Partie 3 : La corrélation

La notion de corrélation se rapporte au degré de liaison qui unit deux ou plusieurs variables. Selon la nature et le nombre de variables impliquées, on utilise une terminologie propre correspondant à des définitions différentes du même concept.

+ Liaison entre deux variables quantitatives distribuées normalement ⇒ corrélation linéaire simple.

+ Intensité de la relation liant une variable dépendante à un ensemble de variables indépendantes quantitatives ⇒ corrélation multiple.

+ Lien entre deux ensembles de variables quantitatives ⇒ corrélation canonique.

+ Relation entre deux variables semi quantitatives ⇒ corrélation de rang.

+ Relation entre deux variables qualitatives ⇒ association.

+ Relation entre deux variables qualitatives binaires ⇒ corrélation de point.

1. Corrélation entre deux variables quantitatives : corrélation de Pearson

La corrélation de Pearson ou de Bravais – Pearson est une mesure de la liaison linéaire existant entre deux variables quantitatives normales.

$$r_{xy} = \frac{\text{Cov. xy}}{\text{Sx} * \text{Sy}} = \frac{\sum \text{xy} - \frac{\sum \text{x} \sum \text{y}}{\text{n}}}{\sqrt{\text{SCEx} * \text{SCEy}}}$$

Toutes les valeurs de $r_{x,y}$ sont comprises entre -1 et +1. r = -1 ou +1 si tous les points du diagramme de dispersion sont situés sur une ligne droite et r = 0 lorsque le nuage de dispersion ne montre aucune tendance de relation entre les deux variables. Aussi, si r est positif ⇒ les deux variables augmentent au même temps et si r est négatif ⇒ l'une des variables augmente quand l'autre diminue.

Application : On désire vérifier la corrélation entre la taille (en cm) et le poids (en kg) des enfants de 2 ans sur un échantillon de 15 individus.

Taille (X)	82.9	83.4	82.4	82.1	84.8	86.7	84.0	89.0	85.0	85.4	87.7	87.7	86.4	86.4	86.9
Poids (Y)	8.7	9.2	9.5	10.1	10.4	10.5	10.8	11.0	11.5	11.6	12.4	13.6	13.8	13.9	14.6

Calcul de $r_{x,y}$

XY	721.23	767.28	782.8	829.21	881.92	910.35	907.2	979	977.5	990.64	1087.5	1192.3	1192.3	1201	1268.7

∑x = 1280.8	∑y = 171.6
SCEx = 62.1	SCEy = 47.9

∑xy = 14689.4

$$r_{xy} = \frac{\text{Cov.xy}}{\text{Sx}*\text{Sy}} = \frac{\sum \text{xy} - \frac{\sum \text{x} \sum \text{y}}{\text{n}}}{\sqrt{\text{SCEx} * \text{SCEy}}} = \frac{14689.4 - \frac{1280.8 * 171.6}{15}}{\sqrt{62.1 * 47.9}} = \frac{37}{54.5} = 0.68$$

2. Test de signification du coefficient de corrélation de Pearson

A partir d'un échantillon de n sujets sur lesquels on relève les couples de valeurs (X, Y), on estime r et on vérifie si l'estimation obtenue est suffisamment distante de 0 pour rejeter l'hypothèse d'indépendance (r = 0). H0 : r = 0 Les variables ne sont pas corrélées

Le r peut être comparé directement à la valeur critique donnée par la table de signification du coefficient de corrélation pour un ddl = au nombre de couples d'observation (x, y) diminué de 2 soit : (n-2) ddl, soit lorsque celle-ci est insuffisante (cas ou ddl > 100 et $\alpha < 1\%$), en calculant :

$$\text{tr} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ et comparer cette valeur à $t_{1-\alpha/2}$ pour un ddl = n-2.

Si $t_r > t_{1-\alpha/2} \Rightarrow$ H0 est rejetée $\Rightarrow$ la liaison est significative.

Si $t_r \leq t_{1-\alpha/2} \Rightarrow$ H0 est acceptée $\Rightarrow$ la liaison n'est pas significative.

