(7)

# Measure of Central Tendency.

A measure of CT is a single value that attempts to describe a set of data by identifying the central position.

The three main types :-

1) Mean -(Average) The mean is the Sum of all values divided by the no. of values.

formula

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{No. of values}}$$

ex ⇒ x = 70, 75, 80, 85, 90

$= (70+75+80+85+90)/5$

$= \frac{400}{5} = 80$

- Mean wrt population ($\mu$-mu) N

The entire group you're studying,

formula

$$\text{population Mean}(\mu) = \sum_{i=1}^{N} \frac{x_i}{N}$$

ex ⇒ N = (24, 23, 2, 1, 28, 27)

$\mu = \frac{24+23+2+1+28+27}{6}$

$\mu = 17.5$

- Mean wrt Sample (n) $\bar{x}$

A part of population you actually collect data from,

formula

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

ex ⇒ n = (23, 2, 28, 27)

$\bar{x} = \frac{23+2+28+27}{4}$

$\bar{x} = \frac{80}{4} = 20$

Mean tells you the "typical or "central" value in a data set.

ex # Data Set (for mean eg.)

| Age | Salary | family size |
|---|---|---|
| 24 | | |
| 26 | | |
| NAN | | |
| 21 | | |
| 20 | | |
| 18 | | |

* NAN → not a number → Empty

How to handle NAN/Empty

1) Delete → loss of info.
2) ignore
3) fill it (Avg.
4) Drop if too many nans.

In this example we ~~are~~ fill it with Avg.

$$Avg = \frac{24+26+21+20+18}{5} = 21.8$$

Now the NAN values replaced by 21.8

• We use Mean where there is no outlier in dataset

2> Median - (Middle value) The median is the middle value of a dataset when it's arranged in order

ex- $\underbrace{\{1,2,3,4,5\}}_{}$ Range = (5-1) = 4 $\Longleftrightarrow$ ex- $\{1,2,3,4,5,100\}$ Range = (100-1) = 99 — outlier

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

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

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

Outlier => It is a number that is complete different than the entire distribution.

* Steps to find the median. {1,2,3,5,4,100}

1) sort the no. (1,2,3,4,5,100)
2) find the central no. (1,2,3,4,5,100)
   if the no. of elements are even we find the avg of central no.s. $\frac{3+4}{2} = \frac{7}{2} = 3.5$

   if the no. of elements are odd we find central no. ex=> (1,2,~~3~~,4,5,100)
   => 4

We used Median when there is outlier in dataset

(9)

3) Mode - Most frequent appeard or occured element.

example- Dataset # Categorical variable

Types of flowers

LiLy

Sunflower

Rose

LiLy

Rose

NAN

Rose

Rose

LiLy

LiLy

⇓

So nomean and median used

⇓

Mode => LiLy (4) or Rose (4)

⇓

Replace with Something.

* ~~we used Mode when the data have categorical replacement,~~

---

# Measure of Dispersion.

It Shows how much the data varies or Spreads out from the center (like mean or median)

Why is it imp in DA => Becaus avg. alone can lie, you need dispersion to know if the data is stable or chaotic.

ex => $X = \{1, 1, 2, 2, 4\}$ $\quad Y = \{0, 2, 2, 3, 3\}$

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

$= 2$

these are same? No, but how it will Know by the technique variance.

# Types of Measure of Dispersion.

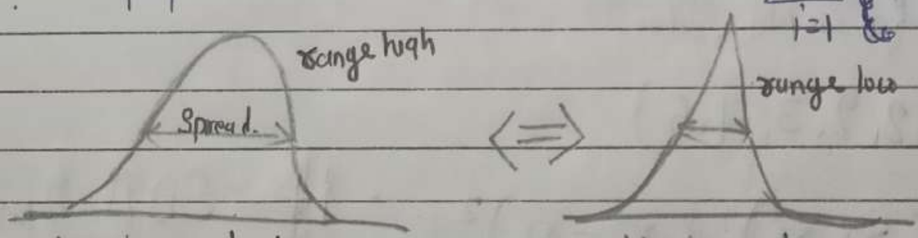
## 1) Variance

It measures the avg. of the squared difference from the mean. (spread of distribution).

formula —

population mean ($\mu$)

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




variance higher — variance low

- if range is high variance will high and vice versa.

formula —

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

→ Bessel's correction, degree of freedom

**Bessel's correction** = Correcting the bias in estimating population variance from a sample.

Use n-1 because when you work with a sample you are guessing — so you adjust for that guess with a tiny correction.

## 2) Standard Deviation (SD)

It tells us how much data values differ from the mean — in the same units as the data.

