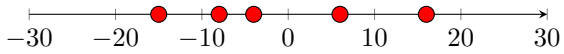
# Chapter 12
# Variance, Covariance and Correlation

## 1 Variance

Consider some numbers on 1 Dimension marked as points.



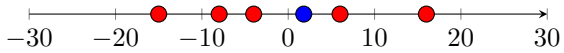

These points are:

| $x$ | Value |
|---|---|
| $x_1$ | -1 |
| $x_2$ | -8 |
| $x_3$ | -4 |
| $x_4$ | 6 |
| $x_5$ | 16 |

Now the mean of this data would be:

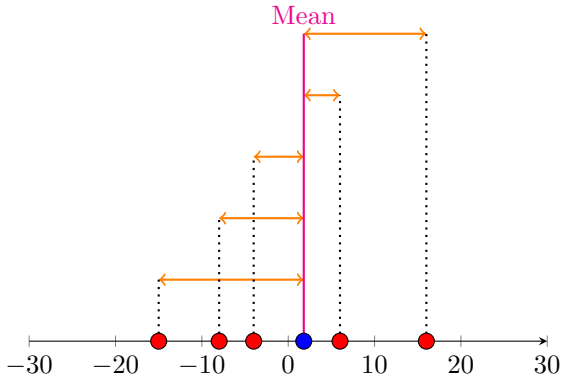
$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{-1 - 8 - 4 + 6 + 16}{5} = \frac{9}{5} = 1.8$$

Let's plot this mean on our number line




Here mean is represented using blue color and all the data points are red colored

Now, what if we need to find how much separated are these points from the mean
or, the average distance of all these points from the mean




We can do this by calculating each point's distance from the mean and the finding average of all those distances
Note that distances can be negative and might result in making the overall average 0, so distances would be squared.
$D_{x_1,\bar{x}} = (\bar{x} - x_1)^2 = (1.5 - (-15))^2 = 272$
$D_{x_2,\bar{x}} = (\bar{x} - x_2)^2 = (1.5 - (-8))^2 = 90.25$
$D_{x_3,\bar{x}} = (\bar{x} - x_3)^2 = (1.5 - (-4))^2 = 30.25$
$D_{x_4,\bar{x}} = (\bar{x} - x_4)^2 = (1.5 - 6)^2 = 20.25$
$D_{x_5,\bar{x}} = (\bar{x} - x_5)^2 = (1.5 - 16)^2 = 210.25$
Now,
$D_{Average} = \frac{D_{x_1,\bar{x}} + D_{x_2,\bar{x}} + D_{x_3,\bar{x}} + D_{x_4,\bar{x}} + D_{x_5,\bar{x}} + D_{x_6,\bar{x}}}{6}$

$= \frac{272 + 90.25 + 30.25 + 20.25 + 210.25}{5} = 124.6$

This average distance of points with mean is actually called "Variance"
Hence, Sample Variance is given by

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

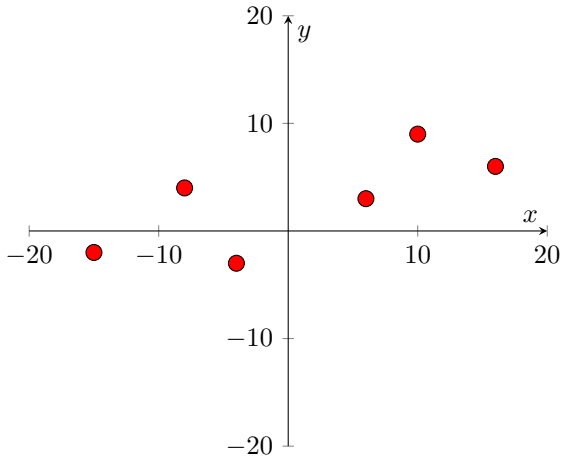
## 2 Covariance

Now, what if instead of just 1 Dimension there were 2 Dimensions and the data had two parameters x and y?

