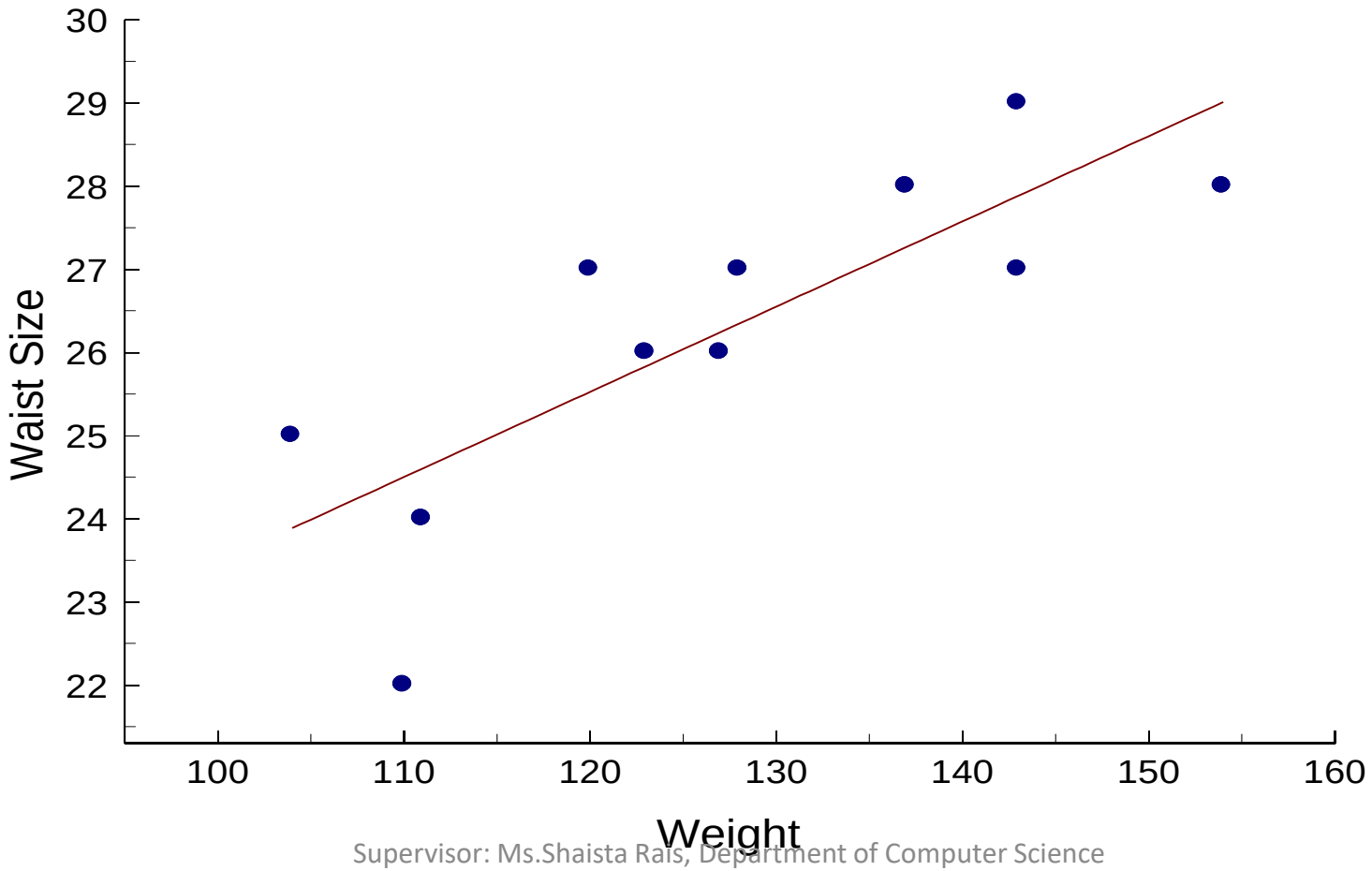
Linear Correlation
Regression Analysis
Covariance

SUPERVISOR:
Ms. Shaista Rais
Department of Computer Science
University of Karachi

Linear Correlation and Regression Analysis



Linear Correlation Analysis

- The **coefficient of linear correlation**, *r*, is a measure of the strength of a linear relationship.
- Consider another measure of dependence: **covariance**.
- Recall: **bivariate data** - ordered pairs of numerical values.

