Lecture 4

Measures of Central Tendency
Mode

The mode, if one exists, is the **most frequently occurring value**.

Mode for Individual Observation

EXAMPLE

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77, 82, **74**, 81, 79, 84, **74**, 78

Find the mode.

Solution In this data set, 74 occurs twice, and each of the remaining values occurs only once. Because 74 occurs with the highest frequency, it is the mode.

Therefore,

Mode =**74 miles per hour**

EXAMPLE

The following data give the speeds (in miles per hour) of eight cars that were stopped on I-95 for speeding violations.

77, **82, 74**, 81, 79, 84, **74**, 78, **82**

Find the mode.

Solution In this data set, 74 and 82 occur twice, and each of the remaining values occurs only once. Because 74, 82 occur with the highest frequency, these are the modes.

Therefore,

Modes =**74 and 82 miles per hour**

Note:

- A major shortcoming of the mode is that a data set **may have none or may have more than one mode**, whereas it will have only one mean and only one median.
- For instance, a data set with each value occurring only once or equal number of times has no mode.
- A data set with only one value occurring with the highest frequency has only one mode. The data set in this case is called **unimodal**.
- A data set with two values that occur with the same (highest) frequency has two modes. The distribution, in this case, is said to be **bimodal**.
- If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be **multimodal**.

Calculating the Mode from Ungrouped Frequency Distribution

EXAMPLE: The following table shows the number of family members of 30 families.

Number of family members	Number of families
2	2
3	8
4	10
5	6
6	3
7	1

Find the mode for these data.

Solution: In this data set 4 occurs with highest frequency (10), so the mode of this data set is 4. Thus, the mode number of family members is 4 and it is unimodal distribution.

Calculating the Mode from Grouped Frequency Distribution

When data are already grouped in a frequency distribution, we must assume that the mode is located in the class with the most frequently occurred items, that is, the class with the highest frequency. To determine a single value for the mode from this class, we use the following equation:

$$\text{Mode} = L_{M_0} + \left(\frac{d_1}{d_1 + d_2}\right)w$$

where

L_{M_0} = lower limit of the modal class

d_1 = frequency of the modal class minus the frequency of the class *preceding* the modal class

d_2 = frequency of the modal class minus the frequency of the class following the modal class.

w = width of the modal class interval

EXAMPLE: Calculate the mode days of maturity of 40-short term investments.

Table:

Class interval	Frequency (f_i)	Cumulative frequency
30—39	3	3
40—49	1	4
50—59	8	12
60—69	10	**22**
70—79	7	29
80—89	7	36
90—99	4	40
Total	40	

Solution:

The modal class is 60—69, L_{M_0} =60, d_1= 10-8= 2, d_2= 10 – 7 = 3, w = 10.

$$\text{Mode} = L_{M_0} + \left(\frac{d_1}{d_1 + d_2}\right)w = 60 + \left(\frac{2}{2+3}\right)10 = 60 + 4 = 64.0$$

Thus, the mode number of days to mature the short-term investment is 64 days.

Overall Observations

- To sum up, we cannot say for sure which of the three measures of central tendency is a better measure overall.
- Each of them may be better under different situations.

- Probably the mean is the most-used measure of central tendency for a data without outliers, followed by the median.
- The mean has the advantage that its calculation includes each value of the data set.
- The median is a better measure when a data set includes outliers.
- The mode is simple to locate, but it is not of much use in practical applications.
- The mean and median can be used only for quantitative data set but mode can be used both quantitative and qualitative data set.

Relationships Among the Mean, Median, and Mode

- A histogram or a frequency distribution curve can assume a symmetric and skewed.
- Now we describe the relationships among the mean, median, and mode for three such histograms and frequency distribution curves.

- Knowing the values of the mean, median, and mode can give us some idea about the shape of a frequency distribution curve.

1. For a **symmetric histogram** and frequency distribution curve with one peak, the values of **the mean, median, and mode are identical**, and they lie at the **center of the distribution**.

