Covariance, Outliers

Presented By

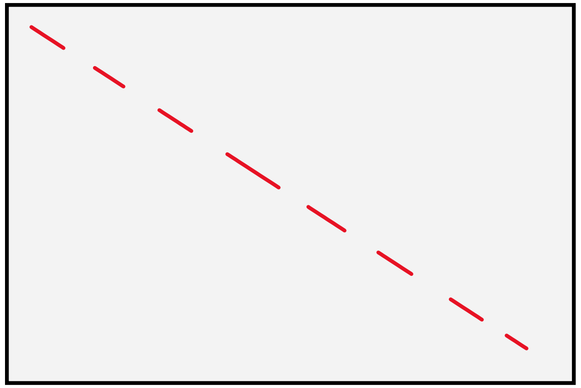
Aditya Joshi

Assistant Professor

Graphic Era Deemed to be University

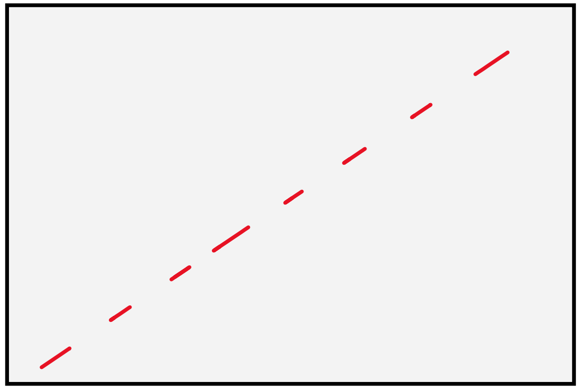
Covariance

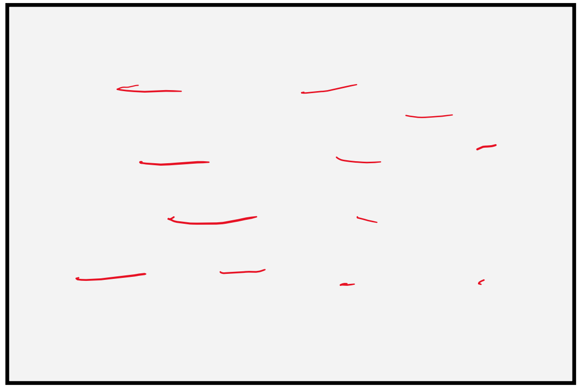
- It is a measure of how much two random variable vary together.
- It is similar to variance but variance tells you how a single variable varies. how far the data is spread out.
- Covariance tells you how two variable vary together.



