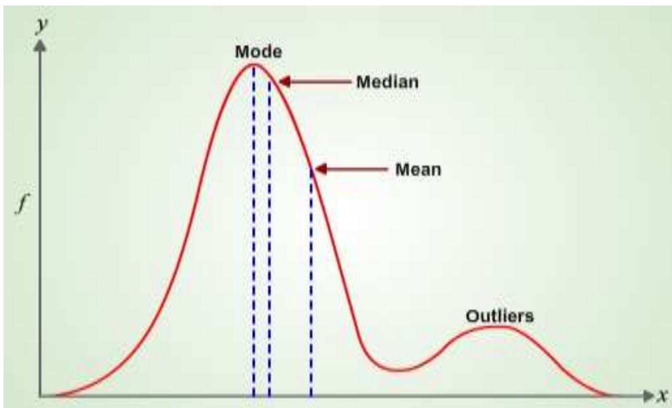
Learning Outcome

By the end of this lecture, you will be able to:

- Compute the range, the variance, and the standard deviation for a set of a data
- Recognize the value of each of the variation measures
- Reproduce and interpret a box-and-whiskers plot

Introduction

Recall that measures of central tendency are used to find central value of the given data.

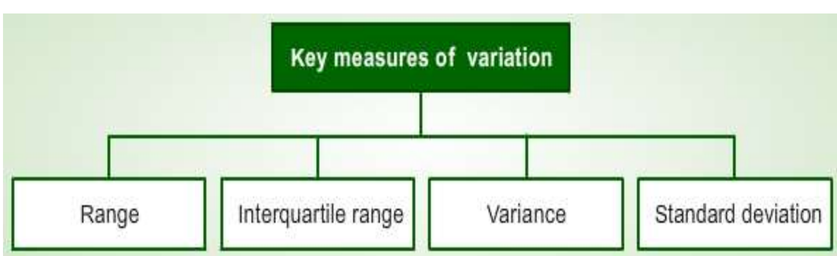


Measures of central tendency

However, other types of analysis are essential to understand variation of the given data. Measures of variation can help conduct detailed of analysis of the data. In this lecture, we will explore measures of variation.

Measures of Variation

Variation happens when not all the data are the same. Measures of variation give information on the spread or variability of the data values.



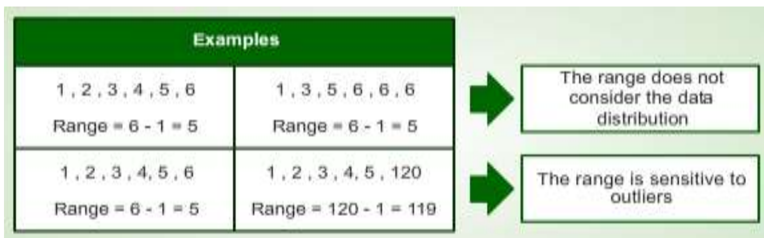
Key measures of variation

Range

The range is the simplest measure of variation. It is defined as the difference between the largest and the smallest observations.

$$Range = Maximum - Minimum$$

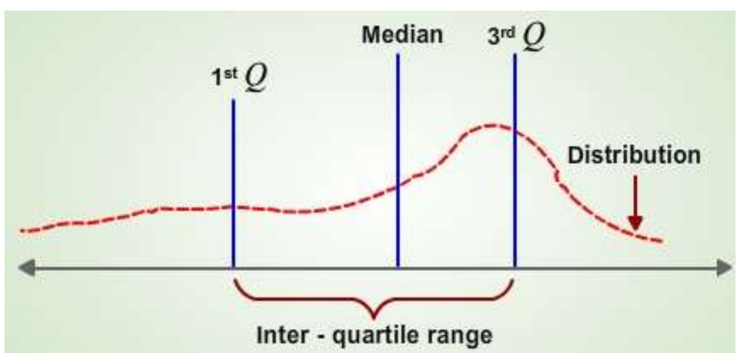
However, the range is not the best value to measure the variation because it does not consider how the data is distributed and is sensitive to outliers.



Interquartile Range

The interquartile range is a measure of variation.

$$Interquartile\ Range = 3rd\ quartile - 1st\ quartile$$



Interquartile range

The interquartile range will not change if the maximum and the minimum values are changed and we can eliminate some outlier problems by using the interquartile range.

Variance

The variance is the average of squared deviations of the values from the mean, for a population, it is the population variance and for a sample, it is the sample variance.

- **Population variance (σ^2) :**

If the population is finite, the population variance is calculated using the formula.

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

N is the population size
μ is mean of the population

- **Sample variance (S^2) :**

The sample variance is calculated using the given formula.

