Applied Statistics I
(IMT224β/AMT224β)

Department of Mathematics
University of Ruhuna

A.W.L. Pubudu Thilan

Measures of variation

- Measures of variation determine the range of the distribution, relative to the measures of central tendency.
- The **range, mean deviation, standard deviation** and **variance** are the commonly used in measure of variation.

The range

- Range is defined as the difference between the maximum and the minimum observation of the given data.
- If x_m denotes the maximum observation x_0 denotes the minimum observation then the range is defined as
$$x_m - x_0.$$
- The range is based on the two extreme observations.

Cont...

- It gives no weight to the central values of the data.
- It is a poor measure of dispersion.
- It does not give a good picture of the overall spread of the observations with respect to the center of the observations.

Example

Find the range of the following three group.

Group A: 30, 40, 40, 40, 40, 40, 50
Group B: 30, 30, 30, 40, 50, 50, 50
Group C: 30, 35, 40, 40, 40, 45, 50

Solution

- In all the three groups the range is 50 -30 = 20.
- In group A there is concentration of observations in the center.
- In group B the observations are friendly with the extreme corner and in group C the observations are almost equally distributed in the interval from 30 to 50.
- The range fails to explain these differences in the three groups of data.

Mean deviation about mean

- The mean deviation is defined as the mean of the absolute deviations of observations from the arithmetic mean.
- Let $x_1, x_2, ..., x_n$ denote n observations. The mean deviation about mean is defined as,

$$\frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} = \frac{\sum_{i=1}^{n} |d_i|}{n}$$

where,
$\overline{x}$=maen of the data
d_i=deviation of i^{th} observation from the mean.

Example

Calculate the mean deviation form mean in respect of the marks obtained by nine students 7, 4, 10, 9, 15, 12, 7, 9, 7.

Solution

$$
\begin{aligned}
\overline{x} &= \frac{7+4+10+9+15+12+7+9+7}{9} \\
&= \frac{80}{9} \\
&= 8.89 \\
\mathrm{M.D} &= \frac{|4-8.89|+|7-8.89|+\ldots+|15-8.89|}{9} \\
&= \frac{21.11}{9} \\
&= 2.35
\end{aligned}
$$

Mean deviation about mean for summarized data

For a summarized data set with values the mean deviation about the mean

$$= \frac{\sum_{i=1}^{k} f_i \, |d_i|}{n}$$

$$\begin{aligned} n = \sum_{i=1}^{k} f_i \quad & - \quad \text{total number of observations} \\ k \quad & - \quad \text{number of different values} \\ d_i \quad & - \quad \text{deviation from the mean.} \end{aligned}$$

Example

Calculate the mean deviation of the following summarized data set.

x_i	f_i
2	1
4	4
6	6
8	4
10	1

Cont...

$$
\begin{aligned}
\text{mean} &= \frac{\sum_{i=1}^{5} f_i x_i}{\sum_{i=1}^{5} f_i} \\
&= \frac{2+16+36+32+10}{16} \\
&= \frac{96}{16} \\
&= 6
\end{aligned}
$$

Cont...

x_i	f_i	d_i	$\lvert d_i \rvert$	$f_i \lvert d_i \rvert$
2	1	-4	4	4
4	4	-2	2	8
6	6	0	0	0
8	4	2	2	8
10	1	4	4	4

The mean deviation about the mean $= \dfrac{24}{16} = 1.5$

Mean deviation about mean for summarized with classes

For a summarized data set with classes the mean deviation about the mean

$$= \frac{\sum_{i=1}^{k} f_i \left| m_i - \overline{x} \right|}{n}$$

$$\begin{aligned}
n = \sum_{i=1}^{k} f_i \quad &- \quad \text{total number of observations} \\
m_i \quad &- \quad \text{mid value of } i^{th} \text{ class} \\
k \quad &- \quad \text{number of classes} \\
f_i \quad &- \quad \text{frequency of } i^{th} \text{ class} \\
\overline{x} \quad &- \quad \text{mean.}
\end{aligned}$$

Example

Calculate the mean deviation from mean from the following data.

