UNIT 2

BASIC STATISTICAL DESCRIPTIONS OF DATA

APEC

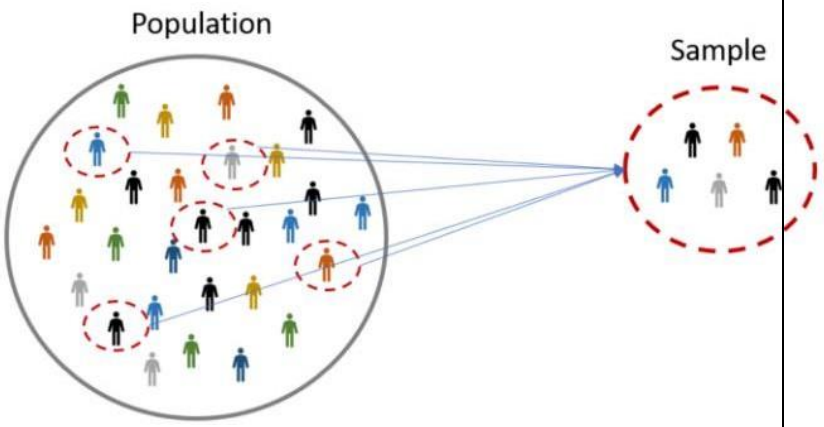
Define Statistics

- Statistics simply means numerical data, and it is a field of mathematics that generally deals with collection of data, tabulation, and interpretation of numerical data.
  - It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life.

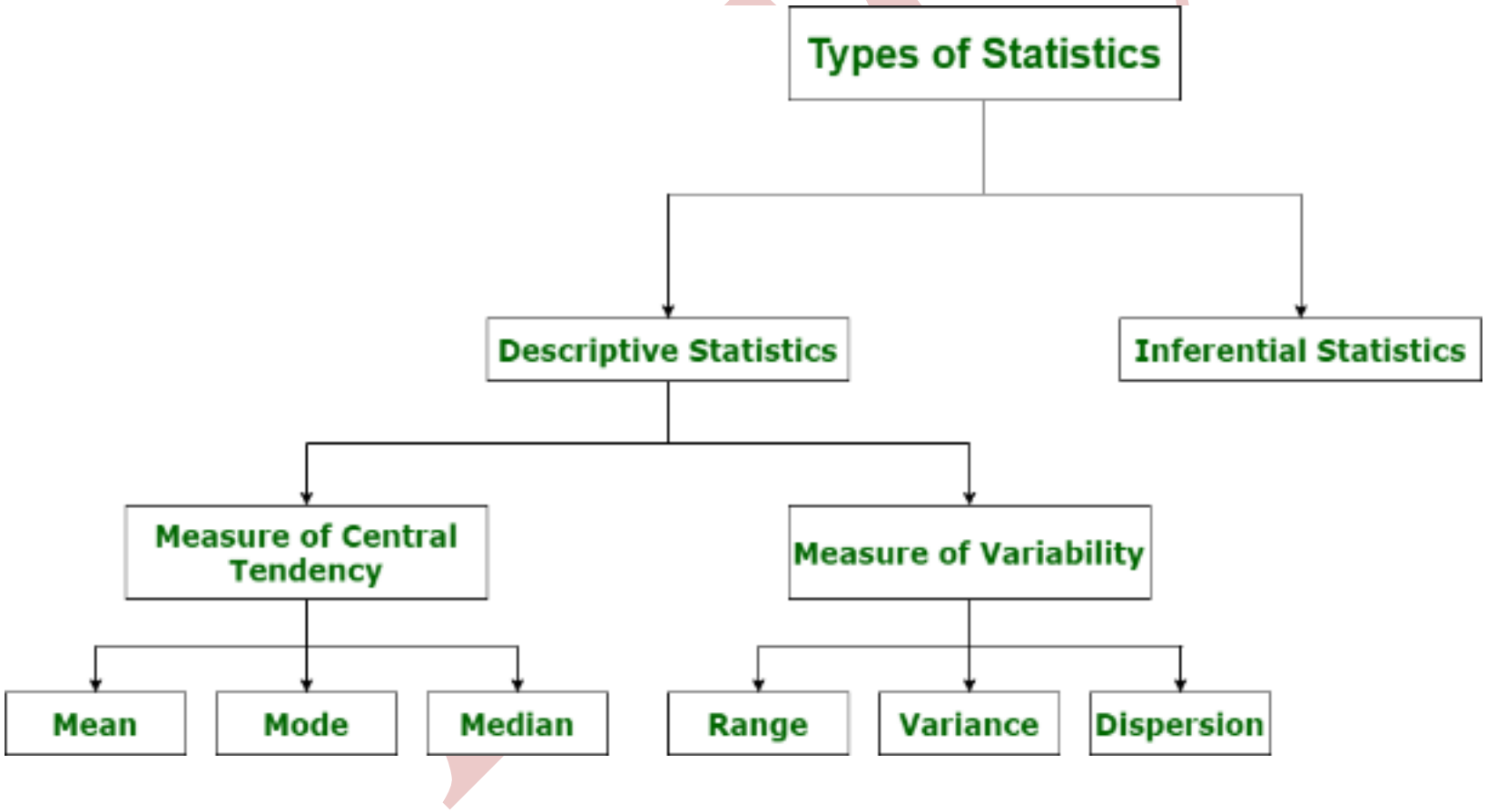
Basic terminology of Statistics :

Population -
It is a collection of individuals whose properties are to be analyzed.

Sample -
It is the subset of a population.