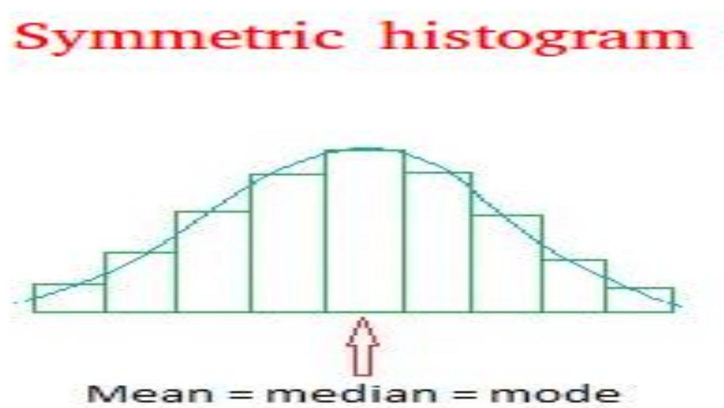
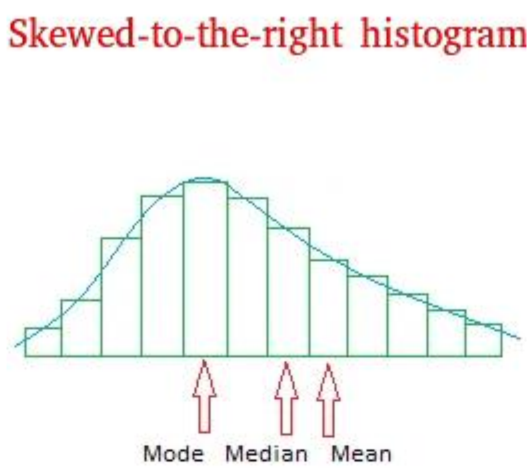


Figure: Mean, median, and mode for a symmetric histogram and frequency distribution curve.

2. For a histogram and a frequency distribution curve **skewed to the right** (see the following Figure), the **value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two**.

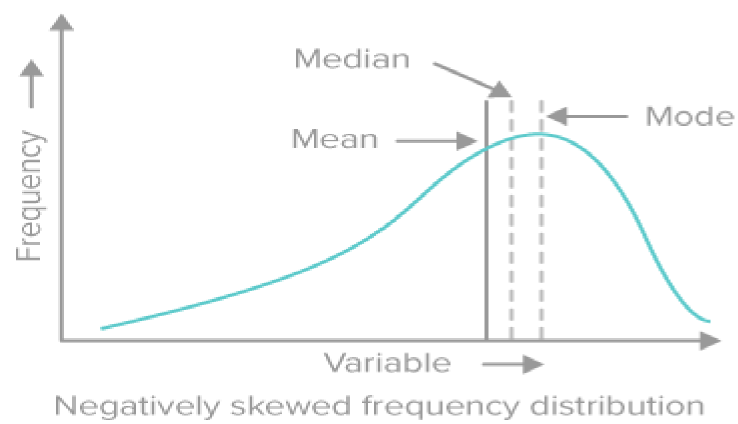
(Notice that the mode always occurs at the peak point.) The value of the **mean is the largest in this case because it is sensitive to outliers** that occur in the right tail. These outliers pull the mean to the right.



Mode < Median< Mean

Figure: Mean, median, and mode for a histogram and frequency distribution curve skewed to the right.

3. If a histogram and a frequency distribution curve are **skewed to the left** (see the following Figure), the value of the **mean is the smallest and that of the mode is the largest, with the value of the median lying between these two.** In this case, the outliers in the left tail pull the mean to the left.



Mean < Median < Mode

Figure: Mean, median, and mode for a frequency distribution curve skewed to the left.

Geometric mean

- The **Geometric Mean (GM)** is the **average value or mean** which signifies the central tendency of the set of numbers by taking the **root of the product of their values**.
- Basically, we **multiply the 'n' values altogether** and **take out the n^{th} root** of the numbers, where n is the total number of values.
- For example: for a given set of two numbers such as 8 and 1, the geometric mean is equal to $\sqrt{(8\times1)} = \sqrt{8} = 2\sqrt{2}$.

- Note that this is **different from the arithmetic mean**.
  - ✓ In the arithmetic mean, data values are added and then divided by the total number of values.
  - ✓ But in geometric mean, the given data values are multiplied, and then you take the root with the radical index for the final product of data values.
- Applications of the geometric mean in finance include
  - ✓ compound interest over several years,
  - ✓ total sales growth, and
  - ✓ population growth.
- An important question concerns the **average growth each year that will result in a certain total growth over several years**.

