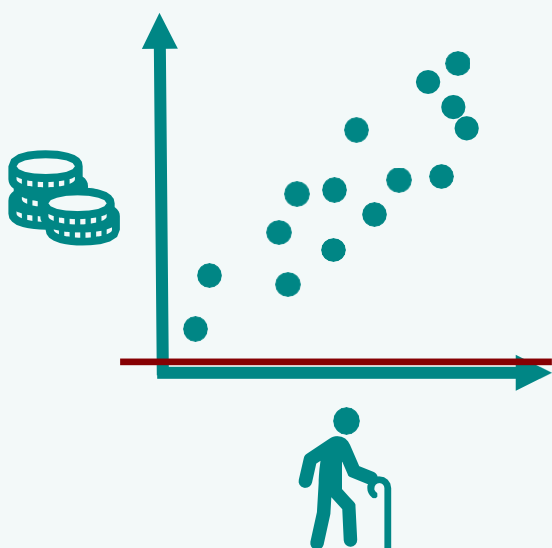


CORRÉLATION DE PEARSON

Playbook

Théorie et exemple



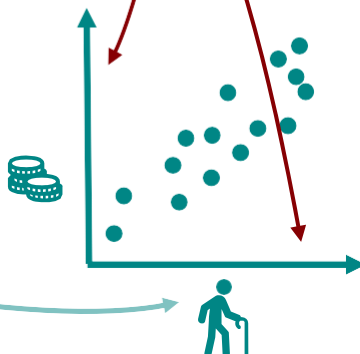


Qu'est-ce que la corrélation de Pearson ?

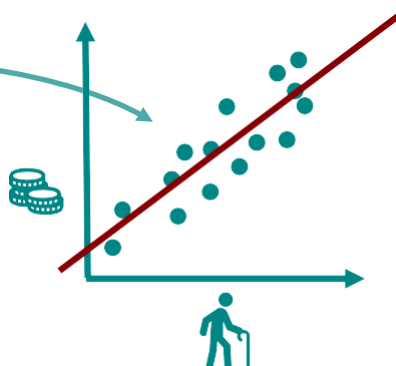
Le corrélation de Pearson analyse le **relation**
entre deux **Variables**.

Par exemple :

Y a-t-il une relation entre
le **salaire** et l'**âge** d'une
personne ?



Dans ce **nuage de points**,
chaque point est une
personne.



Si la relation est confirmée dans cet
exemple, le **salaire** peut être **prédit** en
fonction de l'âge, à l'aide d'une
régression.



Mais faire attention!

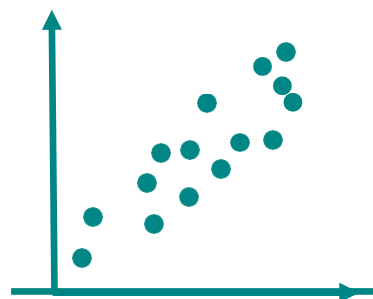
Ce n'est pas toujours facile.
Il doit y avoir une **relation**
de cause à effet claire.



Ce n'est pas parce qu'il y a une
corrélation que vous pouvez dire
dans quelle direction va la
relation .

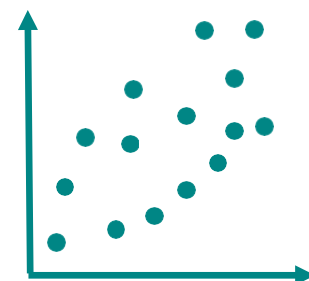
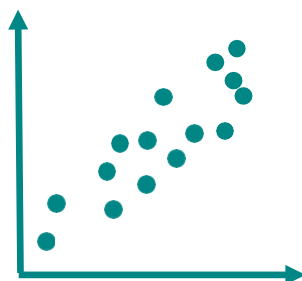


Ainsi, à l'aide de la **corrélation** de **Pearson**, nous pouvons mesurer la **relation linéaire** entre deux variables.

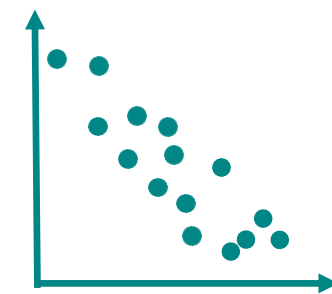
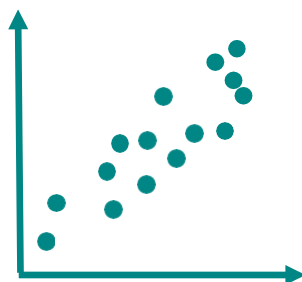


Ici, nous pouvons déterminer :

1 Comment **fort** le **corrélation** est



2 et dans **quel** **direction** le **corrélation** va.



