# Measure of Dispersion

## (Measure of variability)

MD. MAHFUZUR RAHMAN

SENIOR LECTURER, STATISTICS

# Contents

- **Range**
- **Variance & Standard deviation (for grouped and ungrouped data)**
- **Coefficient of Variation (CV)**
- **Shape characteristics:** Skewness & Kurtosis
- **Exploratory data analysis:** Boxplot

# Measure of dispersion

Measures of dispersion measure how spread out a set of data is, how much variability there has in the data.

# Measure of dispersion



- Statistics deals with data that has some variability
- Measure of location (Central tendency) can not always adequately describe a set of observations or performance of a group of individuals
- Two data with same mean, can have different variability (i.e. can disperse differently)

# Measure of dispersion



- Consider two data sets-

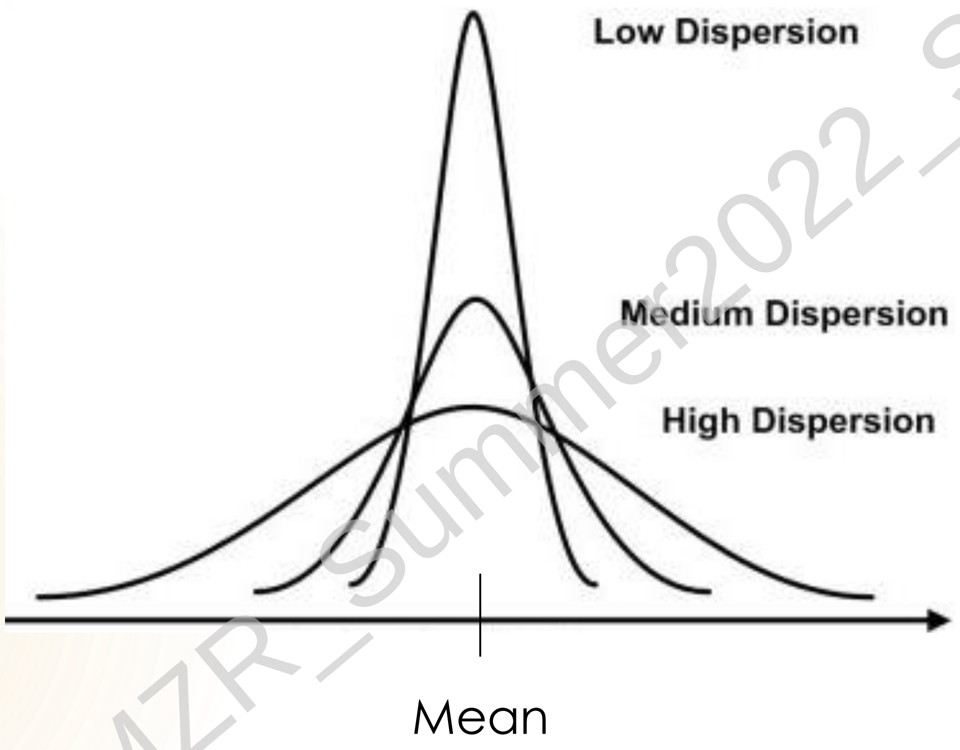
Data 1: 30, 40, 60, 80, 90

Data 2: 50, 55, 60, 65, 70

| Measure | Data 1 | Data 2 |
|---|---|---|
| **Mean** | 60 | 60 |
| **Range** | 90-30=60 | 70-50=20 |

# Measure of dispersion





# Measure of dispersion





**Characteristics of a good measure of variation or dispersion:**

The following are the characteristics of an ideal measure of variation or dispersion

- It should be easy to understand.
- It should be easy to calculate.
- It should be based upon all observations.
- It should be rigidly defined.
- It should be unduly affected by extreme values.
- It should be suitable for further algebraic treatment.
- It should be less affected by sampling fluctuation.

# Measure of dispersion



**Purpose of measure of dispersion or variation:**

Measure of dispersion is important for the following purpose.

- To determine the reliability of an average.
- To compare the variability.
- To compare two or more series with regard to their variability.
- To facilitate the use of other statistical measures.
- It is one of the most important quantities used to characterize a frequency distribution.

