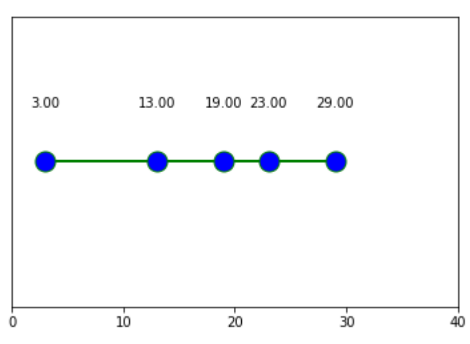
COVARIANCE

RECAP OF VARIANCE

Assume that we counted number of Green shirts in 5 stores Also we counted number of Black shirts in same 5 stores

For Green shirts we can plot it as

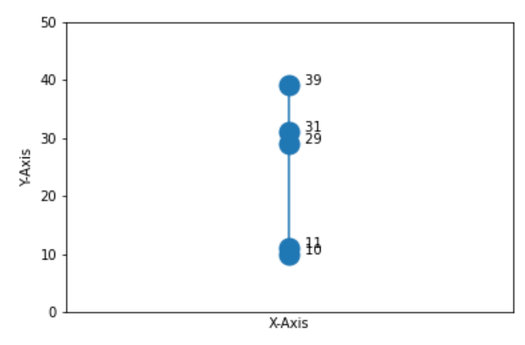


Then we estimate the mean $\bar{x} = 17.6$ and then we claculate the variance

$\frac{\sum(x-\bar{x})^2}{n-1}$ = 101.8

Now assume we counted number of Black shirts in same 5 stores

Ploting it and calculating mean , variance we get

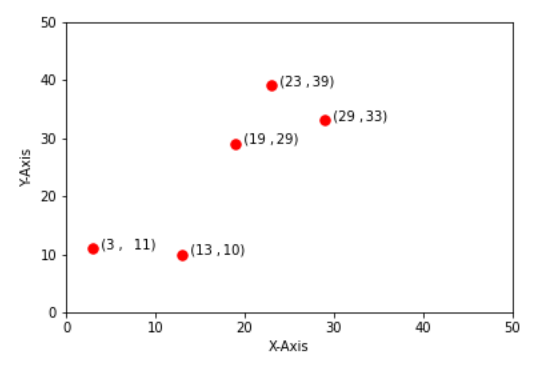


$\bar{y} = 24.4$

$\frac{\sum(y-\bar{y})^2}{n-1}$ = 160.3

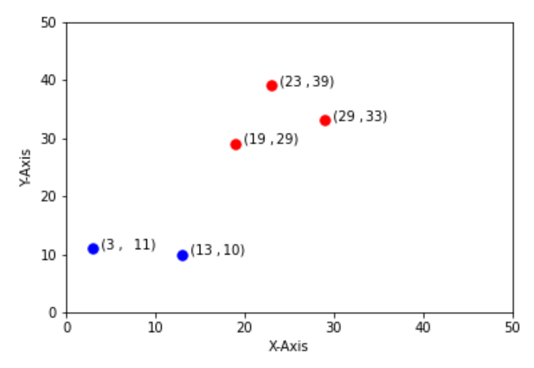
so far we have estimated the mean and variance for two different shirts in same store

Since the measurement came from same stores we can plot each pair as single dot by combining the values on x and y axis



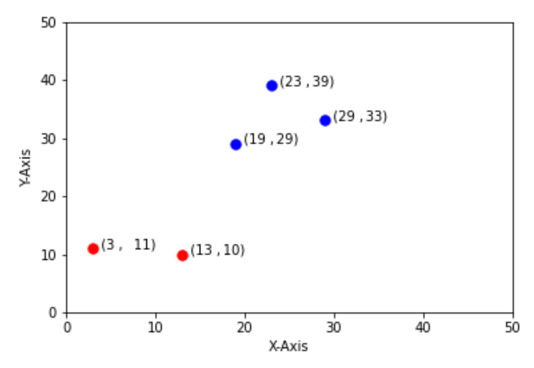
we could see that stores with low number of Green shirts in also have low number of Black shirts

[Blue marked dots]

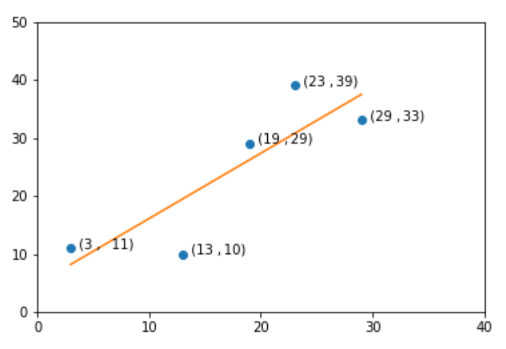


same way stores with high number of Green shirts have high number of black shirts

[Blue marked dots]

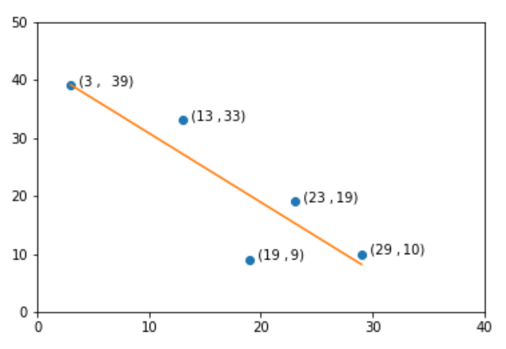


This relationship of low measurement for both low Shirt count in some store and high measurement for both high shirt count in some store can be summerized by a line

