Descriptive Statistics

Numerical Summaries of data:→

Mean :–

$$\bar{x} = \frac{\text{Sum of observations}}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \text{ – Sample}$$

$$\mu = \frac{x_1 + \cdots + x_N}{N} \text{ – population.}$$

Median – A no. that splits the data in half, so that half of the data values are less than the median and half of the data values are greater than the median.

If n is odd – Median is $\left(\frac{n+1}{2}\right)^{th}$ term.

If n is Even – Median is Avg of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ terms.

– A **Statistic** is resistant if its value is not affected much by extreme values (large or small) in the data set.

– Median is resistant but the mean is not. that is mean is more Influenced by the Extreme values.

Describing data using mean and median:→

– Shape of the histogram — Relation b/w mean & median

Skewed to the right	Mean is greater than the median
Skewed to the left	Mean is less than the median
Appr. Symmetric	Mean = Median.

– In a skewed distribution, mean is closer to the tail.

- The Median is usually better than the mean for describing a skewed distribution or a dist^n with outliers.
- Use the mean for reasonably symm. dist^n that don't have outliers.

Range:→

Range = largest value − smallest value.
is the measure of spread of data.

Variance:→ $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$ — population variance.
where μ − pop^n mean

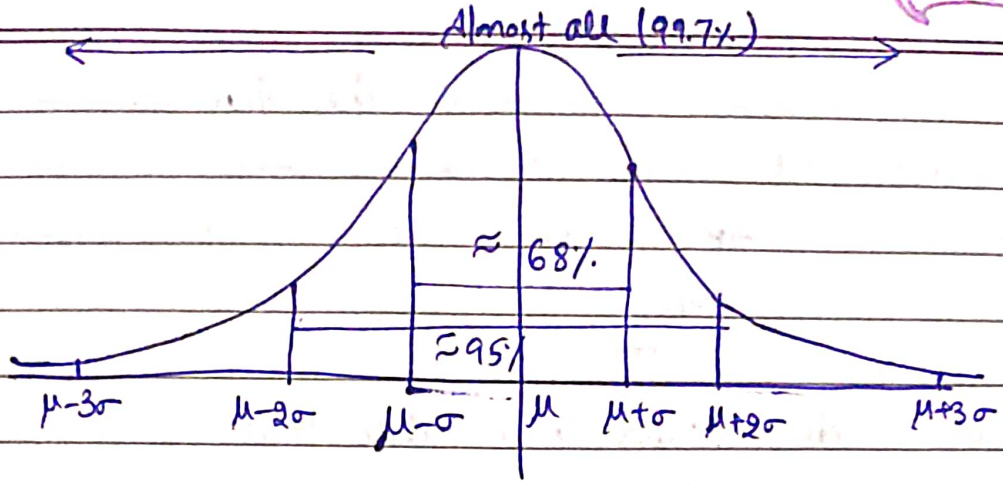
$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ — Sample variance.

where $\bar{x}$ − Sample mean.

Standard deviation:→ $S = \sqrt{s^2}$ — Sample std. dev.
$\sigma = \sqrt{\sigma^2}$ − pop^n std. dev.

The Empirical Rule:→ When a pop^n has a histogram that is approximately bell shaped then

(1) Appx. 68% of the data lie between $\mu-\sigma$ and $\mu+\sigma$ i.e. within one std. deviation of the mean
(2) Appx. 95% of the data lie b/w $\mu-2\sigma$ and $\mu+2\sigma$ i.e. within two std. dev. of the mean.
(3) Appx. all of the data (99.7%) lie b/w $\mu-3\sigma$ and $\mu+3\sigma$ i.e. within 3 std. dev. of the mean.



Chebyshev's Inequality:→ In any data set, the proportion of the data that will be within K std. deviations of the mean is at least $1 - \frac{1}{K^2}$.

Eg. If K=2, then atleast 3/4 (75%) of the data will lie within two Std. dev. of the mean.
If K=3; then atleast 8/9 (88.9%) of the data will be within three std. dev. of the mean.