# Measure of dispersion



Important and most commonly used measures of dispersion-

- **Absolute Measures**
  1. **The Range**
  2. The Mean Deviation (MD) or Average Deviation
  3. The Interquartile Range (IQR)or Quartile Deviation (QD)
  4. **The Variance**
  5. **The Standard Deviation (SD)**

# Measure of dispersion



Important and most commonly used measures of dispersion-

- **Relative Measure:**
  1. **Coefficient of Variation (CV)**
  2. Coefficient of range
  3. Coefficient of quartile deviation
  4. Coefficient of mean deviation

# Range





Difference between highest and lowest value.

**Range**= Highest value (H)- Lowest value (L)

Notes:

1. The unit of the range is the same as the unit of the data.
2. The usefulness of the range is limited. The range is poor measure of the dispersion because it only takes into account two of the values; however, it plays a significant role in many application.

# Range



**Example:**

Below given the weight of 10 newly born babies (in pounds)-

7.5, 4.5, 10.1, 9.6, 5.5, 6.6, 7.8, 5.9, 6.0, 5.5

# Range



**Example:**

Below given the weight of 10 newly born babies (in pounds)-

7.5, 4.5, 10.1, 9.6, 5.5, 6.6, 7.8, 5.9, 6.0, 5.5

$$Range = Highest\ value\ - Lowest\ value$$
$$= 10.1\ - 4.5\ = 5.6\ pounds$$

**Interpretation:** The difference of weights between the healthiest baby and leanest baby is 5.6 pounds

# Mean Deviation (MD) or Average Deviation

The mean of the absolute deviations of each individual value from the average of a set of values, is called the average deviations.

Let x1, x2,... ... ...xn be a set of n values, then its mean deviation is denoted by,

$$A.D = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \text{ ; for raw data}$$

$$A.D = \frac{1}{n}\sum_{i=1}^{k} f_i|x_i - \bar{x}| \text{ ; for group data}$$

# Variance



Calculates variability or dispersion from mean.

# Variance



**Formulas:**

**For raw or ungrouped data-**

**For Population:** let, $X_1, X_2, \ldots, X_N$ are values of a variable from a population of size N. Then,

$$Population\ variance, \sigma^2 = Var(X)$$

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

(Parameter)

**For Sample:** let, $x_1, x_2, \ldots, x_n$ are values of a variable from a sample of size n. Then,

$$Sample\ variance, s^2 = var(X)$$

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

(Statistic)

# Variance



**Formulas:**

**For grouped data-**

**For Population:** let, $X_1, X_2, \ldots, X_K$ are values of a variable from a population of size N and they occurred $f_1, f_2, \ldots, f_K$ times respectively. Then,

$$Population\ variance, \sigma^2 = Var(X)$$

$$= \frac{\sum_{i=1}^{K} f_i (X_i - \mu)^2}{N}$$

(Parameter)

**For Sample:** let, $x_1, x_2, \ldots, x_k$ are values of a variable from a sample of size n and they occurred $f_1, f_2, \ldots, f_k$ times respectively. Then,

$$Sample\ variance, s^2 = var(X)$$

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

(Statistic)

# Standard Deviation (SD)



- ✓ Average variation of the data or observations from mean
- ✓ Can be obtained by taking square root of variance.

# Standard Deviation (SD)





**Formulas:**

**For raw or ungrouped data-**

**For Population:** let, $X_1, X_2, \ldots, X_N$ are values of a variable from a population of size N. Then,

$$Population\ SD, \sigma = SD(X) = \sqrt{Var(X)}$$

$$= \sqrt{\left(\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}\right)}\ unit$$

(Parameter)

**For Sample:** let, $x_1, x_2, \ldots, x_n$ are values of a variable from a sample of size n. Then,

$$Sample\ SD, s = sd(X) = \sqrt{var(X)}$$

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

(Statistic)

# Standard Deviation (SD)





**Formulas:**

**For grouped data-**

**For Population:** let, $\mathbf{X_1, X_2, \ldots, X_K}$ are values of a variable from a population of size N and they occurred $\mathbf{f_1, f_2, \ldots, f_K}$ times respectively. Then,

$$Population\ SD, \sigma = SD(X) = \sqrt{Var(X)}$$

$$= \sqrt{\frac{\sum_{i=1}^{K} f_i (X_i - \mu)^2}{N}}\ unit$$

(Parameter)