- The geometric mean, $\bar{x}_g$ is the nth root of the product of n numbers:

$$\bar{x}_g = \sqrt[n]{x_1 \times x_2 \times \ldots\ldots\times x_n} = (x_1 \times x_2 \times \ldots\ldots\times x_n)^{\frac{1}{n}}$$

- The **geometric mean is used** to obtain **mean growth over several periods, given compounded growth from each period**:

For example, the geometric mean of

1.05 1.02 1.10 1.06 is

$$\bar{x}_g = [(1.05)(1.02)(1.10)(1.06)]^{\frac{1}{4}} = 1.0571$$

➢ When the **numbers are large,** we can make easy calculation by **taking logarithm on both sides** as

$$log\bar{x}_g = \frac{1}{n}[logx_1 + logx_2 + \cdots \dots \dots \dots \dots . + logx_n] = \frac{\sum logx_i}{n}$$

Then $\bar{x}_g$ =Antilog $\frac{\sum logx_i}{n}$

Note

Geometric mean **cannot be used** when the data set **contains 0 or negative** values.

Geometric mean for Frequency Distribution

For a frequency distribution (both group and ungrouped), the geometric mean G.M. is

$$GM = (x_1^{f_1} . x_2^{f_2} \dots \dots \dots \dots x_k^{f_k})^{1/n}, \quad \text{where } n = \sum f_i$$

Taking logarithms on both sides, we get

$$\log GM = \frac{1}{n}(f_1 \log x_1 + f_2 \log x_2 + \cdots \dots \dots . . + f_k \log x_k) = \frac{1}{n}\sum_{i=1}^{k} f_i \log x_i$$

Thus, $GM = Antilog\ \frac{1}{n}\sum_{i=1}^{k} f_i \log x_i$

Example: Find the geometric mean of the following grouped data for the frequency distribution of weights.

Weights of ear heads (g)	No of ear heads (f)
60-80	22
80-100	38
100-120	45
120-140	35
140-160	20
Total	160

Solution:

Weights of ear heads (g)	No of ear heads (f)	Mid x	Log x	f log x
60-80	22	70	1.845	40.59
80-100	38	90	1.954	74.25
100-120	45	110	2.041	91.85
120-140	35	130	2.114	73.99
140-160	20	150	2.176	43.52
Total	160			324.2

From the given data, $n = 160$

We know that the G.M for the grouped data is

$$GM = Antilog\ \frac{1}{n}\sum_{i=1}^{k} f_i \log x_i$$

$$GM = Antilog\ (324.2/160)$$

$$GM = 106.23$$

Advantages of Geometric Mean

- A geometric mean is based upon all the observations
- It is rigidly defined
- The fluctuations of the observations do not affect the geometric mean

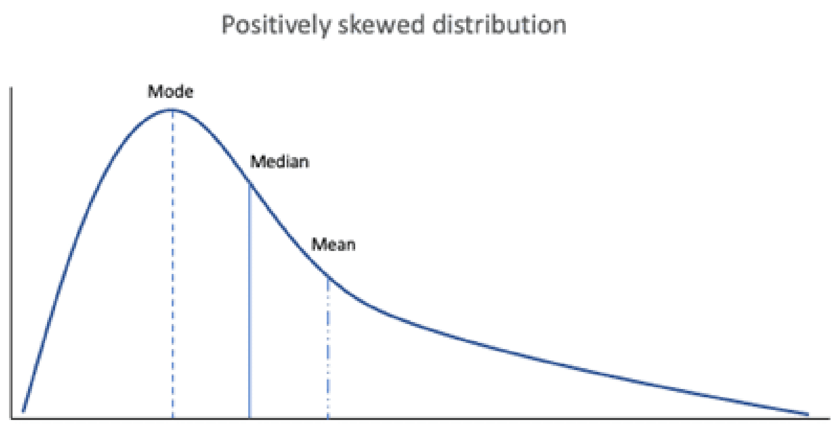
Disadvantages of Geometric Mean

- A geometric mean is not easily understandable by a non-mathematical person
- If any of the observations is zero, the geometric mean becomes zero
- If any of the observation is negative, the geometric mean becomes imaginary

When is the geometric mean better than the arithmetic mean?

