I. Corrélation et régression linéaire : (Liaison entre deux variables)

On constate très souvent dans la pratique qu'il existe une liaison entre deux (ou plusieurs) variables. Par exemple : la relation entre le poids et la taille des hommes adultes sont liés d'une certaine façon ; entre le revenus et les dépenses en nourriture.

Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points. La forme générale de ce graphique indique s'il existe ou non une liaison entre les deux variables.

Pour préciser les choses, on calcule ensuite un indicateur de liaison. Pour cela, il faut d'abord introduire la covariance, généralisation bidimensionnelle de la variance. Comme elle dépend des unités de mesure des deux variables considérées, on doit la rendre intrinsèque en la divisant par le produit des écarts-types. On définit ainsi le coefficient de corrélation linéaire, indicateur de liaison cherche. Il est toujours compris entre 1 et +1, son signe indique le sens de la liaison, tandis que sa valeur absolue en indique l'intensité.

En complément, on explique ce qu'est la régression linéaire d'une variable sur une autre. Lorsqu'il existe une liaison causale entre les deux variables considérées, la régression linéaire permet d'approcher la variable réponse par une fonction de la variable causale.

Le test de corrélation (contrairement à la régression simple) ne propose pas d'identifier une variable dépendante et une variable indépendante. On ne cherche qu'à déterminer l'absence ou la présence d'une relation linéaire significative entre les variables.

1) Détermination de coefficient de corrélation de Person: Pour le calcul du coefficient de corrélation noté ***r***, nous utilisons la formule la plus simple :

$$r = \frac{Cov(xy)}{\delta_x \delta_y}$$

Où :

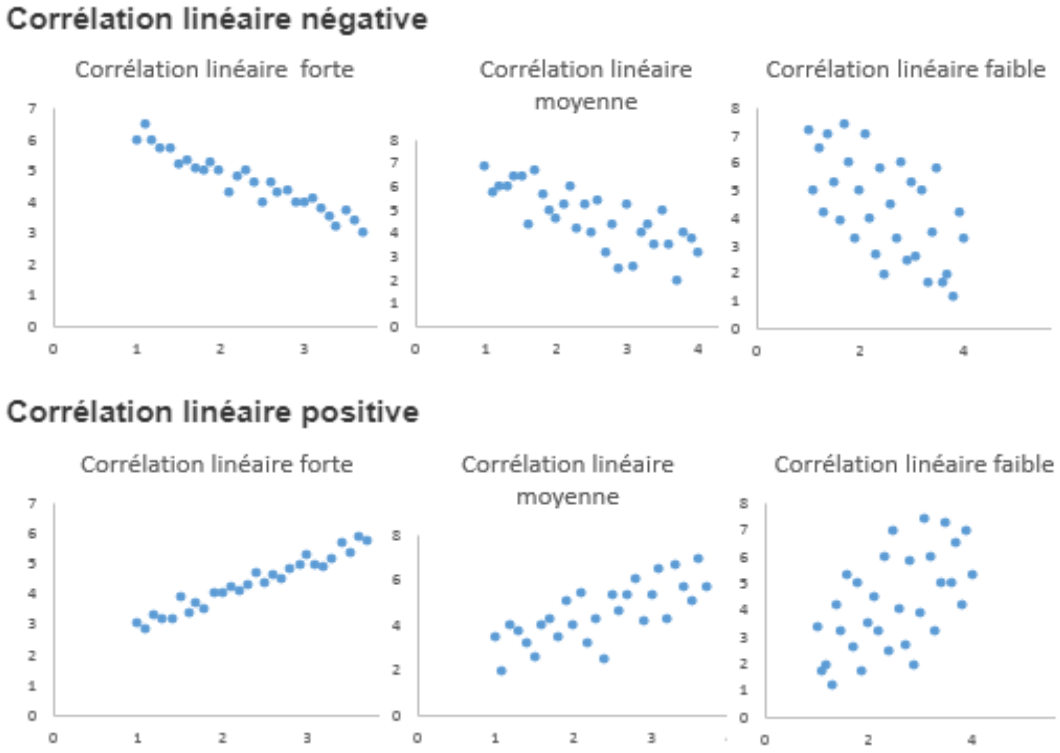
- $\bar{\mathbf{X}}$: la moyenne de la première distribution
- $\bar{\mathbf{Y}}$: la moyenne de la deuxième distribution
- $\boldsymbol{\delta}$ *(x)* : l'écart-type de la variable x.
- $\boldsymbol{\delta}$ *(y)* : l'écart-type de la variable y.
- ***Cov*** : la covariance.

Généralement, les valeurs suivantes seront utilisées pour qualifier la corrélation linéaire:

Valeur de r	**Force du lien linéaire**
Près de 0	Nulle
0< r ≤ 0.2	Faible
0.2< r≤ 0.5	Moyenne
0.5< r ≤ 0.8	Forte
r >0.8	Très forte
1	parfaite

Afin de bien voir la différence entre chacun des qualificatifs de corrélation, voici des nuages de points qui les représentent:

Figue N°01 : les types de corrélation entre deux variables.



Source : Philippe Tassi : Méthodes statistiques 2*e* Eddition , Economica, 1989,p.85

Qu'est-ce que la covariance :

La corrélation est une quantification de la relation linéaire entre des variables continues. Le calcul du coefficient de corrélation de Pearson repose sur le calcul de la covariance entre deux variables continues. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation.