Large Negative Covariance

Large positive Covariance

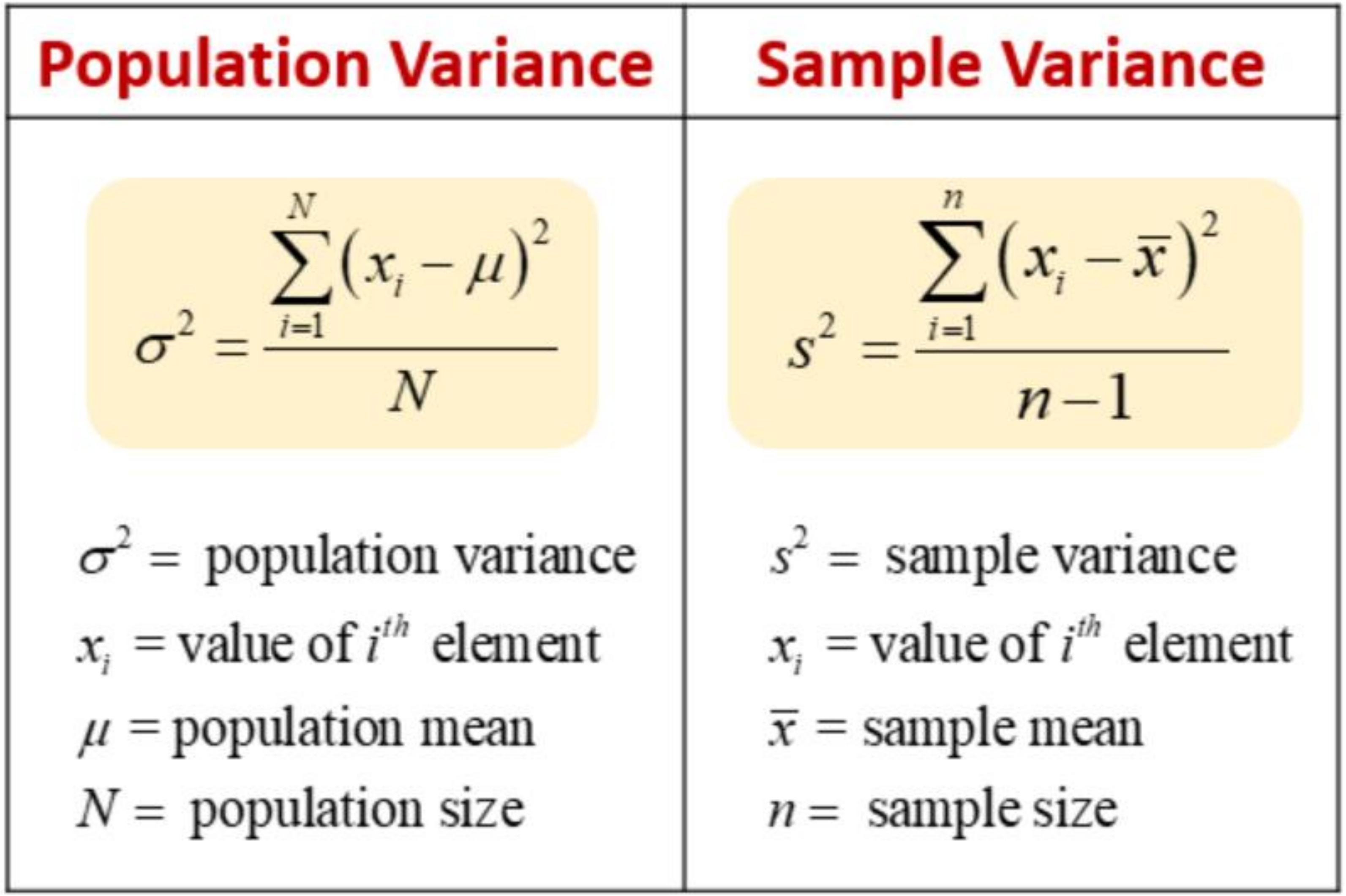




Near Zero Covariance

Variance: Average of the squared difference from the mean.

Variance smallest possible value is 0



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

Q. Find the population variance from the data below σ^2 σ

Years of experience @ Graphic Era
1
3
5
7
14

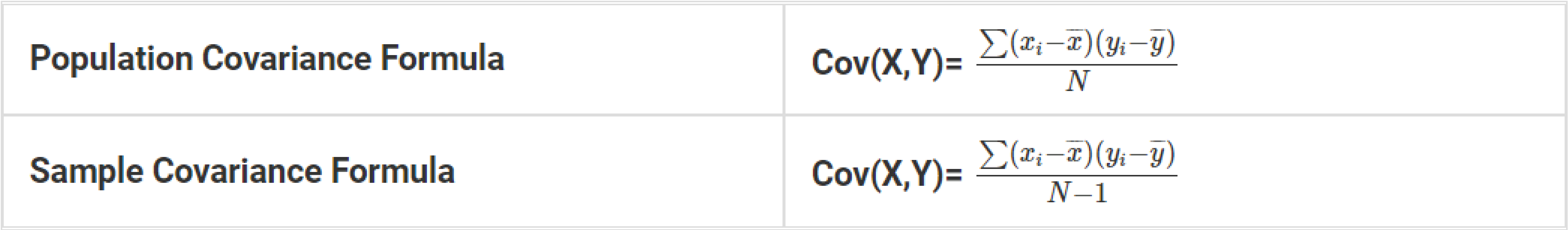
$\mu = 6$

$$\frac{(1-6)^2 + (3-6)^2 + (5-6)^2 + (7-6)^2 + (14-6)^2}{5}$$

$$= \frac{25 + 9 + 1 + 1 + 64}{5} = \frac{100}{5} = 20$$

Q. Consider the sample 6,3,8,5,3 find the sample variance s^2 & s

The Covariance formula



Population Covariance Formula	**Cov(X,Y)=** $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N}$
Sample Covariance Formula	**Cov(X,Y)=** $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$

Notations in Covariance Formulas

- x_i = data value of x
- y_i = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

Q. Calculate covariance for the following sample data set:
x: 2.1, 2.5, 3.6, 4.0
y: 8, 10, 12, 14

$\bar{x} = 3.05$

$\bar{y} = 11$

$$\sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

$$\frac{(2.1 - 3.05)(8-11) + (2.5 - 3.05)(10-11) + (3.6 - 3.05)(12-11) + (4.0 - 3.05)(14-11)}{3}$$

Outliers

An outliers is a piece of data that is an abnormal distance from other points. The data that lies outside the other values.

