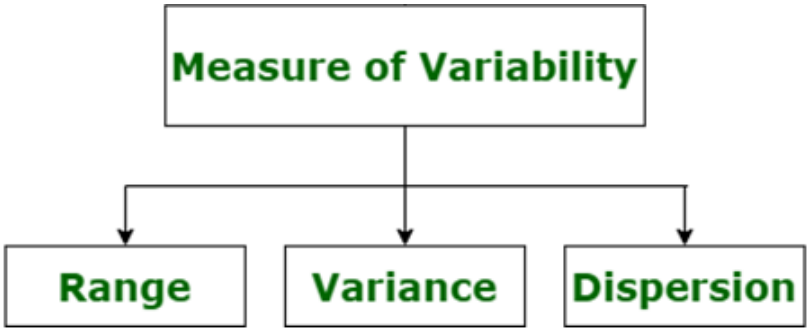
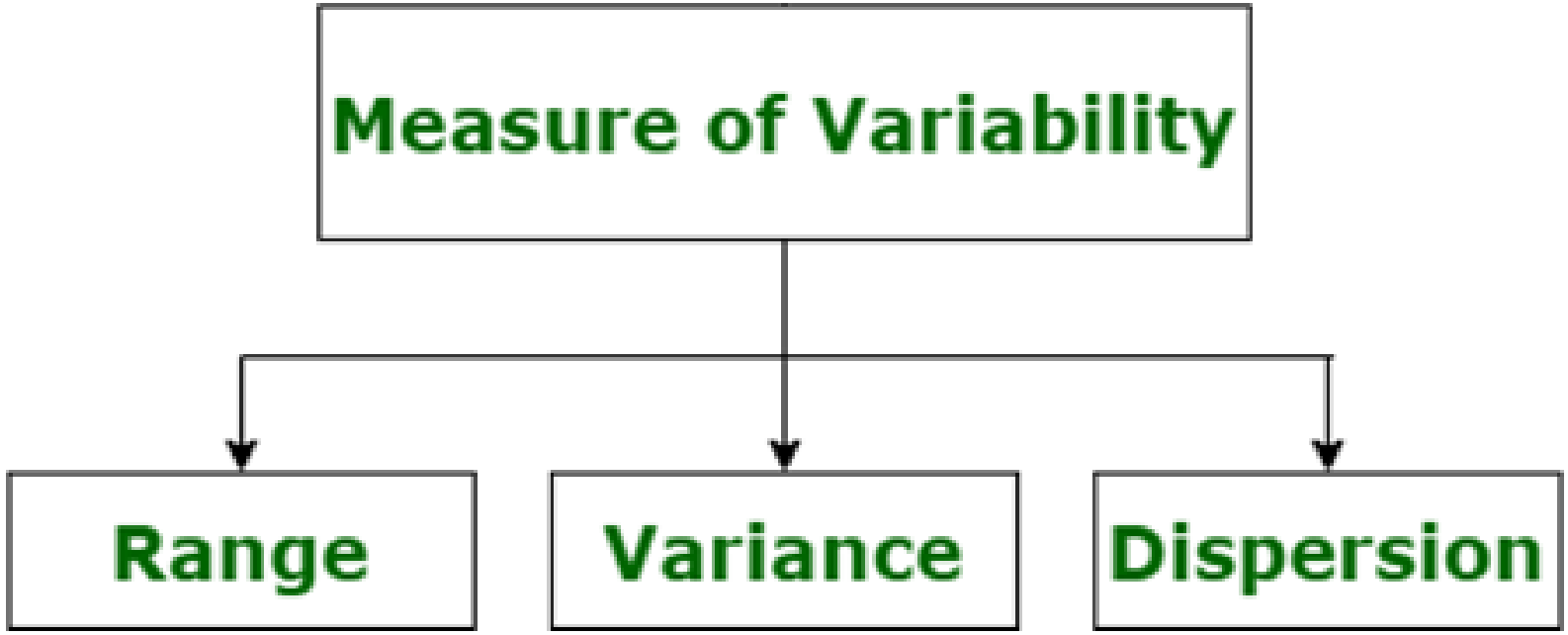
Variability Measures

In the framework of the course:
"Applied Statistics"



Prof. Dr. Mohamed Samer
Engineering in Biosystems, Energy and Environment
Department of Agricultural Engineering
Faculty of Agriculture, Cairo University
E-Mails: msamer@agr.cu.edu.eg; samer@cu.edu.eg
Website: http://scholar.cu.edu.eg/samer/biocv

Introduction


Measure of Variability
Range
Variance
Dispersion

Dispersion:

Dispersion of data used to understand the distribution of data.