- Even though it's less commonly used, the **geometric** mean is more accurate than the arithmetic mean for **positively skewed data**.
- In a **<u>positively skewed</u> distribution**, there's a **cluster of lower scores** and a **spread-out tail on the right**.
- Income distribution is a common example of a skewed dataset.

- While **most values tend to be low**, the arithmetic mean is often **pulled upward** (or rightward) by high values or <u>outliers</u> in a positively skewed dataset.



- Because the **geometric mean tends to be lower than the arithmetic mean,** it represents smaller values better than the arithmetic mean.

- The geometric mean is most appropriate for <u>ratio</u> levels of measurement, where <u>variables</u> have a true zero and don't take on any negative values.

Harmonic Mean (H.M.)

Harmonic Mean is defined as the reciprocal of the arithmetic mean of reciprocals of the observations.

H.M. for Raw data

Let $x_1, x_2, \ldots, xn$ be the n observations then the harmonic mean is defined as

$$HM = \frac{1}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

Example:

A man travels from Dhaka to Rajshahi by a car and takes 4 hours to cover the whole distance. In the first hour he travels at a speed of 50 km/hr, in the second hour his speed is 64 km/hr, in third hour his speed is 80 km/hr and in the fourth hour he travels at the speed of 55 km/hr. Find the average speed of the motorist.

Solution:

x	50	65	80	55	**Total**
$1/x$	0.0200	0.0154	0.0125	0.0182	**0.0661**

$$\text{H. M.} = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$$

$$= \frac{4}{0.0661} = 60.5 \text{ km/hr}$$

Average speed of the motorist is 60.5km/hr

H.M. for Ungrouped Frequency Distribution:

For a frequency distribution

$$\text{H. M.} = \frac{N}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)}$$

Example

The following data is obtained from the survey. Compute H.M

Speed of the car	130	135	140	145	150
No of cars	3	4	8	9	2

Solution:

x_i	f_i	$\frac{f_i}{x_i}$
130	3	0.0231
135	4	0.0091
140	8	0.0571
145	9	0.0621
150	2	0.0133
Total	N = 26	0.1648

$$\text{H. M.} = \frac{N}{\sum_{i=l}^{n} f_i\left(\frac{1}{x_i}\right)}$$

$$= \frac{26}{0.1648}$$

$$\text{H.M} = 157.77$$

H.M. for Grouped Frequency Distribution:

$$\text{The Harmonic mean H.M.} = \frac{N}{\sum_{i=l}^{n} f_i\left(\frac{1}{x_i}\right)}$$

Where *xi* is the mid-point of the class interval

Example 5.13
Find the harmonic mean of the following distribution of data

Dividend yield (percent)	2 – 6	6 – 10	10 – 14
No. of companies	10	12	18

Solution:

Class Intervals	Mid-value (x_i)	No. of companies (f_i)	Reciprocal ($1/x_i$)	f_i ($1/x_i$)
2 – 6	4	10	¼	2.5
6 – 10	8	12	1/8	1.5
10 – 14	12	18	1/12	1.5
Total		N = 40		5.5

The harmonic mean is $\text{H.M.} = \dfrac{N}{\sum_{i=l}^{n} f_i\left(\dfrac{1}{x_i}\right)} = \dfrac{40}{5.5} = 7.27$

Limitations of H.M:

- It is difficult to calculate and is not understandable
- All the values must be available for computation
- It is not popular due to its complex calculation.
- It is usually a value which does not exist in series

When to use?

- *Harmonic mean is used to calculate the average value when the values are expressed as value/unit.*
- *Since the speed is expressed as km/hour, harmonic mean is used for the calculation of average speed.*

Relationship among the averages:
In any distribution when the original items are different the A.M., G.M. and H.M would also differ and will be in the following order:

A.M. ≥ G.M ≥ H.M

Measures of Dispersion

- The measures of central tendency, such as the mean, median, and mode, do not reveal the whole picture of the distribution of a data set.
- Two data sets with the **same mean may have completely different spreads.**
- The variation among the values of observations for one data set may be much larger or smaller than for the other data set.

