SHAPE OF THE DISTRIBUTION

Santosh chhatkuli

SKEWNESS

The graphical features of a distribution are most commonly described in terms of skewness and kurtosis. The frequency curve representing the observed frequency distribution may be either symmetrical or skewed or irregular shapes. An asymmetrical distribution is called skewed distribution. Skewness means lack of symmetry.

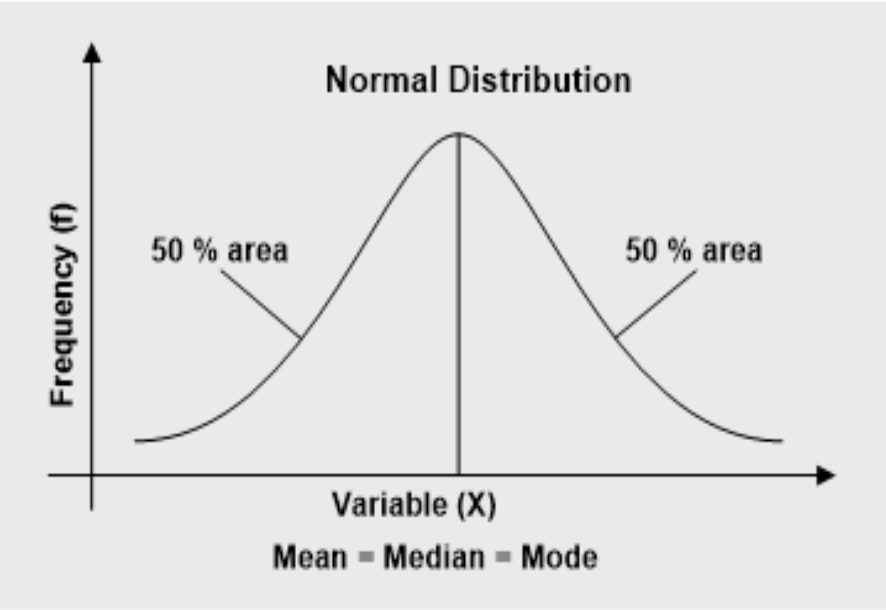
Types of Skewness:

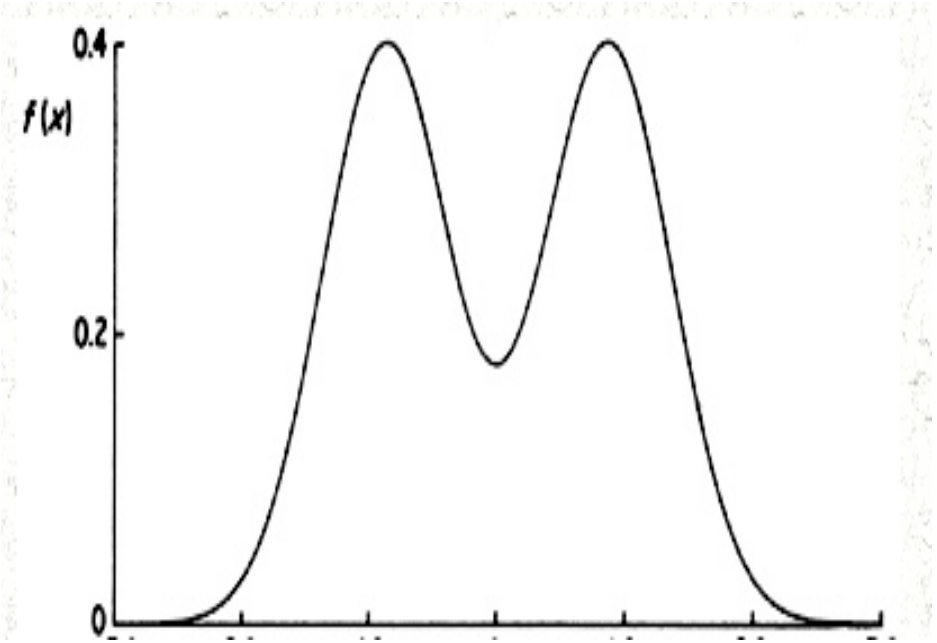
1. Zero skewness (Symmetrical or Normal distribution)
2. Right Skewness or Positive Skewness
3. Left Skewness or Negative Skewness

SYMMETRICAL OR ZERO SKEWED DISTRIBUTION

A distribution is said to be **symmetrical** or **zero skewed** if a vertical line drawn from the center of the frequency curve of distribution to the axis divides the area of the curve into two equal parts such that each part is mirror image of the other.