For example here are the points:

| $x$ | $y$ |
|---|---|
| -15 | -2 |
| -8 | 4 |
| -4 | -3 |
| 6 | 3 |
| 10 | 9 |
| 16 | 6 |

The plot would look like this

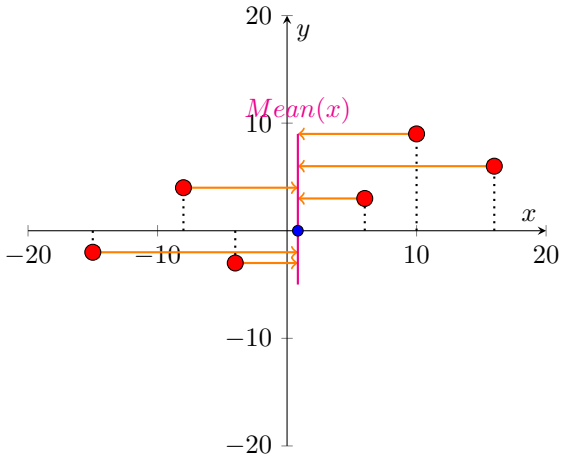



Let's also plot the mean of x and y

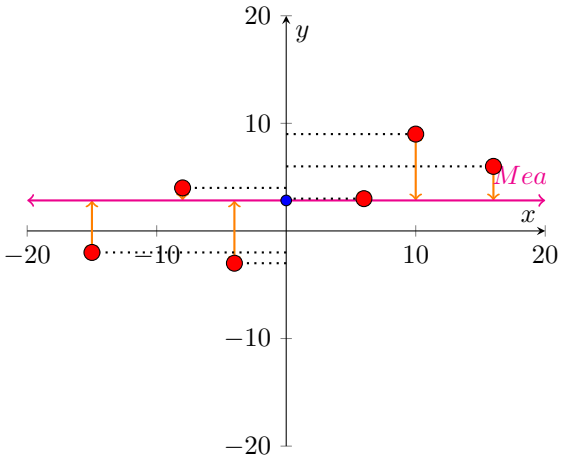
$$\bar{x} = 0.834$$

$$\bar{y} = 2.834$$

Distances of x points from Mean(x)




Distances of y points from Mean(y)




Here we can calculate

Variance along x direction

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

$= \frac{(-15-0.834)^2+(-8-0.834)^2+(-4-0.834)^2+(6-0.834)^2+(10-0.834)^2+(16-0.834)^2}{5} = 138.5666672$

Variance along y direction

$Var(y) = \frac{\sum_{i=1}^{n}(\bar{y}-y_i)^2}{1-n}$

$= \frac{(-2-2.834)^2+(4-2.834)^2+(-3-2.834)^2+(3-2.834)^2+(9-2.834)^2+(6-2.834)^2}{5} = 21.3666672$

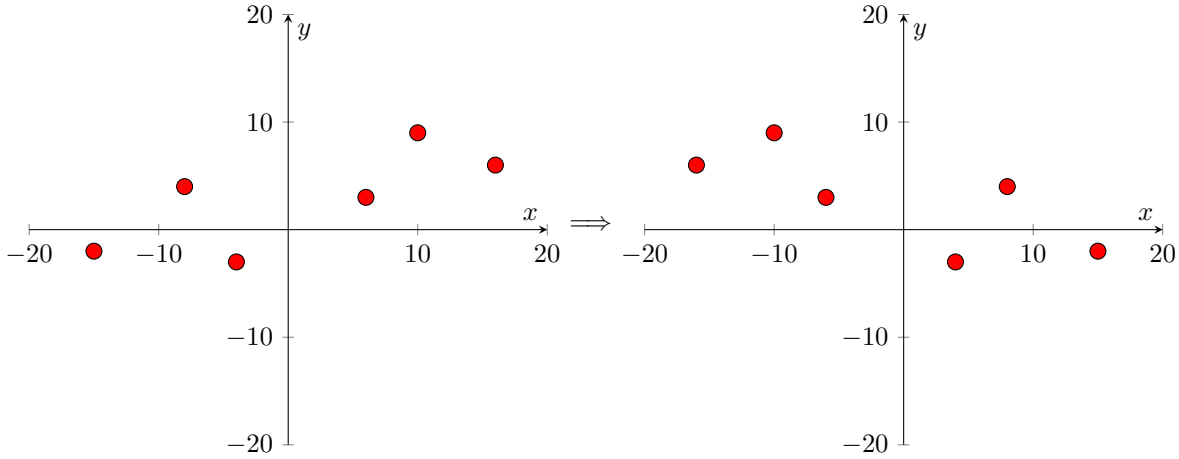
And it's really evident from the plot that values of x are more spread from mean than y
and thus $Var(x) > Var(y)$