Types of Statistics



Descriptive Statistics :

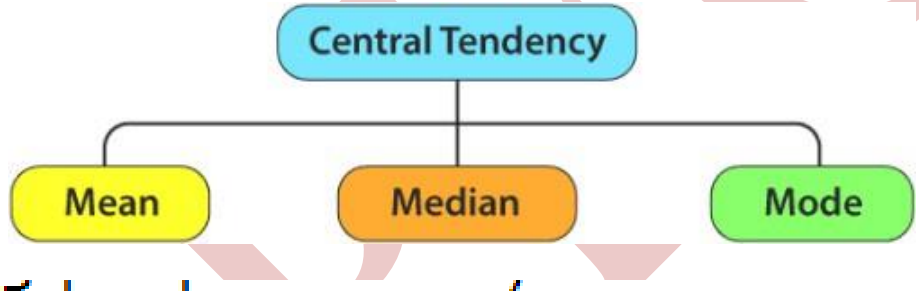
- Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table.
  - It provides a graphical summary of data.
  - It is simply used for summarizing objects, etc.

- The two categories are as follows:
- **Measure of central tendency**
  - Measure of central tendency is also known as summary statistics that is used to represents the centre point or a particular value of a data set or sample set.
- **Measure of Variability**
  - Measure of Variability is also known as measure of dispersion and used to describe variability in a sample or population.

APEC

Measure of central tendency

- Measure of central tendency is also known as **summary statistics** that is used to represents the centre point or a particular value of a data set or sample set.



$$Mean = \frac{sum\ of\ all\ values}{total\ number\ of\ values}$$

$$Median = middle\ value\ (when\ the\ data\ are\ arranged\ in\ order)$$

$$Mode = most\ common\ value$$

Mean

Add all the numbers then divide by the amount of numbers

9, 3, 1, 8, 3, 6

$9 + 3 + 1 + 8 + 3 + 6 = 30$

$30 \div 6 = 5$

The mean is 5

Median

Order the set of numbers, the median is the middle number

9, 3, 1, 8, 3, 6

1, 3, 3, 6, 8, 9

The median is 4.5

Mode

The most common number

9, 3, 1, 8, 3, 6

The mode is 3

Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

mean = (2+4+4+6+8+24)/6 = 8	average of all values
median = (4+6)/2 = 5	avg. of two middle values
mode = 4	most frequent item

attr2 = {2,4,7,10,12}

mean = (2+4+7+10+12)/5 = 7	average of all values
median = 7	middle value
mode = any of them (no mode)	all of them has same freq.

attr3 = {xs,s,s,s,m,m,l}

mean is meaningless for categorical attributes.

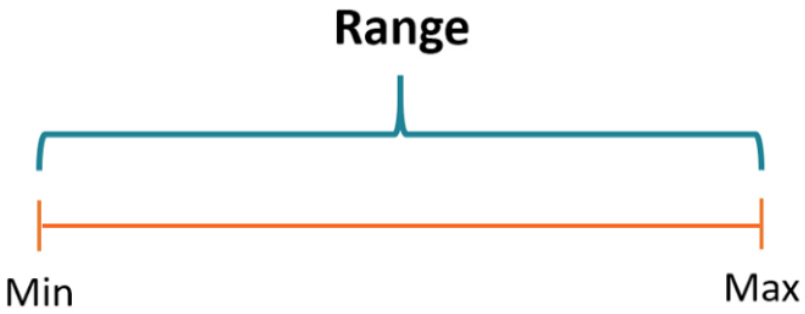
median = s	middle value
mode = s	most frequent item

Measuring Dispersion of Data

- The degree to which **numerical data tend to spread** is called the **dispersion, or variance** of the data.
- The measures include
  - **range,**
  - **quantiles,**
  - **quartiles,**
  - **percentiles, and**
  - **the interquartile range.**
- The **five-number summary,** which can be displayed as a **boxplot**, is useful in identifying outliers.
- **Variance and standard deviation** also indicate the spread of a data distribution.

Range

- Range is the difference between the largest and smallest values in a dataset.
  - It is one of the method in Measures of Dispersion/Variability.



Range = Maximum value - Minimum value

Example:
In **{8, 11, 5, 9, 7, 6, 2500}**
Lowest value is 5, Highest value is 2500,
Range = 2500 – 5 = **2495**.

quartiles

Quartiles divide a sorted data set into quarters by finding the median of all the scores, then finding the median of the lower and upper halves of the scores.

If the number of scores is even,
the median is the average of the two middle scores.

finding the quartiles

1. Arrange the data scores in ascending order.
2. Find the median of the data set (the number in the middle) … Q2.
3. Find the median of the lower half of the scores … Q1.
4. Find the median of the upper half of the scores … Q3.

