Measure of Dispersion

(Measure of variability)

F. M. ARIFUR RAHMAN

SENIOR LECTURER, DEPARTMENT OF MATHEMATICAL & PHYSICAL SCIENCES

Contents

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

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

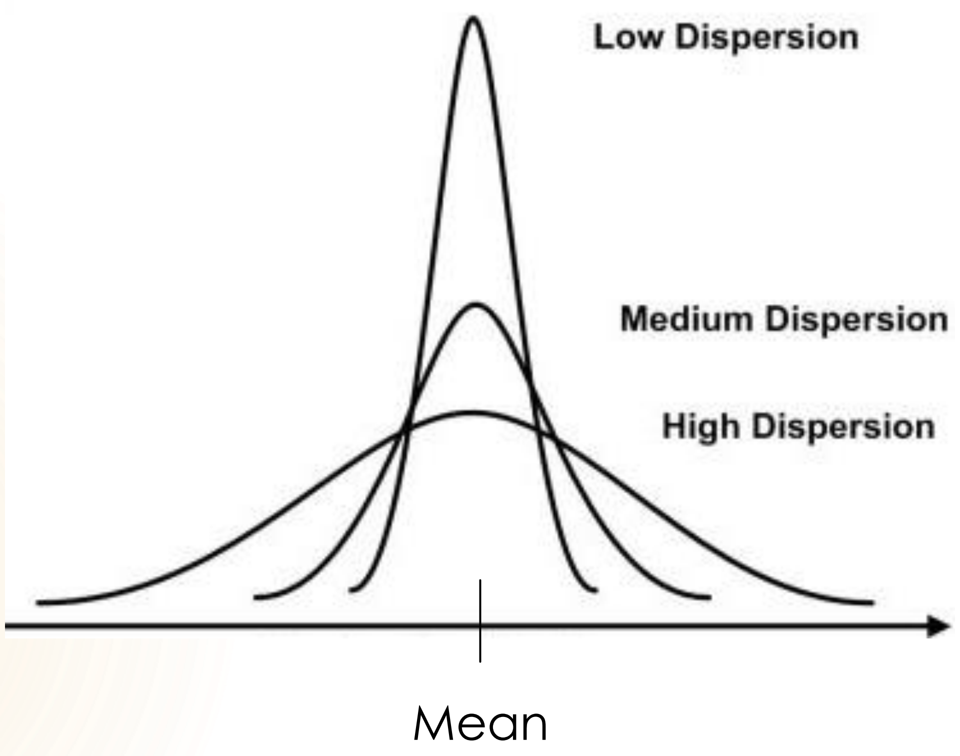
- Consider two data sets-

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

Important and most commonly used measures of dispersion-

1. **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)**
2. **Relative Measure: Coefficient of Variation (CV)**

Range

Difference between highest and lowest value.

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

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

$$\begin{aligned} Range &= Highest\ value\ - Lowest\ value \\ &= 10.1\ - 4.5\ = 5.6\ pounds \end{aligned}$$

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

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

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

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

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

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

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

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 σ to the mean μ:

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

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

- 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

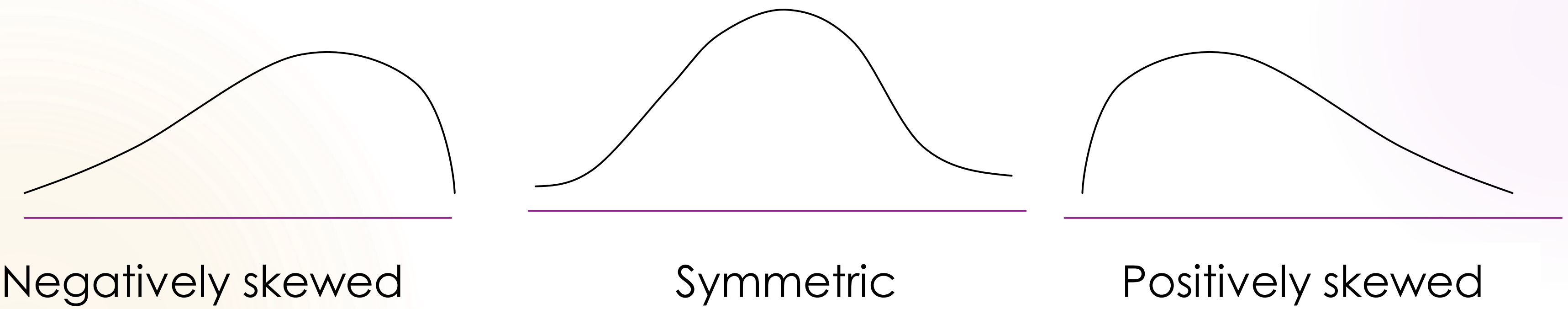
Shape characteristics

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



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

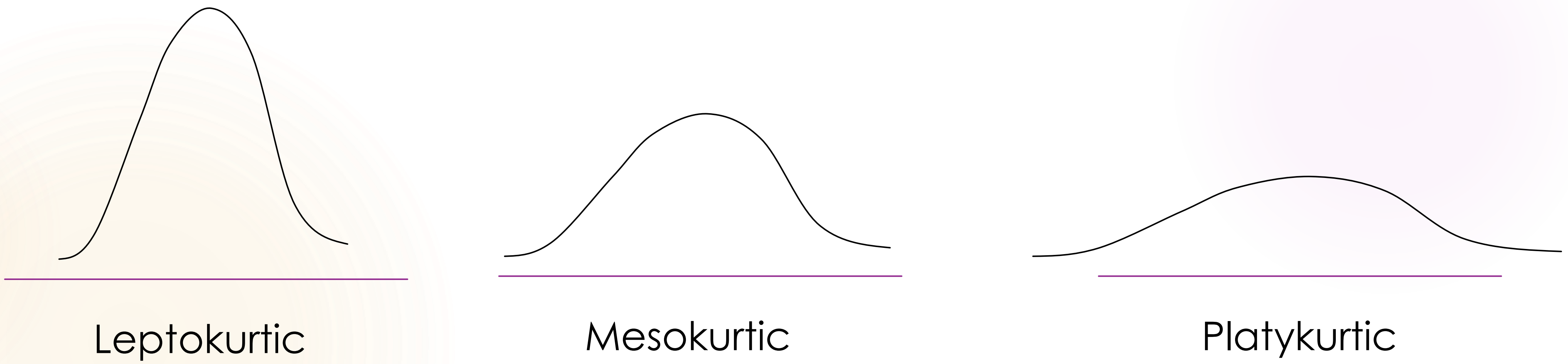
$$Pearson's\ coefficient\ of\ skewness = \frac{3(mean\ \ - median)}{Standard\ Deviation} = \frac{3(30.892 - 30.58)}{2.219}$$
$$= 0.42$$

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

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.



Kurtosis

$$Kurtosis = \frac{\frac{\sum(x_i - \bar{x})^4}{n}}{\left(\frac{\sum(x_i - \bar{x})^2}{n}\right)^2} - 3$$

- If kurtosis=0, then Mesokurtic
- If kurtosis>0, then Leptokurtic
- If kurtosis<0, then Platykurtic.

Kurtosis

Example:

Consider the following data-

4, 2, 4, 3, 3, 5, 4, 4, 3, 4, 4, 4, 5, 6, 4

$$Kurtosis = \frac{\frac{\sum(x_i - \bar{x})^4}{n}}{\left(\frac{\sum(x_i - \bar{x})^2}{n}\right)^2} - 3 = \ 0.32506, which\ is\ greater\ than\ 0.$$

So, the distribution is leptokurtic.

Box & Whisker plot:

Box & Whisker plot:

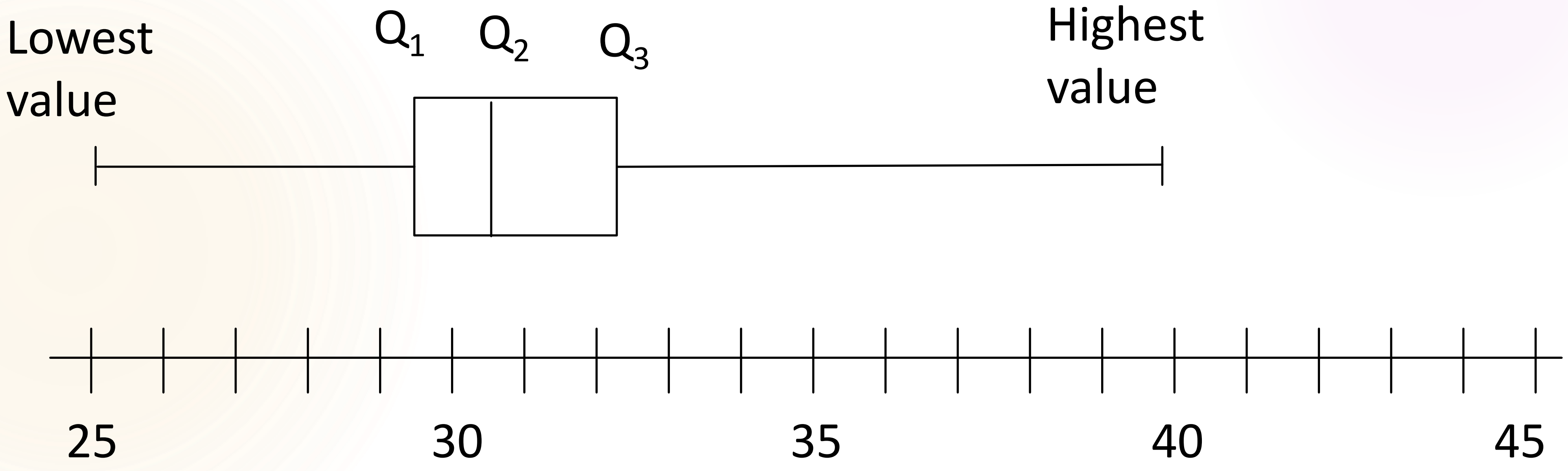
Five number summary-

1. Lowest value
2. Q1
3. Median (Q2)
4. Q3
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:

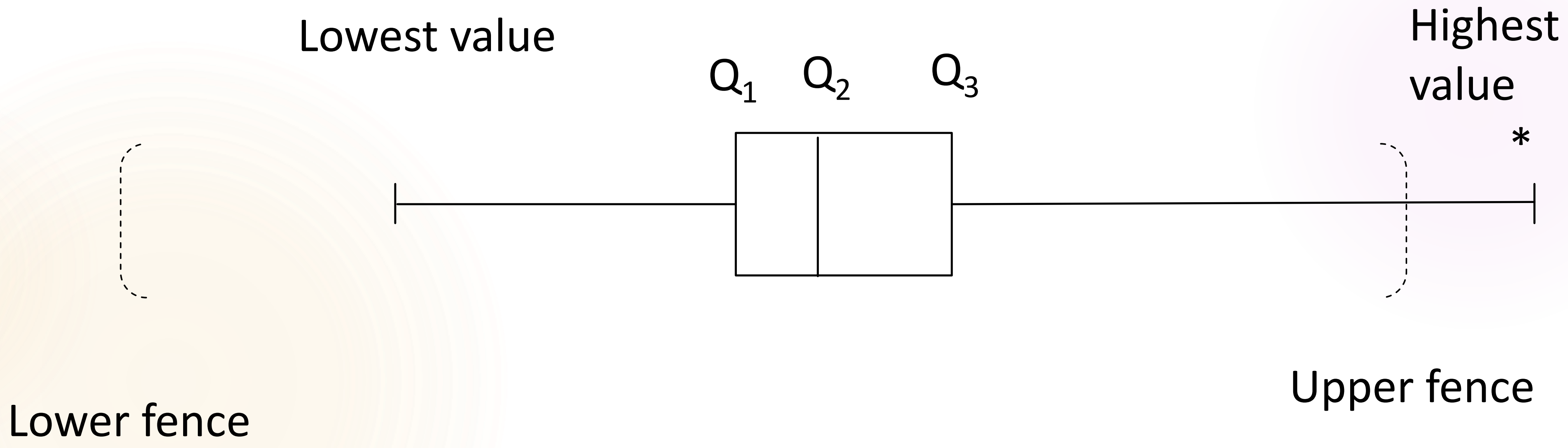
Outliers:

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

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:



Stem and Leaf plot:

Example:

Show the following data in a stem & leaf plot.

44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106

Stem and Leaf plot:

Example:

Key: 6 | 3=63

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

Stem and Leaf plot:

Example:

Show the following data in a stem & leaf plot.

4.4, 4.6, 4.7, 4.9, 6.3, 6.4, 6.6, 6.8, 6.8, 7.2, 7.2, 7.5, 7.6, 8.1, 8.4, 8.8, 10.6

Stem and Leaf plot:

Example:

Key: 6 | 3=6.3

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