It helps to understand the variation of data and provides a piece of information about the distribution data.

Range, IQR, Variance, and Standard Deviation are the methods used to understand the distribution data.

Dispersion of data helps to identify outliers in a given dataset.

Range:

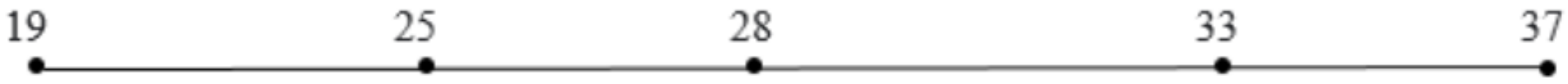
The range is the easiest dispersion of data or measure of variability. The range can measure by subtracting the lowest value from the massive Number.

The wide range indicates high variability, and the small range specifies low variability in the distribution.

To calculate a range, prepare all the values in ascending order, then subtract the lowest value from the highest value.

Range = Highest value – Lowest value

Student_id	1	2	3	4	5
Marks	37	33	19	25	28

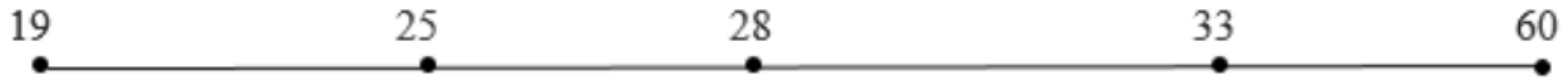


Range = H – L

= 37-19 ⟹ 18 The range of marks is 18.

The range can influence by outliers. If there is one extreme value that can change the value of a range.

Student_id	1	2	3	4	5
Marks	60	33	19	25	28



Range = H – L

= 60-19 ⟹ 41 Now range of marks is 41.

Variance: Variance is a simple measure of dispersion. Variance measures how far each number in the dataset from the mean. To compute variance first, calculate the mean and squared deviations from a mean.

Population variance

$$\text{Variance} = \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$$

Sample variance

$$\text{Variance} = s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n - 1}$$

Observation near to mean value gets the lower result and far from means gets higher value.

Description of a variance formula with example.

Let's say we have values: 5, 7, 9, and 3.

1. Calculate the mean

$$\overline{x} = \frac{\sum x}{n} \Rightarrow \overline{x} = \frac{x1+x2+x3.....xn}{n}$$

$$\textbf{\textit{Mean}} = \frac{5+7+9+3}{4}$$

$$= 6$$

2. Subtract mean from all observation to find the distance of all observation from mean.

$$\textbf{Variance} = \frac{(5-6)^2+(7-6)^2+(9-6)^2+(3-6)^2}{4}$$

$$= \frac{1+1+9+9}{4} \Rightarrow 5$$

Degrees of Freedom (DF)

The number of **degrees of freedom (DF) or (DOF)** is the number of values in the final calculation of a statistic that are free to vary.

The number of independent ways by which a dynamic system can move without violating any constraint imposed on it, is called degree of freedom.