***Application* :**

1) Soit un échantillon de 27 couples (x, y) et r = 0,4. Cette valeur est-elle significativement différente de 0 pour $\alpha = 0.05$.

Hypothèse H0 : les variables X et Y sont indépendantes, r = 0

Conclusion : La valeur lue dans la table du coefficient de corrélation pour $\alpha = 0.05$ et un ddl = 25 est : 0.3809.

$r > 0.3809 \Rightarrow$ r est significativement différent de 0

2) Soit un échantillon de 150 couples (x, y) et r = 0,2. Cette valeur est-elle significativement différente de 0 pour $\alpha = 0.05$.

Valeur de p pour r = 0,2 et ddl=148 : la table des valeurs critiques du coefficient de corrélation est inutilisable, on calcule :

$$\text{tr} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.2\sqrt{148}}{\sqrt{1-0.04}} = \frac{2.43}{0.98} = 2.48$$

$t_{1-\alpha/2}$ pour un ddl = 148 = 1.978

$t_r > t_{1-\alpha/2} \Rightarrow$ H0 est rejetée $\Rightarrow$ la liaison est significative. r est significativement différent de 0.

3) Soit un échantillon de 15 couples (x, y) et un r = 0.5: test de signification ($\alpha = 0.05$).

Valeur critique données par la table des valeurs critiques du coefficient de corrélation pour un ddl de 13 = 0.5139. $r < 0.5139 \Rightarrow$ r est significativement égale à 0. En utilisant la table de Student :

$$tr = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.50\sqrt{13}}{\sqrt{1-0.50^2}} = \frac{1.80}{0.87} = 2.07$$

$t_{1-\alpha/2}$ pour un ddl = 13 = 2.16

$t_r < t_{1-\alpha/2} \Rightarrow$ H0 est acceptée $\Rightarrow$ la liaison n'est pas significative et r est significativement égale à 0.

Commentaire : A travers les 3 exemples présentés, nous comprenons que la significativité d'un coefficient de corrélation ne dépend pas seulement de sa valeur, mais aussi de la taille de l'échantillon. En effet, nous venons de démontrer qu'un coefficient de 0.2 est significatif avec un échantillon de grande taille (n = 150) et qu'un autre de 0.5 ne l'est pas du fait de la faible taille de l'échantillon (n = 15).

3. Coefficient de corrélation de Rang de Spearman

L'estimation et le test de signification du coefficient de corrélation linéaire de Pearson reposaient sur deux conditions : le caractère quantitatif des deux variables et la normalité de leurs distributions. Il arrive fréquemment, en biologie ou dans le domaine médical, qu'une des variables ou même les deux soient semi-quantitatives, ou encore que l'une ou l'autre des deux distributions ne soit pas normale. En pareil cas, l'utilisation du coefficient de corrélation de rang est plus appropriée.

Le coefficient de corrélation de rang de Spearman " r_s " indique le degré de liaison entre le classement des éléments selon la variable x et celui selon la variable y.

- Si $r_s = 1 \Rightarrow$ les classements selon x et y sont identiques.
- Si $r_s = -1 \Rightarrow$ les classements selon x et y sont inverses.
- Si $r_s = 0 \Rightarrow$ les deux variables sont indépendantes et l'ordre de classements selon x et y devient aléatoire.

Pour calculer le coefficient de corrélation de Spearman :

- Indiquer le rang de chaque observation pour chaque variable séparément. Dans le cas où deux ou plusieurs éléments occupent le même rang, il faut leur attribuer le rang moyen.
- Calculer la différence d_i de classement pour chaque paire d'observations
- Calculer la quantité suivante :

$$r_s = 1 - \frac{6 \sum di^2}{n(n^2 - 1)}$$

***Application* :** Etapes de calcul du coefficient de corrélation de Spearman

Elément	Valeur de x	Valeur de y	Rang sur x	Rang sur y	Différences d_i sur les rangs	d_i^2
1	12	14	4	6	2	4
2	15	7	5	3	2	4
3	18	20	7	9	2	4
4	22	18	9	7	2	4
5	3	8	1	4	3	9
6	7	3	3	1	2	4
7	4	6	2	2	0	0
8	17	12	6	5	1	1
9	20	19	8	8	0	0
						$\sum d_i^2 = 30$

$$r_s = 1 - \frac{6 \sum di^2}{n(n^2 - 1)} = 1 - \frac{6*30}{9(81 - 1)} = 0.75$$

4. Test de signification du coefficient de corrélation de Rang de Spearman

+ <u>Si n ≥ 10</u> : Tester la signification de r_s par les méthodes utilisées pour le coefficient de corrélation linéaire de Pearson.