For symmetrical distribution, Mean = Median = Mode



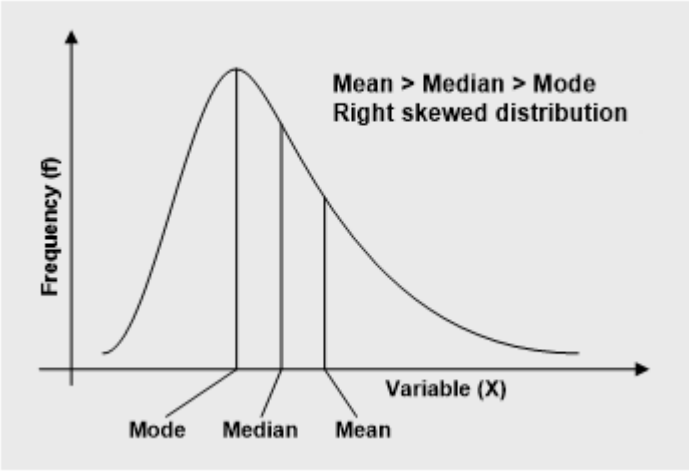


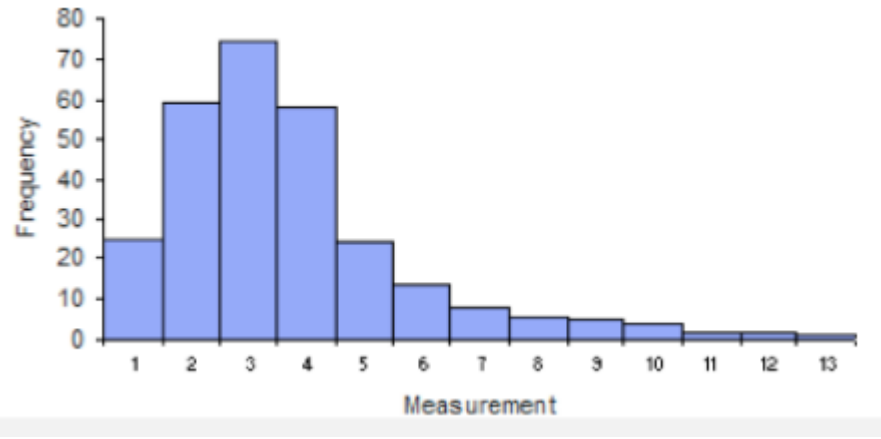
The distribution in the right is symmetrical but not normal. All symmetrical distributions are not normal.

POSITIVE SKEWNESS

If more observations lie to the left of the mean, the distribution is balanced at the mean by a long tail to the right then the distribution is said to be **positively skewed** or **right skewed**. Right tail is heavier than the left tail. The positive skewness arises when mean is increased by some unusually high value. These extremely high values pull the mean upward and frequency curve is distorted to the right.

Example: Wealth of people in the country, Marks of students in difficult exam, Salary of employee of a company, No. of children in family etc