Size of items	frequency
3-4	3
4-5	7
5-6	22
6-7	60
7-8	85
8-9	32
9-10	8

Cont...

Size of items	f_i	m_i	$f_i m_i$	$\lvert m_i - \overline{x} \rvert$	$f_i \lvert m_i - \overline{x} \rvert$
3-4	3	3.5	10.5	3.59	10.77
4-5	7	4.5	31.5	2.59	18.13
5-6	22	5.5	121.0	1.59	34.98
6-7	60	6.5	390.0	0.59	35.40
7-8	85	7.5	637.5	0.41	34.85
8-9	32	8.5	272.0	1.41	45.12
9-10	8	9.5	76.0	2.41	19.28
Total	217		1538.5		198.5

Cont...

$$\text{mean} = \frac{\sum f_i m_i}{\sum f_i} = \frac{1538.5}{217} = 7.09$$

$$\text{M.D} = \frac{\sum_{i=1}^{7} f_i \left| m_i - \overline{x} \right|}{n} = \frac{198.53}{217} = 0.915$$

Variance

- Variance is another absolute measure of dispersion.
- It is defined as the average of the squared difference between each of the observations in a set of data and the mean.
- For a sample data the variance is denoted by S^2 and the population variance is denoted by σ^2.

Sample variance

Let $x_1, x_2, ..., x_n$ denote n observation of the sample and let $\overline{x}$ denote the sample mean. Then the sample variance S^2 is given by,

$$\begin{aligned} S^2 &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \\ S^2 &= \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}. \end{aligned}$$

Note: The positive square root of sample variance is called as sample standard deviation and it is denoted by S.

Population variance

Let $x_1, x_2, ..., x_N$ denote N observation of the population and let μ denote the population mean. Then the population variance σ^2 is given by,

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

Note: The positive square root of population variance is called as population standard deviation and it is denoted by σ.

Example 1

Calculate the variance for the following sample data: 2, 4, 8, 6, 10, and 12.

Solution

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{6} x_i}{6} = \frac{42}{6} = 7 \\
S^2 &= \frac{\sum_{i=1}^{6} (x_i - 7)^2}{6-1} \\
&= \frac{(2-7)^2 + (4-7)^2 + \ldots + (12-7)^2}{5} \\
&= \frac{70}{5} \\
&= 14
\end{aligned}
$$

Example 2

Consider the following distribution of data 10, 18, 18, 12, 11, 15, 14.

Solution

Population variation σ^2 is equal to

$$
\begin{aligned}
\sigma^2 &= \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \text{ where } N = 7 \text{ and} \\
\mu &= \frac{10+18+12+11+15+14+13}{7} \\
&= 13.29 \\
\sigma^2 &= \frac{(10-13.29)^2+(18-13.29)^2+\ldots+(14-13.29)^2}{7} \\
&= 6.204 \\
\sigma &= 2.491
\end{aligned}
$$

Sample variance for summarized data with values

The sample variance for summarized data with values is given by

$$S^2 = \sum_{i=1}^{k} \frac{f_i(x_i - \overline{x})^2}{n-1};$$

$$\begin{aligned} k \quad &- \quad \text{number of values} \\ n = \sum_{i=1}^{k} f_i \quad &- \quad \text{number of observations} \end{aligned}$$

Note: S=sample standard deviation.

Population variance for summarized data with values

The population variance for summarized data with values is given by

$$\sigma^2 = \sum_{i=1}^{k} \frac{f_i(x_i - \overline{x})^2}{N};$$

$$\begin{aligned} k &\quad - \quad \text{number of values} \\ N = \sum_{i=1}^{k} f_i &\quad - \quad \text{population size} \end{aligned}$$

Note: σ=population standard deviation.

Example

Consider the number of children in 128 families which is summarized as follows.

x_i	f_i
0	20
1	15
2	25
3	30
4	18
5	10
6	6
7	3
8	1

Find the sample variance.