It is literally Jus the square root of variance

formula -

for population (N)

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

for Sample (n)

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

Example :— for sample data

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

| X | $\bar{x}$ | $(x-\bar{x})$ | $(x-\bar{x})^2$ |
|---|---|---|---|
| 1 | 2.83 | −1.83 | 3.34 |
| 2 | = | −0.83 | 0.6889 |
| 2 | | −0.83 | 0.6889 |
| 3 | | 0.17 | 0.03 |
| 4 | | 1.17 | 1.37 |
| 5 | | 2.17 | 4.71 |

- if SD↑ then dispersion or spread↑
- if SD↓ spread↓

Small SD ⇒ data is close to mean (consistent)

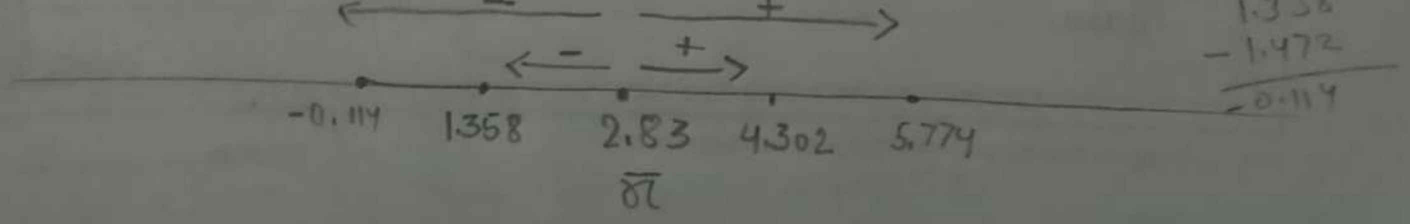
Large SD ⇒ data is spread out (more varient)

$$S^2 = \frac{10.84}{5} = 2.168$$

$$S = \sqrt{2.168} = 1.472.$$

$$\begin{array}{r} 2.83\,(\bar{x}) \\ +1.472\,(S) \\ \hline 4.302 \\ +1.472 \\ \hline 5.774 \end{array}$$

$$\begin{array}{r} 2.83 \\ -1.472 \\ \hline 1.358 \\ -1.472 \\ \hline -0.114 \end{array}$$



## Range –

Range = Highest value - lowest value

It tells you the total spread of data

ex – $X = \{12, 18, 25, 30, 40\}$

Max = 40   Min = 12

Range = 40 - 12 = 28

So the data values stretch across 28 units

- Range is super sensitive with outliers. One extreme value can totally mess it up.

# Percentile And Quartile

percentage – {1, 2, 3, 4, 5, 6, 7, 8}

percentage of even no. $= \frac{\text{No. of even no.}}{\text{total no. of nos.}} = \frac{4}{8} = \frac{1}{2}$

$= 0.5 = 50\%$

percentage of odd no. $= \frac{4}{8} = \frac{1}{2} = 0.5 = 50\%$

13

Percentile - we have seen Gate, SAT CAT JEE etc gave result in percentile.

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

example - if A person in 99 percentile, It means the person has got better marks then 99% of the entire Students.

- Dataset Example (Sorted)

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

$\underbrace{11, 11}_{\text{same}}$

*What is percentile ranking of 10?

$$\text{percentage Rank of } x = \frac{\text{\# of values below } x}{n} \times 100$$

$$10 = \frac{\cancel{16}^{8}}{\cancel{20}} \times 1\cancel{0}0 = 80 \text{ percentile}$$

10 is greater than 80% of entire distribution.

for 11,

$$11 = \frac{17}{\cancel{20}} \times \cancel{100}^{5} = 85 \text{ percentile}$$

*What is the value that exists at 25 percentile

$$\text{Value} = \frac{\text{percentile}}{100} \times (n)$$

$$= \frac{\cancel{25}^{1}}{4\,\cancel{100}} \times (20)$$

$$= \frac{20}{4} = 5 = \text{Index}$$

between $5^{th}$ and $6^{th}$ position

for 55 percentile

$= \frac{\cancel{55}^{11}}{\cancel{100}_{20}} \times (20) = \frac{220}{20} = 11^{th}$ position

for 40 percentile

$= \frac{\cancel{40}^{8}}{\cancel{100}_{5}} \times 20 = 8^{th}$ position.

What if your index is in decimal value.

$5.5^{th}$ position
b/w then $5^{th}$ and $6^{th}$ position i.e 5 and 5 in ques.
Avg. $\frac{5+5}{2} = \frac{10}{2} = 5$.

# Five Numbers Summary.

1) Minimum
2) Q1 first quartile (25th %)
3) Median (Q2 / 50th %)
4) Q3 third quartile (75th %)
5) Maximum

$\}$ used to remove outliers

15

Dataset

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

$> -3$ ? $< 13$

Quartile - divide a sorted dataset into 4 equal parts.