a set of test scores

60 60 40 74 64 65 88 41 70 42 57 30 58 66 66 68

scores sorted in ascending order

30 40 41 42 57 58 60 60 64 65 66 66 68 70 74 88

finding the medians

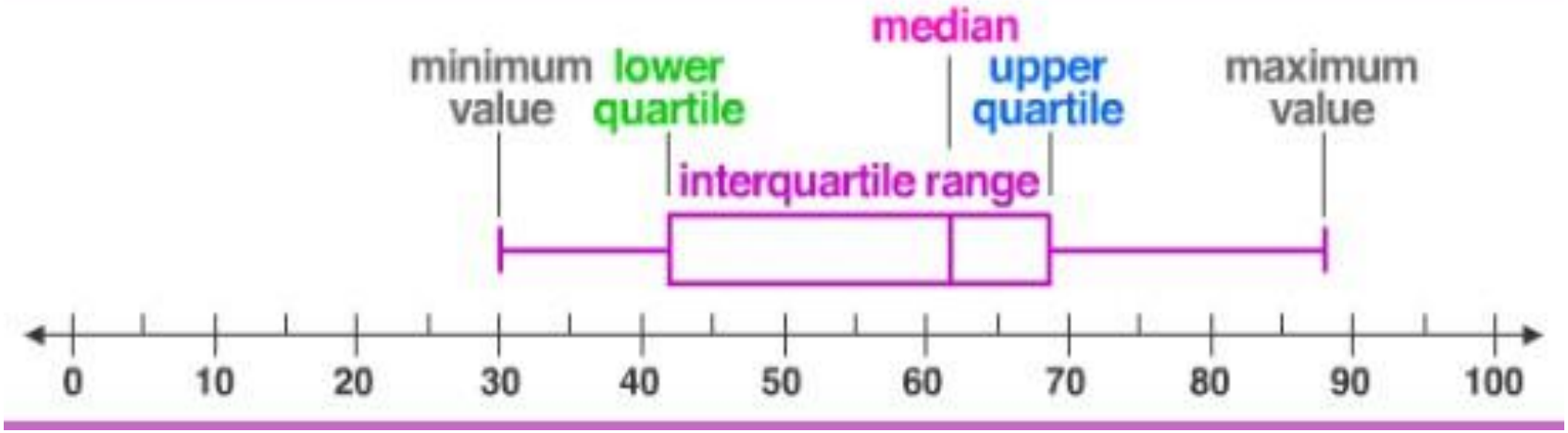
30 40 41 42 57 58 60 60 64 65 66 66 68 70 74 88

lower quartile 62 upper quartile

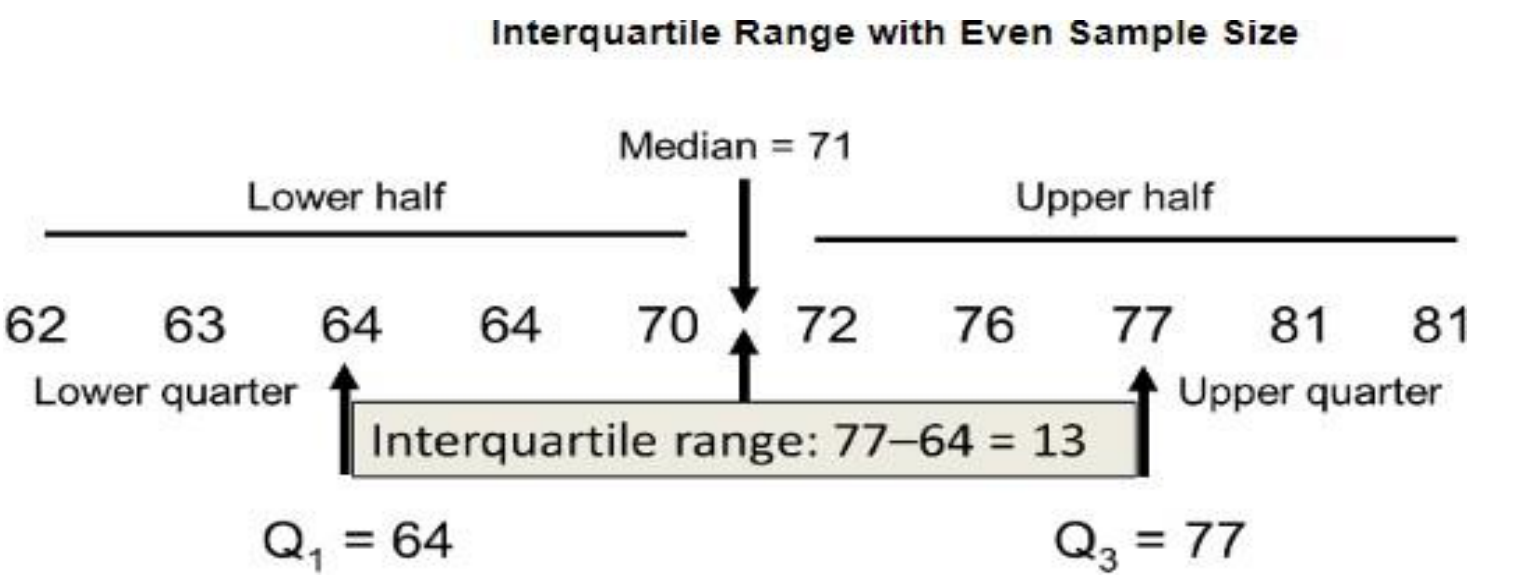
Q1 Q2 Q3

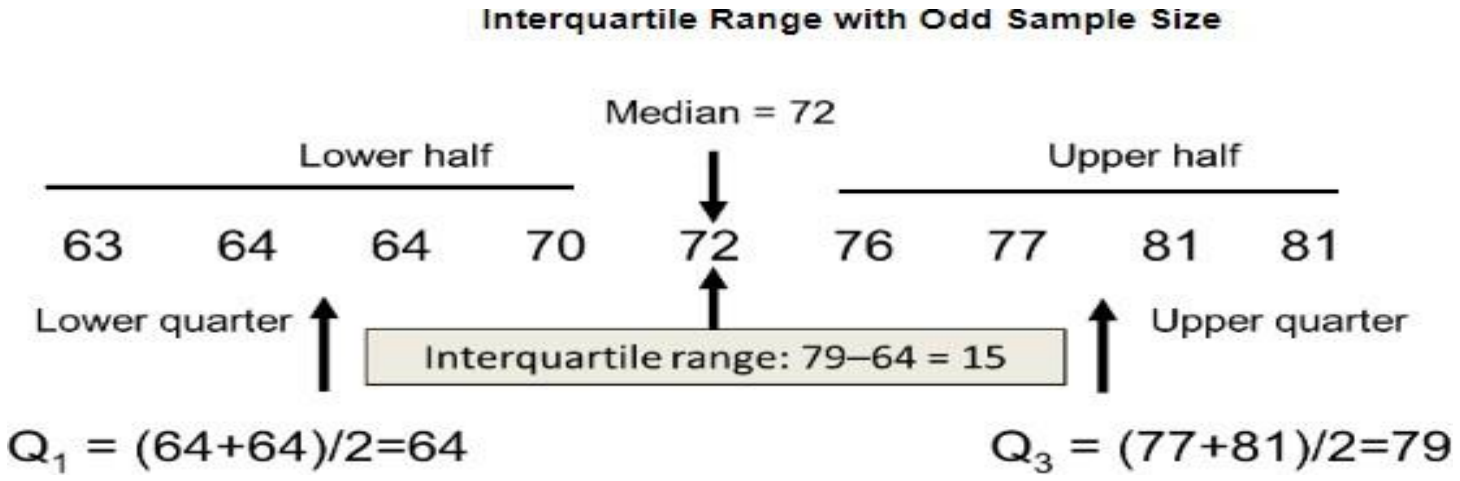
to make a box-and-whisker plot (box plot) of the data set

