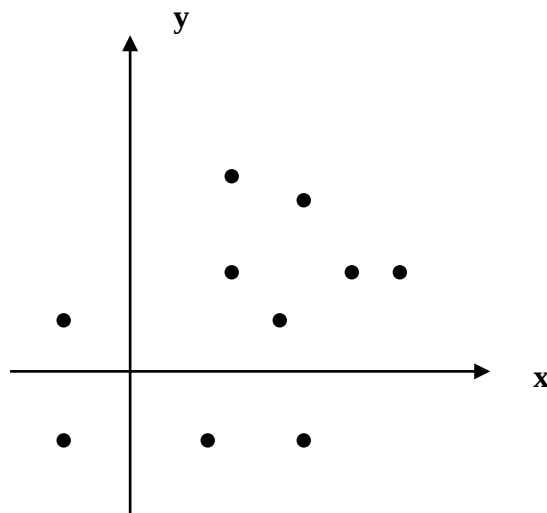


Le calcul de la variance et du coefficient de corrélation linéaire

Le but de ce cours est de voir réellement le principe de la covariance et pourquoi on utilise le coefficient de corrélation linéaire « r ».

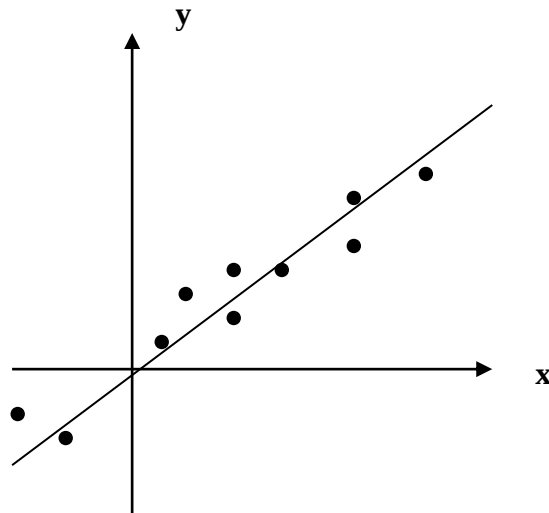
On peut prendre un exemple : nous avons une série double $x_i y_i$, la covariance entre x et y est donnée par la formule suivante : $\text{Cov}(x,y) = \Sigma (x_i - \bar{x}) (y_i - \bar{y}) / n$ sachant que « n » c'est l'effectif total. On peut aussi écrire $\sigma_{xy} = \Sigma x_i y_{i/n} - \bar{x} \bar{y}$. Lorsqu'on a une série statistique à deux variables, cette série peut par exemple (selon les valeurs) être représentée par un nuage de points de la manière suivante



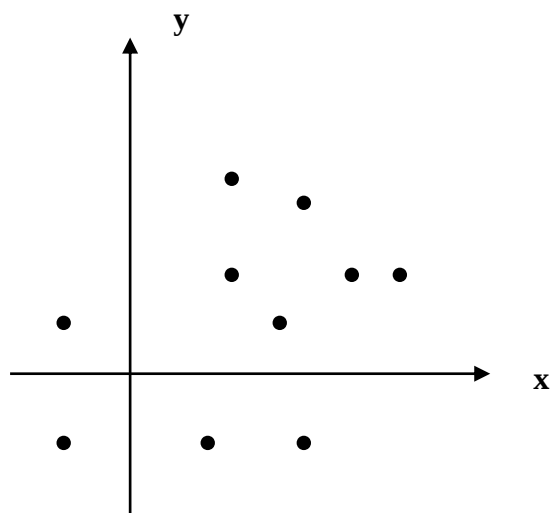
L'idée de la covariance est d'interpréter le degré de dépendance entre x et y . Autrement dit, existe-t-il une liaison entre x et y ? S'il n'existe pas de lien entre x et y , on dira que les variables x et y sont indépendantes et s'il existe un lien entre les deux variables, il peut être de type affine, exponentiel ou de type puissance. On va donc expliquer pourquoi la covariance permet de mesurer ce degré de dépendance entre les x et les y et on dit « plus la covariance est élevée plus la dépendance entre x et y est forte et plus la corrélation entre x et y est élevée », autrement dit, si la covariance est quasiment proche de 0 on dira que les variables x et y sont indépendantes. Dans ce cas le 0 est synonyme de variables indépendantes.