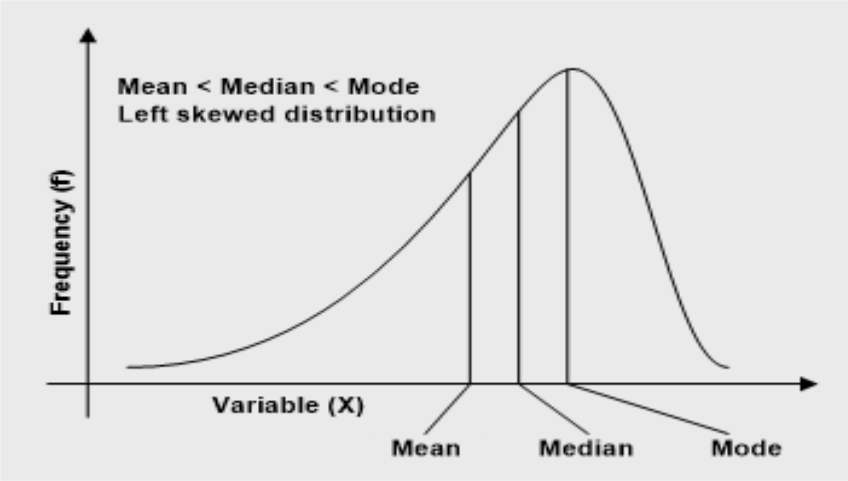


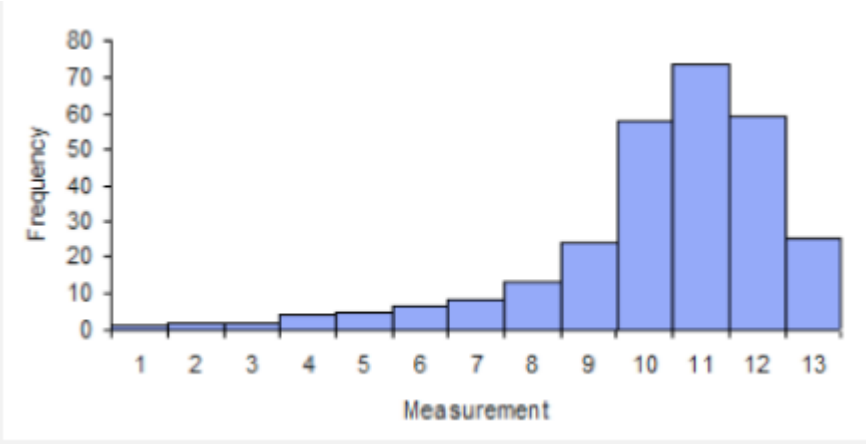


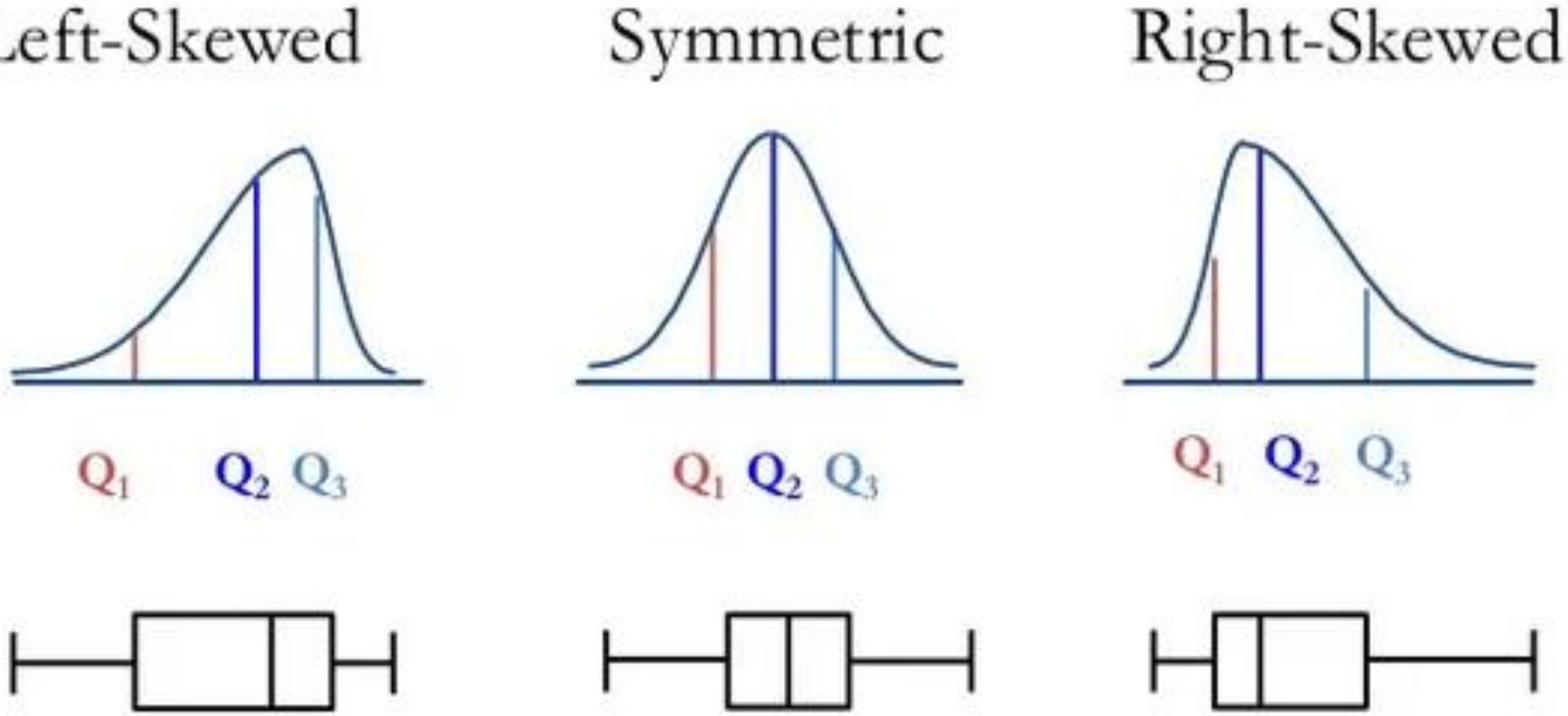
NEGATIVE SKEWNESS

If more observations lie to the right of the mean, the larger tail extends to the left then the distribution is said to be **negatively skewed** or **left skewed**. The left tail is heavier than right tail. Negative skewness occurs when mean is reduced by some extremely low values. These extremely small values pull the mean downward and the frequency curve is distorted to the left.

Example: Marks of students in easier exam, Household income of rich country, age at death etc.