1. Draw a box - from the lower to the upper quartile.
2. Draw a line in the box to show the median.
3. Draw a line from the left of the box to the lowest score.
4. Draw a line from the right of the box to the highest score.



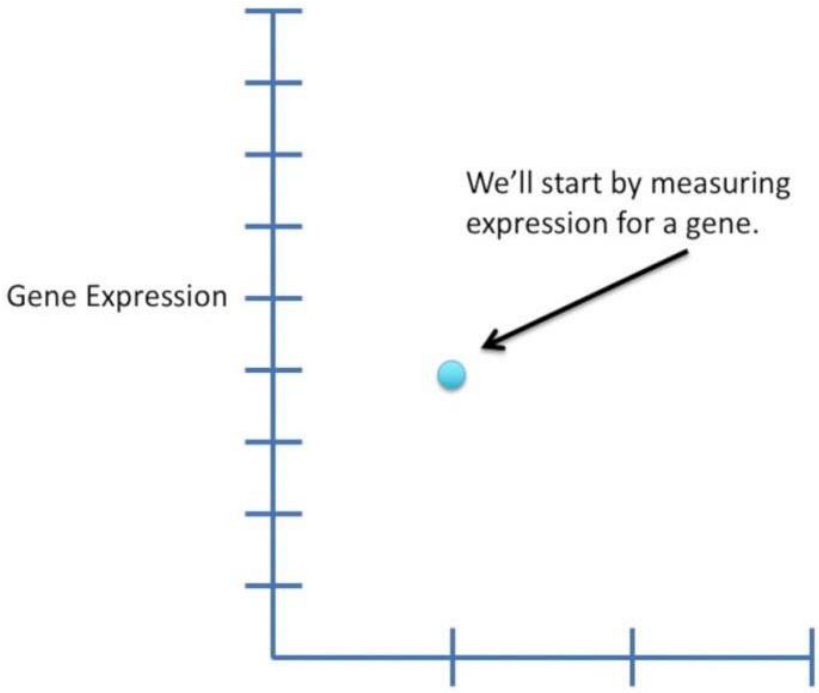
- **Formula for Interquartile Range = Q_3-Q_1**

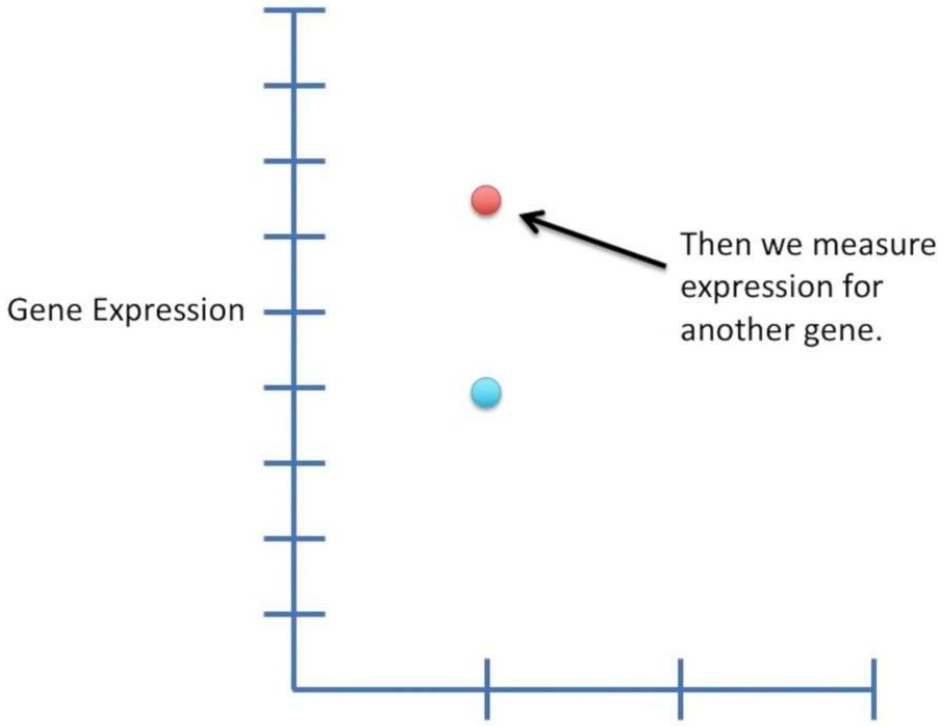

Interquartile Range with Even Sample Size
Median = 71
Lower half
Upper half
62 63 64 64 70 72 76 77 81 81
Lower quarter
Upper quarter
Interquartile range: 77–64 = 13
$Q_1 = 64$
$Q_3 = 77$