Example 1:

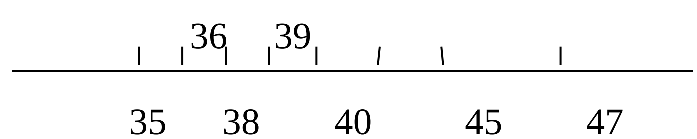
Consider the following two data sets on the ages (in years) of all workers working for each of two small companies.

Company 1: 47, 38, 35, 40, 36, 45, 39
Mean =280/7 = 40

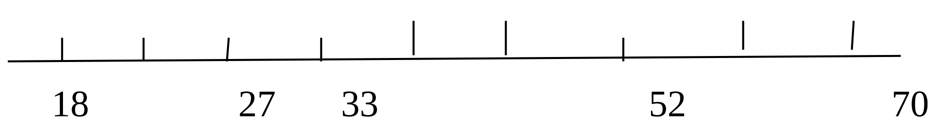
Company 2: 70, 33, 18, 52, 27
Mean = 200/5=40

- The **mean age of workers** in both these companies is the **same**, 40 years.
- If we do not know the ages of individual workers at these two companies and are told only that the mean age of the workers at both companies is the same, we may deduce that the workers at these two companies have a similar age distribution.
- As we can observe, however, the variation in the workers' ages for each of these two companies is very different.

Company 1



Company 2



Example 2:

The platelet counts (×109 / L) of 12 patients in two wards were measured. The resultant values were as follows:

Ward A: 186, 191, 199, 200, 209, and 215 $\bar{x} = 200$
Ward B: 160, 185, 190, 204, 217, and 244 $\bar{x} = 200$

- Both of the groups had a mean platelet count of 200×109 / L.
- However, there was a large difference between the two groups in terms of the dispersion of the data. That is, the platelet counts for patients in ward B were more widely spread out compared with those for patients in ward A.

- Thus, the mean, median, or mode by itself is usually not a sufficient measure to reveal the shape of the distribution of a data set.
- We also need a measure that can provide some information about the variation among data values.
- The measures that help us learn about the spread of a data set are called the **measures of dispersion**.
- **The measures of central tendency and dispersion taken together give a better picture of a data set than the measures of central tendency alone.**

Different Measures of Dispersion:

(i) Range
(ii) Interquartile range
(iii) Variance
(iv) Standard deviation and
(v) Coefficient of variation

Range

The **range** is the simplest measure of dispersion to calculate. It is obtained by taking the difference between the largest and the smallest values in a data set.

Finding the Range for Ungrouped Data

Range = Largest value - smallest value

Thus, the ranges ages of workers in Company 1 and Company 2 in example 1 are:

Range of ages for company 1 in example 1: R1 = 47-35 =12
Range of ages for company 2 in example 1: R2 = 70-18 = 52

The results indicate that ages of workers in company 1 spread over a range of 12 years and the ages of workers in company 2 spread over a range of 52 years.

Again, ranges of platelet count of patients in ward A and B are as follows:

Range in example 2 for ward A: R_A= 215 - 186 = 29

Range in example 2 for ward B: R_B= 244 – 160 = 84

The results indicate that the range of platelet count in ward B was larger than that in ward A, although the two wards had the same mean.

Advantages

- A prime advantage of range is that it is easy to calculate and easy to understand.
- Moreover, range is measured in the same units as the original data; thus, range has a direct interpretation.

Disadvantages:

- The range, like the mean, has the disadvantage of being influenced by outliers. Consequently, the range is not a good measure of dispersion to use for a data set that contains outliers.
- Another disadvantage of using the range as a measure of dispersion is that its calculation is based on two values only: the largest and the smallest. All other values in a data set are ignored when calculating the range. Thus, the range is not a very satisfactory measure of dispersion.

ii. Interquartile Range

- Before calculating interquartile range, we have to know about **Quartile, Decile and Percentile.**
- The median divides all ordered values equally into two parts. Similarly, **the values can be equally divided into a larger number of parts if desired**.

Quartile, Decile and Percentile

- **Quartiles: distribution is divided into quarters (into four parts)**.
- **Deciles: distribution is divided into tenths (into ten equal parts).**
- **Percentile: distribution is divided into hundredths (into hundred equal parts)**.