MEASURES OF SKEWNESS

Karl Pearson's first coefficient of skewness

Coefficient of skewness $S_k = \frac{\bar{X} - m_o}{s}$

Notation :

$\bar{X}$ = Sample mean, m_o = Sample mode and s = sample standard deviation.

Karl Pearson's second coefficient of skewness

Coefficient of skewness $S_k = \frac{3\,(\bar{X} - m_d)}{s}$

Notation: $\bar{X}$ = Sample mean, m_d = Sample median and s = sample standard deviation

The value of S_k in the second equation lies between

-3 and +3.

i.e. $-3 \leq S_k \leq +3$

Interpretation:

1. If $S_k = 0$, then distribution is symmetrical. Any symmetrical distribution should have a coefficient of skewness near zero.
2. If $S_k > 0$, then distribution is right skewed and
3. If $S_k < 0$, then distribution is left skewed.

The sign of Sk indicates the direction of the skewness. The negative sign indicates negative skewness and positive sign indicates positive skewness. The magnitude of Sk indicates the degree of skewness like mildly, moderately, highly, extremely etc.

Example: The following are the ages of 48 patients admitted to the emergency room of a hospital. Compute coefficient of skewness . How would you describe the shape of these data?

32	63	33	57	35	54	38	53	42	51
42	48	43	46	61	53	09	13	16	16
31	30	28	28	25	23	23	22	21	17
13	30	14	29	16	28	17	27	21	24
22	23	61	55	34	42	13	26		

Solution:

Variable X = Age of patients admitted to the emergency room of a hospital

Sample size (n) = 48

Computation of sample mean

Here,

$$\sum X = 1548$$

$$\sum X^2 = 60446$$

Hence,

$$\bar{X} = \frac{\sum X}{n} = \frac{1548}{48} = 32.25 \text{ years}$$

Computation of sample SD

$$s = \sqrt{\frac{1}{n-1}\{\sum X^2 - n\,.\,\bar{X}^2\}} = \sqrt{\frac{1}{48-1}\{60446 - 48\,x\,32.25^2\}} = 14.9631 \text{ years}$$

Computation of sample median

Data Array

9	13	13	13	14	16	16	16	17	17
21	21	22	22	23	23	23	24	25	26
27	28	28	**28**	**29**	30	30	31	32	33
34	35	38	42	42	42	43	46	48	51
53	53	54	55	57	60	61	63		

Median = The value of $\frac{(n+1)}{2}$ the ordered data

= The value of 24.5 the ordered data

$$= \frac{Value\ of\ 24th\ ordered\ data + Value\ of\ 25the\ oredered\ data}{2}$$

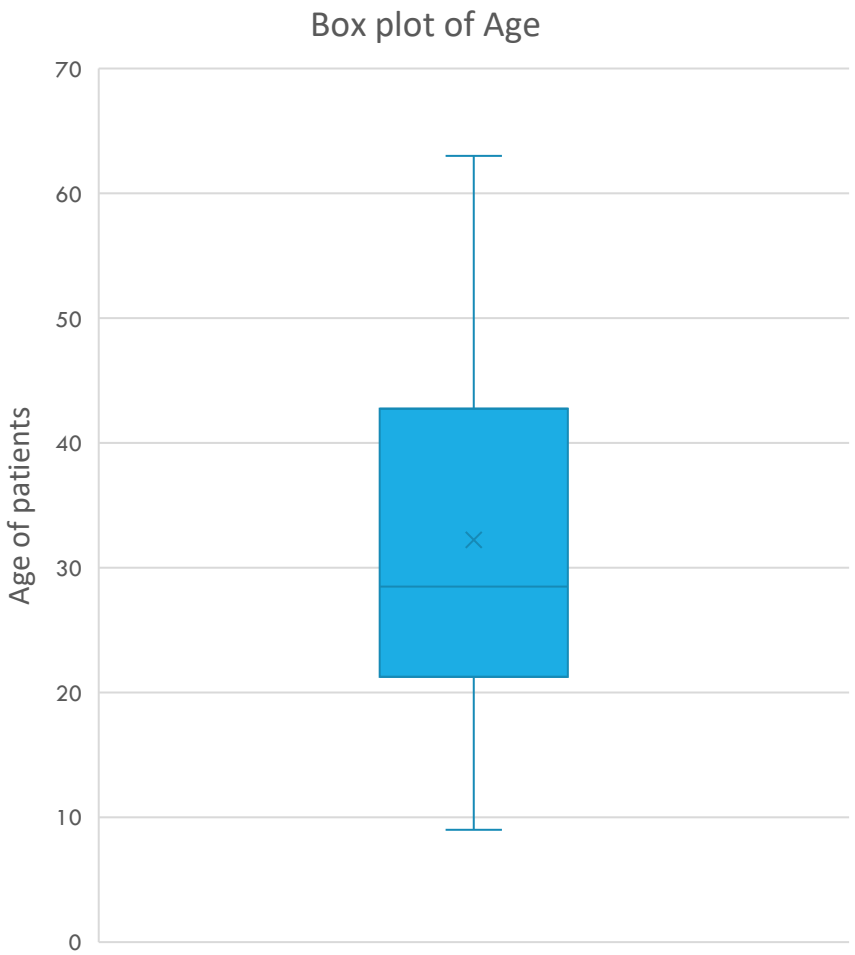
$$= \frac{28+29}{2} = 28.5 \text{ years}$$