Solution

x_i	f_i	$f_i x_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
0	20	0	-2.7	7.29	145.8
1	15	15	-1.7	2.89	43.35
2	25	50	-0.7	0.49	12.25
3	30	90	0.3	0.09	2.7
4	18	72	1.3	1.69	30.42
5	10	50	2.3	5.29	52.9
6	6	36	3.3	10.89	65.34
7	3	21	4.3	18.49	55.47
8	1	8	5.3	28.09	28.09
		342			436.321

Cont...

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{9} f_i x_i}{\sum_{i=1}^{9} f_i} \\
&= \frac{342}{128} \\
&= 2.67 \\
&\simeq 2.7 \\
S^2 &= \frac{\sum_{i=1}^{9} f_i (x_i - \overline{x})^2}{n-1} \\
&= \frac{436.32}{128-1} \\
&= 3.435
\end{aligned}
$$

Example

Consider the following summarized distribution of data and find the variance.

x_i	f_i
2	3
3	2
9	4
12	9
15	1
19	1

Solution

x_i	f_i	$f_i x_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$f_i(x_i - \overline{x})^2$
2	3	6	-7.5	56.25	168.75
3	2	6	-6.5	42.25	84.30
9	4	36	-0.5	0.25	1.00
12	9	108	2.5	6.25	56.25
15	1	15	5.5	30.25	30.25
19	1	19	9.5	90.25	90.25
	20				431

Cont...

$$
\begin{aligned}
\overline{x} &= \frac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \frac{190}{20} = 9.5 \\
\sigma^2 &= \frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})^2}{N} \\
&= \frac{431}{20} \\
&= 21.55
\end{aligned}
$$

For a summarized sample data with classes

$$\text{Sample variance} = S^2 = \frac{\sum_{i=1}^{k} f_i(m_i - \overline{x})^2}{n-1}$$

$$\begin{aligned} k &- \text{number of class} \\ m_i &- \text{mid value of } i^{th} \text{ class} \\ \overline{x} &- \text{mean} \\ n &= \sum_{i=1}^{k} f_i \end{aligned}$$

For a summarized sample data with classes

$$S^2 = \left[\frac{\sum_{i=1}^{k} f_i d_i^2 - \frac{\left(\sum_{i=1}^{k} f_i d_i\right)^2}{n}}{(n-1)} \right] w^2$$

w – class width of the class contains the assumed mean

f_i – frequency of i^{th} class

d_i – deviation of i^{th} class from the class that contains assumed mean.

Note: S=sample standard deviation

For a summarized distribution with classes

$$\text{Population variance} = \sigma^2 = \frac{\sum_{i=1}^{k} f_i(m_i - \overline{x})^2}{N}$$

$$\begin{aligned} k &\quad - \quad \text{number of class} \\ m_i &\quad - \quad \text{mid value of } i^{th} \text{ class} \\ \overline{x} &\quad - \quad \text{mean} \\ n &\quad = \quad \sum_{i=1}^{k} f_i \end{aligned}$$

For a summarized distribution with classes

$$\sigma^2 = \left[\frac{\sum_{i=1}^{k} f_i d_i^2 - \frac{\left(\sum_{i=1}^{k} f_i d_i\right)^2}{N}}{N} \right] w^2$$

w – class width of the class contains the assumed mean

f_i – frequency of i^{th} class

d_i – deviation of i^{th} class from the class that contains assumed mean.

Note: σ=population standard deviation

Example

The following table illustrates the pocket money given to a sample of students on a particular day of school. Calculate the variance.

