = 5

④ Q_3 = 7.5

⑤ maximum = 9



⇓

TO treat outliers