Now, the coefficient of skewness is given by,

$$S_k = \frac{3\,(\bar{X} - m_d)}{s}$$
$$= \frac{3\,(32.25 - 28.5)}{14.9631}$$
$$= 0.7519$$

Conclusion:

The distribution of age of patients is right skewed



MEASURE OF SKEWNESS BASED ON PARTITIONING VALUES

Bowley's coefficient of skewness

$$\text{Coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2\,Q_2 + Q_1}{Q_3 - Q_1}$$

The Bowley's coefficient of skewness ranges from -1 to +1 i.e. $-1 \leq S_k(B) \leq +1$.

Kelly's coefficient of skewness

$$\text{Coefficient of skewness} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} = \frac{P_{90} - 2\,P_{50} + P_{10}}{P_{90} - P_{10}}$$

The Kelly's coefficient of skewness also ranges from $-1 \leq S_k(K) \leq +1$.

Interpretation:

In both formula $S_k = 0$ implies absence of skewness or symmetrical distribution. The value of S_k greater than 0 indicates positive skewness and value of S_k lesser than 0 indicates negative skewness.

KURTOSIS

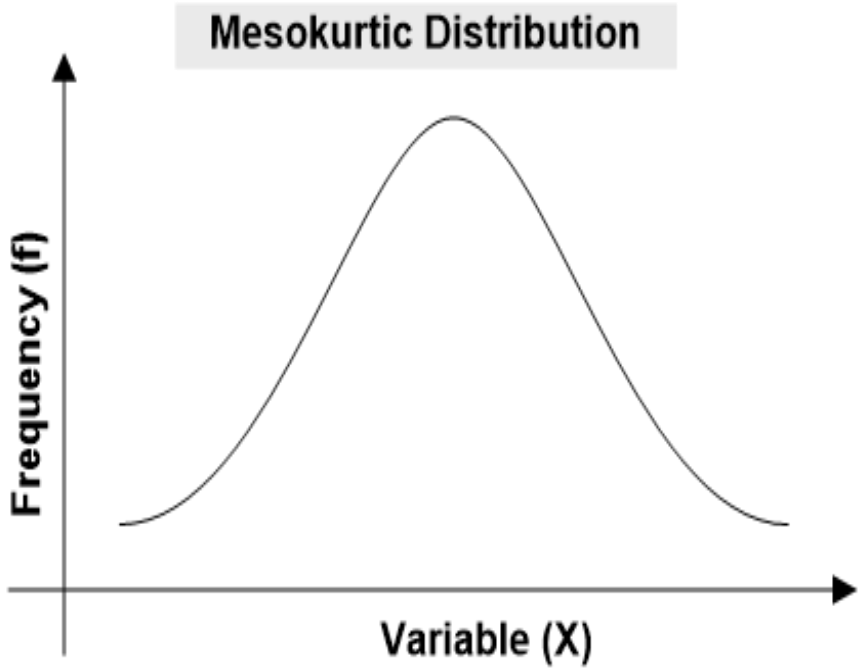
Kurtosis refers to ‘flatness’ or ‘peakedness’ of a frequency curve in comparison to normal curve. It is the fourth characteristic of a set of observations which shows degree of departure form normal distribution and it related to the size of the curve not its shape. Data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak. Kurtosis is a measure of how outlier-prone a distribution is.