Inter Quartile Range IQR - how spread out the 50% of data b/w Q3 and Q1

$$IQR = Q_3 - Q_1$$

$$\text{Lower Bound} = Q_1 - 1.5 \times (IQR)$$

$$\text{Higher Band} = Q_3 + 1.5 \times (IQR)$$

$Q_1 = (25\%) = \frac{\cancel{25}}{\cancel{100}} \times \cancel{20}^{5} = 5^{th}$ position $= 3$

$Q_3 = (75\%) = \frac{\cancel{75}^{15}}{\cancel{100}} \times \cancel{20} = 15^{th}$ position $= 7$

$$IQR = Q_3 - Q_1 = 7 - 3 = \underline{4}$$

$$LB = 3 - 1.5 \times 4 = -3$$
$$HB = 7 + 1.5 \times 4 = 13$$

After removing outlier, the remaining value will be

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

1) Minimum = 1 2) $Q_1 = 3$ 3) Median/$Q_2 = 5$

4) $Q_3 = 7$ 5) Maximum = 9

Page No. 16

Date

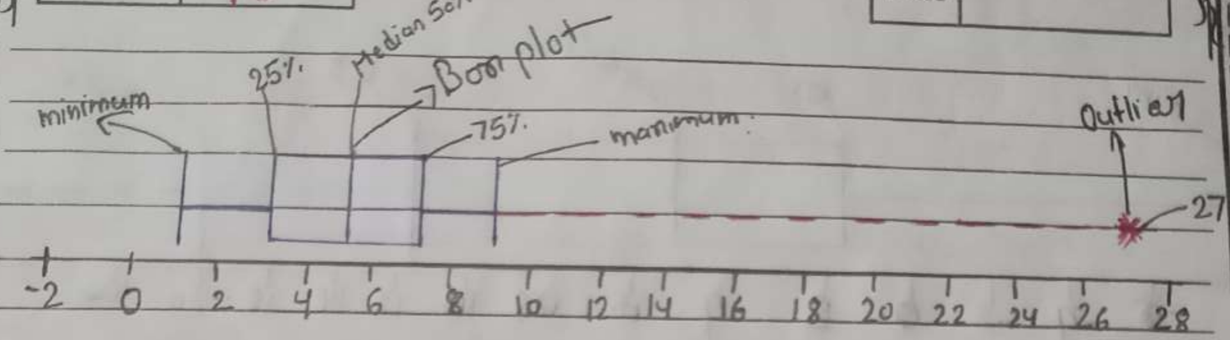



fig Box plot (identifies outliers).

Dataset

$\{-8, 1, 2, 4, 5, 6, 8, 15, 20, 120\}$

$> -13.5$ $< 26.5$

$Q_1 = (25\%) = \frac{\cancel{25}^{5}}{\cancel{100}_{2\,\cancel{10}}} \times \cancel{10} = \frac{5}{2} = 2.5^{th}$ position

$= 2^{nd} + 3^{rd}$

$= \frac{1+2}{2} = \frac{3}{2} = 1.5$

$Q_3 = (75\%) = \frac{\cancel{75}^{15}}{\cancel{100}_{2}} \times \cancel{10} = 7.5^{th}$

$= 7^{th} + 8^{th} = \frac{8+15}{2} = \frac{23}{2} = 11.5$

$IQR = Q_3 - Q_1 = 11.5 - 1.5 = 10$

$LB = 1.5 - 1.5 \times 10 = -13.5$

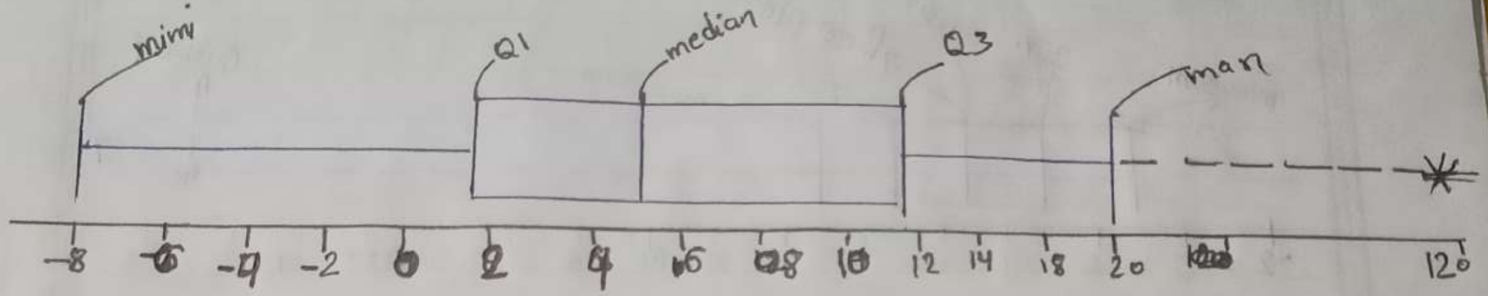
$HB = 11.5 + 1.5 \times 10 = 26.5$

After removing outliers

$\{-8, 1, 2, 4, 5, 6, 8, 15, 20\}$

17

1) Minimum = $-8$ 2) $Q1 = 1.5$ 3) Median = $5$ 4) $Q3 = 11.5$ 5) Max = $20$




—||0||—