Covariance and Correlation

The terms covariance and correlation are very similar to each other in probability theory and statistics. **Both the terms describe the extent to which a random variable or a set of random variables can deviate from the expected value. But what is the difference between the terms?**

These two ideas are similar, but not the same. Both are used to determine the linear relationship and measure the dependency between two random variables. But are they the same? **Not really.**

Covariance and Correlation

- **Covariance** is when two variables vary with each other, whereas **Correlation** is when the change in one variable results in the change in another variable.
- Before having the difference between **Covariance** and **Correlation** , we must understand **Variance** and **Standard Deviation**

Variance

Variance is the expectation of the squared deviation of a random variable from its mean. Informally, it measures how far a set of numbers are spread out from their average value.

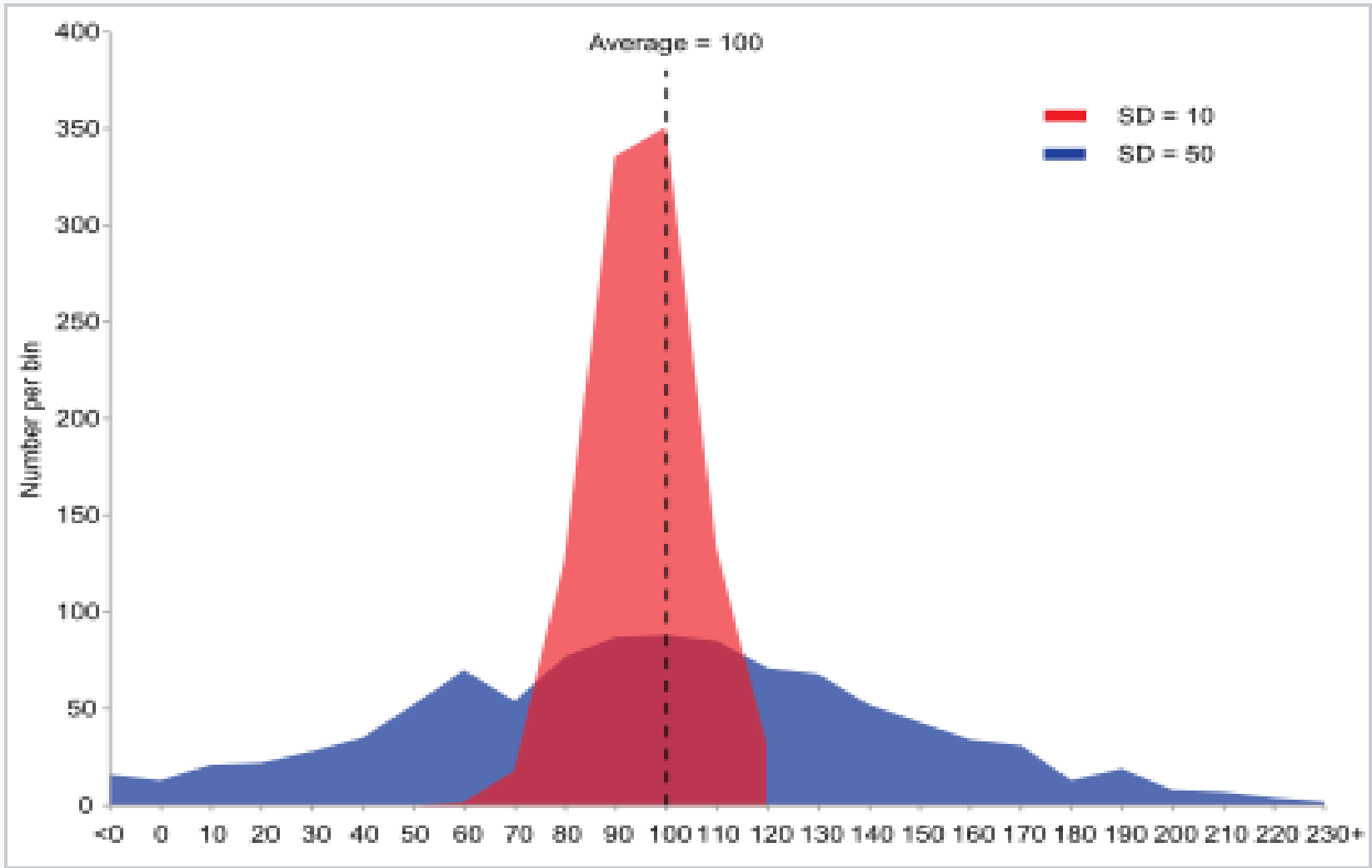
$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