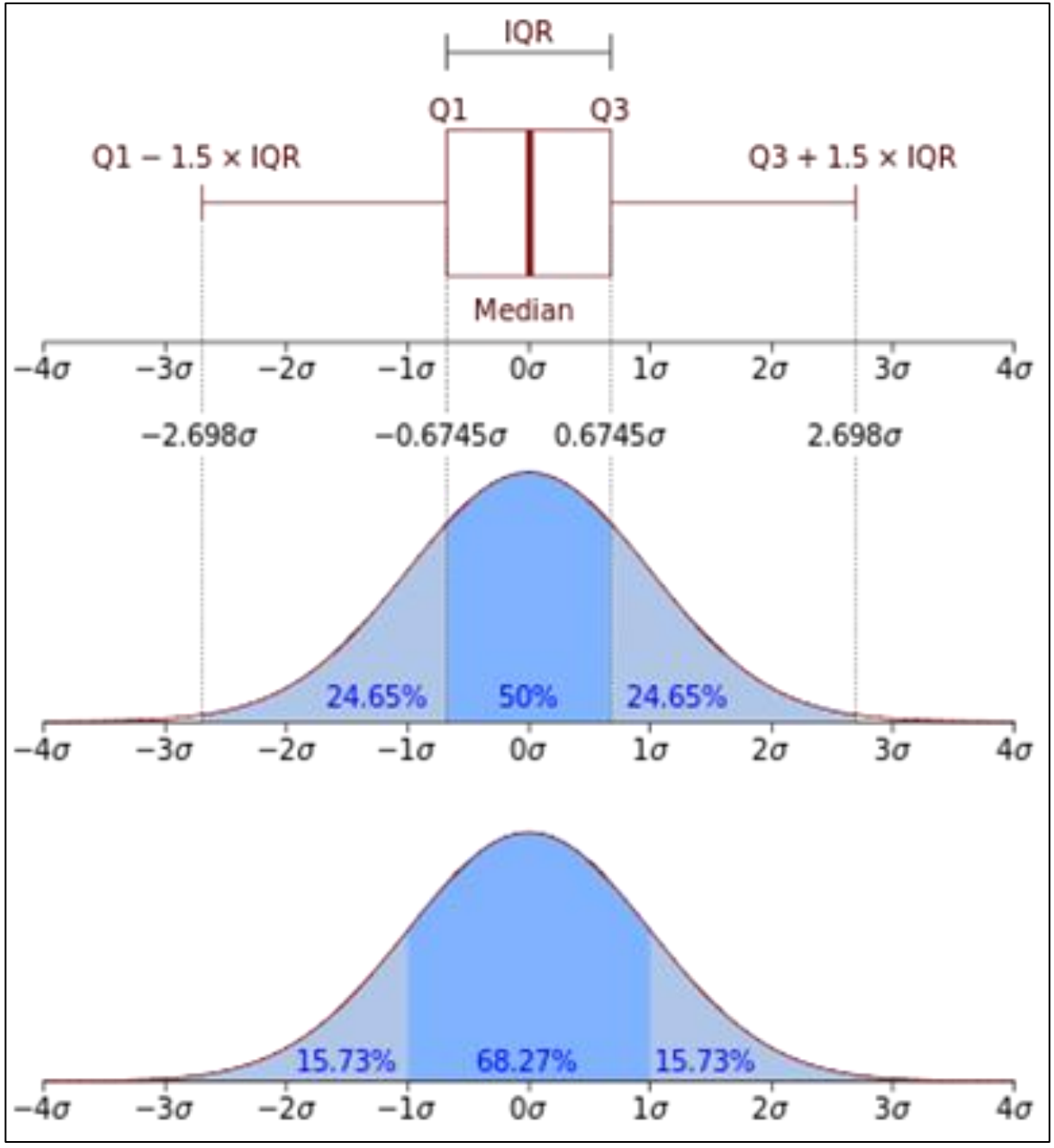
DF can be defined as the minimum number of independent coordinates that can specify the position of the system completely.

Estimates of statistical parameters can be based upon different amounts of information or data. The number of independent pieces of information that go into the estimate of a parameter is called the degrees of freedom.

DF of an estimate of a parameter is equal to the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself (which, in sample variance, is one, since the sample mean is the only intermediate step).

Interquartile Range (IQR)

In descriptive statistics, the **Interquartile Range (IQR)**, also called the mid-spread or middle fifty, is a measure of statistical dispersion, being equal to the difference between the upper and lower quartiles, IQR = $Q_3 - Q_1$.

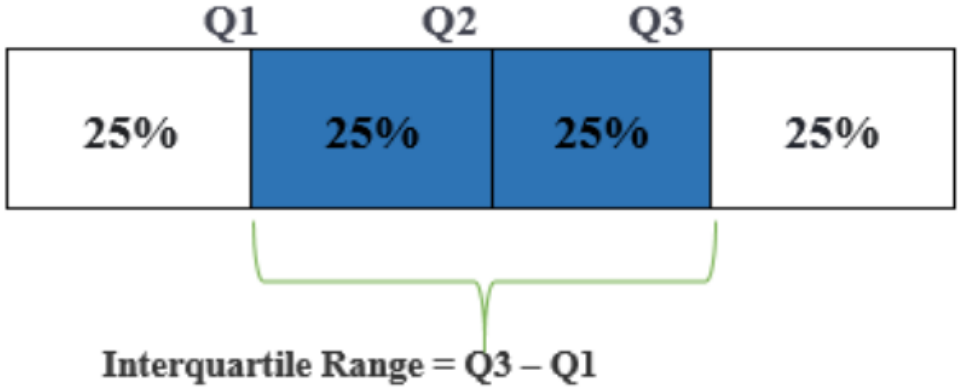


Boxplot (with an interquartile range) and a probability density function (pdf) of a Normal N (0, σ^2) Population

Interquartile Range (IQR)

IQR is a range (*the boundary between the first and second quartile*) **and Q3** (**the boundary between the third and fourth quartile**).**IQR is preferred over a range as**, *like a range, IQR does not influence by outliers.*

IQR is used to **measure variability** by splitting a data set into four equal quartiles.