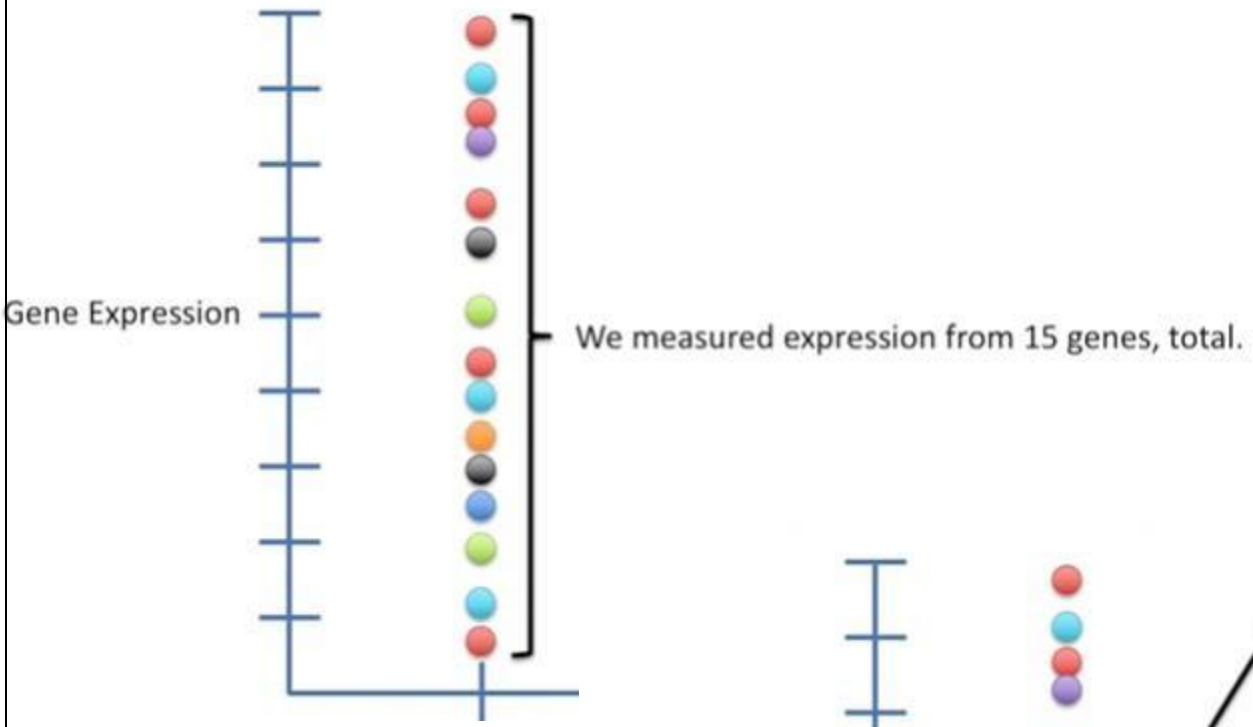

Interquartile Range with Odd Sample Size
Median = 72
Lower half
Upper half
63 64 64 70 72 76 77 81 81
Lower quarter
Upper quarter
Interquartile range: 79–64 = 15
$Q_1 = (64+64)/2=64$
$Q_3 = (77+81)/2=79$

