24/2/24

APRIL 2023

| SUN | MON | TUE | WED | THU | FRI | SAT |
|---|---|---|---|---|---|---|
| | | | | | | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| 30 | | | | | | |

MARCH'23
SATURDAY
18
11th Week • 077-288

# Stats - 2.

## data dispersion (describes the spread of the data)

1. Range
2. variance / standard deviation
3. mean absolute deviation
4. Inter Quartile Range (Quartile)

→ we can divide out data in to 4 parts.
min, $Q_1$, $Q_2$, $Q_3$, max (also called 5 no summary)

**Range** ⇒ maximum - minimum (consider 2 pts in the estimation)

Eg) $\boxed{1}$, 1, 1, 1, 2, 2, 2, 3, 3, 4, 4, 4, 4, $\boxed{5}$

= 5 - 1
= 4

Eg) 1, 1, 1, 2, 2, 2, 2, 3, 4, 4, 5, 120 → outlier

Range = ?

= 120 - 1
= 119 (we cannot consider the range as there is outlier).

**variance**

it indicates how close or far the data points from the mean.

population variance :- $\cancel{\sum_{i=1}^{n}} (x_i - \bar{x})^2$

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

Health lies in labor, and there is no royal road to it but through toil.

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

N = population data set

n = sample data set

Sample variance

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

$$\Downarrow$$

$$\boxed{s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$ (sample variance with the bassal's correction)

Eg) 10, 12, 14, 15, 17, 18, 18, 24

Ans) $N = 8$

population mean $(\mu) = \frac{128}{8} = 16$

population variance =

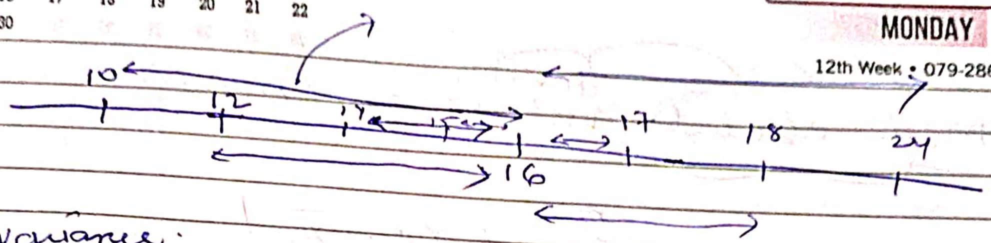
$\frac{1}{8} \times (10-16)^2 + (12-16)^2 + (14-16)^2 + (15-16)^2 + (17-16)^2 + (18-16)^2 + (18-16)^2 + (24-16)^2$

$= \frac{130}{8} = 16.25$

Note:-

If my variance is ~~low~~ less, my data distribution is fine.

If variance is high, my data distribution can go here and there data distribution




sample variance.

let $n=6$

$\frac{130}{6}$

bessel correction $= \frac{130}{8-1} = \frac{130}{7}$

standard deviation $\Rightarrow \sqrt{\text{variance}}$

population standard deviation $= \sqrt{16.25} = 4.03$

→ variance is giving high value as we perform square & then submission of data.

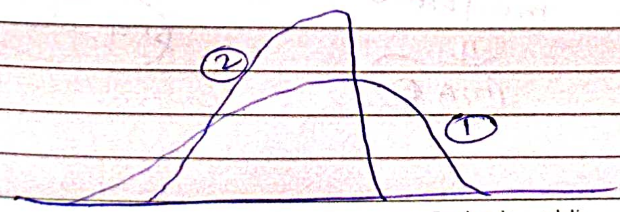
so we have to compress the value, so we use squareroot.

less standard, very close to mean & data set is good.

Note:

→ The more the data are spread out, the greater the range, variance and standard deviation.

→ the more the data are concentrated, the smaller the range, variance and standard deviation.




2nd one is best bcoz it is concentrated

mean absolute deviation ~~value~~ ⇒ (MAD)

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

Formulas

MSE = variance = $(x_i - \bar{x})^2$ ⇒ very high value ⇒ cancelling the -ve effect around the mean.

MAD ~~RMSE~~ = MAD = $|x_i - \bar{x}|$ ⇒ absolute value ⇒ cancelling ~~the~~ -ve effect around the mean.

RMSE ~~MAD~~ stand dev = $\sqrt{\text{variance}}$

In ML, MSE, MAD, RMSE are used to calculate loss.

Q) 3-5 diffs betn variance, std dev, MAD when to use.

note :-

Data point ⇒ mean ⇒ spread → compress ⇒ std dev

spread → very high ⇒ variance

spread → moderate ⇒ MAD

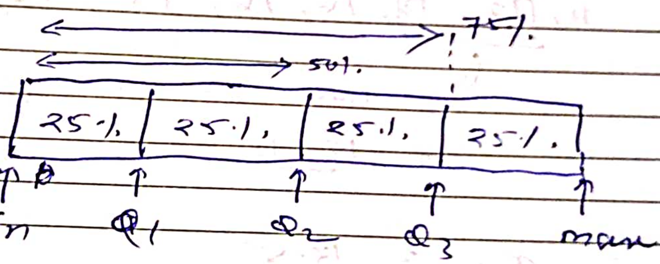
⇒ If we want very high data value we will use variance
" " " " low " " " " " stepler.