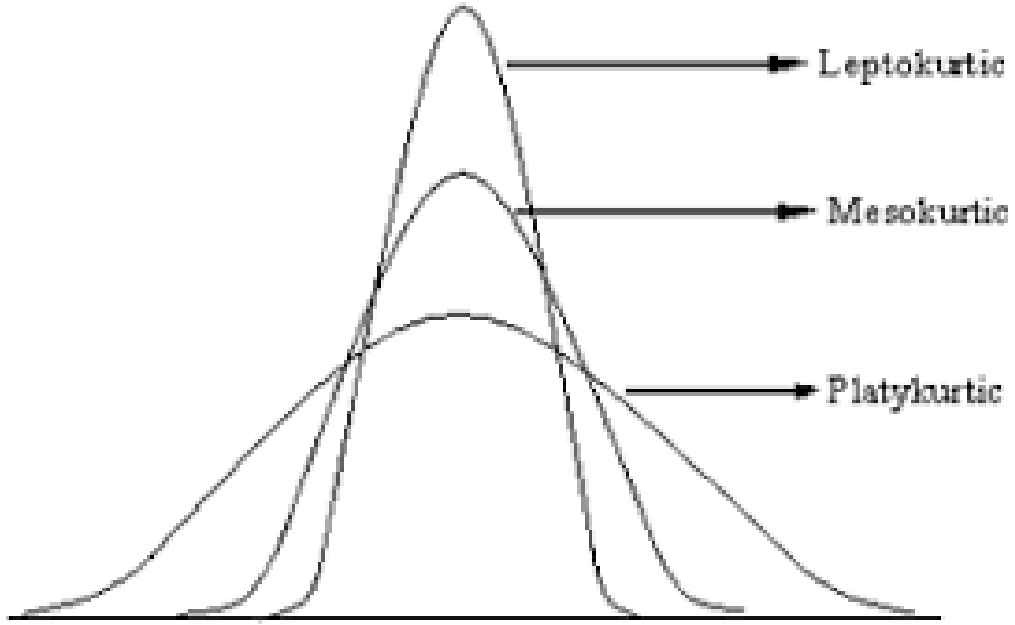
Types of Kurtosis

1. Mesokurtic Distribution (Zero kurtosis or Normal distribution)
2. Leptokurtic Distribution (High Kurtosis)
3. Platykurtic Distribution (Low Kurtosis)

MESOKURTIC DISTRIBUTION

The distribution which is not very peaked or very flat-topped is called ***Mesokurtic***. In this distribution more observations lie in the 'shoulders' and around mean. Distribution is bell shaped and is called normal distribution.

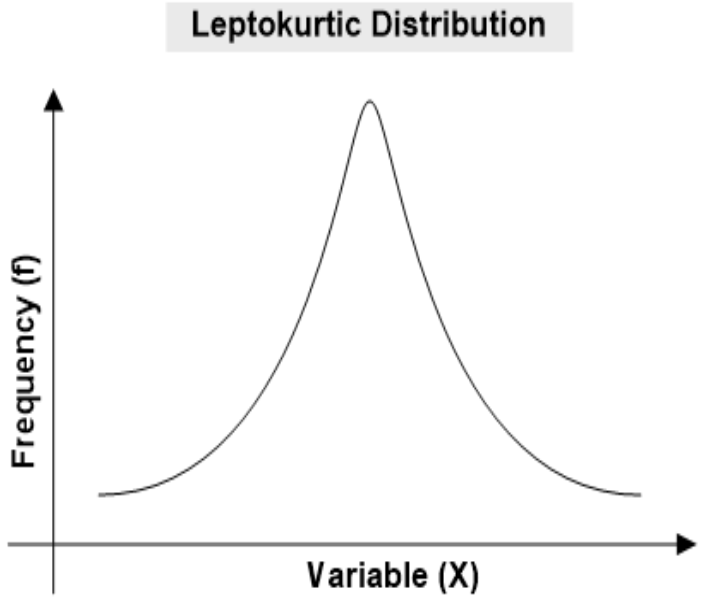




LEPTOKURTIC DISTRIBUTION (POSITIVE KURTOSIS)

"Lepto" means "slender". A distribution is said to be ***Leptokurtic*** if most of observations are concentrated near the mean and in the tails. The distribution has relatively higher peak than normal curve and has larger tails on both sides. Such a distribution might be the composite of two normal populations with the same variance but different means.

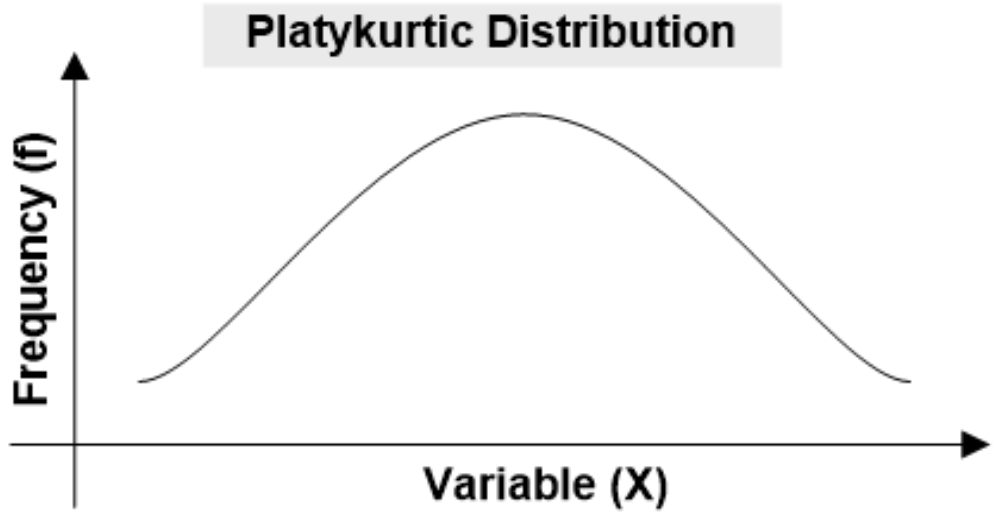
(Sharp peak and more heavier tail than normal distribution)