Calculation of quartile:

If the observations are **sorted from smallest to largest** and equally divided into 4 parts, the corresponding value of the proportion is called the quartiles, which is denoted by the symbol Q_i *(i=1,2,3)*. For frequency-table data, Q_i is calculated as

$$Q_i = l + \frac{h}{f}\left(\frac{in}{4} - c\right) \; ; \qquad i = 1,2,3$$

where,

l = Lower limit of the ith quartile class,
h = width of the ith quartile class,
f = frequency of the ith quartile class
c = cumulative frequency of the class preceding the ith quartile class
n = total frequency

Calculate quartiles of 40-short term investments.

Table 1:

Class interval	Frequency (f_i)	Cumulative frequency
30—39	3	3
40—49	1	4
50—59	8	12
60—69	10	22
70—79	7	29
80—89	7	36
90—99	4	40
Total	40	

First, calculate $\frac{i(n+1)}{4}$.

For first quartile Q_1, $\frac{(n+1)}{4} = \frac{41}{4} = 10.25$. As shown in Column 3 of Table, the cumulative frequency of the group that the first quartile lies in group 50-59. Therefore, l=50, h = 10, f =8, c = 4.

Thus

$$Q_1 = 50 + \frac{10}{8}(10 - 4) = 50 + 7.5 = 57.5$$

For third quartile Q_3, $\frac{3(n+1)}{4} = \frac{3\times 41}{4} = 30.75$. As shown in Column 3 of Table, the cumulative frequency of the group that the third quartile lies in group 80-89. Therefore, l=80, h = 10, f =7, c = 29.

Thus

$$Q_3 = 80 + \frac{10}{7}(30 - 29) = 80 + 1.4 = 81.4$$

ii. Interquartile Range

The interquartile range is defined as

IQR = $Q_3 - Q_1 = 81.4 - 57.5 = 23.9$

Calculation of Decile:

If the observations are sorted from smallest to largest and equally divided into 10 parts, the corresponding value of the proportion is called the deciles, which is denoted by the symbol D_i. For frequency-table data, D_i is calculated as

$$D_i = l + \frac{h}{f}\left(\frac{in}{10} - c\right)$$

where

l = Lower limit of the ith decile class,

h = width of the ith decile class,

f = frequency of the ith decile class

c = cumulative frequency of the class preceding the ith decile class

n = total frequency

Calculate third and sixth deciles

First, calculate $\frac{i(n+1)}{10}$.

For third decile D_3, $\frac{3(n+1)}{10} = \frac{3\times 41}{10} = 12.3$. As shown in Column 3 of Table, the cumulative frequency of the group that the third decile lies in group 50-59. Therefore, l=50, h = 10, f =8, c = 4.

Thus,

$$D_3 = 50 + \frac{10}{8}(12 - 4) = 50 + 10 = 60$$

For sixth quartile D_6, $\frac{6(n+1)}{10} = \frac{6\times41}{10} = 24.6$. As shown in Column 3 of Table, the cumulative frequency of the group that the sixth decile lies in group 70-79. Therefore, l=70, h = 10, f =7, c = 22.

Thus,

$$Q_6 = 70 + \frac{10}{7}(24 - 22) = 70 + 2.9 = 72.9$$

Calculation of Percentile:

If the observations are sorted from smallest to largest and equally divided into 100 parts, the corresponding value of the proportion is called the percentile, which is denoted by the symbol P_i. For frequency-table data, P_i is calculated as

$$P_i = l + \frac{h}{f}\left(\frac{in}{100} - c\right)$$

where
l = Lower limit of the ith percentile class,
h = width of the ith percentile class,
f = frequency of the ith percentile class
c = cumulative frequency of the class preceding the ith percentile class
n = total frequency

Calculate 25th and 75th percentile and find interquartile range

First, calculate $\frac{i(n+1)}{100}$.

For 25th percentile P_{25}, $\frac{25(n+1)}{100} = \frac{25\times41}{100} = 10.25$. As shown in Column 3 of Table, the cumulative frequency of the group that the 25th percentile lies in group 50-59. Therefore, l=50, h = 10, f =8, c = 4.