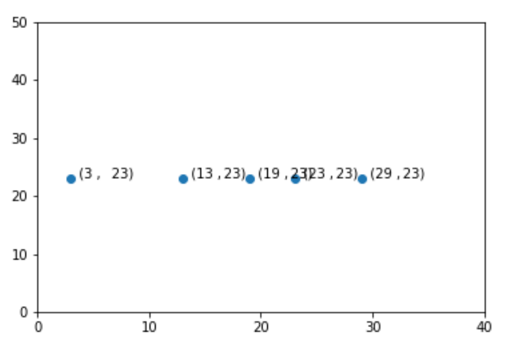


This line represents the positive slope which means both Green shirt count and Black shirt count increases together in a positive trend

Similar way if we plot something like this we get negative slop which means both Green shirt count Black shirt count decreases together in negative trend



And if data looked something like this every value for Green shirt is paired with same value for Black shirt then there would be no trend (Positive/Negative)



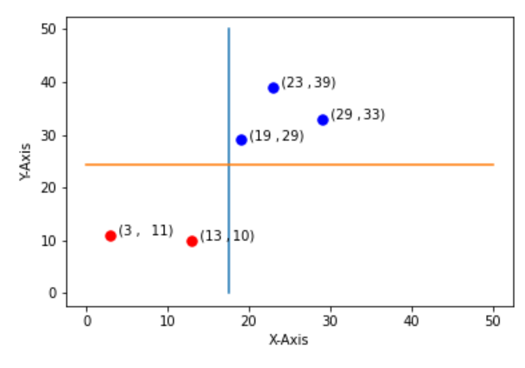
Covariance can be used to get an idea on these 3 types of relations

$$Covariance = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

Now let Plot a graph with a straight lines indicating mean for both axis

Blue line indicate Green shirts mean

Orange line indicate Black shirts mean



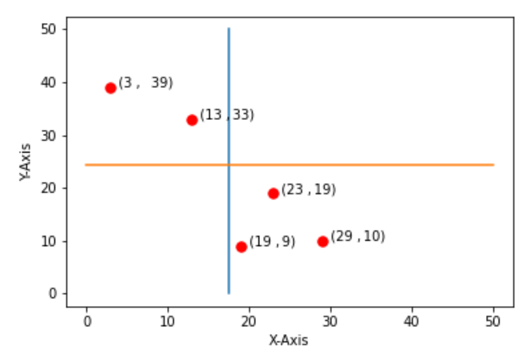
No Lets calculate the difference

$$Covariance = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{(3-17.6)(12-24.4)+...}{5-1} = \frac{181+66.2+6.4+55+155}{5-1} = 116$$

Since the covariance value is 116 is positive which means slope of the relationship is positive

NOTE : The covariance value doesn't tells us the slope of the line is steep or not steep it just tells us the slope is positive. It won't tell us if the points are close enough to the line or far from it

Now if we look at the other case and calculate the difference from mean to the points

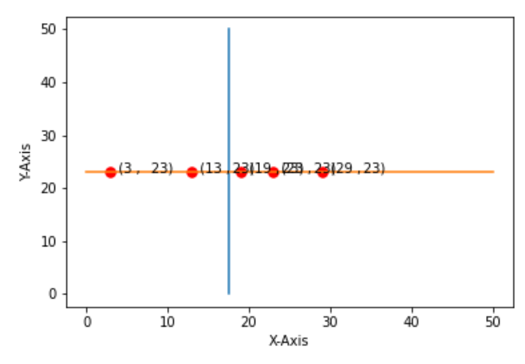


$$Covariance = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{(3-17.6)(13-17.6)+...}{5-1} = \frac{(-216.1)+(-54.3)+(-18.5)+(-26.9)+(-104.9)}{5-1} = -105.15$$

covariance is negative which means we will be having a negative trend.

Slope of the relationship is negative

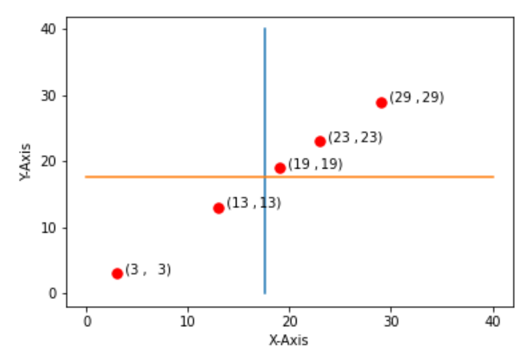
Now lets look at the case where there is no trend and calculate the covariance



$$Covariance = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{(3-17.6)(23-23)+...}{5-1} = \frac{(0)+(0)+(0)+(0)+(0)}{5-1} = 0$$

Covariance is hard to interpret why ?

Assume we have data where X and Y are same and after plotting data we get a plot like this



meanX = 17.6 meanY = 17.6

Now lets calculate the covariance

$$Covariance = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$$

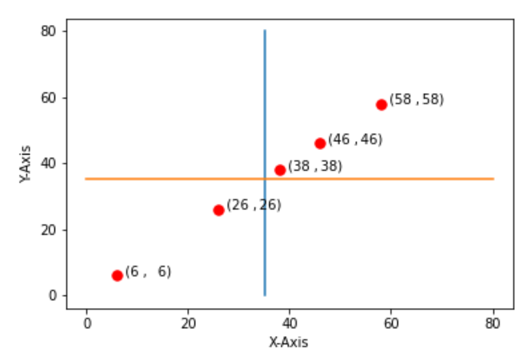
NOTE: Here the X and Y are same so we can substitute X=Y and $\bar{X} = \bar{Y}$

and we get the Covariance as

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

when we put the point in to the formulae we get 102

Now lets see what happens if we multiply the data by 2 we get a plot like



The position of the data is same and the point falls on same st.line Only change happended was the scale that the data is on

Covariance for this would be 408

we could see covariance value changes even when the relationship doesn't

Covariance values are sensitive to scale of the data and thismakes them difficult to interpret