S^2 = sample variance

x_i = the value of the one observation

$\bar{x}$ = the mean value of all observations

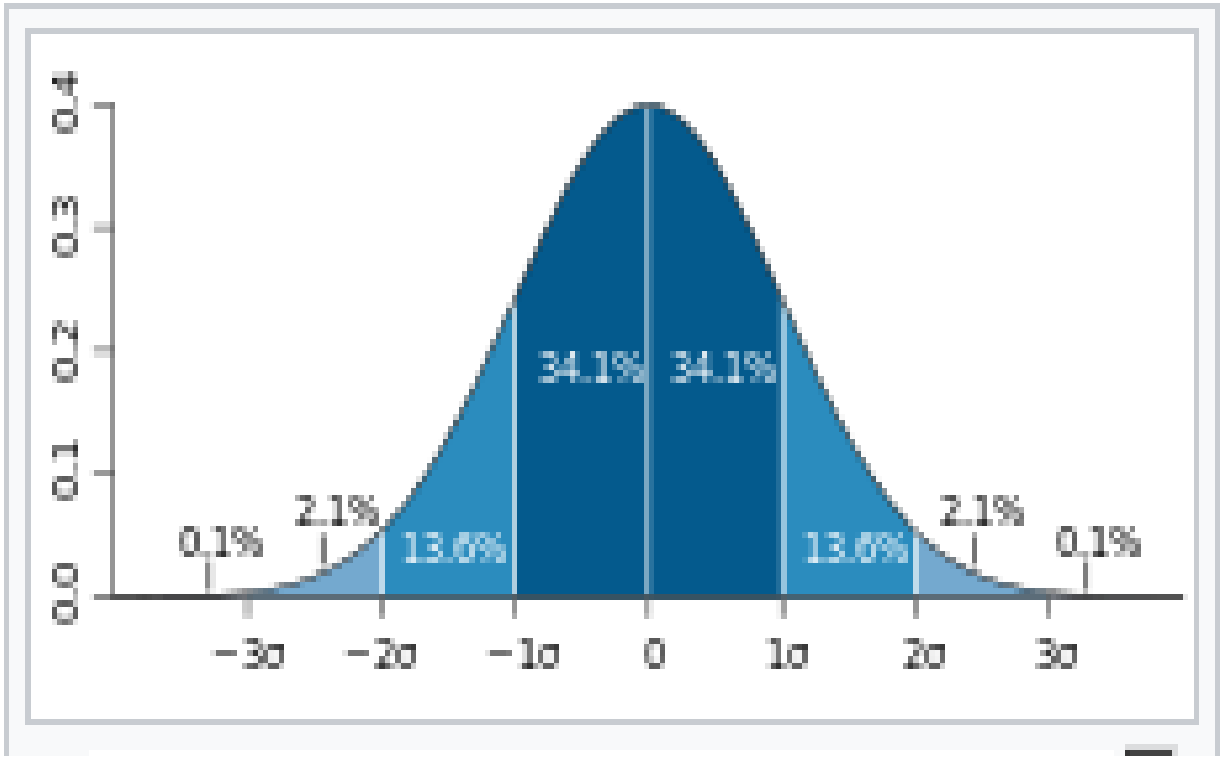
n = the number of observations



Example of samples from two populations with the same mean but different variances. The red population has mean 100 and variance 100 (SD=10) while the blue population has mean 100 and variance 2500 (SD=50).

Standard Deviation

- Standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range. It essentially measures the absolute variability of a random variable.



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

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

Covariance

- Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).
- The values of covariance can be any number between the two opposite infinities. Also, it’s important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.