! Nous pouvons lire les deux dans le **Pearson corrélation coefficient r** , quel est **entre -1 et 1**.



Comment lire la **force** de la **corrélation** ?

La **force de la corrélation** peut être lue dans un **tableau**.

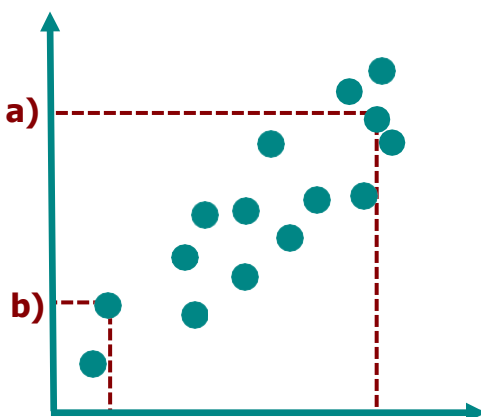
Montant de r	Force de la corrélation	
$0,0 < 0,1$	Pas de corrélation	Si r est entre 0 et 0.1 , nous parler de Non corrélation .
$0,1 < 0,3$	Faible corrélation	
$0,3 < 0,5$	Corrélation moyenne	Si r est compris entre 0,7 et 1 , on parle de corrélation très forte .
$0,5 < 0,7$	Corrélation élevée	
$0,7 < 1$	Corrélation très élevée	

D'après Kuckartz et al. : Statistics, An Understandable Introduction, 2013, p. 11. 213 images

U Positif corrélation

Il existe une corrélation positive

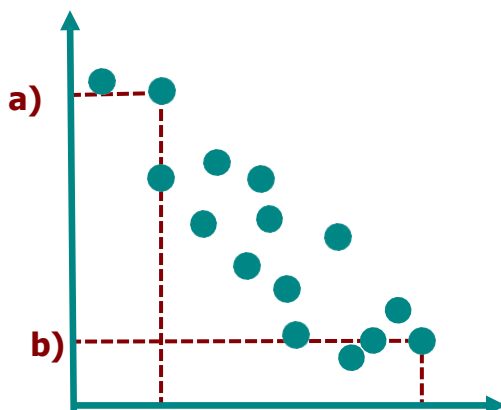
- a) lorsque **les valeurs d'une variable** vont de pair avec les **grandes valeurs** de l'**autre variable**
- b) ou lorsque de **petites valeurs** de **une variable** va de pair avec **Petites valeurs** de l'**autre variable**



B Négatif corrélation

Il existe une corrélation négative

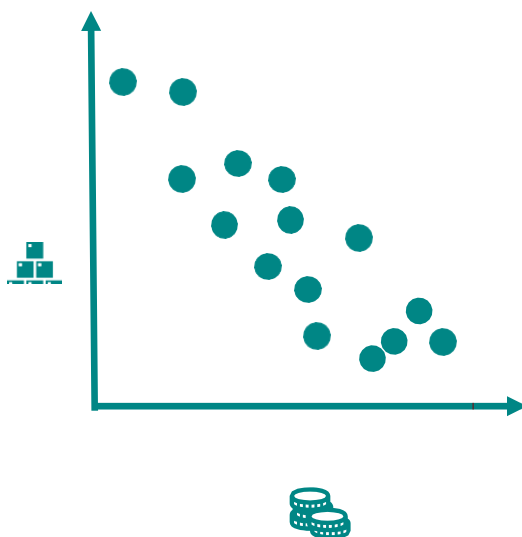
- a) lorsque **les valeurs d'une variable** vont de pair avec les **petites valeurs** de **l'autre variable**
- b) ou lorsque de **petites valeurs** de **une variable** va de pair avec **grandes valeurs** de **l'autre variable**



Il existe généralement une **corrélation négative** entre **le prix** du produit et **les ventes** **le volume**.

Il en résulte un **coefficient** de **corrélation négatif**.

$$r < 0$$





Comment la **corrélation** de **Pearson** est-elle calculée ?