Money	f_i
$12.5-<17.5$	2
$17.5-<22.5$	22
$22.5-<27.5$	19
$27.5-<32.5$	14
$32.5-<37.5$	13
$37.5-<42.5$	4
$42.5-<47.5$	6
$47.5-<52.5$	1
$52.5-<57.5$	1

Solution

Money	f_i	m_i	d_i	$f_i d_i$	$f_i d_i^2$
$12.5- <17.5$	2	15	-4	-8	32
$17.5- <22.5$	22	20	-3	-66	198
$22.5- <27.5$	19	25	-2	-38	76
$27.5- <32.5$	14	30	-1	-14	14
$32.5- <37.5$	13	35	0	0	0
$37.5- <42.5$	4	40	1	4	4
$42.5- <47.5$	6	45	2	12	24
$47.5- <52.5$	1	50	3	3	9
$52.5- <57.5$	1	55	4	4	16
	72			-103	373

Cont...

$$
\begin{aligned}
S^2 &= \left[\frac{\sum_{i=1}^{k} f_i d_i^2 - \dfrac{\left(\sum_{i=1}^{k} f_i d_i\right)^2}{n}}{(n-1)}\right] w^2 \\
&= \left[\frac{373 - \dfrac{(-103)^2}{72}}{71}\right] \times 5^2 \\
&= \left[\frac{373 - \dfrac{10609}{72}}{71}\right] \times 25 \\
&= 79.455
\end{aligned}
$$

Inter Quartile Range (IQR)

- The interquartile range (IQR) is a descriptive statistic used to summarize the extent of the spread of your data.
- The IQR is the distance between the 1^{st} quartile (25^{th} percentile) and 3^{rd} quartile (75^{th} percentile),

$$\mathrm{IQR} \;=\; Q_3 - Q_1.$$

- Fifty per cent of the measurements are between the lower quartile and the upper quartile.

Advantage and disadvantage

- **Advantage**
  More stable estimator of spread since they use two values closer to middle of the distribution that vary less from sample to sample than more extreme values.
- **Disadvantage**
  These measures are totally dependent on just two values and ignore all other observations in a data set.

Semi Inter Quartile Range

$$\begin{aligned}\text{Semi Inter Quartile Range} &= \text{Quartile deviation} \\ &= \frac{Q_1 - Q_1}{2}.\end{aligned}$$

- The semi Inter quartile range is a slightly better measure of absolute dispersion than the range.
- But it ignores the observation on the tails.

Coefficient of Quartile Deviation

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as,

$$\begin{aligned}\text{Coefficient of Quartile Deviation} &= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} \\ &= \frac{Q_3 - Q_1}{Q_3 + Q_1}\end{aligned}$$

It is pure number free of any units of measurement. It can be used for comparing the dispersion in two or more than two sets of data.

Example

The wheat production (in Kg) of 20 acres is given as:

1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, 1885.

Find the IQR, quartile deviation and coefficient of quartile deviation.

Solution

After arranging the observations in ascending order, we get

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470,
1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$
\begin{aligned}
Q_1 &= \text{Value of } \left[\frac{n+1}{4}\right]^{th} \text{ item} \\
Q_1 &= \text{Value of } \left[\frac{20+1}{4}\right]^{th} \text{ item}
\end{aligned}
$$

Cont...

$$
\begin{aligned}
Q_1 &= \text{Value of } [5.25]^{th} \text{ item} \\
Q_1 &= 5^{th} \text{ item} + 0.25(6^{th} \text{ item} - 5^{th} \text{ item}) \\
&= 1240 + 0.25(1320 - 1240) \\
&= 1240 + 20 \\
&= 1260
\end{aligned}
$$

Cont...

$$
\begin{aligned}
Q_3 &= \text{Value of } \ 3\left[\frac{n+1}{4}\right]^{th} \text{ item} \\
Q_3 &= \text{Value of } \ 3\left[\frac{20+1}{4}\right]^{th} \text{ item} \\
Q_3 &= \text{Value of } \ [15.75]^{th} \text{ item} \\
&= 15^{th} \text{ item} + 0.75(16^{th} \text{ item} - 15^{th} \text{ item}) \\
&= 1750 + 0.75(1755 - 1750) \\
&= 1753.75
\end{aligned}
$$

Cont...

$$
\begin{aligned}
\text{IQR} &= Q_3 - Q_1 = 1753.75 - 1260 = 492.75 \\
\text{QD} &= \frac{Q_3 - Q_1}{2} = \frac{492.75}{2} = 246.875 \\
\text{Coefficient of QD} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\
&= \frac{1753.75 - 1260.00}{1753.75 + 1260.00} \\
&= 0.164
\end{aligned}
$$

Five number summaries

The five number summary of a set of observations on a single variable consists of the following statistics:

- Maximum (max)
- Upper Quartile (Q3)
- Median (M)
- Lower Quartile (Q1)
- Minimum (min)

Example

Compute the five number summary for the following observations:

19 11 7 24 13 15 10 3 10 20.