Quartiles:→ Every data has three Quartiles.

(1) The first Quartile - Q_1, is the 25th percentile, i.e. Q_1 separates the lowest 25% of the data from the highest 75%.
(2) The Second Quartile - Q_2 is the 50th percentile. i.e. Q_2 separates the lower 50% of the data from the upper 50%. Q_2 is same as median.
(3) The third Quartile - Q_3 is the 75th percentile, i.e. Q_3 separates the lowest 75% of the data from the highest 25%.

The five-number summary:→ The five number summary of a data set consists of the following quantities.

Minimum | First Quartile (Q_1) | Median (Q_2) | Third Quartiles (Q_3) | Maximum

Outliers:→ An outlier is a value that is considerably larger or smaller than most of the values in a data set.

Inter Quartile range:→ (IQR)

$$IQR = Q_3 - Q_1$$

Stem and leaf plot:→ It is a method of sorting the data in such a way that each number is divided into two parts called stem and a leaf.

- A stem is a leading digit of each no. and is used in sorting.
- A leaf is the rest of the number or trailing digits.

Eg. If no. is 25

Stem (leading digit)	Trailing digit
2	5

If no. is 2.5

Stem	leaf
2	5 (decimal)

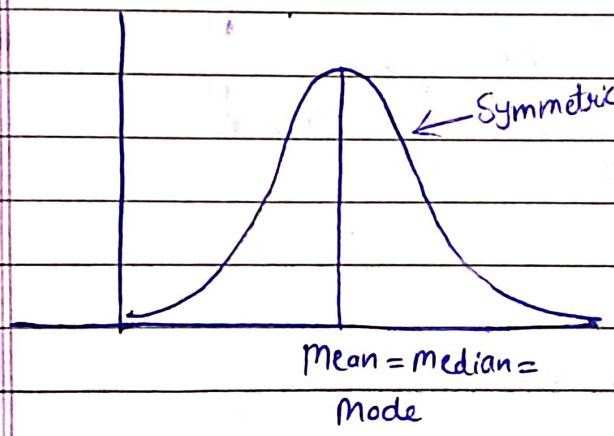
Ex Marks of the students are given below. Create a stem and leaf plot.

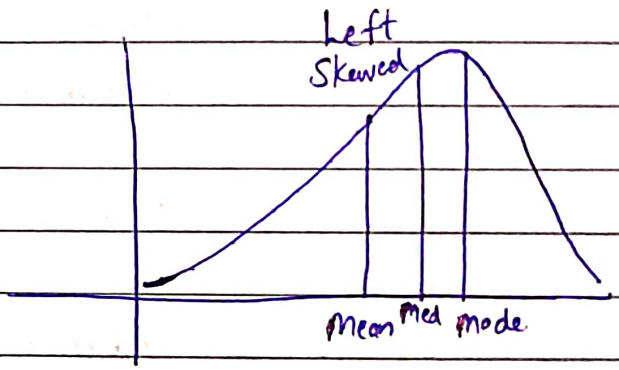
37 62 68 70 70 73 77 81 83
84 86 88 89 92 94 97 98 100 100

Stem	Leaf
3	7
4	
5	
6	2 8
7	0 0 3 7
8	1 3 4 6 8 9
9	2 4 7 8
10	0 0

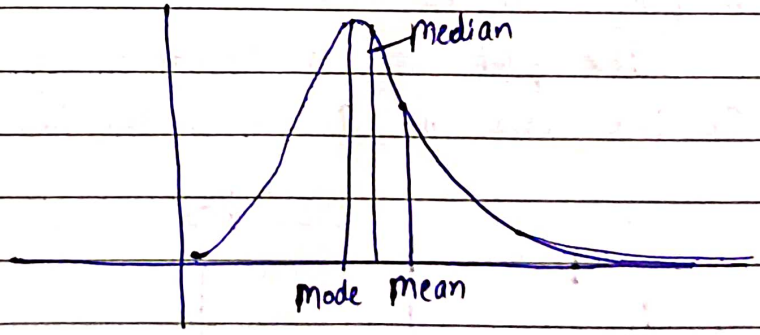
Key 3|7 = 37 ~~test~~ marks

Symmetric





(Negative Skewed)



Right Skewed (Positive Skewed)

Ex:→ 250 254 220 223 183 249 242 257 199 232
207 238 245 240 253 250 258 232 243 256
258 235 241 247

Create step and leaf plot.