Covariance

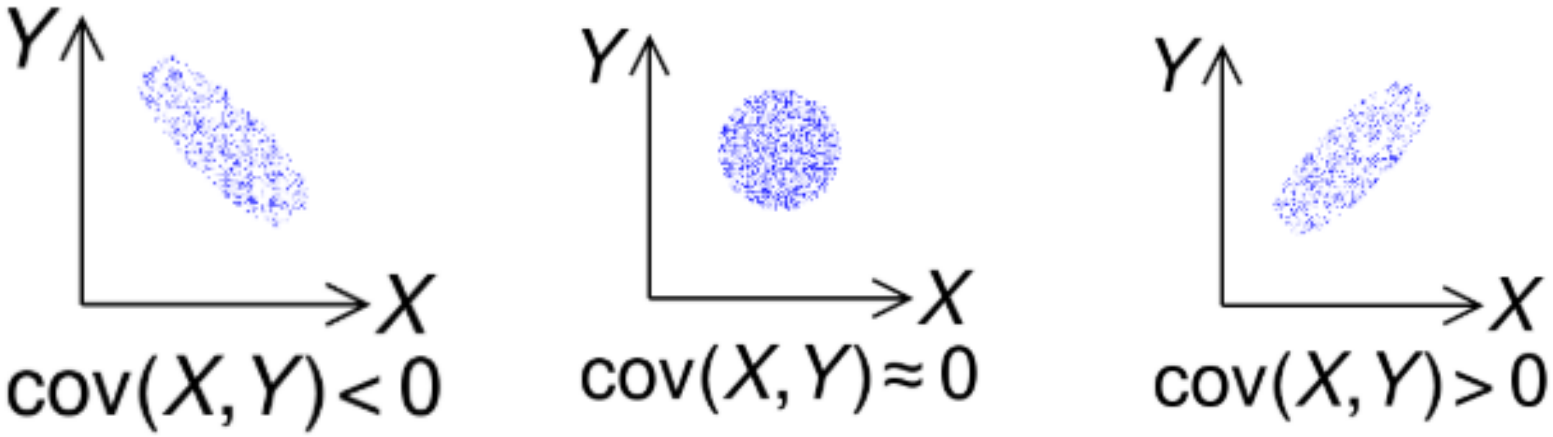
The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Covariance

- The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables.
- Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.

Covariance