Thus,

$$P_{25} = 50 + \frac{10}{8}(10 - 4) = 50 + 7.5 = 57.5$$

For 75th percentile P_{75}, $\frac{75(n+1)}{100} = \frac{75\times41}{100} = 30.75$. As shown in Column 3 of Table, the cumulative frequency of the group that the sixth decile lies in group 80-89. Therefore, l=80, h = 10, f =7, c = 29.

Thus,

$$P_{75} = 80 + \frac{10}{7}(30 - 29) = 80 + 1.4 = 81.4$$

ii. Interquartile Range

The interquartile range is defined as

IQR = $P_{75} - P_{25} = 81.4 - 57.5 = 23.9$

Variance and Standard Deviation

- The **standard deviation** is the most-used measure of dispersion.
- The value of the standard deviation tells how closely the values of a data set are clustered around the mean.
- In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively smaller range around the mean.
- In contrast, a larger value of the standard deviation for a data set indicates that the values of that data set are spread over a relatively larger range around the mean.

- The *standard deviation is obtained by taking the positive square root of the* **variance**.
- The variance calculated for population data is denoted by σ^2 and the variance calculated for sample data is denoted by s^2.
- Consequently, the standard deviation calculated for population data is denoted by σ and the standard deviation calculated for sample data is denoted by s.
- Therefore, it would be reasonable to measure the dispersion based on the degree of spread of values around their mean.
- Such a measure is realized in what is known as variance and standard deviation.

Calculation of Variance and Standard Deviation for Individual Observation

Calculating the variance in an original dataset

Let $x_1, x_2, \dots \dots \dots \dots .., x_N$ be a set of N measurements.

Following are what we will call the *basic formulas* that are used to calculate the population variance

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

Following are what we will call the *basic formulas* that are used to calculate the sample variance

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

- Variance reflects the average degree of dispersion of the data.
- Obviously, greater dispersion in the observed data corresponds to greater variance.
- We divide by n−1 instead of by n in our definition of sample variance, s^2.
- The theoretical reason for this choice of divisor is n – 1 instead of n is that it provides a "better" estimator of the true population variance σ^2.

Standard Deviation

With respect to standard deviation, the population standard deviation, σ, is the (positive) square root of the population variance and is defined as

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

The sample standard deviation, s, is

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

Example: A professor teaches two large sections of basic statistics and randomly selects a sample of test scores from both sections. Find the range and standard deviation for each sample:

Section 1	50	60	70	80	90
Section 2	72	68	70	74	66

Solution: The Mean of test score of section 1 $= \overline{x_1} = \frac{\sum x_i}{n} = \frac{350}{5} = 70$ and

The Mean of test score of section 2$= \overline{x_2} = \frac{\sum x_i}{n} = \frac{350}{5} = 70$

Range of test scores of section 1 = 90 – 50 = 40
Range of test scores of section 2 = 74 – 66 = 8

- Although the average grade for both sections is 70, we notice that the grades in section 2 are closer to the mean, 70, than are grades in section 1.
- And just as we would expect, the range of section 1, 40, is larger than the range of section 2, which is 8.

Similarly, we would expect the standard deviation for section 1 to be greater than the standard deviation for section 2.

Section 1 x_i	$(x_i - \bar{x})^2$	Section 2 x_i	$(x_i - \bar{x})^2$
50	400	72	4
60	100	68	4
70	0	70	0
80	100	74	16
90	400	66	16
Total	$\sum_{i=1}^{n}(x_i - \bar{x})^2$ =1000		$\sum_{i=1}^{n}(x_i - \bar{x})^2$ =40

$$s_1 = \sqrt{s_1^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1000}{4}} = \sqrt{250} = 15.8$$

$$s_2 = \sqrt{s_2^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{40}{4}} = \sqrt{10} = 3.16$$

<u>Two Observations</u>

1. The values of the variance and the standard deviation are never negative.
2. The measurement units of variance are always the square of the measurement units of the original data and the measurement units of standard deviation are always same as the original data.