**For Sample:** let, $\mathbf{x_1, x_2, \ldots, x_k}$ are values of a variable from a sample of size n and they occurred $\mathbf{f_1, f_2, \ldots, f_k}$ times respectively. Then,

$$Sample\ SD, s = sd(X) = \sqrt{var(X)}$$

$$= \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n-1}}\ unit$$

(Statistic)

# Example 1



Below given the weight of 10 newly born babies (in pounds)-

7.5, 4.5, 10.1, 9.6, 5.5, 6.6, 7.8, 5.9, 6.0, 5.5

Find SD for the above data. Interpret the result.

# Example 1



Below given the weight of 10 newly born babies (in pounds)-

7.5, 4.5, 10.1, 9.6, 5.5, 6.6, 7.8, 5.9, 6.0, 5.5

Find SD for the above data. Interpret the result.

$$\boldsymbol{mean}, \bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{7.5 + 4.5 + 10.1 + 9.6 + 5.5 + 6.6 + 7.8 + 5.9 + 6.0 + 5.5}{10} = 6.9$$

# Example 1



$$\boldsymbol{variance}, var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$= \frac{\begin{array}{c}(7.5-6.9)^2+(4.5-6.9)^2+(10.1-6.9)^2+(9.6-6.9)^2+(5.5-6.9)^2\\+(6.6-6.9)^2+(7.8-6.9)^2+(5.9-6.9)^2+(6.0-6.9)^2+(5.5-6.9)^2\end{array}}{10-1}$$

$$= \frac{\begin{array}{c}(.6)^2+(-2.4)^2+(3.2)^2+(2.7)^2+(-1.4)^2\\+(-0.3)^2+(0.9)^2+(-1)^2+(-0.9)^2+(-1.4)^2\end{array}}{9}$$

$$= \frac{0.36+5.76+10.24+7.29+1.96+0.09+0.81+1+0.81+1.96}{9}$$

$$= \frac{30.28}{9} = 3.36$$

# Example 1



$$\boldsymbol{sd}, s = \sqrt{var(X)} = \sqrt{3.36} = 1.83\; pounds$$

**Interpretation:** The average variation of the weights of the newly born babies from the mean weight is 1.83 pounds

# Example 2



Consider the following data-

| Monthly income ('000 tk) | No. of respondents ($f_i$) |
|---|---|
| 5-30 | 7 |
| 30-55 | 10 |
| 55-80 | 6 |
| 80-105 | 4 |
| 105-130 | 3 |
| **Total** | **30** |

Find SD and interpret the result.

# Example 2



| Monthly income ('000 tk) | No. of respondents ($f_i$) | Class Midpoint ($x_i$) | $f_i x_i$ | $(x_i - \bar{x})$ | $f_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 5-30 | 7 | 17.5 | 122.5 | -38.33 | 10284.32 |
| 30-55 | 10 | 42.5 | 425 | -13.33 | 1776.89 |
| 55-80 | 6 | 67.5 | 405 | 11.67 | 817.13 |
| 80-105 | 4 | 92.5 | 370 | 36.67 | 5378.76 |
| 105-130 | 3 | 117.5 | 352.5 | 61.67 | 11409.57 |
| **Total** | **30** | | **1675** | | **29666.67** |

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{1675}{30} = 55.83\ thousand\ taka$$

# Example 2



| Monthly income ('000 tk) | No. of respondents ($f_i$) | Class Midpoint ($x_i$) | $f_i x_i$ | $(x_i - \bar{x})$ | $f_i(x_i - \bar{x})^2$ |
|---|---|---|---|---|---|
| 5-30 | 7 | 17.5 | 122.5 | -38.33 | 10284.32 |
| 30-55 | 10 | 42.5 | 425 | -13.33 | 1776.89 |
| 55-80 | 6 | 67.5 | 405 | 11.67 | 817.13 |
| 80-105 | 4 | 92.5 | 370 | 36.67 | 5378.76 |
| 105-130 | 3 | 117.5 | 352.5 | 61.67 | 11409.57 |
| **Total** | **30** | | **1675** | | **29666.67** |

$$\boldsymbol{SD}, s = \sqrt{\frac{\sum f_i(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{29666.67}{30-1}} = 31.98\ thousand\ taka$$