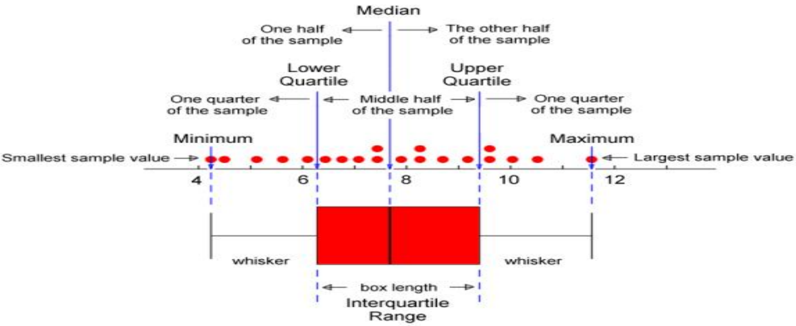
Solution

1. We order the observations 3 7 10 10 11 13 15 19 20 24.
2. The minimum and maximum are 3 and 24, respectively.
3. The median is $(11 + 13)/2 = 12$ because 11 and 13 are the two observations in the middle of the list.
4. The lower quartile is 9.25.
5. The upper quartile is 19.25.

Box-and-whisker plot

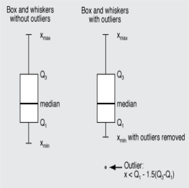
- Using box-and-whisker plot we can represent five-number summary visually.
- The length of the box is the interquartile range of the sample.
- A line is drawn across the box at the sample median.
- Whiskers sprout from the two ends of the box until they reach the sample maximum and minimum.

Cont...


Median
One half of the sample
The other half of the sample
Lower Quartile
Upper Quartile
One quarter of the sample
Middle half of the sample
One quarter of the sample
Minimum
Maximum
Smallest sample value
Largest sample value
4
6
8
10
12
whisker
whisker
box length
Interquartile Range

Outlier

- An outlier is any value that lies more than one and a half times the length of the box from either end of the box.
- That is, if a data point is below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times$ IQR, it is viewed as being too far from the central values to be reasonable.



Example

Find the outliers, if any, for the following data set:

10.2, 14.1, 14.4. 14.4, 14.4, 14.5, 14.5, 14.6, 14.7, 14.7, 14.7, 14.9, 15.1, 15.9, 16.4.

Solution

1. The median will be at position $(15 + 1) \div 2 = 8$. Then $Q_2 = 14.6$.

2. $Q_1 = 14.4$ and $Q_3 = 14.9$.

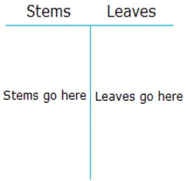
3. Then $\text{IQR} = 14.9 - 14.4 = 0.5$.

4. Outliers will be any points below Q_1 - $1.5 \times \text{IQR} = 14.4$ - $0.75 = 13.65$ or above $Q_3 + 1.5 \times \text{IQR} = 14.9 + 0.75 = 15.65$.

5. Then the outliers are at 10.2, 15.9, and 16.4.

Stem and leaf plot

- A stem and leaf plot organizes data by showing the items in order using stems and leaves.
- The leaf is the last digit on the right or the ones digits. The stem is the remaining digit or digits.
- For 12, 2 is the leaf and 1 is the stem.
- For 45.7, 7 is the leaf and 45 is the stem.