Calculation of Variance and Standard Deviation using Frequency Distribution:

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

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

$$N = \sum f_i$$

Sample Variance and Standard Deviation for Frequency Distribution

The sample standard deviation, s, is

$$Smaple\ variance = s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^2 - \frac{\left(\sum_{i=1}^{n} f_i x_i\right)^2}{n}}{n-1}}$$

$$n = \sum f_i$$

Example: Calculate variance and standard deviation from the following table.
Solution: Direct Method

Class interval	Frequency f_i	Midpoint (x_i)	$(x_i - \bar{x})^2$	$f_i (x_i - \bar{x})^2$
30—39	3	34.5	1122.25	3366.75
40—49	1	44.5	552.25	552.25
50—59	8	54.5	182.25	1458.00
60—69	10	64.5	12.25	122.50
70—79	7	74.5	42.25	295.75
80—89	7	84.5	272.25	1905.75
90—99	4	94.5	702.25	2809.00
Total	40			10510.00

$$s^2 = \frac{\sum f_i (x_i - \bar{x})^2}{n-1} = \frac{10510.00}{39} = 269.48$$

$$s = \sqrt{s^2} = \sqrt{269.48} = 16.42$$

Alternatively

Class interval	Frequency f_i	Midpoint (x_i)	$f_i x_i$	$f_i x_i^2$
30—39	3	34.5	103.5	3570.75
40—49	1	44.5	44.5	1980.25
50—59	8	54.5	436	23762
60—69	10	64.5	645	41602.5
70—79	7	74.5	521.5	38851.75
80—89	7	84.5	591.5	49981.75
90—99	4	94.5	378	35721
Total	40		2720	195470

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i x_i^2 - \frac{\left(\sum_{i=1}^{n} f_i x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{195470 - \frac{(2720)^2}{40}}{40-1}}$$

$$= \sqrt{\frac{195470 - \frac{7398400}{40}}{39}} = \sqrt{\frac{195470 - 184960}{39}} = \sqrt{\frac{10510}{39}} = 16.42$$

Coefficient of Variation

- Although the standard deviation is useful for measuring the variability within a sample dataset, when a **comparison of variability between datasets is needed**, it might be inappropriate to use standard deviation to directly compare the degree of dispersion between two groups under the following conditions:

  (1) The two means are quite different. For example, if the means of two samples are 100 and 1000, but the standard deviation is 10 in both samples, how can variability be compared between the two samples?

  (2) The two indicators are measured in different units. For example, in the measurement of human physiological indicators, the unit of height

is usually centimeters, whereas the unit of weight is usually kilograms. How, then, can height and weight be compared?

- In both cases, we can use the coefficient of variation.
- The coefficient of variation, referred as CV, is a **quantity jointly determined by the mean and the standard deviation.**

The calculation formula for CV is as follows

The population coefficient of variation is

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad if \quad \mu > 0$$

The sample coefficient of variation is

$$CV = \frac{s}{\bar{x}} \times 100\% \qquad if \quad \bar{x} > 0$$

It can be seen in above Formula that CV is a **unit-free measure** because the standard deviation is standardized by the mean. CV is thus **more appropriate for comparing the dispersion of data with different units** or with a considerable difference in the means.

Example

In a survey on the physiological characteristics of school-age children in a certain region in 2020, 140 10-year-old boys were also randomly selected. The mean and standard deviation of the height of these boys were 140.8 cm and 7.0 cm, respectively, and the mean and standard deviation for body weight were 35.6 kg and 7.0 kg, respectively. Compare the degree of variation of height and body weight.

Solution

Height: $\bar{x}_1 = 140.8, \;\; s_1 = 7.0$ weight: $\bar{x}_2 = 35.6, \;\; s_1 = 7.0$

The CV of the height of the boys is

$$CV_{Height} = \frac{s_1}{\bar{x}_1} \times 100\% = \frac{7}{140.8} \times 100\% = 5.0\%$$

The CV of the weight of the boys is

$$CV_{Weight} = \frac{s_2}{\bar{x}_2} \times 100\% = \frac{7}{35.6} \times 100\% = 19.7\%$$

Therefore, the variation of body weight is greater than the variation of height among 10-year-old boys in this sample.

- When using the coefficient of variation, it should be noted that it is only meaningful to compare related indicators.
- Additionally, when the mean is less than the standard deviation, the practical application value of the coefficient of variation should be carefully considered.
- In this case, the coefficient of variation will be more than 100%, especially when the mean is close to 0, and it should not be used.