**Interpretation**: Average variation of the monthly incomes of the respondents from mean income is 31.98 thousand taka

# Coefficient of Variation (CV)

The coefficient of variation (CV) is defined as the ratio of the standard deviation $\sigma$ to the mean $\mu$:

$$Population\ CV, \quad C_v = \frac{\sigma}{\mu}$$

$$Sample\ CV, \quad c_v = \frac{s}{\bar{x}}$$

- It shows the extent of variability in relation to the mean of the population
- The coefficient of variation should be computed only for data measured on a ratio scale
- For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation

# Relative Measures (Others)



- Coefficient of range = $\frac{L-S}{L+S} \times 100$
- Coefficient of mean deviation from A = $\frac{MD(A)}{A} \times 100$
- Coefficient of quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100$

- **Inter-relationship:**

$$Mean\ deviation = \frac{4}{5} \times standard\ deviation$$

$$Quartile\ Deviation = \frac{2}{3} \times standard\ deviation$$

# Shape characteristics

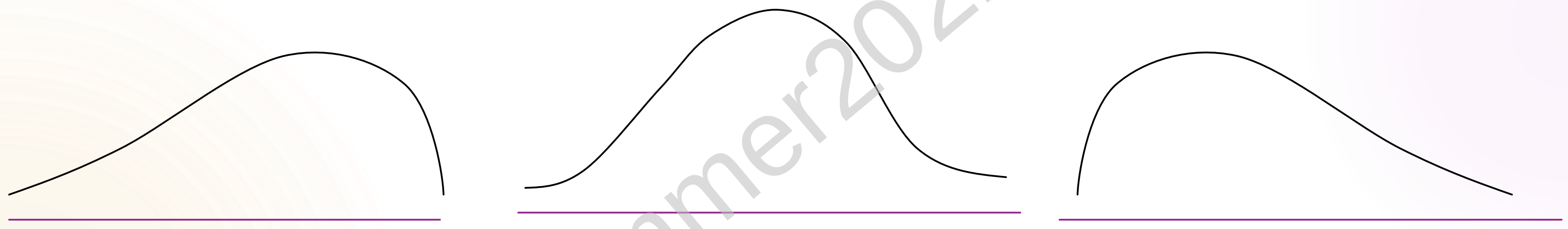


Shape of a distribution can be identified by using two characteristics-

1. **Skewness**
2. **Kurtosis**

# Skewness

A measure of the asymmetry (lack of symmetry) of a distribution



Negatively skewed Symmetric Positively skewed

# Skewness



**Note:**

- The normal distribution is **symmetric** and has a **skewness = 0**. Here, **Mean=Median=Mode**
- A distribution with a significant **positive skewness** has a long right tail and has **skewness>0**. Here, **Mean>Median>Mode**
- A distribution with a significant **negative skewness** has a long left tail and has **skewness<0**. Here, **Mean<Median<Mode**

# Skewness



**Formulas:**

$$1.\ Pearson's\ coefficient\ of\ skewness = \frac{3(mean - median)}{Standard\ Deviation} = \frac{mean - mode}{Standard\ Deviation}$$

$$2.\ Bowley's\ coefficient\ of\ skewness = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$

# Skewness



**Example:**

For a distribution we have-

mean= 30.892, median= 30.58, SD= 2.219, $Q_1$= 29.50, $Q_3$= 32.1

Is the distribution is positively skewed? How? What is the value of coefficient of skewness?