La covariance est une mesure de l'association ou du lien qui existe entre deux variables. Pour comprendre la covariance, revenons à la notion de variance. La variance d'une variable est une mesure qui quantifie la dispersion moyenne des valeurs prises par cette variable autour de sa moyenne.

Deux variables covariants ensemble lorsqu'un écart à la moyenne d'une variable est accompagné par un écart dans le même sens ou dans le sens opposé de l'autre pour le même sujet. Plus ce pattern est présent dans l'ensemble des

observations, plus les deux variables semblent partager une association entre elles. Autrement dit, deux variables covariants lorsque la variation d'une des variables autour de sa moyenne semble influencer la manière dont l'autre variable varie autour de sa moyenne. La covariance exprime donc une quantité de variance partagée entre deux variables. En effet, tout comme la variance, la covariance peut se quantifier. Plus la valeur de la covariance est élevée, plus les deux variables partagent une portion importante de variance.

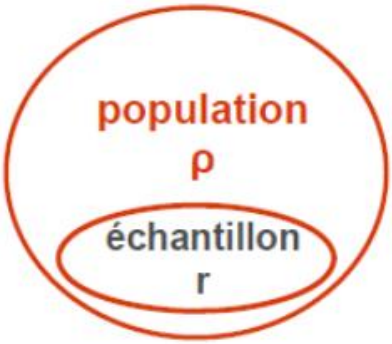
Voici la formule permettant de calculer la covariance entre deux variables continues :

$$Cov(X,Y) = \frac{\sum[(x_i - \overline{X})(y_i - \overline{Y})]}{N} \quad = \quad Cov(X,Y) = \frac{\sum(x_i y_i)}{N} - (\overline{XY})$$

Test du coefficient de corrélation :

Après le calcul du coefficient de corrélation il faut déterminer si le coefficient de corrélation ρ de 0**.** $r \approx \rho$

Figure N°01 :



- L'hypothèse nulle **(H0) : ρ=0** (il n'y a pas une relation entre **X** et **Y**).
- L'hypothèse alternative (bilatérale) **(H1) : $\rho \neq 0$** (il y'a une relation entre **X** et **Y**).

Le test du coefficient de corrélation consiste à calculer la grandeur $\mathbf{t_0}$ et à la comparer à la valeur seuil $\mathbf{t\alpha}$ sur la table de la loi de Student à **(*n*-2)** degrés de libertés.

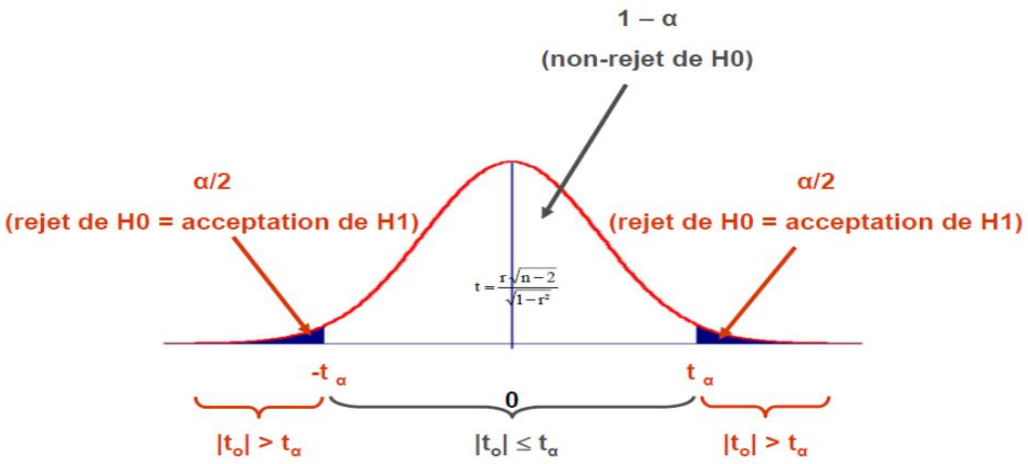
$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- **tα:** tirez de la table du Student.
- **t_0:** Valeur observée/calculée de **t** sur l'échantillon.
- ***r*** : coefficient de corrélation.
- ***n*** : la taille de l'échantillon.

Conditions d'application :

- Indépendance des observations.
- Liaison linéaire entre X et Y.
- Distribution conditionnelle normale et de variance constante.

Figure N°02 :



Abscisses : valeurs possibles de **t** sous H_0 ($\rho = 0$)

Détermination du degré de signification associé à t_o (*P* - value) :

ρ -value= probabilité d'observer une valeur plus grande que $\mathbf{t_o}$ sous l'hypothèse nulle $\mathbf{H_0}$.

Exemple01 :

$\mathbf{t_o}$ =2.12 ***n***= 20 (*n*-2) = 18 *ddl*

0.02 < ρ <0.05 → ***P*** **< α** → rejet de H_0

Remarque :

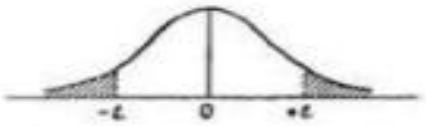
On accepte l'hypothèse alternative $\mathbf{H_1}$ si seulement si $\mathbf{H_0}$ est inferieur de **0.05**.

On rejette l'hypothèse alternative $\mathbf{H_1}$ si seulement si $\mathbf{H_0}$ est supérieur de **0.05**.

Figure N°03 : Table du Student.

Table de t (*).

La table donne la probabilité α pour que t égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291