Le **coefficient** de **corrélation de Pearson** est obtenu route cette **équation** :

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

x_i sont les **individues Valeurs** d'une **variable**, par exemple l'âge

y_i sont les **valeurs individuelles** de l'autre **variable**, par exemple le salaire

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Où **r** est le coefficient de corrélation de Pearson,

x et **y** sont respectivement les **valeurs moyennes** des deux variables.



Dans l'**équation**, nous pouvons voir que la respective **valeur** moyenne est d'abord **soustraite** des deux variables.

Exemple:

Dans notre exemple, nous calculons les **valeurs moyennes** de l'âge et du **salaire**.

Nous soustrayons ensuite les **valeurs moyennes** de l'âge et du **salaire** de chaque personne.

On **multiplie** ensuite les deux valeurs.

Ensuite, nous **additionnons** les résultats individuels de la multiplication.

$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$



$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$



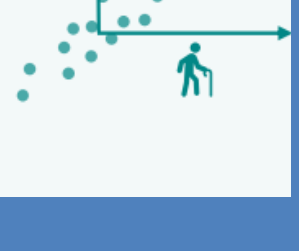
$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$



$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$



$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$

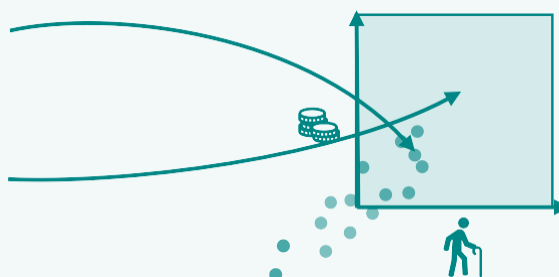


L'expression au **dénominateur** permet de s'assurer que le **coefficient** de **corrélation** est **mis à l'échelle** entre **-1 et 1**.

$$r = \frac{\sum (Age_i - \overline{Age}) \cdot (salary_i - \overline{salary})}{\sqrt{\sum (Age_i - \overline{Age})^2 \cdot \sum (salary_i - \overline{salary})^2}}$$

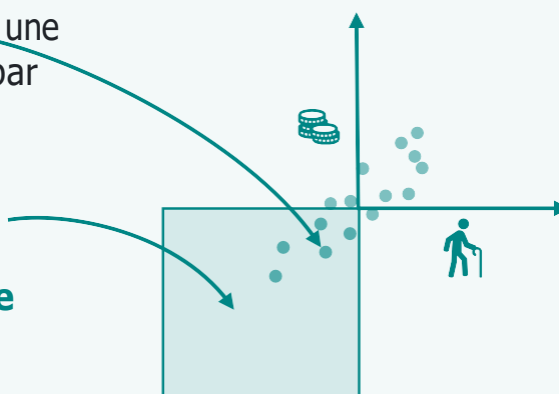
Si nous **multiplions maintenant deux valeurs positives**, nous obtenons une **valeur positive**.

Ainsi, toutes les valeurs qui se trouvent dans cette zone ont une **influence positive** sur le coefficient de corrélation.



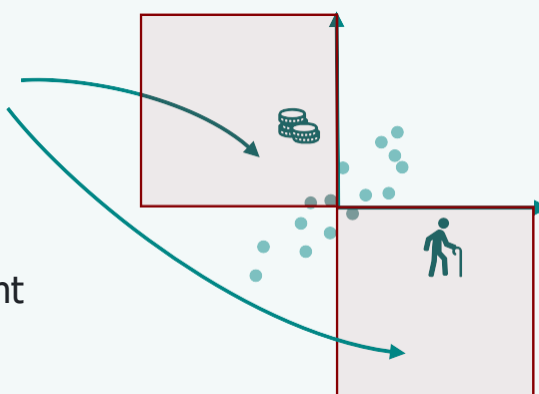
Si nous **multiplions maintenant deux valeurs négatives**, nous obtenons une **valeur positive**. (Moins multiplié par moins est plus).

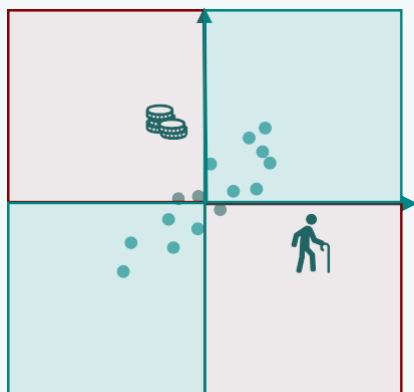
Ainsi, toutes les valeurs qui se trouvent dans cette zone ont également une **influence positive** sur le coefficient de corrélation.