PLATYKURTIC DISTRIBUTION (NEGATIVE KURTOSIS)

A distribution is said to be ***Platykurtic*** if large number of observations have low frequency and spread in the mid of class-interval. Frequency curve of Platykurtic distribution is more flat-topped than normal curve. Inter-quartile range is high. Such a distribution might be the composite of two normal populations with the same mean but with different standard deviations. A uniform distribution would be the extreme case.

(Flat peak, more dispersed observations and more thinner or lighter tail than normal distribution)



MEASURES OF KURTOSIS

The kurtosis is measured with the help of Quartiles and Percentiles. Measure of kurtosis based on Quartiles and Percentiles is known as percentile coefficient of kurtosis, and is given by,

$$\text{Coefficient of Kurtosis (k)} = \frac{Semi\ Interquartile\ range}{Middle\ 80\ \%\ range} = \frac{(Q_3 - Q_1)/2}{P_{90} - P_{10}}$$

Notations:

Q_1 = First Quartile, Q_3 = Third Quartile, P_{10} = Tenth Percentile, P_{90} = Ninetieth Percentile.

Interpretation:

1. If k = 0.263, the distribution is Normal
2. If k > 0.263, the distribution is Leptokurtic
3. if k < 0.263, the distribution is Platykurtic

Example: The following are the ages of 48 patients admitted to the emergency room of a hospital. Compute coefficient of kurtosis . How would you describe the size of these data?

32	63	33	57	35	54	38	53	42	51	42
48	43	46	61	53	09	13	16	16	31	30
28	28	25	23	23	22	21	17	13	30	14
29	16	28	17	27	21	24	22	23	61	55
34	42	13	26							

Solution:

Variable X = Age of patients admitted to the emergency room of a hospital

Sample size (n) = 48

Data Array

9	13	13	13	14	16	16	16	17	17
21	21	22	22	23	23	23	24	25	26
27	28	28	28	29	30	30	31	32	33
34	35	38	42	42	42	43	46	48	51
53	53	54	55	57	60	61	63		

P_{10} = The value of $\frac{10(n+1)}{100}$ the ordered data

= The value of 4.9 th ordered data

= The value of 5^{th} ordered data = 14 yrs

$P_{25} = Q_1$ = The value of $\frac{25(n+1)}{100}$ the ordered data

= The value of 12.25 th ordered data

= The value of 12^{th} ordered data = 21 yrs

$P_{75} = Q_3$ = The value of $\frac{75(n+1)}{100}$ th ordered data

= The value of 36.75 th ordered data

= The value of 37^{th} ordered data = 43 yrs

P_{90} = The value of $\frac{90(n+1)}{100}$ th ordered data

= The value of 44.10 th ordered data

= The value or 44^{th} ordered data = 55 yrs

$$\text{Coefficient of kurtosis (k)} = \frac{Semi-interquartile\ range}{Middle\ 80\ \%\ range}$$

$$= \frac{\frac{(Q_3 - Q_1)}{2}}{P_{90} - P_{10}}$$

$$= \frac{(43-21)/2}{55-14}$$

$$= 11/41$$

$$= 0.268$$

Conclusion: Since coefficient of kurtosis k = 0.268 is slightly greater than 0.263, the distribution ages of patients is almost symmetrical.

Example: The grades of students in Statistics paper is given below:

Grade	No. of students
52 – 58	2
58 – 64	12
64 – 70	10
70 – 76	19
76 – 82	16
82 – 88	9
88 – 94	7
94 – 100	5
Total	**80**

(a) Compute coefficient of kurtosis

(b) Comment on the shape of the distribution

Solution:

Here, X = Grades of students

The coefficient of kurtosis (k) is given by,

Coefficient of kurtosis (k)

$$= \frac{Semi-interquartile\ range}{Middle\ 80\ \%\ range}$$

$$= \frac{\frac{(Q_3 - Q_1)}{2}}{P_{90} - P_{10}}$$

We need to compute P_{10}, P_{25}, P_{75} and P_{90} in order to compute coefficient of kurtosis K.