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

n is sample size
$\bar{x}$ is mean of sample

Standard Deviation

Standard deviation:

- Is the most commonly used measure of variation
- Is the square root of variance
- Shows variation about the mean
- Has the same unit as the original data

Population standard deviation (σ) :

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

N is the population size
μ is the population mean

Sample standard deviation (S) :

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

n is the sample size
$\bar{X}$ is the sample mean

Example

Compute the variance and the standard deviation from the sample data.

10, 12, 14, 15, 17, 18, 18, 24

Solution:

$$n = 8$$

$$\sum_{i=1}^{8} x_i = 128$$

$$\sum_{i=1}^{8} x_i^2 = 2178$$

$$Variance = S^2 = \frac{2178 - \frac{128^2}{8}}{7} = 18.5714$$

$$Standard\ Deviation = S = \sqrt{18.5714} = 4.3095$$

Example

The data shows grades of a population of ten students on a certain quiz.

4, 6, 9, 4, 5, 7, 8, 5, 6, 4

1. Compute the range of the population of grades.
2. Compute the variance and the standard deviation of the population of grades.
3. Assume that these data represent a sample rather than a population. Compute the variance and the standard deviation of the sample of grades.

Solution:

$$N = n = 10$$
$$\sum_{i=1}^{10} x_i = 67$$
$$\sum_{i=1}^{10} x_i^{\,2} = 364$$

$$\begin{aligned} Range &= Maximum - Minimum \\ &= 9 - 4 \\ &= 5\ points \end{aligned}$$

$$\sigma^2 = \frac{364 - \frac{67^2}{10}}{10} = 2.76$$
$$\sigma = \sqrt{2.76} = 1.6613$$

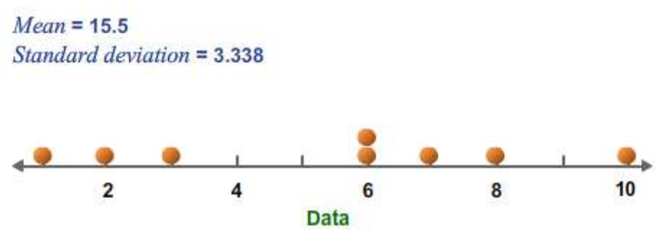
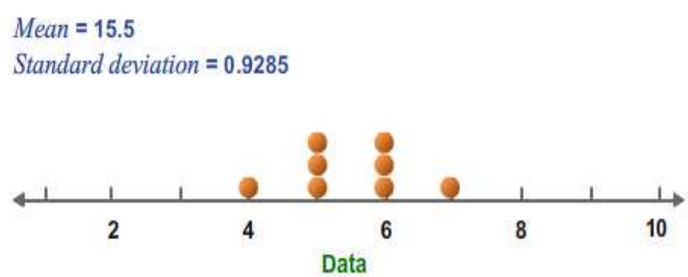
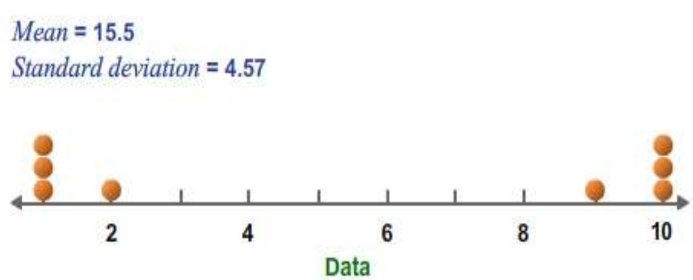
$$s^2 = \frac{364 - \frac{67^2}{10}}{9} = 3.0667$$
$$s = \sqrt{3.0667} = 1.7512$$

Mean and Standard Deviation

Consider three samples A, B and C that have equal means.

Sample A	Mean = 15.5	Standard deviation = 3.338
Sample B	Mean = 15.5	Standard deviation = 0.9285
Sample C	Mean = 15.5	Standard deviation = 4.57

Sample 'B' has the smallest standard deviation, which means that the observations are closer to each other more than the other samples.

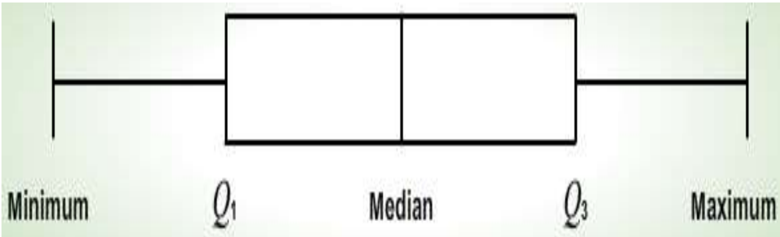
Sample A	
Sample B	
Sample C	

Box and Whiskers Plot

Box and whiskers plot is a graphical display of data. It uses five values:

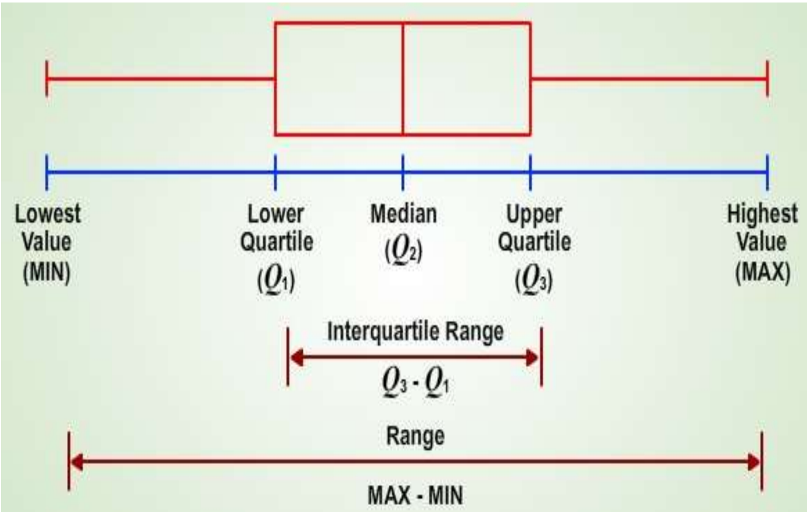
1. Minimum
2. First quartile - $Q1$
3. Median
4. Third quartile - $Q3$
5. Maximum

A sample box and whiskers plot is shown.



Box and Whiskers plot

The formula for calculating the interquartile range, inner fences and outer fences are given.



Box and Whiskers plot

Interquartile range(IQR)	$Q_3 - Q_1$
Inner fences	$Lower\ Inner\ Fence\ (LIF) = Q_1 - 1.5\ IQR$ $Upper\ Inner\ Fence\ (UIF) = Q_3 - 1.5\ IQR$
Outer fences	$Lower\ Outer\ Fence\ (LOF) = Q_1 - 3\ IQR$ $Upper\ Outer\ Fence\ (UOF) = Q_3 - 3\ IQR$

Box and Whiskers Plot > Steps

The four steps to construct a box and whiskers plot are given.

Steps to construct a box and Whiskers plot

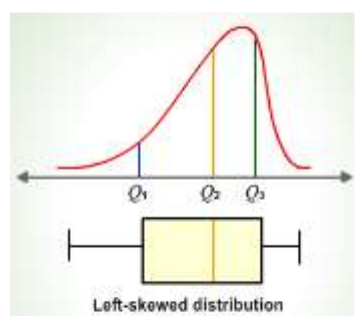
Step 1	Find $Q1$, $Q2$, and $Q3$
Step 2	Extend the line in both directions down to the minimum and up to the maximum.
Step 3	Any value that falls between the inner and the outer fences is considered as a possible or mild outlier.
Step 4	Any value that falls outside the outer fences is considered as a probable or extreme outlier. Note that you might find out that what you thought is a minimum or a maximum is actually a mild or an extreme outlier.

Distribution Shape and Box and Whiskers Plot

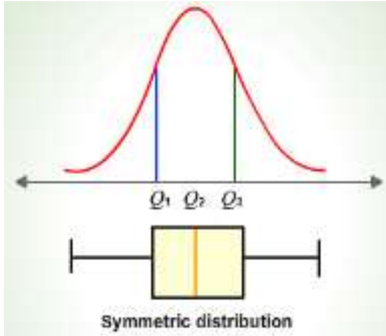
Depending on the box and whiskers plot, we can determine the shape of the data

Left-skewed distribution

If the distance between $Q1$ and $Q2$ is greater than the distance between $Q3$ and $Q2$, the data skewed to the left.

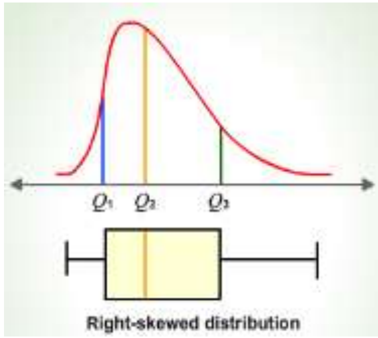


Symmetric distribution



If the distance between $Q1$ and $Q2$ equals the distance between $Q3$ and $Q2$, the data symmetric

Right-skewed distribution



If the distance between $Q1$ and $Q2$ is less than the distance between $Q3$ and $Q2$, the data skewed to the right

Recap

In this lecture, you have learned that:

- The key measures of variation are:
    - The range
    - The variance
    - The standard deviation
- The range is the difference between the maximum value and the minimum value
- The variance is the average square deviation from the mean
- The standard deviation is the square root of variance
- The box and whiskers plot is a graphical display of data. It uses five values:
    - Minimum
    - First quartile
    - Median
    - Third quartile
    - Maximum