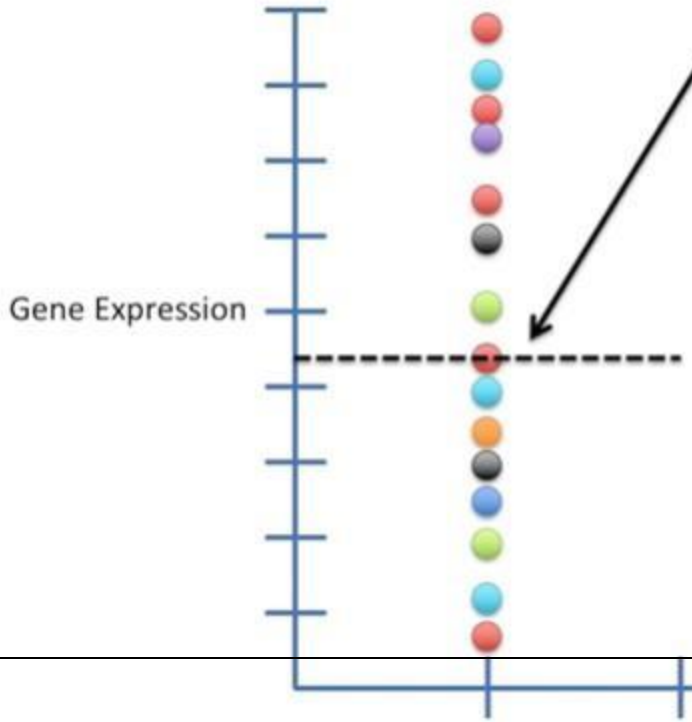
Quantiles

- **Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities**





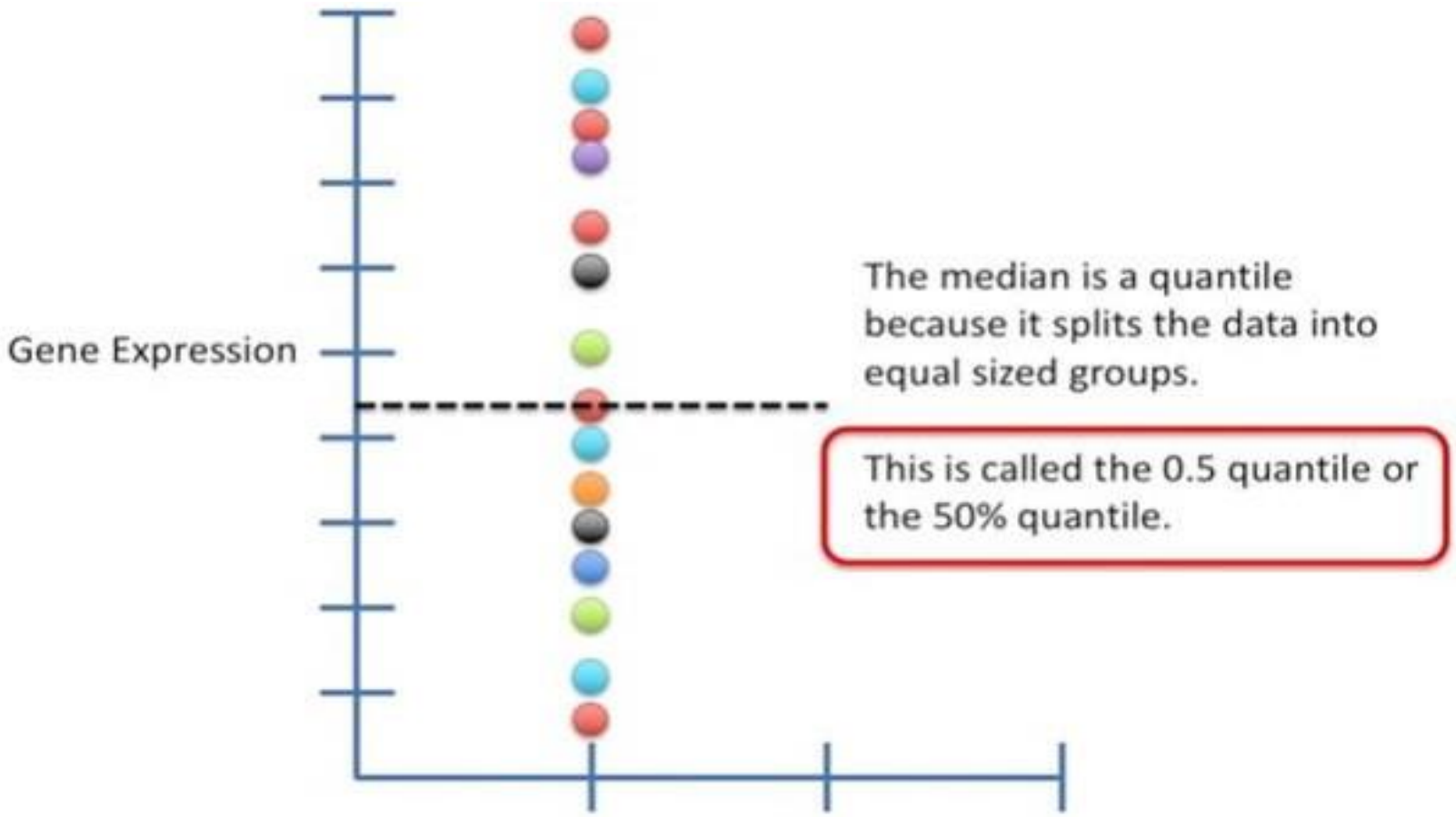


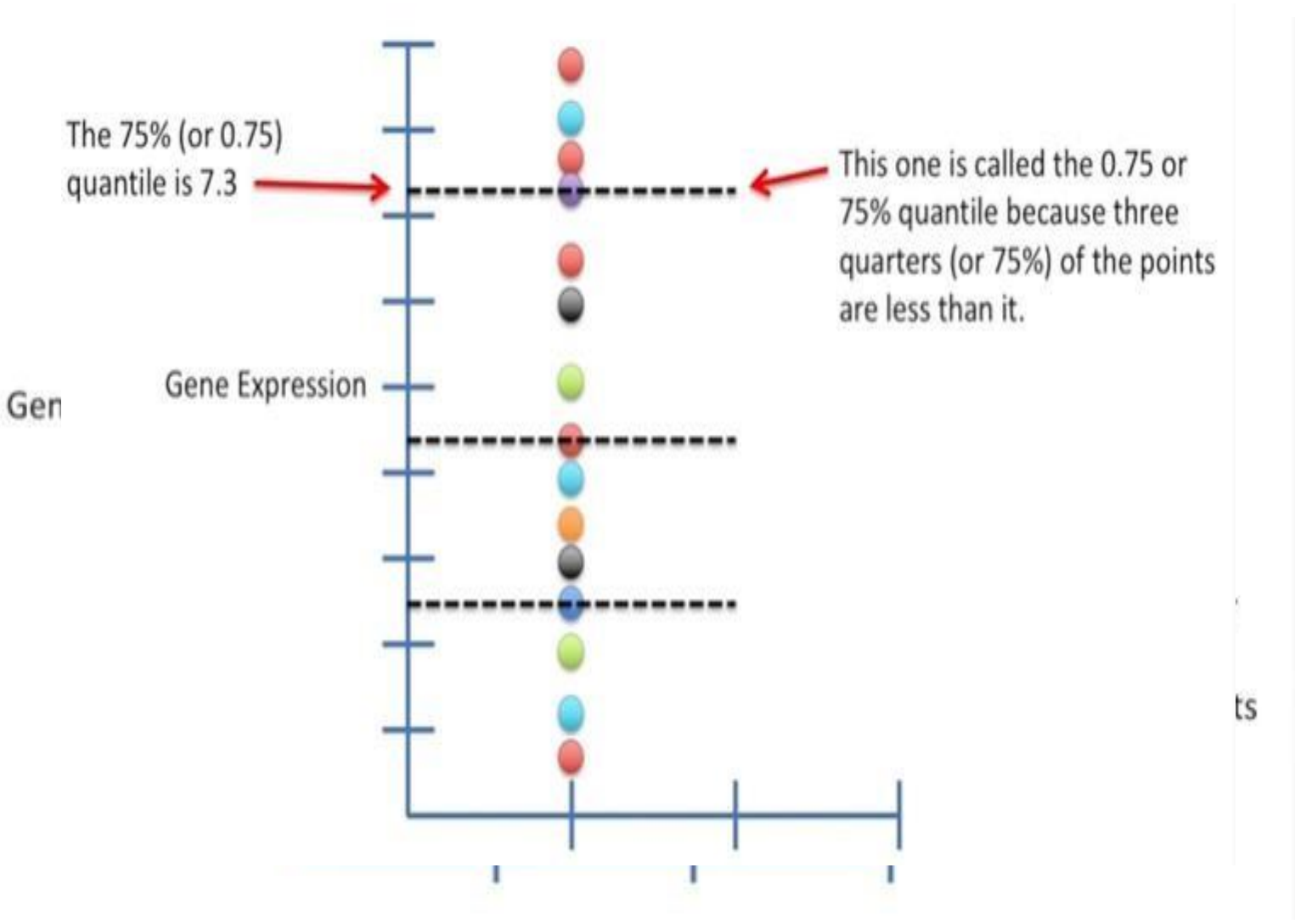


This is the **median** value; 50% of the genes have higher expression and 50% of the genes have lower expression.

Technically speaking, **the median is a quantile** because it splits the data into groups that contain the same number of data points.

Sometimes this quantile (the median) is labeled **0.5** (since it splits the data in half) and sometimes it is labeled **50%** (since 50% of the data is above it and 50% is below it.)


Gene Expression
The median is a quantile because it splits the data into equal sized groups.
This is called the 0.5 quantile or the 50% quantile.


The 75% (or 0.75) quantile is 7.3
This one is called the 0.75 or 75% quantile because three quarters (or 75%) of the points are less than it.
Gene Expression

What is a Quantile?

- The word “quantile” comes from the word **quantity**. In simple terms, a quantile is where a sample is divided into equal-sized, adjacent, subgroups. It can also refer to dividing a probability distribution into areas of equal probability.

  - The median is a quantile;
    - the median is placed in a probability distribution so that exactly half of the data is lower than the median and half of the data is above the median.
    - The median cuts a distribution into two equal areas and so it is sometimes called 2-quantile.

  - Quartiles are also quantiles;
    - they divide the distribution into four equal parts.
    - Percentiles are quantiles that divide a distribution into 100 equal parts
    - Deciles are quantiles that divide a distribution into 10 equal parts.

Quick view:

- Here's a simple definition of each:
  - **Percentiles:** Range from 0 to 100.
  - **Quartiles:** Range from 0 to 4.
  - **Quantiles:** Range from any value to any other value.

- Note that percentiles and quartiles are simply *types* of quantiles.
  - 4-quantiles are called *quartiles*.
  - 5-quantiles are called *quintiles*.
  - 8-quantiles are called *octiles*.
  - 10-quantiles are called *deciles*.
  - 100-quantiles are called *percentiles*.

- Note that percentiles and quartiles share the following relationship:
  - 0 percentile = 0 quartile (also called the minimum)
  - 25th percentile = 1st quartile
  - 50th percentile = 2nd quartile (also called the median)
  - 75th percentile = 3rd quartile
  - 100th percentile = 4th quartile (also called the maximum)

Data
3
4
4
6
7
9
12
13
14
16
17
19
22
23
23
25
28
29
34
37

- Here's how to interpret these values:
  - The 0 percentile and 0 quartile is **3**.
  - The 25th percentile and 1st quartile is **8.5**.
  - The 50th percentile and 2nd quartile is **16.5**.
  - The 75th percentile and 3rd quartile is **23.5**.
  - The 100th percentile and 4th quartile is **37**.

HOW TO FIND QUANTILES?

- Find the number in the following set of data where 20 percent of values fall below it, and 80 percent fall above:
-

1 3 5 6 9 11 12 13 19 21 22 32 35 36 45 44 55 68 79 80 81 88 90 91 92 100 112 113 114 120 121 132 145 146 149 150 155 180 189 190

Solution

- Step 1: Order the data from smallest to largest.
  Step 2: Count how many observations you have in your data set. this particular data set has 40 items.
  Step 3: Convert any percentage to a decimal for "q". We are looking for the number where 20 percent of the values fall below it, so convert that to .2.
  Step 4: Insert your values into the formula:
  ith observation = q (n + 1)
  ith observation = .2 (40 + 1) = 8.2
  Answer:
- The ith observation is at 8.2, so we round down to 8. The 8th number in the set is 13, which is the number where 20 percent of the values fall below it.

Variance

- It simply describes how much a random variable defers from expected value and it is also computed as square of standard deviation.

$$\sigma^2 = \frac{\sum(xi - \bar{x})^2}{N}$$

- Example:
  - Heights : 600mm, 470mm, 170mm, 430mm and 300mm.
  - Find out the Mean, the Variance, and the Standard Deviation.

- Answer:
- Step 1 : Find the Mean:

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5}$$

$$= \frac{1970}{5}$$

$$= 394$$

- Step 2: Calculate the Variance. Calculate each height's difference from the Mean. Take each difference, square it, and then average the result.

Variance

$$\sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

$$= \frac{42436 + 5776 + 50176 + 1296 + 8836}{5}$$

$$= \frac{108520}{5}$$

$$= 21704$$

- Step 3: Standard Deviation is just the square root of Variance, so:

Standard Deviation

$$\sigma = \sqrt{21704}$$

$$= 147.32...$$

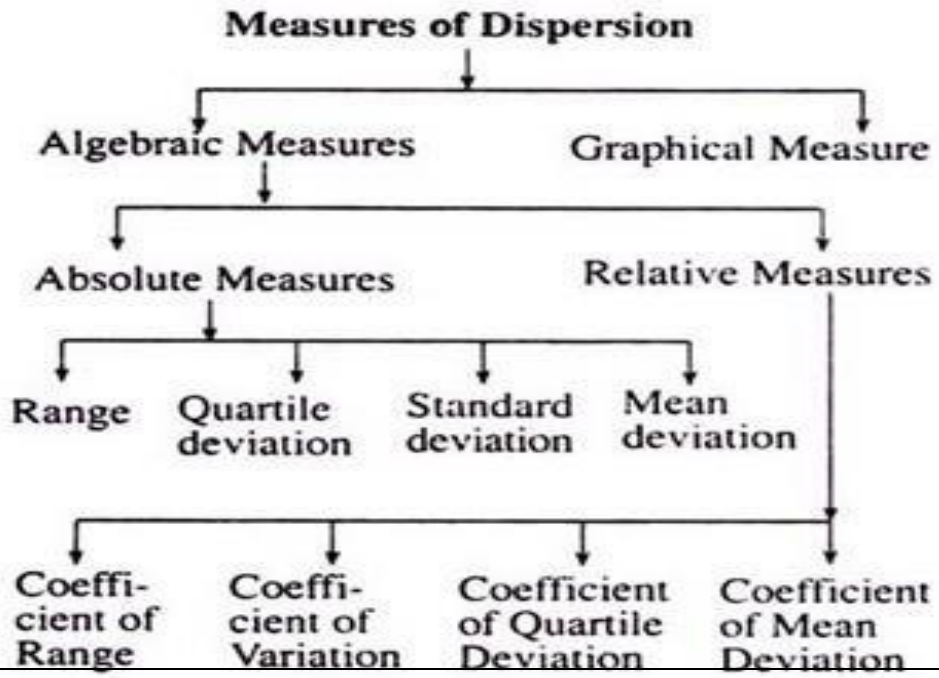
$$= \mathbf{147}\ \textit{(to the nearest mm)}$$

- **Population Variance**
  - The variance is the average of the squared deviations about the mean for a set of numbers. The population variance is denoted by σ^2.
- **Sample Variance**
  - The sample variance is denoted by s2. The main use for sample variances is as estimators of population variances. The computation of the sample variance differs slightly from computation of the population variance. The sample variance uses n – 1.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

What is Dispersion?

- Dispersion in statistics is a way of describing how spread out a set of data is.
  - When a data set has a large value, the values in the set are widely scattered;
  - when it is small the items in the set are tightly clustered.



Coefficient of variation

- The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the CV, the greater the dispersion.

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

σ = population standard deviation

μ = population mean

- **Question**

Given : The mean of a data is 25.6 and its coefficient of variation is 18.75.

To find : The standard deviation ?

- **Answer**

The standard deviation is 4.8.

Example

The following table gives the values of mean and variance of heights and weights of the 10th standard students of a school.

	Height	Weight
Mean	155 cm	46.50 kg^2
Variance	72.25 cm^2	28.09 kg^2

Which is more varying than the other?

Solution

For comparing two data, first we have to find their coefficient of variations

Mean $\overline{x}$ 1= 155cm, variance $\sigma_1^2 = 72.25\ cm^2$

Therefore standard deviation $\sigma_1 = 8.5$

Coefficient of variation

$$C.V_1 = \frac{\sigma_1}{\overline{x}_1} \times 100\%$$

$$C.V_1 = \frac{8.5}{155} \times 100\% = 5.48\% \quad \text{(for heights)}$$

Mean $\overline{x}$ 2 = 46.50 kg, Variance $\sigma_2^2 = 28.09\ kg2$

Standard deviation $\sigma_2 = 5.3kg$

Coefficient of variation

$$C.V_2 = \frac{\sigma_2}{\overline{x}_2} \times 100\%$$

$$C.V_2 = \frac{5.3}{46.50} \times 100\%$$

= 11.40% (for weights)

$C.V_1 = 5.48\%$ and $C.V_2 = 11.40\%$

Since $C.V_2 > C.V_1$, the weight of the students is more varying than the height.

Example

The consumption of number of guava and orange on a particular week by a family are given below.

Number of Guavas	3	5	6	4	3	5	4
Number of Oranges	1	3	7	9	2	6	2

Which fruit is consistently consumed by the family?

Solution

First we find the coefficient of variation for guavas and oranges separately.

x_i	x_i^2
3	9
5	25
6	36
4	16
3	9
5	25
4	16
$\Sigma x_i = 30$	$\Sigma x_i^2 = 136$

Number of guavas, $n = 7$

$$\text{Mean } \bar{x}_1 = \frac{30}{7} = 4.29$$

$$\text{Standard deviation } \sigma_1 = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

$$\sigma_1 = \sqrt{\frac{136}{7} - \left(\frac{30}{7}\right)^2} = \sqrt{19.43 - 18.40} \simeq 1.01$$

Coefficient of variation for guavas

$$\text{C.V}_1 = \frac{\sigma_1}{\bar{x}_1} \times 100\% = \frac{1.01}{4.29} \times 100\% = 23.54\%$$

x_i	x_i^2
1	1
3	9
7	49
9	81
2	4
6	36
2	4
$\Sigma x_i = 30$	$\Sigma x_i^2 = 184$

Number of oranges $n = 7$

$$\text{Mean } \bar{x}_2 = \frac{30}{7} = 4.29$$

$$\text{Standard deviation } \sigma_2 = \sqrt{\frac{\Sigma x_i^2}{n} - \left(\frac{\Sigma x_i}{n}\right)^2}$$

$$\sigma_2 = \sqrt{\frac{184}{7} - \left(\frac{30}{7}\right)^2} = \sqrt{26.29 - 18.40} = 2.81$$

Coefficient of variation for oranges

$$C.V_2 = \frac{\sigma_2}{\bar{x}_2} \times 100\% = \frac{2.81}{4.29} \times 100\% = 65.50\%$$

$C.V_1 = 23.54\%$, $C.V_2 = 65.50\%$ Since, $C.V_1 < C.V_2$, we can conclude that the consumption of guavas is more consistent than oranges.