# Skewness



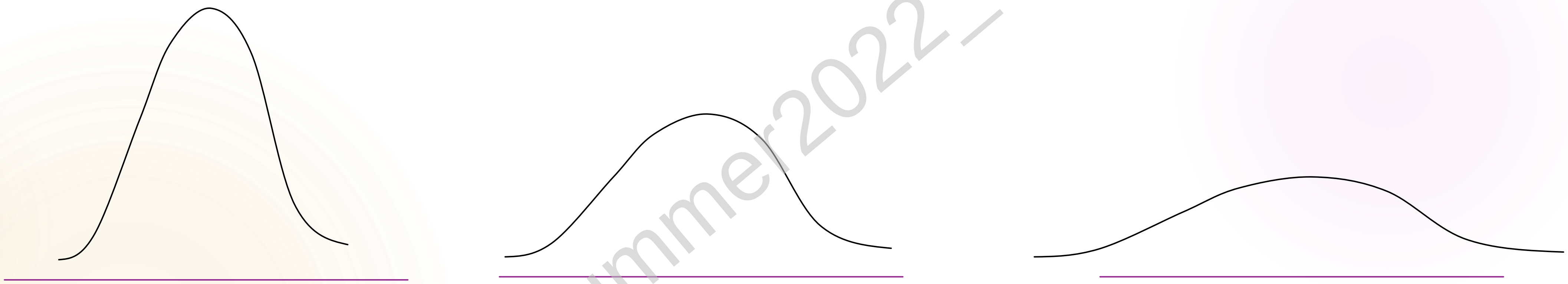
$$Pearson's\ coefficient\ of\ skewness = \frac{3(mean\ - median)}{Standard\ Deviation} = \frac{3(30.892 - 30.58)}{2.219}$$
$$= 0.42$$

$$Bowley's\ coefficient\ of\ skewness = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}$$
$$= \frac{(32.1 - 30.58) - (30.58 - 29.50)}{32.1 - 29.50}$$
$$= 0.17$$

Yes, the distribution is positively skewed. Because the coefficient of skewness is greater than 0. The value of skewness is 0.42.

# Kurtosis

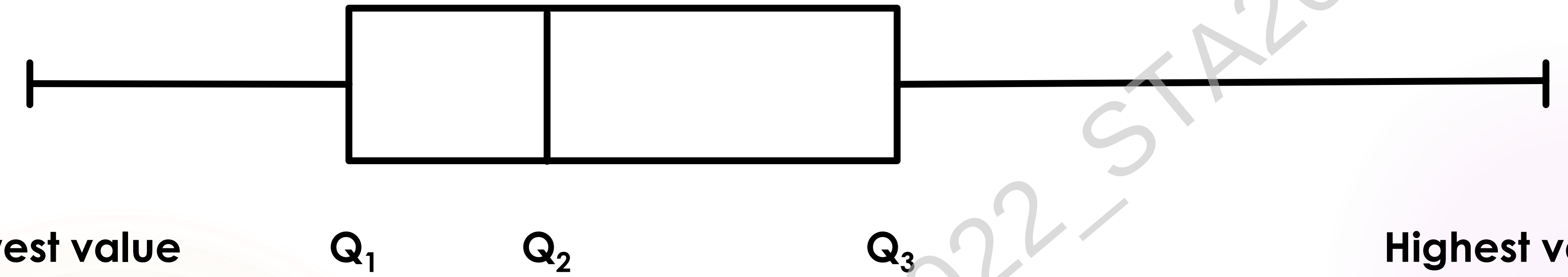
A measure of the extent to which observations cluster around a central point. A provides a measure of peakedness i.e. how peak the distribution is.