Example:

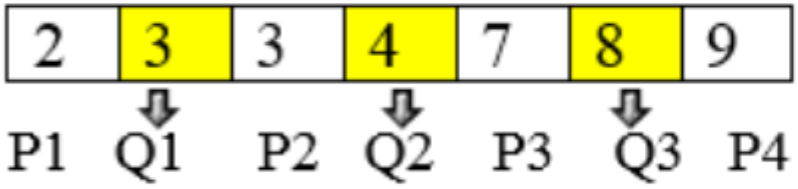
a) Odd number of elements

4, 8, 3, 3, 7, 2, 9

Sort the values

2, 3, 3, 4, 7, 8, 9

Divide into four equal parts



Quartile 1 (Q1) = **3**

Quartile 2 (Q2) or Median = **4**

Quartile 3 (Q3) = **8**

P1, P2, P3, P4 are four parts.

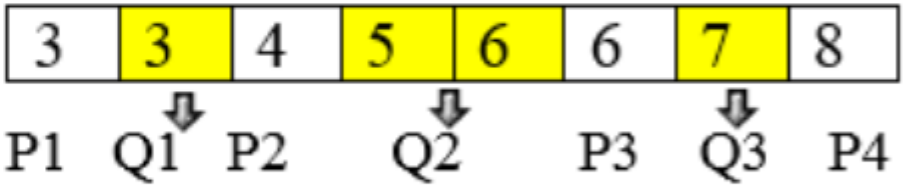
b) Even number of elements

7, 8, 4, 6, 5, 6, 3, 3

Sort the values

3, 3, 4, 5, 6, 6, 7, 8

Divide into four equal parts



Quartile 1 (Q1) = **3**

Quartile 2 (Q2) or Median =5+6/2 ⟹ **5.5**

Quartile 3 (Q3) = 7

P1, P2, P3, P4 are four parts.

IQR uses a box plot to find the outliers. *"To estimating IQR, all the values form (sort) in the ascending order else it will provide a negative value, and that influences to find the outliers."*

Formula to find outliers

[Q1 – 1.5 * IQR, Q3 + 1.5 * IQR]

If the value does not fall in the above range it considers outliers.

Example: 5, 8, 15, 20, 10, 18, 3, 12, 6, 14, 11. Check 20 is outlier or not

Values sort in ascending order

3	5	6	8	10	11	12	14	15	18	20
		↓ Q1			↓ Q2			↓ Q3		

IQR = Q3 – Q1

15- 6 ⟹ 9

[6 – 1.5 * 9, 15 + 1.5 * 9]

↓

[(-7.5), 28.5]

20 is not outlier as fall in above range

Values sort in descending order

20	18	15	14	12	11	10	8	6	5	3
		↓ Q1			↓ Q2			↓ Q3		

IQR = Q3 – Q1

6 – 15 ⟹ -9

[6 – 1.5 * -9, 15 + 1.5 * -9]

↓

[(+7.5), -28.5]

20 is outlier as not fall in above range

Standard Deviation (SD)