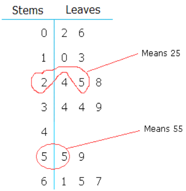


Example 1

Draw the stem and leaf plot for following data.
24, 10, 13, 2, 28, 34, 65, 67, 55, 34, 25, 59, 6, 39, 61.

Solution

1. First, put this data in order
   02, 06, 10, 13, 24, 25, 28, 34, 34, 39, 55, 59, 61, 65, 67.
2. We will use 0, 1, 2, 3, 4, 5, and 6 as stems.



Example 2

Draw the stem and leaf plot for following data.
104, 107, 112, 115, 115, 116, 123, 130, 134, 145, 147.

Solution

1. This time, the data is already in order.
2. We will use 10, 11, 12, 13, and 14 as stems.

Stems	Leaves
10	4 7
11	2 5 5 6
12	3
13	0 4
14	5 7

Means 145

Example 3

Draw the stem and leaf plot for following two groups.

$$\text{Grade for class A} \quad : \quad 60, 68, 70, 75, 84, 86, 90,$$
$$91, 92, 94, 94, 96, 100, 100$$
$$\text{Grade for class B} \quad : \quad 60, 60, 70, 71, 73, 73, 75,$$
$$76, 77, 84, 85, 86, 91, 92$$

The plot is displayed as:

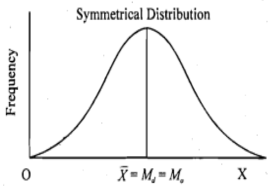
Class A Leaves	Stems	Class B Leaves
8 0	6	0 0
5 0	7	0 1 3 3 5 6 7
6 4	8	4 5 6
6 4 4 2 1 0	9	1 2
0 0	10	

Measures of skewness

- The need to study these concepts arises from the fact that the measures of central tendency and dispersion fail to describe a distribution completely.
- It is possible to have frequency distributions which differ widely in their nature and composition and yet may have same central tendency and dispersion.
- Thus, there is need to supplement the measures of central tendency and dispersion.

Concept of skewness

- The skewness of a distribution is defined as the lack of symmetry.
- In a symmetrical distribution, the Mean, Median and Mode are equal to each other.

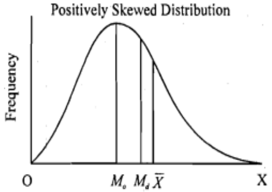


Cont...

- The presence of extreme observations on the right hand side of a distribution makes it positively skewed.
- We shall in fact have

$$\text{Mean} > \text{Median} > \text{Mode}$$

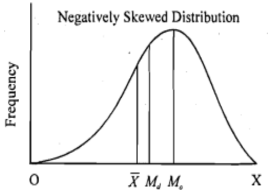
when a distribution is positively skewed.



Cont...

- On the other hand, the presence of extreme observations to the left hand side of a distribution make it negatively skewed and the relationship between mean, median and mode is:

$$\text{Mean} < \text{Median} < \text{Mode}.$$



Different measures of skewness

The direction and extent of skewness can be measured in various ways. We shall discuss three measures of skewness in this section.

1. Karl Pearson's coefficient of skewness
2. Bowley's coefficient of skewness
3. Kelly's coefficient of skewness

Karl Pearson's coefficient of skewness

- The mean, median and mode are not equal in a skewed distribution.
- The Karl Pearson's measure of skewness is based upon the divergence of mean from mode in a skewed distribution.

$$S_{kp_1} = \frac{\text{mean - mode}}{\text{standard deviation}} \quad \text{or} \quad S_{kp_2} = \frac{3(\text{mean - median})}{\text{standard deviation}}$$

Properties

1. $-1 \leq S_{kp} \leq 1$.
2. $S_{kp} = 0 \Rightarrow$ distribution is symmetrical about mean.
3. $S_{kp} > 0 \Rightarrow$ distribution is skewed to the right.
4. $S_{kp} < 0 \Rightarrow$ distribution is skewed to the left.

Advantage and disadvantage

Advantage

S_{kp} is independent of the scale. Because (mean-mode) and standard deviation have same scale and it will be canceled out when taking the ratio.