+ <u>Si n<10</u> : Comparer r_s à la valeur théorique tirée de la table des valeurs critiques du coefficient de corrélation de Spearman.

Exemples :

1. Reprenant les données de l'exemple précédent : $r_s = 0.75$ et $n = 9$

   Valeur critique pour ($\alpha = 0.05$et $n = 9$) = 0.68.

   $r_s > 0.68 \Rightarrow$ la corrélation est significative et r_s est différent de 0.

2. Soit les données suivantes : $r_s = 0.45$ et $n = 27$

   Valeur critique pour ($\alpha = 0.05$et $n-2 = 25$) = 0.3809.

   $r_s > 0.3809 \Rightarrow$ la corrélation est significative et r_s est différent de 0.

Table des valeurs critiques du coefficient de corrélation linéaire de Pearson.

La table fournit, pour différents effectifs d'échantillons (de 3 à 50), la valeur critique r_α correspondant à divers seuils de signification (0.05, 0.01). La probabilité α se rapporte à un test bilatéral.

n	α = 0.05	α = 0.01	n	α = 0.05	α = 0.01	n	α = 0.05	α = 0.01
3	0.997	1.000	19	0.456	0.575	35	0.334	0.430
4	0.950	0.990	20	0.444	0.562	36	0.329	0.424
5	0.878	0.959	21	0.433	0.549	37	0.325	0.418
6	0.811	0.917	22	0.423	0.537	38	0.320	0.413
7	0.755	0.875	23	0.413	0.526	39	0.316	0.408
8	0.707	0.834	24	0.404	0.515	40	0.312	0.403
9	0.666	0.798	25	0.396	0.505	41	0.308	0.398
10	0.632	0.765	26	0.388	0.496	42	0.304	0.393
11	0.602	0.735	27	0.381	0.487	43	0.301	0.389
12	0.576	0.708	28	0.374	0.479	44	0.297	0.384
13	0.553	0.684	29	0.367	0.471	45	0.294	0.380
14	0.533	0.661	30	0.361	0.463	46	0.291	0.376
15	0.514	0.641	31	0.355	0.456	47	0.288	0.372
16	0.497	0.623	32	0.349	0.449	48	0.285	0.368
17	0.482	0.606	33	0.344	0.442	49	0.282	0.365
18	0.468	0.590	34	0.339	0.436	50	0.279	0.361

Table des valeurs critiques du coefficient de corrélation de rang de Spearman

La table fournit, pour différents effectifs d'échantillons (de 5 à 30), la valeur critique $r_{s\alpha}$ correspondant à divers seuils de signification (0.05, 0.01). La probabilité α se rapporte à un test bilatéral.

n	**α = 0.05**	**α = 0.01**	**n**	**α = 0.05**	**α = 0.01**
5	0.9000	-	18	0.4716	0.5975
6	0.8286	0.9429	19	0.4579	0.5825
7	0.4750	0.8929	20	0.4451	0.5684
8	0.7143	0.8571	21	0.4351	0.5545
9	0.6833	0.8167	22	0.4241	0.5426
10	0.6364	0.7818	23	0.4150	0.5306
11	0.6091	0.7545	24	0.4061	0.5200
12	0.5804	0.7273	25	0.3977	0.5100
13	0.5549	0.6978	26	0.3894	0.5002
14	0.5341	0.6747	27	0.3822	0.4915
15	0.5179	0.6536	28	0.3749	0.4828
16	0.5000	0.6324	29	0.3685	0.4744
17	0.4853	0.6152	30	0.3620	0.4665