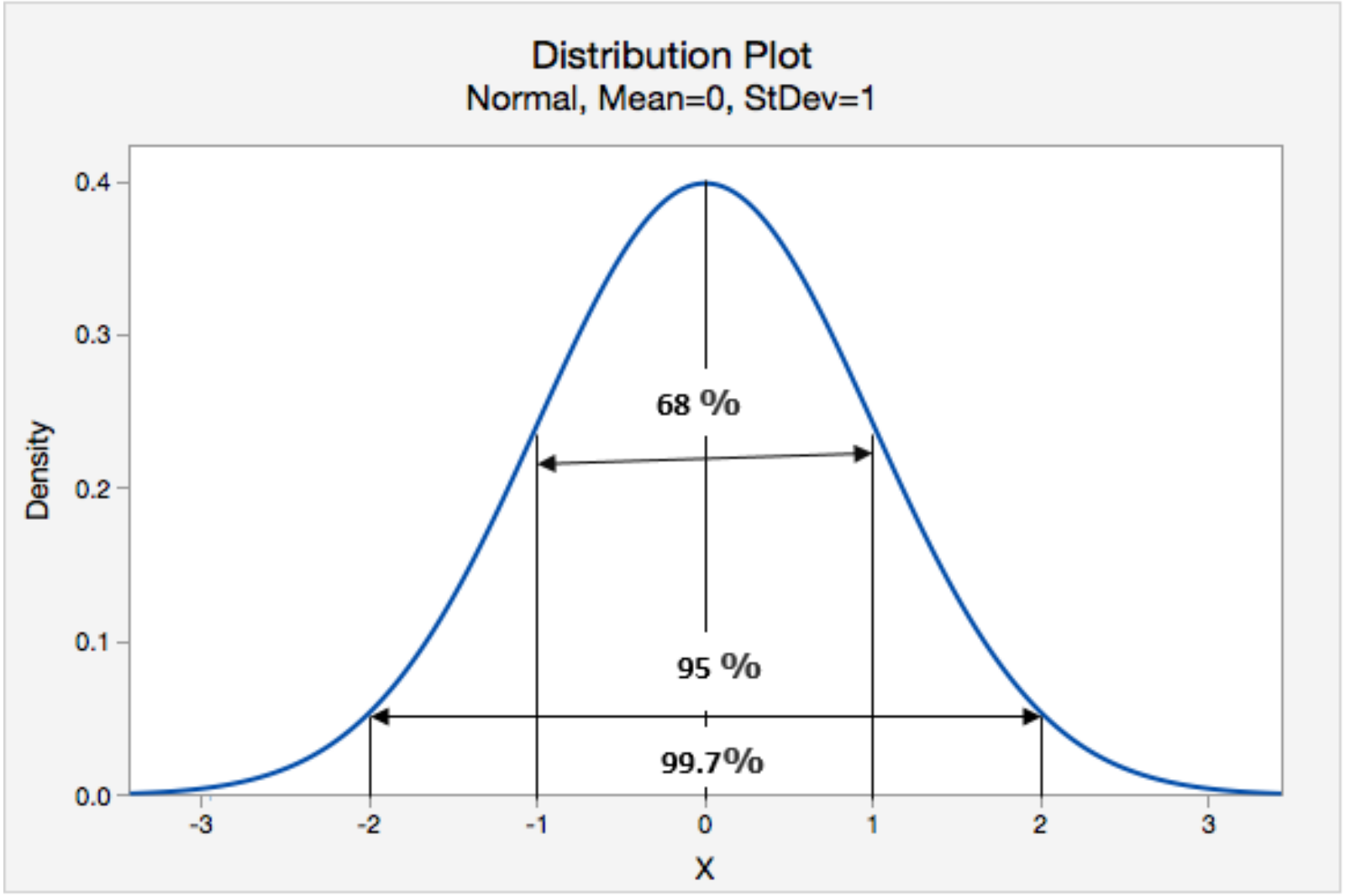
Standard deviation (SD) is a squared root of the variance to get original values. Low standard deviation indicates data points close to mean.

In statistics and probability theory, the **standard deviation** (σ) shows how much variation or dispersion from the average exists.

A low standard deviation indicates that the data points tend to be very close to the mean (also called expected value).