Disadvantage

S_{kp} depends on the extreme values.

Example

Calculate Karl Pearson's coefficient of skewness of the following data set ($S = 1.7$).

Value (x)	1	2	3	4	5	6	7
Frequency (f)	2	3	4	4	6	4	2

Solution

$$
\begin{aligned}
\text{mean} = \overline{x} &= \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i}^{n} f_i} \\
&= \frac{1 \times 2 + 2 \times 3 + \ldots + 7 \times 2}{25} \\
&= \frac{104}{25} \\
&= 4.16 \\
\text{mode} &= 5 \\
S_{kp} &= \frac{\text{mean-mode}}{\text{standard deviation}} \\
&= \frac{4.16 - 5}{1.7} = -0.4941
\end{aligned}
$$

Since $S_{kp} < 0$ distribution is skewed left.

Bowley's coefficient of skewness

- This measure is based on quartiles.
- For a symmetrical distribution, it is seen that Q_1, and Q_3 are equidistant from median (Q_2).
- Thus $(Q_3 - Q_2) - (Q_2 - Q_1)$ can be taken as an absolute measure of skewness.

$$
\begin{aligned}
S_{kq} &= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \\
&= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}
\end{aligned}
$$

Properties

1. $-1 \leq S_{kq} \leq 1$.
2. $S_{kq} = 0 \Rightarrow$ distribution is symmetrical about mean.
3. $S_{kq} > 0 \Rightarrow$ distribution is skewed to the right.
4. $S_{kq} < 0 \Rightarrow$ distribution is skewed to the left.

Advantage and disadvantage

Advantage
S_{kq} does not depend on extreme values.

Disadvantage
S_{kq} does not utilize the data fully.

Example

The following table shows the distribution of 128 families according to the number of children.

No of children	No of families
0	20
1	15
2	25
3	30
4	18
5	10
6	6
7	3
8 or more	1

Find the Bowley's coefficient of skewness.

Cont...

No of children	No of families	Cumulative frequency
0	20	20
1	15	35
2	25	60
3	30	90
4	18	108
5	10	118
6	6	124
7	3	127
8 or more	1	128

Cont...

$$
\begin{aligned}
Q_1 &= \left(\frac{128+1}{4}\right)^{th} \text{ observation} \\
&= (32.25)^{th} \text{ observation} \\
&= 1 \\
Q_2 &= \left(\frac{128+1}{2}\right)^{th} \text{ observation} \\
&= (64.5)^{th} \text{ observation} \\
&= 3 \\
Q_3 &= 3\left(\frac{128+1}{4}\right)^{th} \text{ observation} \\
&= (96.75)^{th} \text{ observation} \\
&= 4
\end{aligned}
$$

Cont...

$$\begin{aligned} S_{kq} &= \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} \\ &= \frac{4 + 1 - 2 \times 3}{4 - 1} \\ &= -\frac{1}{3} \\ &= -0.333 \end{aligned}$$

Since $S_{kq} < 0$ distribution is skewed left.

Kelly's coefficient of skewness

- Bowley's measure of skewness is based on the middle 50% of the observations because it leaves 25% of the observations on each extreme of the distribution.
- As an improvement over Bowley's measure, Kelly has suggested a measure based on P_{10} and, P_{90} so that only 10% of the observations on each extreme are ignored.

$$
\begin{aligned}
S_p &= \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{(P_{90} - P_{50}) + (P_{50} - P_{10})} \\
&= \frac{P_{90} + P_{10} - 2P_{50}}{P_{90} - P_{10}}.
\end{aligned}
$$

Properties

1. $-1 \leq S_p \leq 1$.
2. $S_p = 0 \Rightarrow$ distribution is symmetrical about mean.
3. $S_p > 0 \Rightarrow$ distribution is skewed to the right.
4. $S_p < 0 \Rightarrow$ distribution is skewed to the left.

Remark

It may be noted here that although the coefficient S_{kp}, S_{kq} and S_p are not comparable, however, in the absence of skewness, each of them will be equal to zero.

Thanks