Leptokurtic Mesokurtic Platykurtic

# Box & Whisker plot






**Five number summary-**

1. Lowest value
2. $Q_1$
3. Median ($Q_2$)
4. $Q_3$
5. Highest value

# Box & Whisker plot



**Example:**

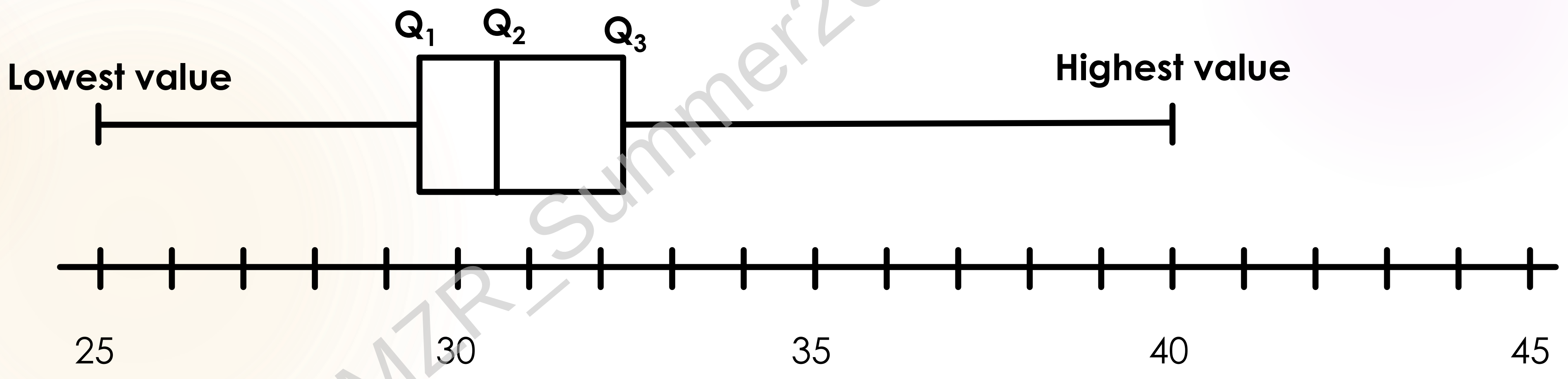
For a distribution, Lowest value= 25, Highest value= 40, Q1= 29.50, Q3= 32.1, and Median= 30.58. Show these information in a boxplot.

# Box & Whisker plot



**Example:**

For a distribution, Lowest value= 25, Highest value= 40, Q1= 29.50, Q3= 32.1, and Median= 30.58. Show these information in a boxplot.



# Box & Whisker plot



**Outliers:**

$$Interquartile\ Range, IQR = Q_3 - Q_1$$

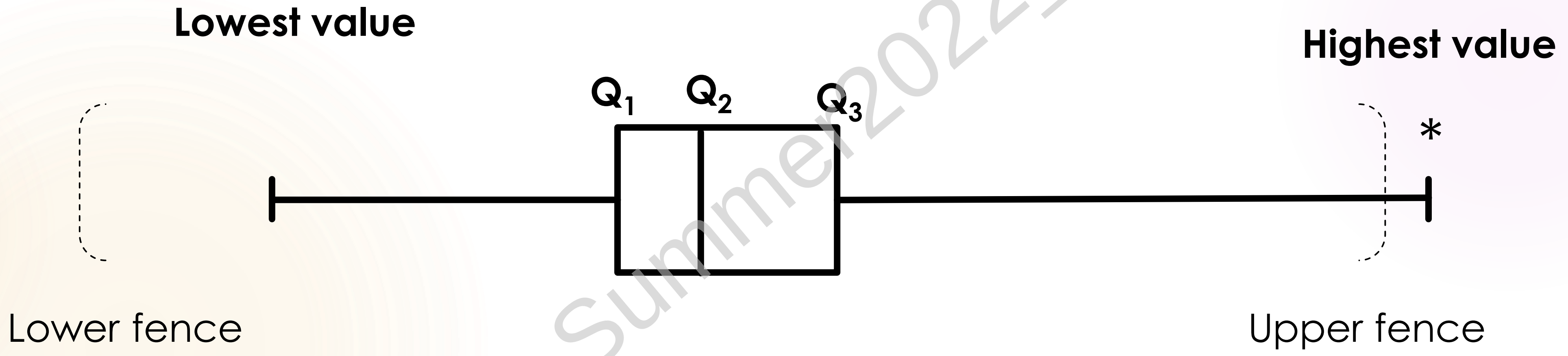
$$Lower\ fence = Q_1 - 1.5 * IQR$$
$$Upper\ fence = Q_3 + 1.5\ * IQR$$

Any observation having value out of (beyond) these two fences is called outliers and represented by '*' sign on the boxplot. (One * for each outlier)

# Box & Whisker plot



**Outliers:**



# Class task





**Question:**

A random sample of 20 people was taken to know the time passed on Facebook during last two weeks (in hours). The recorded data were as follows-

67, 76, 85, 42, 93, 48, 93, 46, 52, 72, 77, 53, 41, 48, 86, 78, 56, 80, 70, 66

Show this data in a boxplot. Measure the coefficient of skewness. Comment on your findings.