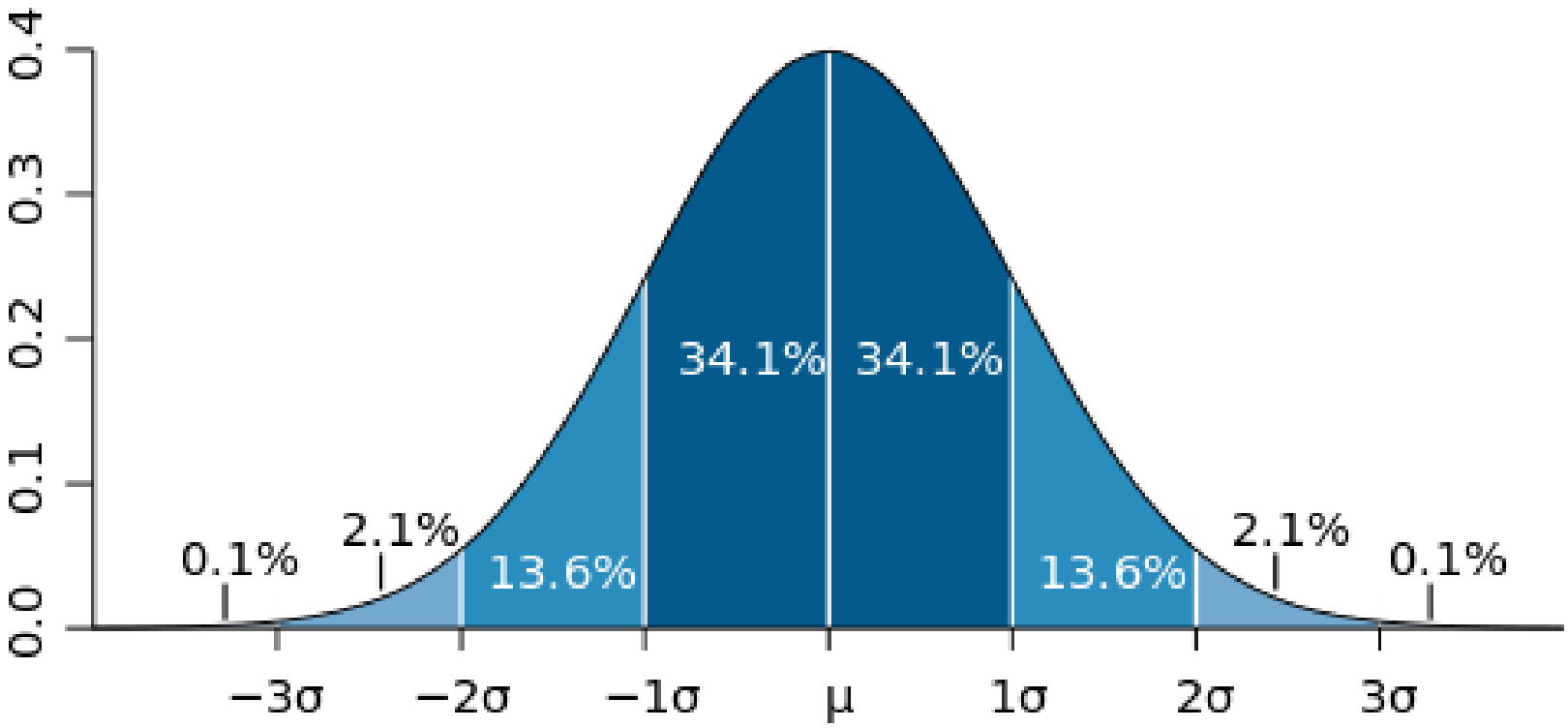
A high standard deviation indicates that the data points are spread out over a large range of values.

For a finite set of numbers, the standard deviation is found by taking the square root of the average of the squared differences of the values from their average value.



X indicates the mean value

68 % of values lie within 1 standard deviation.
95 % of values lies within 2 standard deviation.
99.7 % of values lie within 3 standard deviation.



A plot of a normal distribution (or bell-shaped curve) where each band has a width of 1 standard deviation

For a finite set of numbers, the standard deviation is found by taking the square root of the average of the squared differences of the values from their average value. For example, consider a **population** consisting of the following eight values:

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\frac{2+4+4+4+5+5+7+9}{8} = 5.$$

First, calculate the difference of each data point from the mean, and square the result of each:

$$\begin{aligned}
(2-5)^2 &= (-3)^2 = 9 & \quad (5-5)^2 &= 0^2 = 0 \\
(4-5)^2 &= (-1)^2 = 1 & \quad (5-5)^2 &= 0^2 = 0 \\
(4-5)^2 &= (-1)^2 = 1 & \quad (7-5)^2 &= 2^2 = 4 \\
(4-5)^2 &= (-1)^2 = 1 & \quad (9-5)^2 &= 4^2 = 16.
\end{aligned}$$

Next, calculate the mean of these values, and take the square root:

$$\sqrt{\frac{9+1+1+1+0+0+4+16}{8}} = 2.$$

This quantity is the *population* standard deviation, and is equal to the square root of the variance.

Population SD:

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

Sample SD:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

Standard Error (SE)

The **standard error (SE)** is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.