Si nous **multiplions** une valeur **positive** et une valeur négative, nous obtenons une **valeur négative**. (Moins fois plus est moins).

Ainsi, toutes les valeurs qui se situent dans ces plages ont une **influence négative** sur le coefficient de corrélation.





Par conséquent, si nos valeurs se situent **principalement dans les deux zones vertes**, nous obtenons un **coefficient de corrélation positif**.

Et donc, une **relation positive**.

Si nos valeurs sont **Dans les deux zones rouges**, nous obtenons un **coefficient de corrélation négatif**.

Et donc, une **relation négative**.



Si les points sont **répartis sur les quatre domaines**, les termes positifs et les termes négatifs s'annulent et nous obtenir une **corrélation très faible ou nulle**.



Le **coefficient de corrélation** est généralement calculé à partir des données d'un **échantillon**.



Cependant, nous voulons souvent tester un l'hypothèse sur la **population**.



Échantillonnage



Échantillon de population

Dans le cas de l'**analyse** des **corrélations**, nous voulons alors savoir s'il y a une **corrélation** dans la **population**.

Pour cela, nous vérifions si le **coefficient de corrélation** dans l'échantillon est statistiquement significativement **différent de zéro**.

Le **hypothèse** dans la **corrélation de Pearson**



Hypothèse nulle

H0

Le coefficient de corrélation
ne diffère pas
significativement de **zéro**.



Il **n'y a pas de**
relation.

Attention:

Il est toujours vérifié si le
L'hypothèse nulle est rejetée ou non !

Hypothèse alternative

H1

Le coefficient de corrélation **Diffère**
significativement de **zéro**.



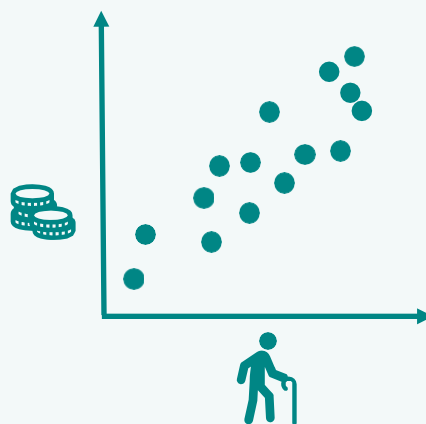
Il **y a une** relation
linéaire.

Dans notre exemple, nous pourrions avoir la **question de recherche** :

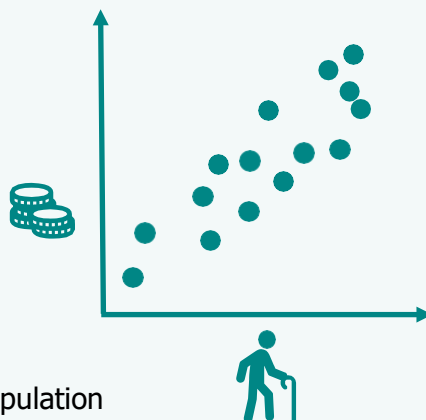
Y a-t-il une **corrélation** entre **l'âge** et **le salaire** ?
dans la **population britannique** ?



Population



Pour le savoir, nous tirons au sort un échantillon et vérifions si, dans cet échantillon, le coefficient de corrélation est significativement **différent de zéro**.



Échantillon de population

L'hypothèse **nulle** est alors :

Il n'y a pas de **corrélation** entre le **salaire** et l'âge dans la population britannique.

Et l'hypothèse **alternative** :

Il existe une **corrélation** entre le **salaire** et l'âge dans la population britannique.

Si le coefficient de corrélation est significativement différent de zéro en fonction de l'échantillon prélevé



Échantillon

peut être vérifié à l'aide d'un **test T**.

$$t = \frac{r \cdot \sqrt{n - 2}}{\sqrt{1 - r^2}}$$

Où **r** est le coefficient de corrélation

et **n** est la taille de l'échantillon .

Une **p-value** peut alors être calculée à partir de la **statistique de test t**.

Si la **p-value** est inférieure au **seuil de signification** spécifié, qui est **généralement** de **5 %**, alors l'hypothèse nulle est rejetée, sinon elle ne l'est pas.

Qu'en est-il de l'hypothèse pour une corrélation de **Pearson** ?



Ici, nous devons **distinguer** si nous voulons simplement calculer le **coefficient de corrélation de Pearson** ou si nous voulons **tester une hypothèse**.