Quantile measure or position measure

25% 50% 75%

Quartile means we have to spread/split ~~part~~ our data in to 4 equal segments

$$\boxed{\begin{aligned} Q_1 &= 25\% \times n \\ Q_2 &= 50\% \times n \\ Q_3 &= 75\% \times n \end{aligned}}$$




$$IQR = Q_3 - Q_1$$
$$= 75 - 25$$
$$= 50\%$$

lower fence ⇒ $\boxed{Q_1 - 1.5 \times IQR}$

upper fence ⇒ $\boxed{Q_3 + 1.5 \times IQR}$

for identification of outliers

Q) 11, 12, 13, 16, 16, 17, 18, 21, 22. Find out the quartile

Ans) $Q_1 = ?$ $n = 9$

$Q_2 = ?$

$Q_3 = ?$

| SUN | MON | TUE | WED | THU | FRI | SAT | SUN | MON | TUE | WED | THU | FRI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| 26 | 27 | 28 | 29 | 30 | 31 | | | | | | | |

09.00 so $Q_1 = 25\% \times n = \frac{25}{100} \times n = \frac{1}{4} \times (n+1)$

10.00

because $n = 9$ (odd)
11.00 so we have to write $n+1$.

12.00

$= \frac{n+1}{4} = \frac{10}{4} = 2.5$ th position
13.00

2.5
14.00 11, $\boxed{12, 13}$, 16, 16, 17, 18, 21, 22

15.00 so take the mean of both
16.00 $= \frac{12+13}{2}$

17.00 $Q_1 = 12.5$

here % = percentile not percentage
18.00

$Q_2 = 50\%$ of $n$

$= \frac{50}{100} \times (9+1)$

$= \frac{1}{2} \times 10 = 5$ position

11, 12, 13, 16, $\boxed{16}$, 17, 18, 21, 22

$Q_2 = 16$

$Q_3 = 75\%$ of $n$

$= \frac{75}{100} \times (n+1)$

$= \frac{3}{4} \times 10 = 7.5$ th position

11, 12, 13, 16, 16, 17, $\boxed{18, 21}$ 22 $= \frac{18+21}{2} = \frac{39}{2} = 19.5$

Q2

$$\boxed{11 \quad 12 \;\}\; 13 \quad 16 \quad 16 \quad 17 \quad 18 \;\}\; 21 \quad 22}$$

Q1 Q3

pos value

Q1 = 25 percentile → 2.5 ⇒ 13.5

Q2 = 50 " → 5 ⇒ 16 ⇒ median (middle value)

Q3 = 75 " → 7.5 ⇒ 19.5

IQR = Q3 − Q1 = 19.5 − 13.5
= 7

Five no summary is

minimum value = 11
Q1 = 13.5
Q2 = 16
Q3 = 19.5
max value = 22

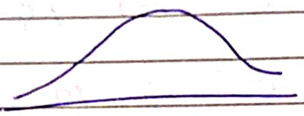
all together is called box plot

lower fence ⇒ Q1 − ~~IQR~~ 1.5 × IQR
⇒ 12.5 − 1.5 × 7
⇒ 12.5 − 10.5
⇒ 2

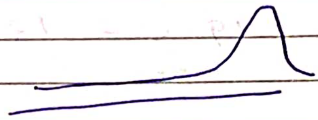
upper fence ⇒ Q3 + 1.5 × IQR
⇒ ~~30~~ 19.5 + 1.5 × 7
⇒ 19.5 + 10.5
⇒ ~~40~~ 30

~~note~~ whatever value comes after upperfence; they are called outliers.

so in this eg our outliers are :-
50, 65.



← normal distribution
(no outlier)

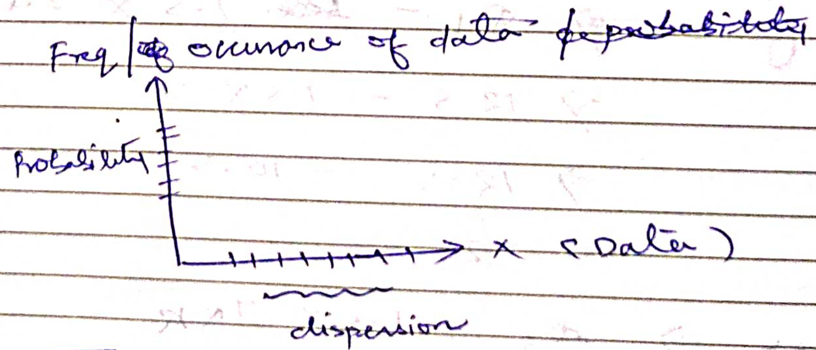


← left skewed
(outlier in left hand side)

← right skewed
(outlier in right hand side)

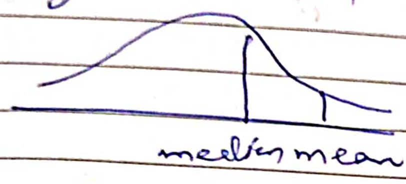
so as per our example, the outliers are in right hand side, so it is right skewed.

Data



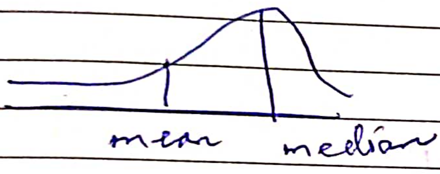

$\boxed{\text{median} < \text{mean}}$

In right skewed




mean > median.

In left skewed.




mean < median.

Probability | PDF | PMS

→ probability density function

→ probability mass function

Random variable ⇒ continuous | Discrete