In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

$$\mathrm{SE} = \frac{\sigma}{\sqrt{n}}$$

SE = standard error of the sample

σ = sample standard deviation

n = number of samples

Coefficient of Variation (CV)

In probability theory and statistics, the **coefficient of variation (CV)** is a normalized measure of dispersion of a probability distribution or frequency distribution.

It is also known as unitized risk or the **variation coefficient**.

The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage.

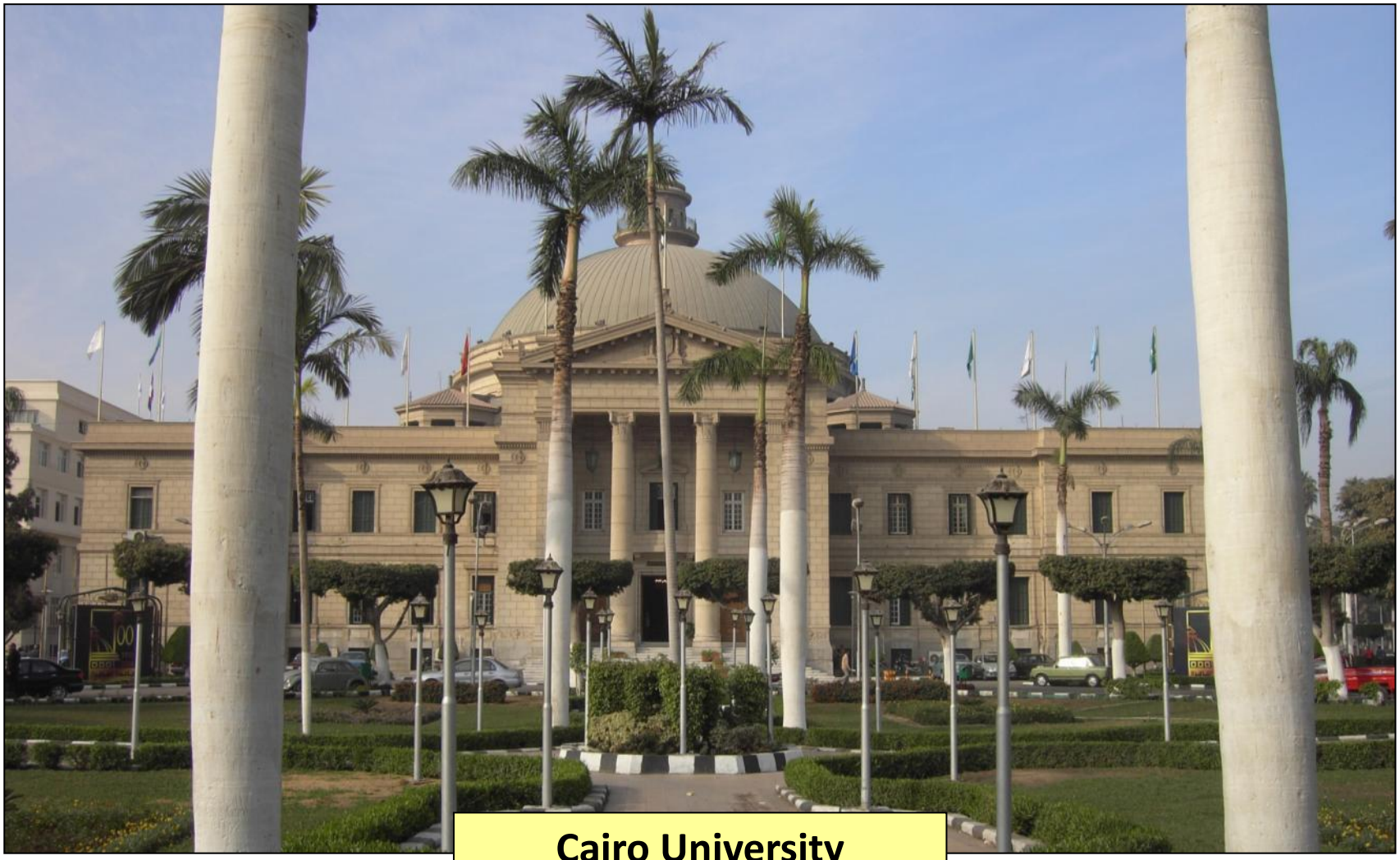
The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ:

$$c_v = \frac{\sigma}{\mu}$$

It shows the extent of variability in relation to mean of the population.

Exercises and Solutions

Thank You!



Cairo University