Derivation of the covariance:

Goal: a measure of the linear relationship between two variables.

Consider the following set of bivariate data:

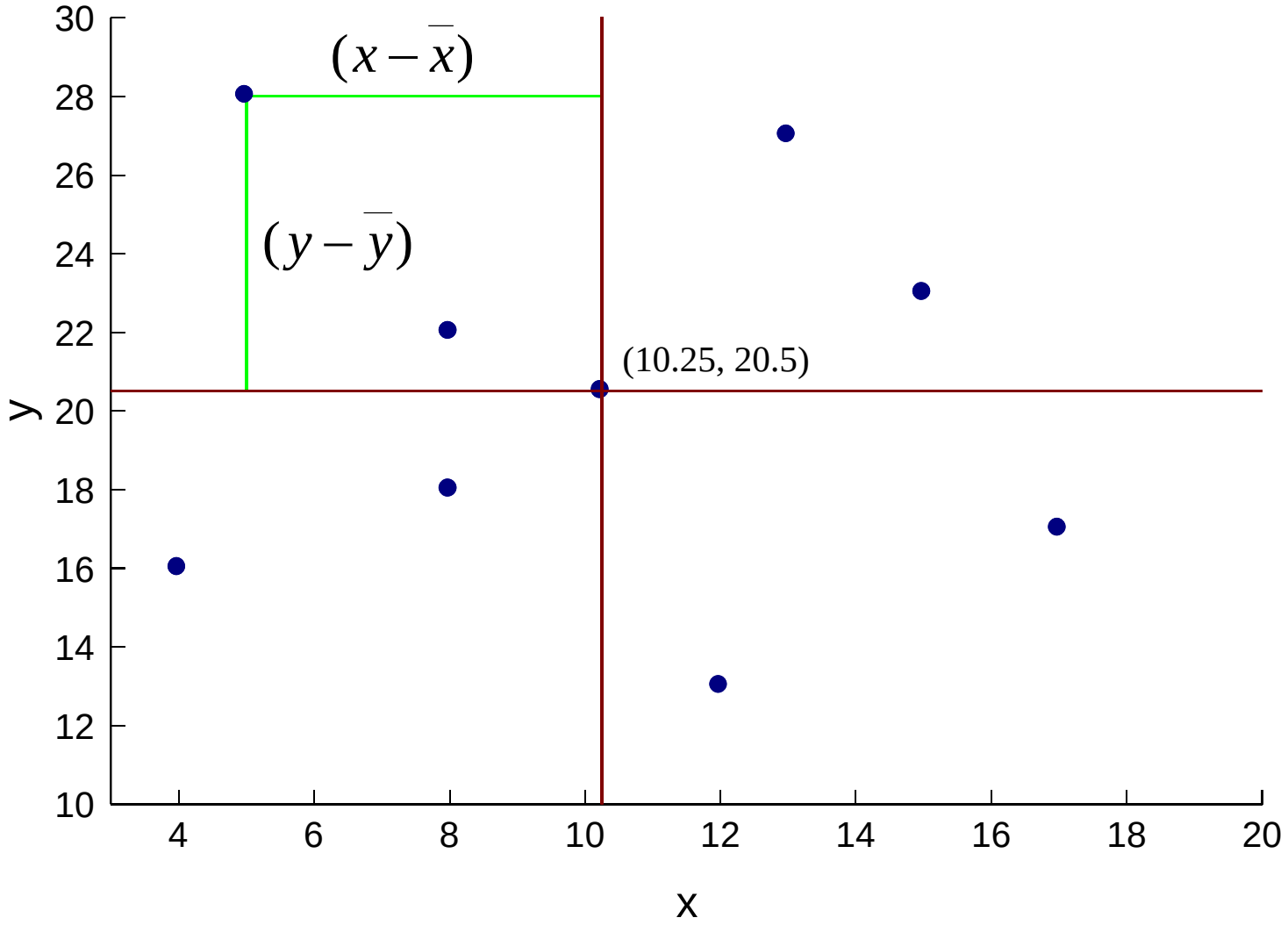
{(8, 22), (5, 28), (8, 18), (4, 16), (13, 27), (15, 23), (17, 17), (12, 13)}

