DAY 2

BASIC TO INTERMEDIATE

(1) Arithmetic mean for population & Sample

Mean (Average)

Population (N)

$x = (1, 1, 2, 2, 3, 3, 4, 5, 5, 6)$

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

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10} = \frac{32}{10} = 3.2$$

Sample (n)

$$\boxed{\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}} = 3.2$$

★ Central Tendency (Central Measure of Tendency)

(1) mean (2) median (3) Mode

Interview Question →

Q:- What is Central Tendency? or.
What is measure of Central Tendency?

Ans. It Refers to the measure used to determine the center of the distribution of data.

Eg:-

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

$$\text{Mean } (\mu) = \sum_{i=1}^{N} \frac{x_i}{N} = \frac{132}{11} = 12$$

$\therefore \boxed{\mu = 12}$

Median:-

$x = \{1, 1, 2, 2, 3, 3, 4, 5, 5, 6, 100\}$

Step 1 in median = Sort the numbers.

Step 2 Take the Central Element.

in above Eg:- Total $N = 11$

means the central Element will be - 3

$1, 1, 2, 2, 3, \boxed{3}, 4, 5, 5, 6, 1000$

This number 3 is Median.

{Median works well with Outlier}

Mode:- = Eg: $x = \{1, \underbrace{2, 2}_{2}, 3, 4, 5, \underbrace{6, 6, 6}_{3}, 7, 8, 100, 200\}$

mode :- {Most Frequent Element}

$\boxed{\text{mode} = 6}$ → measure of Central Tendency.

★ **measure of Dispersion**

① Variance.

② Standard Deviation.

{Dispersion.}

⇓

Spread

① Variance :-

Two Types of variance.

① Population variance ② Sample variance.

① population variance :- Formula

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

② Sample Variation = Formula.

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

Letter	Uppercase	Lowercase.
Alpha	A	α
Beta	B	β.
Gamma	Γ	γ
Delta	Δ	δ
Epsilon	E	ε
Zeta	Z	ζ
Eta	H	η
Theta	Θ	θ
Iota	I	ι
Kappa	K	κ
Lambda	Λ	λ
Mu	M	μ
Nu	N	ν
Xi	Ξ	ξ
Omicron	O	o
Pi	Π	π
Rho	P	ρ
Sigma	Σ	σ

Letter	Uppercase	Lower case
Tau	T	τ
Upsilon	Y	υ
Phi	Φ	ϕ
Chi	X	χ
Psi	Ψ	ψ
Omega	Ω	ω

Eg of population variation.

x	μ	$(x_i - \mu)$	$(x_i - \mu)^2$
1	2.83	-1.83	3.34
2	2.83	-0.83	0.6889
2	2.83	-0.83	0.6889
3	2.83	0.17	0.0289 (0.03)
4	2.83	1.17	1.3689 (1.37)
5	2.83	2.17	4.7089 (4.71)
$\mu = 2.83$			10.725 (10.84)

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} = \frac{10.84}{6} = \boxed{1.81}$$

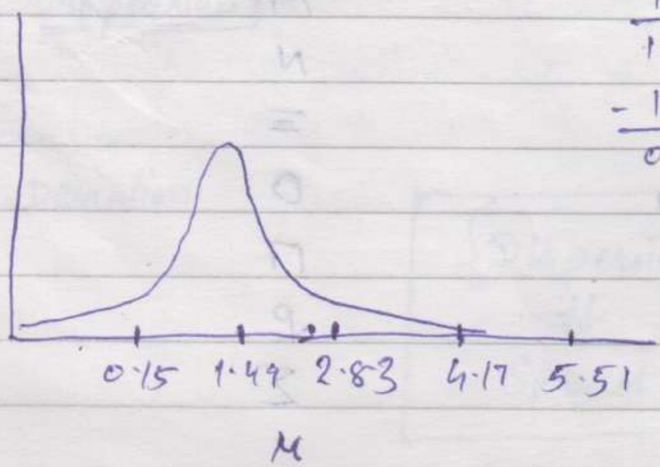
② Standard Deviation :-

$$\boxed{\sigma = \sqrt{\text{Variance}}} = \sqrt{1.81} = 1.34.$$

Graph.

$$\begin{array}{r} 2.83 \\ -1.34 \\ \hline 1.49 \end{array} \quad \begin{array}{r} 2.83 \\ +1.34 \\ \hline 4.17 \end{array}$$

$$\begin{array}{r} -1.34 \\ \hline 0.15 \end{array} \quad \begin{array}{r} +1.34 \\ \hline 5.51 \end{array}$$



Percentiles And Quartiles.

This is the first Step to find outliers.