Note: $P_{25} = Q_1$ and $P_{75} = Q_3$

Grades	Frequency (f)	Cummulative Frequency (c.f.)
52 - 58	2	2
58 - 64	12	14
64 – 70	10	24
70 – 76	19	43
76 – 82	16	59
82 – 88	9	68
88 – 94	7	75
94 – 100	8	80
Total	n = 80	

Tenth Percentile

Rank order for $P_{10} = \frac{10\ x\ n}{100} = 8^{th}$ ordered data

The P_{10} is the 8^{th} ranked data which lies in the class 58 – 64, because cummulative frequency upto this class is 14. The estimated P_{10} is given by,

$$P_{10} = L_{P_{10}} + \frac{\frac{10n}{100} - c.f}{f} * h = 58 + \frac{8-2}{12} * 6 = 61$$

Twenty Fifth Percentile

Rank order for $P_{25} = \frac{25\ x\ n}{100} = 20^{th}$ ordered data

The P_{25} is the 20^{th} ranked data which lies in the class 64 – 70, because cummulative frequency upto this class is 24.

The estimated P_{25} is given by,

$$P_{25} = L_{P_{25}} + \frac{\frac{25n}{100} - c.f}{f} * h = 64 + \frac{20-14}{10} * 6 = 67.6$$

Seventy Fifth Percentile

Rank order for $P_{75} = \frac{75\ x\ n}{100} = 60^{th}$ ordered data

The P_{75} is the 60^{th} ranked data, and looking into the c.f. column the c.f. which is just greater than rank order 60 is 68 corresponding to the class 82 – 88. Hence P_{75} lies somewhere between 82 to 88

The estimated P75 is given by,

$$P_{75} = L_{P_{75}} + \frac{\frac{75n}{100} - c.f}{f} * h = 82 + \frac{60 - 59}{9} * 6 = 82.67$$

Ninetieth Percentile

Rank order for $P_{90} = \frac{90\ x\ n}{100} = 72^{th}$ ordered data

The c.f. which is just greater than rank order value 72 is 75 which is against the class grouping 88 – 94. Hence P90 falls in the class 88 – 94.

The estimated P_{90} is given by,

$$P_{90} = L_{P_{90}} + \frac{\frac{90\ n}{100} - c.f}{f} * h = 88 + \frac{72 - 68}{7} * 6 = 91.43$$

Now, the coefficient of kurtosis is given by,

$$k = \frac{\frac{(Q_3 - Q_1)}{2}}{P_{90} - P_{10}}$$
$$= \frac{(82.67-67.6)/2}{91.43 - 61}$$
$$= \frac{7.535}{30.43}$$
$$= 0.247$$

Conclusion:

Since, coefficient of kurtosis k = 0.247 < 0.263, the distribution of grades is Platykurtic

STEM AND LEAF DISPLAY

Stem-and-leaf display is a graphical method of displaying numerical data and is one of the most useful techniques of exploratory data analysis. It gives the rank order of the numerical observations in the data set as well as the shape of the distribution. It is not suitable for large data set and for continuous random variable.

Stem = The greatest common place value of the data or leading digits

Leaf = The next greatest common place value or trailing digits

Example: The following are the ages of 48 patients admitted to the emergency room of a hospital. Draw a stem-and-leaf display and comment on the size of the distribution.

32	63	33	57	35	54	38	53	42	51	42
48	43	46	61	53	09	13	16	16	31	30
28	28	25	23	23	22	21	17	13	30	14
29	16	28	17	27	21	24	22	23	61	55
34	42	13	26							

Example: The following data represent the length of life in years, measured to the nearest tenth of 30 similar fuel pumps.

2.0	3.0	0.3	3.3	1.3	0.4	0.2	6.0	5.5	6.5
0.2	2.3	1.5	4.0	5.9	1.8	4.7	0.7	4.5	0.3
1.5	0.5	2.5	5.0	1.0	6.0	5.6	6.0	1.2	0.2

Draw a stem-and-leaf display and comment on the size of the distribution.

Stem and Leaf display of Age Data

Ordered Array

9	13	13	13	14	16
16	16	17	17	21	21
22	22	23	23	23	24
25	26	27	28	28	28
29	30	30	31	32	33
34	35	38	42	42	42
43	46	48	51	53	53
54	55	57	61	61	63

Frequency	Stem	Leaf
1	0	9
9	1	333466677
15	2	112233345678889
8	3	00123458
6	4	222368
6	5	133457
3	6	113

Stem and Leaf display of fuel pump life data

Data Array

0.2	0.2	0.2	0.3	0.3	0.4
0.5	0.7	1.0	1.2	1.3	1.5
1.5	1.8	2.0	2.3	2.5	3.0
3.3	4.0	4.5	4.7	5.0	5.5
5.6	5.9	6.0	6.0	6.0	6.5

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

Conclusion: The shape of the distribution is irregular. There seems to be mix of two distribution. There is trend of two cluster; either low life or higher life