Pour montrer que « si la covariance est élevée, alors la dépendance entre x et y est forte », on procède comme suit :

Si on a un nuage de point comme le montre la figure suivante :



On dit qu'il existe entre x et y une relation de type affine c'est-à-dire qui peut être ajuster par une droite et dans ce cas si on calcule la covariance, on va trouver une covariance élevée car la relation entre x et y existe. En revanche, si on a un nuage de points comme ceci (voir figure ci-dessous) :

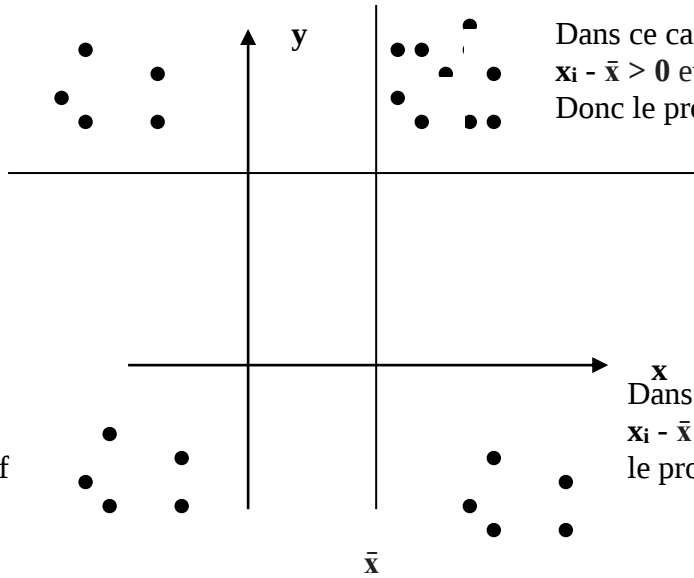


On trouvera une covariance avec une petite valeur et dans ce cas on dira que la relation entre x et y est faible.

Exemple : on reprend la formule de la covariance qui st la suivante :

Cov(x,y) = $\Sigma (x_i - \bar{x}) (y_i - \bar{y}) / n$, on remarque que $(x_i - \bar{x})$ c'est l'écart entre x et sa moyenne, la même chose pour le y. Autrement dit, on va savoir comment ces deux écarts covarient . pour cela on va créer un schéma (voir figure ci-dessous).

Dans ce quadrant on a
 $x_i - \bar{x} < 0$ et $y_i - \bar{y} > 0$
 le produit est négatif

 $\bar{y}$ 

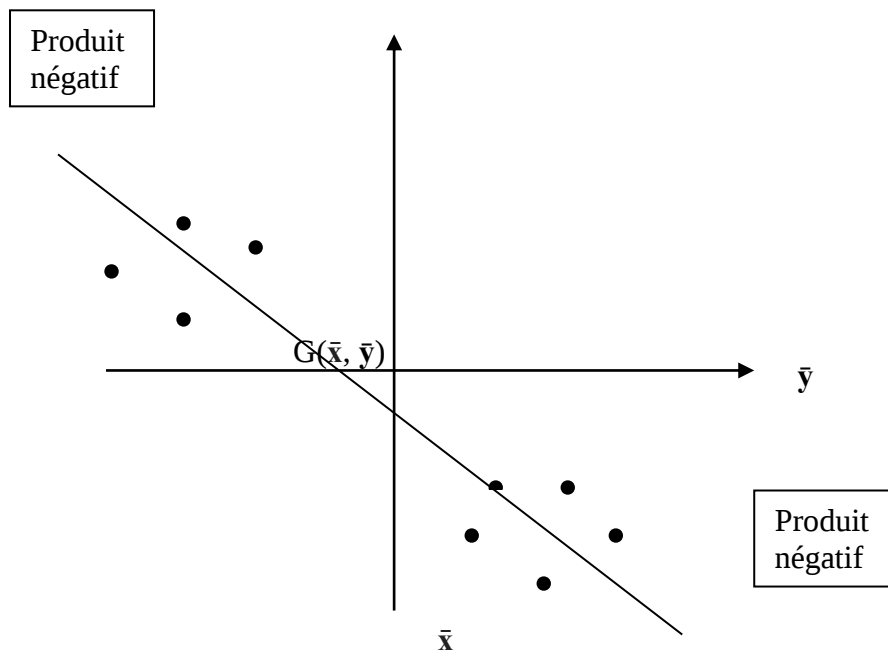
Dans ce quadrant on a
 $x_i - \bar{x} > 0$ et $y_i - \bar{y} > 0$
 Donc le produit est positif

Dans ce quadrant on a
 $x_i - \bar{x} < 0$ et $y_i - \bar{y} < 0$
 Donc le produit est positif