Soln

Stem	Leaf
18	3
19	9
20	7
21	
22	0 3
23	2 2 5 8
24	0 1 2 3 5 7 9
25	0 0 3 4 6 7 8 8

Key 18|3 = 183

Left Skewed

Range = 258 - 183 = 75

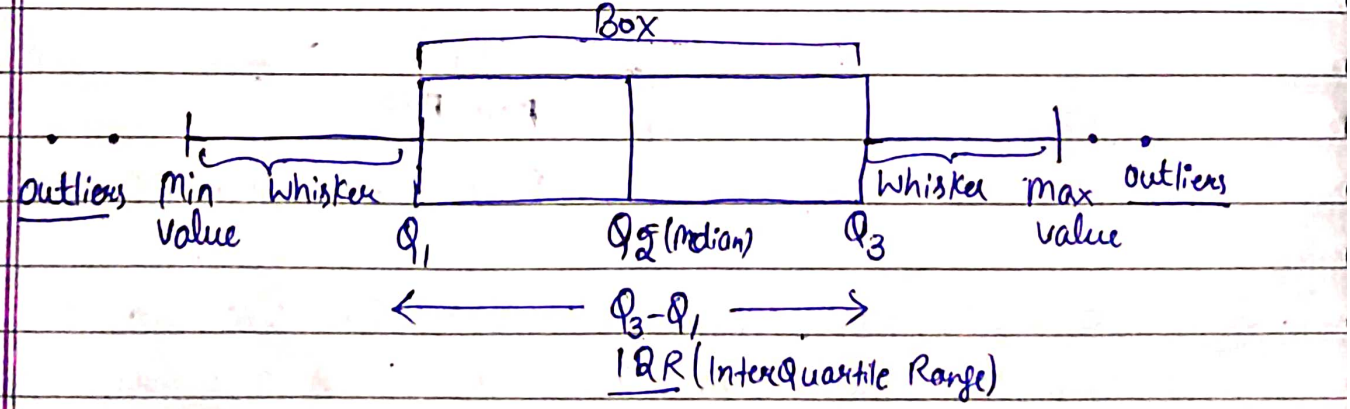
Median = $\frac{242+243}{2} = \frac{485}{2} = 242.5$.

Mean = $\frac{\text{Sum}}{24}$

Mode = 232, 250, 258

Box plot:→ (5 no. summary)

Box plot is a standard way to display the distribution of the data on a 5-number summary.



Outliers Boundary:→

Lower Outlier – $Q_1 - (1.5 * IQR)$

Higher outlier – $Q_3 + (1.5 * IQR)$

Ex The data is given as
22, 25, 17, 19, 33, 64, 23, 17, 20, 18

Soln Arrange the data in stem & leaf

Stem	Leaf
1	7, 7, 8, 9
2	0, 2 3, 5
3	3
4	
5	
6	4

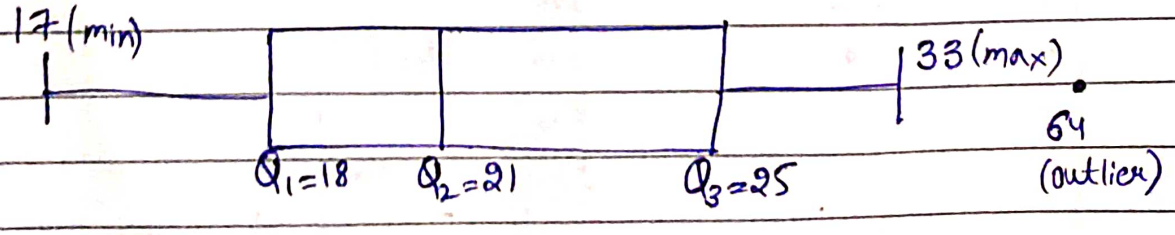
17
17
18 — Q_1
19
20
22 — $\frac{20+22}{2} = 21 = Q_2$ (Median)
23
25 — Q_3
33
64