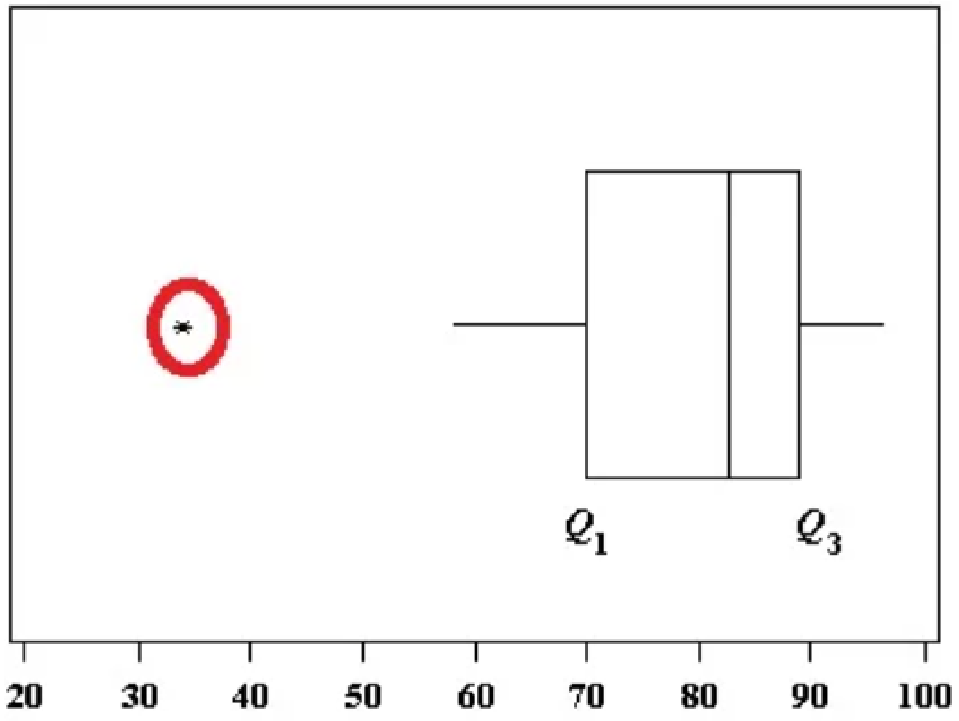
For e.g. in this set of random number

1, 99, 100, 101, 103, 109, 110, 201

1 & 201 are outliers because 1 has extremely low value and 201 has extremely high value.

A box and whiskers chart (boxplot) often shows outliers:

The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR.



An outlier is defined as being any point of data that lies over 1.5 IQRs below the first quartile (Q1) or above the third quartile (Q3)in a data set.

High = (Q3) + 1.5 IQR

Low = (Q1) – 1.5 IQR

$IQR = Q_3 - Q_1 = 36 - 14$

Q_1

$Q_3 = \boxed{22}$

Example Question: Find the outliers for the following data set: 3, 10, 14, 22, 19, 29, 70, 49, 36, 32.

$3, 10, \boxed{14}, 19, 22 \mid 29, 32, \boxed{36}, 49, 70$

Step 1: Find the IQR, Q1(25th percentile) and Q3(75th percentile).

IQR = 22

Q1 = 14

Q3 = 36

Step 2: Multiply the IQR you found in Step 1 by 1.5:

IQR * 1.5 = 22 * 1.5 = 33.

Step 3: Add the amount you found in Step 2 to Q3 from Step 1:

IQR×1.5 Q3

33 + 36 = 69.

This is your **upper limit**. Set this number aside for a moment.

Step 3: Subtract the amount you found in Step 2 from Q1 from Step 1:

Q1 IQR × 1.5

14 – 33 = -19.

This is your **lower limit**. Set this number aside for a moment.

Step 5: Put the numbers from your data set in order:

3, 10, 14, 19, 22, 29, 32, 36, 49, 70

Step 6: Insert your low and high values into your data set, in order:

-19, 3, 10, 14, 19, 22, 29, 32, 36, 49, 69, 70

Step 6: Highlight any number below or above the numbers you inserted in Step 6:

-**19**, 3, 10, 14, 19, 22, 29, 32, 36, 49, 69, **70**