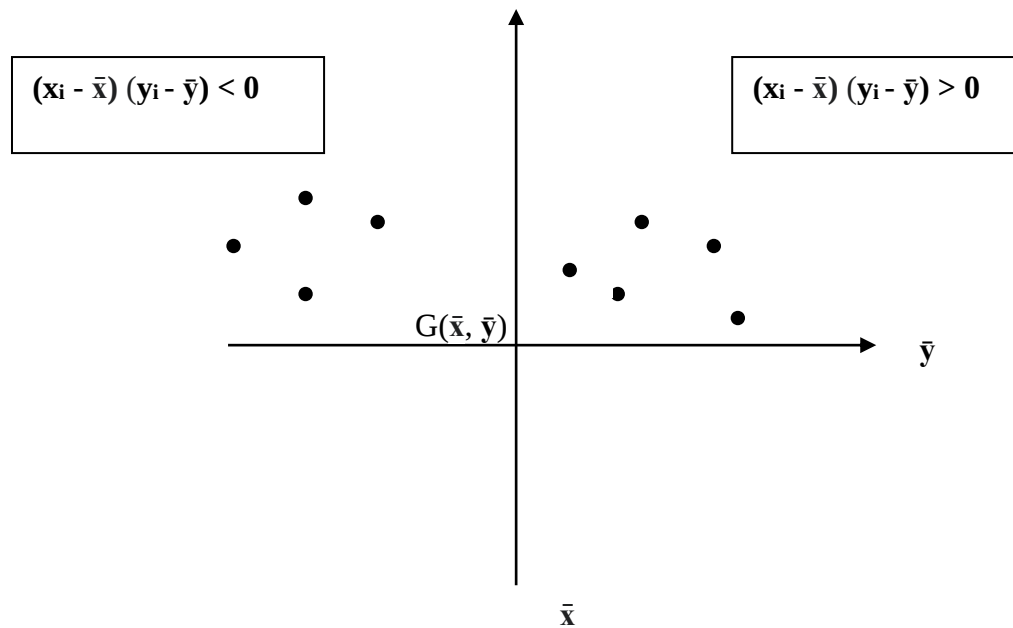
Dans ce quadrant on a
 $x_i - \bar{x} > 0$ et $y_i - \bar{y} < 0$
 le produit est négatif

 $\bar{x}$

A partir du schéma précédent, on constate que si on :



Dans ce cas, on peut dire qu'il y a une bonne corrélation (négative), mais si on tombe sur un nuage comme ci-dessous :



La corrélation dans ce cas est mauvaise, car la valeur de la covariance est tirée par le bas et par cause la dépendance entre x et y est quasiment nulle.

Conclusion :

Le calcul de la covariance nous donne une idée générale sur la dépendance ou la corrélation entre les variables, mais il y a un inconvénient majeur on va l'expliquer dans ce qui suit.

Admettant qu'on a une population représentée dans le tableau suivant :

x_i (représente la taille des individus en m)	y_i (représente le poids en kg)
1,50	45
1,55	47
1,65	50

En général ces deux variables sont corrélées par ce que statistiquement plus l'individu est grand plus il pèse lourd.

Si on calcule la covariance de cette série, on va trouver une valeur et en réalité cette valeur ne veut rien dire. On sait tout simplement que si la valeur de la covariance est élevée, alors la corrélation est forte par ce que les unités jouent un rôle important.

On suppose que $\sigma_{xy} = V$, maintenant si on change l'unité de la taille en cm et on laisse le kg pour le poids (voir tableau ci-dessous) :

x_i (représente la taille des individus en cm)	y_i (représente le poids en kg)
150	45
155	47
165	50

Et on calcule de nouveau la covariance, on trouvera que $\sigma_{xy} = V_1$ et V_1 sera plus grand que V pourtant on a gardé la même série. Normalement si on exprime la taille des individus en « m » ou en « cm », la covariance reste toujours la même et cela représente l'inconvénient de la covariance car elle tient compte des unités et à ce moment, on a plus de repère. Si on dit par exemple un tel étudiant a eu une note de 14, 17 ou 18 dans un module quelconque on ne sait pas exactement si c'est une bonne note ou pas, car il nous faut une échelle (un repère) pour se décider.

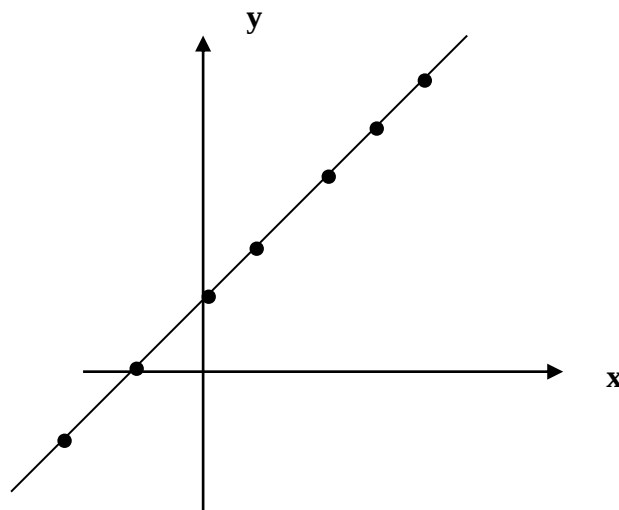
Pour remédier à ce problème, il faut travailler avec des variables centrées réduites c'est-à-dire il faut transformer x_i et y_i selon le schéma suivant : $(x_i - \bar{x}) / \sigma_x$ et $(y_i - \bar{y}) / \sigma_y$ sachant que σ_x et σ_y sont respectivement les écarts type des x_i et des y_i . Le but de cette transformation est d'obtenir un coefficient de corrélation linéaire « r » qui est un réel, sachant que $-1 \leq r \leq +1$ et $r = \sigma_{xy} / (\sigma_x \sigma_y)$ donc r est un réel sans unité, plus il est proche de +1 ou -1 plus la corrélation est forte.