The main observation here is that,
If we change the signs of x values, the plot flips along vertical axis

| $x$ | $y$ |
|---|---|
| -15 | -2 |
| -8 | 4 |
| -4 | -3 |
| 6 | 3 |
| 10 | 9 |
| 16 | 6 |

$\Longrightarrow$

| $x$ | $y$ |
|---|---|
| 15 | -2 |
| 8 | 4 |
| 4 | -3 |
| -6 | 3 |
| -10 | 9 |
| -16 | 6 |




But still, the values of $Var(x)$ and $Var(y)$ remains the same as before

$$Var(x) = 138.5666672$$

$$Var(y) = 21.3666672$$

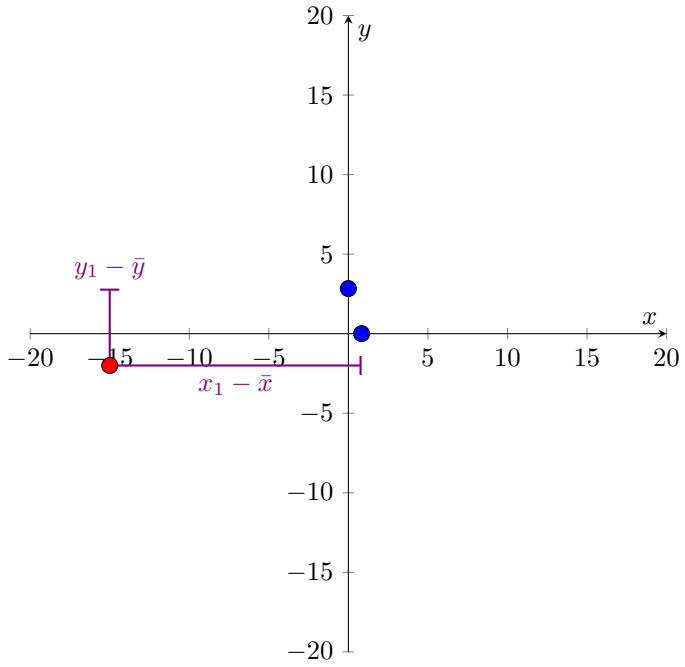
We need some other metric too....

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

Ok,but what does $(x_i - \bar{x})(y_i - \bar{y})$ means?

| $x$ | $y$ | $(x - \bar{x})$ | $(y - \bar{y})$ | $(x - \bar{x})(y - \bar{y})$ |
|---|---|---|---|---|
| -15 | -2 | -15-0.834=-15.834 | -2-2.834=-4.834 | -15.834 * -4.834 = 76.541556 |
| -8 | 4 | -8-0.834=-8.834 | 4-2.834=1.166 | -8.834 * 1.166 = -10.300444 |
| -4 | -3 | -4-0.834=-4.834 | -3-2.834=-5.834 | -4.834 * -5.834 = 28.201556 |
| 6 | 3 | 6-0.834=5.166 | 3-2.834=0.166 | 5.166 * 0.166 = 0.857556 |
| 10 | 9 | 10-0.834=9.166 | 9-2.834=6.166 | 9.166 * 6.166 = 56.517556 |
| 16 | 6 | 16-0.834=15.166 | 6-2.834=3.166 | 15.166 * 3.166 = 48.015556 |