Percentage :-

Eg :- 1, 2, 3, 4, 5

Find the % of the numbers that are odd?

$$\% = \frac{\text{No. of Numbers that are odd}}{\text{Total Numbers}} \times 100$$

$$= \frac{3}{5} \times 100 = 0.6 \times 100 = 60\%$$

Percentile :-

Defination:- A percentile is a value below which a certain percentage of observation lie??

Dataset :- 2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, $\boxed{10}$ 11, 11, 12

What is the percentile ranking of 10?

Eg:- Here $x = 10$

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

Where $n = 20$ as compare to above example.

$$x = \frac{16}{20} \times 100 = 80\%$$

This means that the 80% of the entire distribution is less then 10.

Assignment

What is the percentile Ranking of 11?

$x = 11$

$$\text{Percentile Ranking of } x = \frac{\#\text{ of value below } x}{n} \times 100$$

$$= \frac{17}{20} \times 100 = \boxed{85\%}$$

Now:

(2) What value exists at percentile Ranking of 25%?

Formula.

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

$$= \frac{25}{100} \times 21 = \boxed{5.25} \rightarrow \text{Index position.}$$

So According to our dataset.

2, 2, 3, 4, $\boxed{5, 5}$, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11, 11, 12

5.25 lie between these. ∴ our value will be $\boxed{5}$ for 25%.

Now For 75 percentile. $\frac{75}{100} \times 21 = \boxed{15.75}$

Acc. to our dataset

2, 2, 3, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 8, $\boxed{9, 9,}$ 10, 11, 11, 12.

This is our 15th and 16th number where 15.75 lie.

So take mean of these.

$$\mu = \frac{9+9}{2} = \frac{18}{2} = \boxed{9}$$

9 will be the value where 75 percentile lie.

Five Number Summary

1. Minimum.
2. First Quartile (Q_1)
3. Median.
4. Third Quartile (Q_3)
5. Maximum.

Removing the outliers.

Suppose the dataset is
$\{1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27\}$

First we have to define the {Lower Fence $\leftrightarrow$ Higher fence}

The number that is outside the Lower Fence and Higher Fence is the outlier..

Formula.

$$\text{Lower Fence} = Q_1 - 1.5\,(IQR)$$
$$\text{Higher Fence} = Q_3 + 1.5\,(IQR)$$

Where (IQR) = Interquartile Range.
$IQR = Q_3 - Q_1$

where $Q_3 = 75\%$ (Percentile).
$Q_1 = 25\%$ (Percentile).

For 25% $= \frac{25}{100} \times (19+1) = 5 \Rightarrow$ Index position.

So for above data set $Q_1 (25\%) = 3$.
Similarly $Q_3 [75\%] = 7$

Now Interquartile Range (IQR) = Q3 − Q1
= 7 − 3 = 4

Lower Fence = Q1 − 1.5 (IQR)
= 3 − 1.5 (4)
= 3 − 6 = −3

Higher Fence = Q3 + 1.5 (IQR)
= 7 + (1.5)(4)
= 7 + 6 = 13

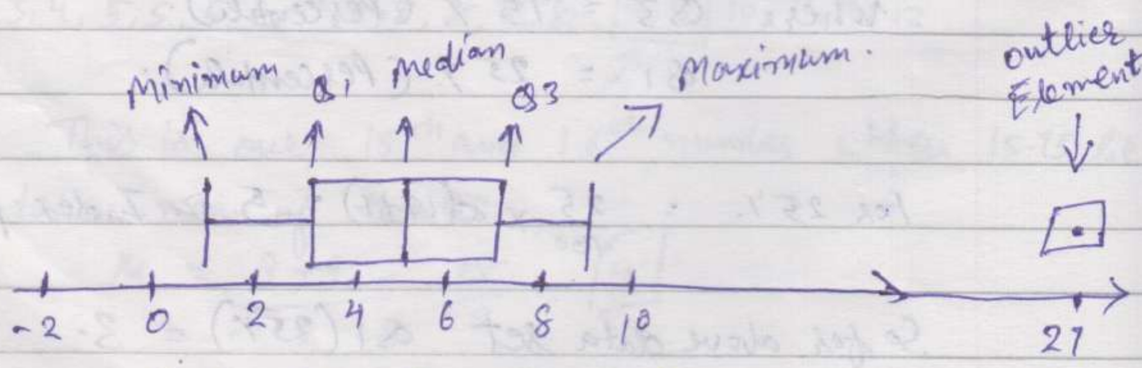
So the Lower Fence ⟷ Higher Fence Range.
[−3 ⟷ 13]

So in our dataset.
1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, ~~27~~

27 is the outlier because $\boxed{27 > 13}$

Now Five Number Summary
(1) Minimum = 1
(2) First Quartile (Q1) = 3
(3) median = 5
(4) Third Quartile (Q3) = 7
(5) maximum = 9

Box plot



Box plots are used to determine the outlier.

★ Sample variance.

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

where $n-1$ = Basal correction.
or Degree of freedom.