Pour calculer le coefficient de corrélation de Pearson, seules deux **variables métriques** doivent être présentes.



le poids

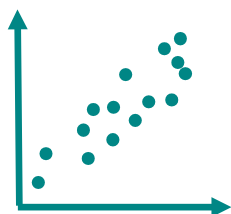


d'une personne ; le salaire ou l'électricité

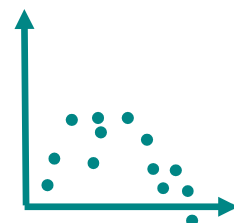


d'une personne ; consommation.

Le coefficient de corrélation de Pearson nous indique alors la taille de la **relation linéaire**.



S'il existe une **relation non linéaire**, nous **ne pouvons pas** le dire à partir du coefficient de corrélation de Pearson.



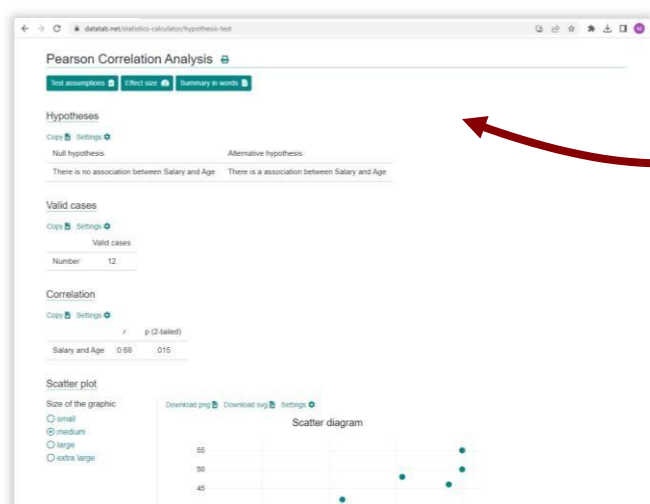
⋮

!

Toutefois si nous voulons **À test** si le **Pearson corrélation coefficient** est **significativement différent de zéro**, les deux **Variables** doivent aussi être **normalement distribués**!

Si ce n'est pas le cas, la **statistique de test t** calculée ou la **valeur de p** ne peuvent pas être interprétées de

Comment calculons-nous une **corrélation de Pearson** avec *l'onglet DATA* ?



Si vous le souhaitez, vous pouvez également calculer le **Corrélation** de Pearson avec *l'onglet DONNÉES*

**TRÈS À L'ONGLET
DONNÉES**

Si vous le souhaitez, vous pouvez bien sûr calculer une analyse de corrélation en ligne avec l'onglet **DONNÉES**.

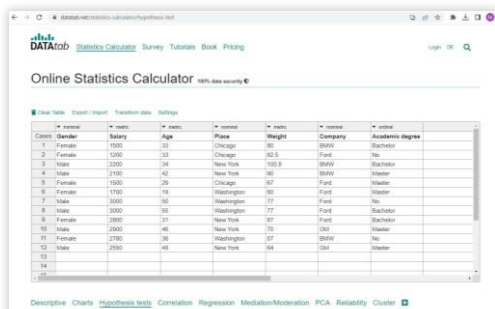
1 Copiez simplement vos données dans ce [table](#) et cliquez sur ou le Hypothèses ou Corrélation onglet.

2

Si vous cliquez maintenant sur deux métriques Variables a Pearson corrélation volonté être calculé automatiquement.

!

Si vous ne savez pas exactement comment faire interpréter le résultats vous pouvez aussi juste cliquer sur Résumé dans mots!



Case#	Gender	Salary	Age	Place	Weight	Company	Academic degree
1	Female	1000	23	Chicago	60	Bank	Bachelor
2	Female	1200	33	Chicago	62.5	Farm	No
3	Male	2000	30	New York	100.5	Bank	Bachelor
4	Male	2100	42	New York	90	Bank	Master
5	Female	1000	25	Chicago	60	Farm	Master
6	Female	1100	35	Washington	67	Farm	Master
7	Male	3000	50	Washington	177	Farm	No
8	Male	3000	50	Washington	177	Farm	Bachelor
9	Female	2000	31	New York	67	Farm	Bachelor
10	Male	2000	40	New York	70	Bank	Master
11	Female	2700	30	Washington	57	Bank	No
12	Male	2700	40	New York	64	Bank	Master

En savoir ^{plus}

CORRÉLATION DE PEARSON

sur notre site web **datatab.net**

Feel free
to share!