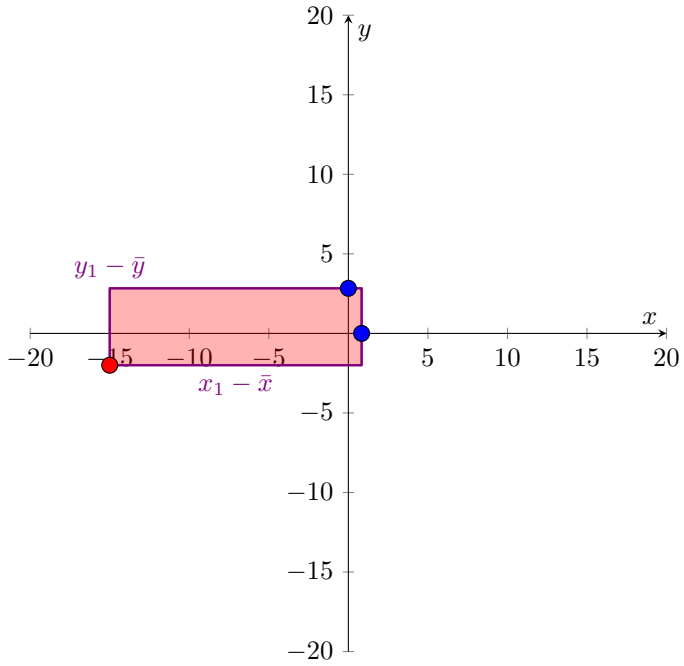
Let's plot this for 1st point



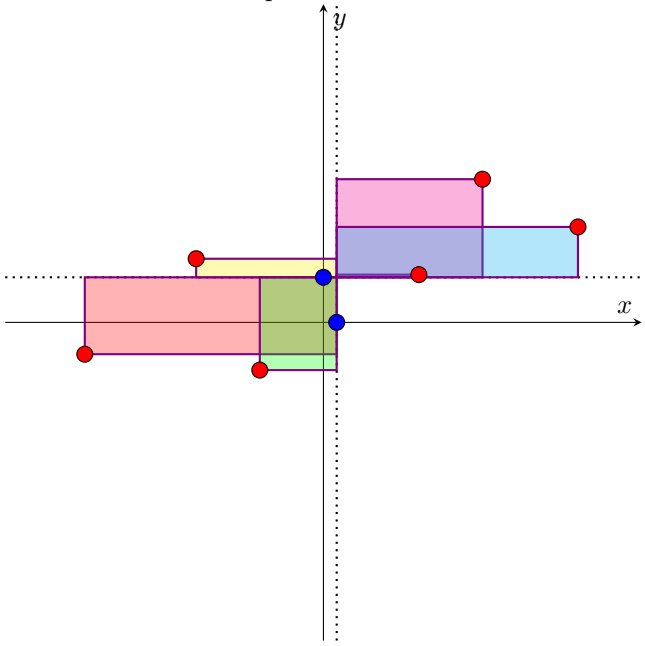

Red color: Point 1

Blue color: Means of x and y

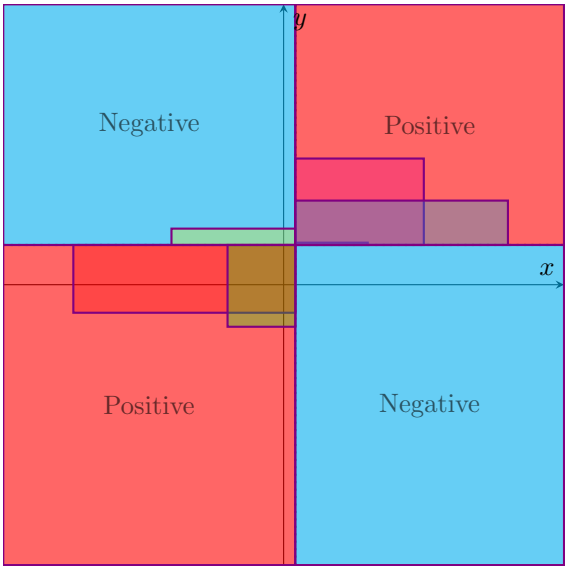
Here, we plotted $(x_1 - \bar{x}) and (y_1 - \bar{y})$, and it's looking really intuitive that $(x_1 - \bar{x})(y_1 - \bar{y})$ is just the area within that rectangle..




Let's draw this for all points




Now, we need to take signed sum of these areas,




that would be

$$Signed\ Sum\ Of\ Areas = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

after that we just need to find average value of these signed areas

$$\therefore Average\ Signed\ Sum = \frac{Signed\ Sum\ Of\ Areas}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n};$$

But what does this average signed sum of areas tell us?
A lot actually,

- if this value is positive, this means the slope of the best fit line would be positive, this means value of y increases with increase in x.
- if this value is negative, this means the slope of the best fit line would be negative, this means value of y decreases with increase in x.
- if this value is large, this means there is on an average a large separation between values of x and y with their means, i.e., values are more spread from mean.
- if this value is small, this means there is on an average very less sepration between values of x and y with their means, i.e, values are less spread from mean.
- if this value is 0, slope is 0, and no linear relationship exist between $x$ and $y$.

and since, this value is giving us an idea about how 2 values x and y are varying that is why it is called "Covariance(x,y)"
Hence, the formula

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

So, this formula tells about relation between 2 variables but still doesnt tell us anything about the "Strength" of this relation. Let us now find that strength.

