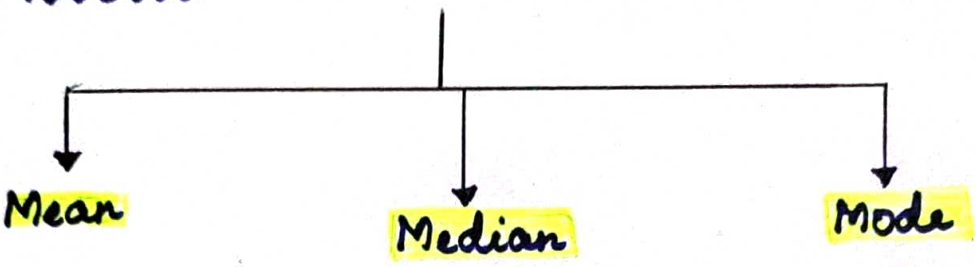
STATISTICS [DAY 2]

* MEASURE OF CENTRAL TENDENCY



Simple definitions :

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

MEAN : (Average)

Suppose if we have a data like

{24, 25, NAN, 21, 20, 18}

* Take the average of the above data
* Replace that value with the NAN

$$\text{Average} = \frac{24+26+21+20+18}{5} = 21.8 = \text{NAN}$$

Formula :-

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

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

MEDIAN :

Suppose if we have the data {1, 2, 4, 100, 5}

* Sort the numbers
* Find the central number

→ {1, 2, 4, 5, 100}
→ 4 = Median

Suppose if number of elements are even, we find the average of central elements

Data : { 1, 2, 3, 4, 5, 100 }

Sorting : { 1, 2, 3, 4, 5, 100 }

Average : $\frac{3+4}{2} = 3.5 \Rightarrow$ Median.

MODE : [Most frequent occuring elements]

If suppose the dataset is about the types of flowers

{Lily, sunflower, Rose, Lily, Rose, NAN, Rose, Rose, sunflower, sunflower, sunflower, sunflower}

→ Replace the NAN with most frequently occuring elements

→ Here Rose and sunflower is frequently occured element

→ Randomly choose one.

EXAMPLE:

(1) $X = \{24, 25, 26, 27, 28, 90, 100, 1000, 1200, 1400, 1400, 1400\}$

Calculate Average, Median, Mode.

Soln:

$$\text{Average (Mean)} = \frac{24+25+26+27+28+90+100+1000+1200+1400+1400+1400}{12}$$

$$\boxed{\mu = 560}$$

$$\text{Median} = \frac{90+100}{2} = 95$$

Mode = Most frequently occurring element = 1400

Measure of Dispersion

* Variance σ^2
* Standard Deviation σ

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

* To find Spread of the distribution, we can use variance formula
* To find range of the distribution, we use range as well as variance.

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

Population Standard deviation

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

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

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

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

(1) Compute variance and standard deviation

for $x = \{23, 21, 20, 19, 24, 27, 28\}$

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

x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
23	23.14	−0.14	0.0196
21	23.14	−2.14	4.5796
20	23.14	−3.14	9.8596
19	23.14	−4.14	17.1396
24	23.14	0.86	0.7396
27	23.14	3.86	14.8996
28	23.14	4.86	23.6196
			70.8572

$$S^2 = \frac{70.8572}{7-1}$$

$S^2 = 11.8095 \Longrightarrow$ variance

$S = 3.4364 \Rightarrow$ standard deviation.

PERCENTILES

Simple definition :-

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

99 percentile → It means the person has got better marks than 99 % of the entire students (say).

Example Questions

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

What is the percentile ranking of 10.

Soln

$$\text{Percentile rank of } x = \frac{\text{no. of values below } x}{n} \times 100$$

Percentile rank of 10 = $\frac{16}{20} \times 100 =$ 80 percentile

∴ 10 is greater than 80 percentage of entire data

What is the value that exist at 25 percentile

Soln

$$\text{value} = \frac{\text{Percentile}}{100} \times (n+1)$$

$$= \frac{25}{100} \times (21) = 5.25 \Rightarrow \text{index}$$

value = 5

FIVE-NUMBER SUMMARY

→ set of descriptive statistics that provides information about a dataset

→ consist of the five most important sample percentiles

1) Minimum
2) First Quartile (25%) Q_1
3) Median
4) Third Quartile (75%) Q_3
5) Maximum

1) Find the outliers {1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

(i) First find the Lower fence and higher fence

$$\text{Lower fence} = Q_1 - 1.5\,(IQR)$$

$$\text{Higher fence} = Q_3 + 1.5\,(IQR)$$

IQR → Inter Quartile Range

$$IQR = Q_3 - Q_1$$

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

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

$$= 5\text{th index}$$

$$\therefore Q_1 = 3$$

Q_3 = (75 percentile) = $\frac{75}{100} \times 20$

= 15th index

= 7

∴ IQR = $Q_3 - Q_1$ = 7 − 3 = 4

∴ Lower fence = 3 − 1.5 (4) = −3

Higher fence = 7 + 1.5 (4) = 13

* Now check do you have the values greater than −3
↳ Yes

check do you have the values lesser than 13
↳ No

∴ From the dataset, 27 > 13

∴ outlier = 27 is removed.

∴ The terms remaining are {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 = 5
(4) Q3 = 7
(5) Maximum = 9

Now construct box plot → identifies the outliers using five-number summary