$IQR = Q_3 - Q_1$
$= 25 - 18 = 7$

Lower outlier $= 18 - (1.5 * 7) = 18 - 10.5 = 7.5$
Higher outlier $= 25 + (1.5 * 7) = 35.5$

→ Any data < 7.5 and > 35.5 will be outlier
$\Rightarrow$ 64 is an outlier.

So Minimum = 17
Maximum = 33



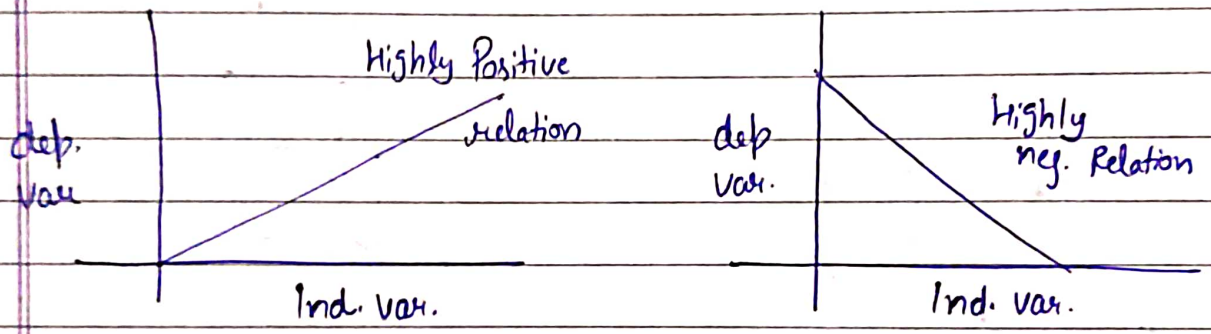
Ex 1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

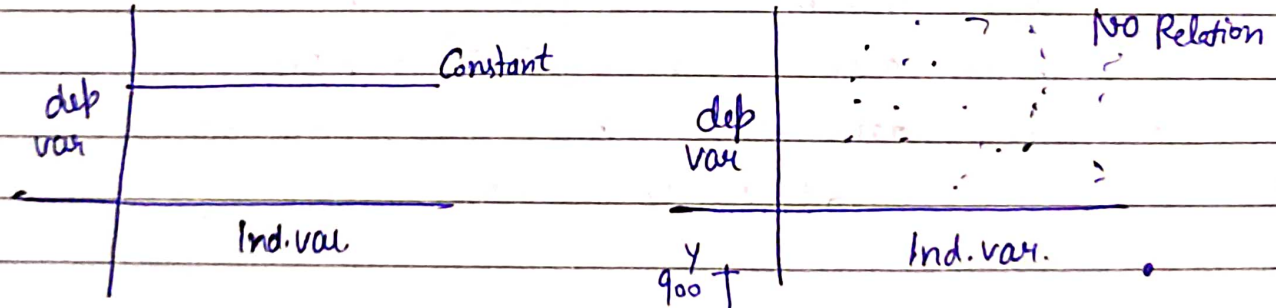
Create a box plot of the above data.

Scatter diagram:→ (Scatter plot)

It is used to analyse the relationship b/w two variables. One is called dependent and another is Independent Variable.

Four types of Scatter diagrams are there ↓





Ex

	Temperature	Sales (Rs.)
1	32	500
2	36	700
3	30	400
4	34	600
5	40	900
6	38	800