# 3 Correlation

If we consider the two variables $x$ and $y$ as vectors.

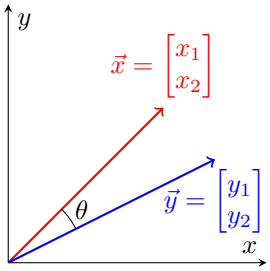
$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

but representing an "n" dimensional vector on a 2D plot is not possible, so let's just consider $n = 2$
$\therefore$

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \in \mathbb{R}^2$$




Here $\theta$ is the angle between $\vec{x}$ and $vecy$

Now, we can find how 'related' or 'close' are 2 vectors, using the dot product.

$$cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}||.||\vec{y}||} = \frac{< \vec{x}, \vec{y} >}{\sqrt{< x, x >}.\sqrt{< y, y >}} = \frac{x_1.y_1 + x_2.y_2}{\sqrt{x_1^2 + x_2^2}.\sqrt{y_1^2 + y_2^2}}$$

for n dimensional vectors

$$cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{||\vec{x}||.||\vec{y}||} = \frac{< \vec{x}, \vec{y} >}{\sqrt{< x, x >}.\sqrt{< y, y >}} = \frac{x_1.y_1 + x_2.y_2 + \cdots + x_n.y_n}{\sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}}$$

$$cos(\theta) = \frac{\sum_{i=1}^{n}(x_i.y_i)}{\sqrt{\sum_{i=1}^{n}(x_i^2)}.\sqrt{\sum_{i=1}^{n}(y_i^2)}}$$

Now consider, this table of values of x and y

| **x** | **y** |
|---|---|
| 1 | 200 |
| 2 | 240 |
| 2.5 | 340 |
| 3 | 500 |
| 4.5 | 620 |

We don't actually need a relation between these 2 variables, we actually want relation between the spreading of these 2 variables from their mean.

We can calculate how far are each points on x and y from their mean $\bar{x}$ and $\bar{y}$

| **x** | **y** | $x - \bar{x}$ | $y - \bar{y}$ |
|---|---|---|---|
| 1 | 200 | -1.6 | -180 |
| 2 | 240 | -0.6 | -140 |
| 2.5 | 340 | -0.1 | -40 |
| 3 | 500 | 0.4 | 220 |
| 4.5 | 620 | 1.9 | 340 |

So, we would consider $x \vec{-} \bar{x}$ and $y \vec{-} \bar{y}$ as the vectors we need to find relation between.

$$\vec{x^*} = x \vec{-} \bar{x} = \begin{bmatrix} -1.6 \\ -0.6 \\ -0.1 \\ 0.4 \\ 1.9 \end{bmatrix}$$

$$\vec{y^*} = y \vec{-} \bar{y} = \begin{bmatrix} -180 \\ -140 \\ -40 \\ 220 \\ 340 \end{bmatrix}$$

$\therefore$

$$cos(\theta) = \frac{\sum_{i=1}^{n}(x_i^*.y_i^*)}{\sqrt{\sum_{i=1}^{n}(x_i^{*2})}.\sqrt{\sum_{i=1}^{n}(y_i^{*2})}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

But,

$$\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{x}) = Cov(x, y).(1 - n)$$

and

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

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = Var(y)(1 - node)$$

Therefore,

$$cos(\theta) = \frac{Cov(x, y)}{\sqrt{Var(x)}.\sqrt{Var(y)}}$$

This $cos(\theta)$ is actually called Pearson's Correlation coefficient between x and y.

$$-1 <= Correlation(x, y) = \rho(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)}.\sqrt{Var(y)}} <= 1$$

Interpretation of Pearson Correlation coefficient

- $\rho(x, y)$ ranges between -1 and 1
- if $\rho(x, y) = 1$ this means $\theta = 0°$, which means vectors are aligned and overlapping each other, this means both are highly correlated. This means the spreading of x and y from mean is very much similar.
- if $\rho(x, y) = 0$ this means $\theta = 90°$, which means vectors are orthogonal from each other, this means both are not that much similar.
- if $\rho(x, y) = -1$ this means $\theta = 180°$, which means vectors are opposite of each other in direction, this means they are not correlated.