$\bar{x} = 10.25 \qquad \bar{y} = 20.50$

Consider a graph of the data:

1. The point $(\bar{x}, \bar{y})$ is the **centroid** of the data.
2. A vertical and horizontal line through the centroid divides the graph into four sections.

Graph of the data, with centroid.



Note:

1. Each point (x, y) lies a certain distance from each of the two lines.
2. $(x - \bar{x})$: the horizontal distance from (x, y) to the vertical line passing through the centroid.
3. $(y - \bar{y})$: the vertical distance from (x, y) to the horizontal line passing through the centroid.
4. The distances may be positive, negative, or zero.
5. Consider the product: $(x - \bar{x})(y - \bar{y})$
   a. If the graph has lots of points to the upper right and lower left of the centroid (positive linear relationship), most products will be positive.
   b. If the graph has lots of points to the upper left and lower right of the centroid (negative linear relationship), most products will be negative.

Covariance:

The **covariance of *x* and *y*** is defined as the sum of the products of the distances of all values x and y from the centroid divided by $n - 1$.

$$\text{covar}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Note:

$\sum (x - \bar{x}) = 0$ and $\sum (y - \bar{y}) = 0$ always!

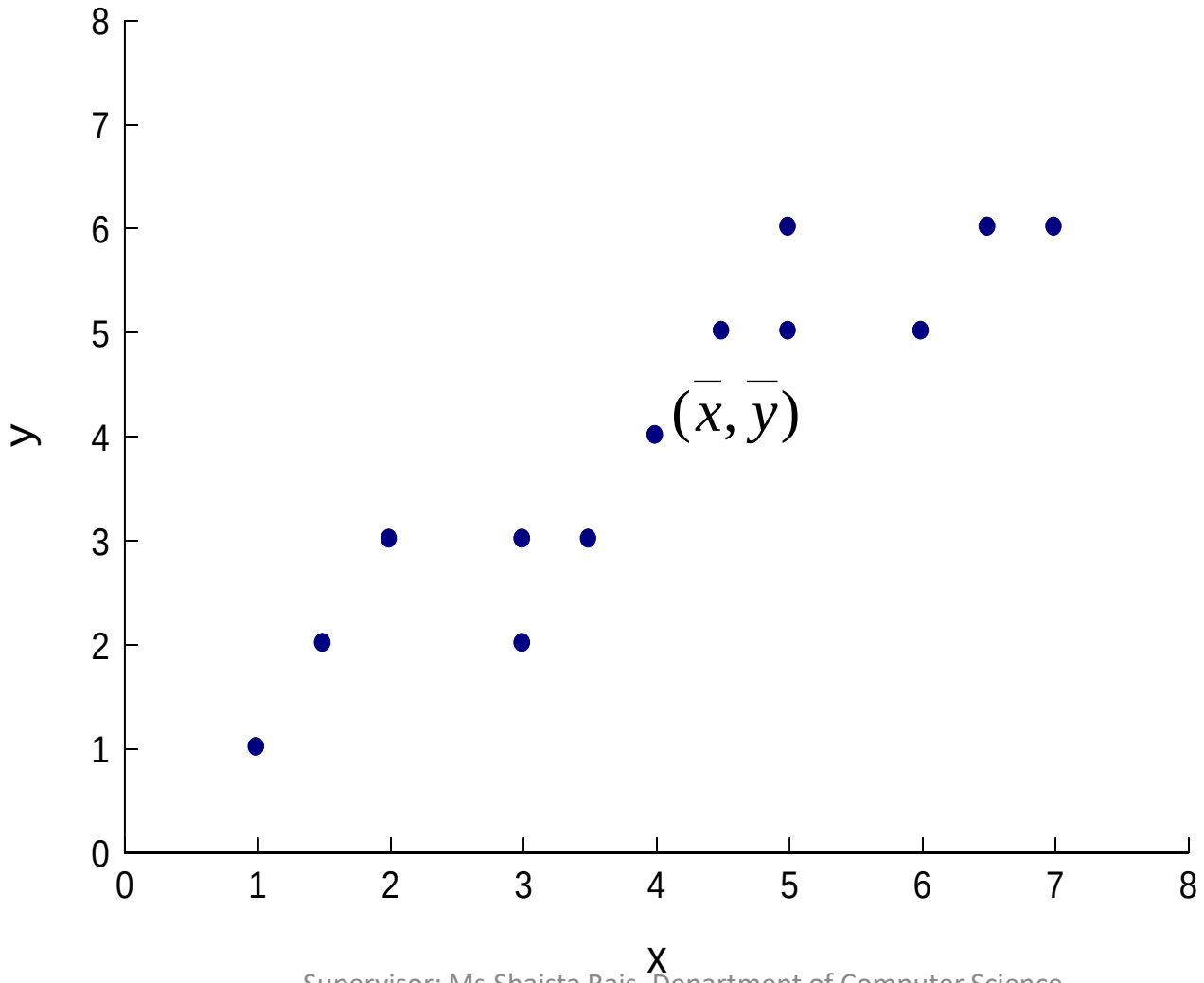
Calculations for finding covar(x, y):

Points	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
(8, 22)	-2.25	1.5	-3.375
(5, 28)	-5.25	7.5	-39.375
(8, 18)	-2.25	-2.5	5.625
(4, 16)	-6.25	-4.5	28.125
(13, 27)	2.75	6.5	17.875
(15, 23)	4.75	2.5	11.875
(17, 17)	6.75	-3.5	-23.625
(12, 13)	1.75	-7.5	-13.125
Total	0.00	0.0	-16.000