Rappel :

$$\sigma_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} .$$

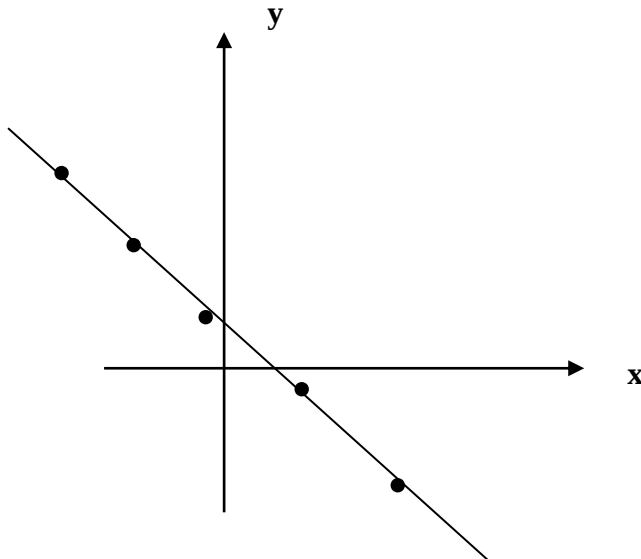
Exemple :

Soit le nuage de points suivant :



Si le nuage des points est parfaitement aligné, alors on est sûr que nous avons une corrélation parfaite c'est-à-dire $r = +1$.

Si on a un nuage de points comme le montre la figure suivante :



Dans ce cas le coefficient de corrélation $r = -1$ et la corrélation est parfaitement négative.

Exemple :

Soit le tableau ci-dessous, calculer le coefficient de corrélation « r ».

x_i	y_i
5	10
10	30
15	50

Réponse :

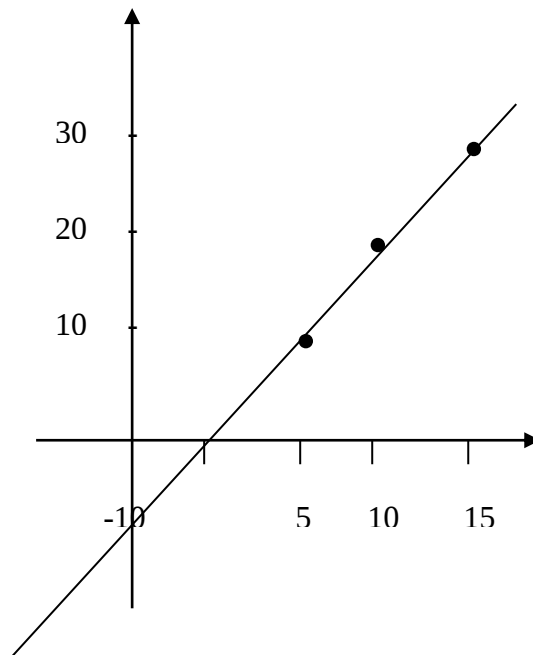
x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
5	10	-5	25	-20	400	100
10	30	0	0	0	0	0
15	50	5	25	20	400	100
$\Sigma = 30$	$\Sigma = 90$		$\Sigma = 50$		$\Sigma = 800$	$\Sigma = 200$
$\bar{x} = 10$	$\bar{y} = 30$					

$$\sigma_{xy} = 200/3, \sigma_x = \sqrt{50/3}, \sigma_y = \sqrt{800/3}$$

$$r = 1$$

Question : Pourquoi la corrélation est parfaite ?

Réponse : Par ce qu'en réalité, on a pris $y = 4x - 10$ puis on a généré les y_i à partir des x_i .



Pour trouver l'équation de cette droite, il faut appliquer la méthode suivante.

$$y - \bar{y} = a (x - \bar{x})$$

$$a = \sigma_{xy} / \sigma_x^2 = (200/3) / (50 / 3) = 4$$