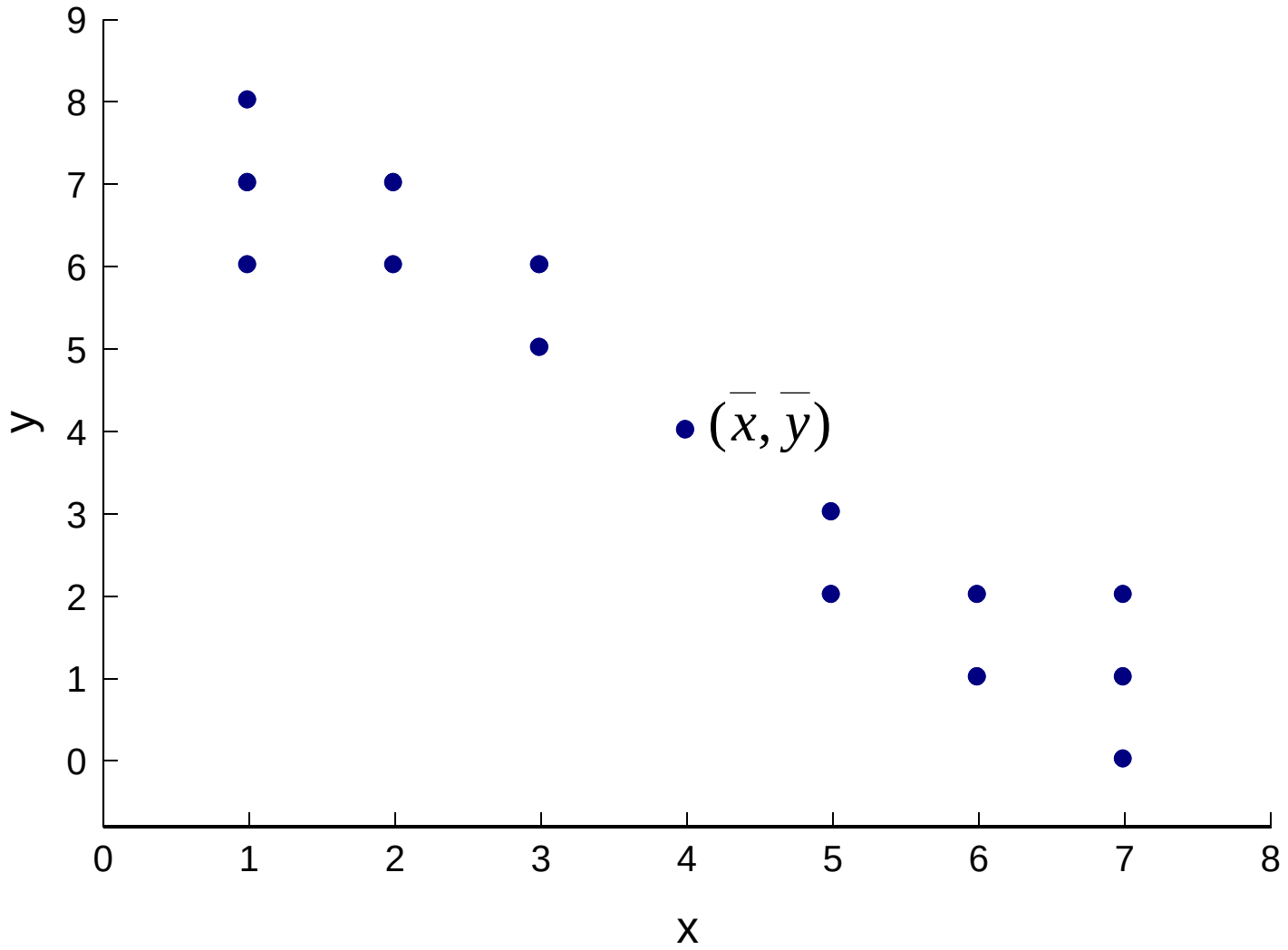
$$\text{covar}(x, y) = \frac{-16}{7} = -2.2857$$

Data and Covariance:

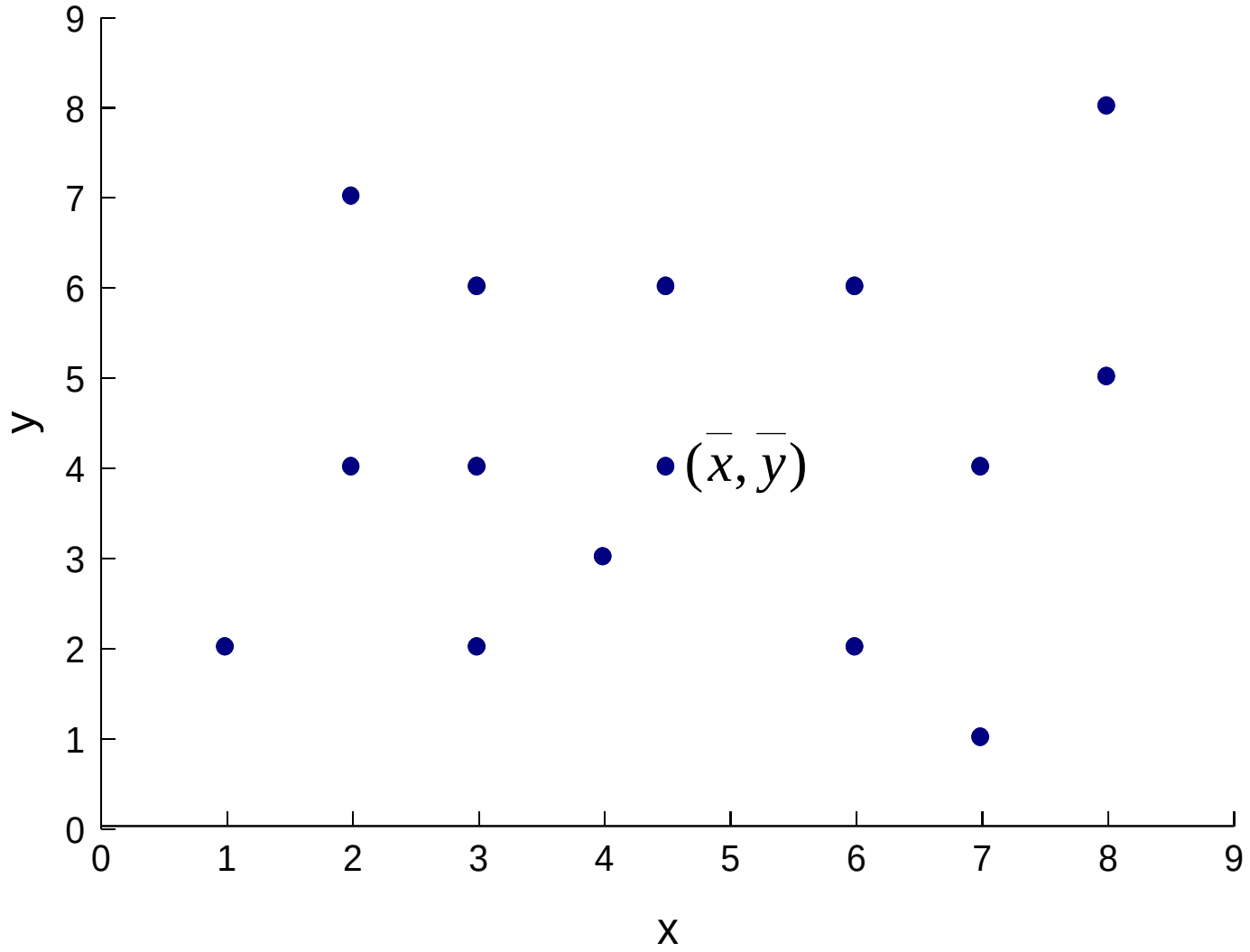
Positive covariance



Negative covariance



Covariance near 0



Problem:

1. The covariance does *not* have a standardized unit of measure.
2. Suppose we multiply each data point in the example in this section by 15.

   The covariance of the new data set is -514.29.
3. The amount of the dependency between *x* and *y* *seems stronger*. But the relationship is really the *same*.
4. We must find a way to eliminate the effect of the spread of the data when we measure the strength of a linear relationship.

Solution:

1. Standardize x and y:

$$x' = \frac{x - \bar{x}}{s_x} \qquad \text{and} \qquad y' = \frac{y - \bar{y}}{s_y}$$

2. Compute the covariance of x' and y'.
3. This covariance is *not* affected by the spread of the data.
4. This is exactly what is accomplished by the **coefficient of linear correlation**:

$$r = \text{covar}(x', y') = \frac{\text{covar}(x, y)}{s_x \cdot s_y}$$

Note:

1. The coefficient of linear correlation standardizes the measure of dependency and allows us to compare the relative strengths of dependency of different sets of data.
2. Also commonly called **Pearson's product moment**, ***r***.

Calculation of *r* (for the data presented in this section):

$s_x = 4.71$ and $s_y = 5.37$

$$r = \frac{\text{covar}(x, y)}{s_x \cdot s_y} = \frac{-2.2857}{(4.71)(5.37)} = -0.0904$$

Alternative (Computational) Formula for *r*:

$$r = \frac{\text{covar}(x, y)}{s_x \cdot s_y} = \frac{\frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}}{s_x \cdot s_y} = \frac{\text{SS}(xy)}{\sqrt{\text{SS}(x) \cdot \text{SS}(y)}}$$

1. This formula avoids the separate calculations of the means, standard deviations, and the deviations from the means.
2. This formula is easier and more accurate: minimizes round-off error.