Introduction

Une corrélation linéaire permet de voir s'il existe un lien entre deux variables. En d'autres mots, on cherche à voir si lorsqu'une d'entre elles varie, l'autre varie aussi. Par exemple, la chaleur (température) et la vente de crème glacée. Lorsqu'il fait plus ou moins chaud (variation de la température), la vente de crème glacée augmente ou diminue (variation de la vente de crème glacée).Vous avez sûrement déjà entendu l'adage « *corrélation ne veut pas dire causalité* ». En effet, une corrélation ne permet de dire que l'une des deux variables provoque l'autre, il peut y avoir une autre explication (par exemple, une troisième variable qui provoque les deux). Une corrélation permet simplement de voir s'il y a un lien entre les deux variables.

On parle de **corrélation positive** quand les deux variables augmentent et/ou diminuent ensemble. Si on reprend la chaleur et la vente de crème glacée, c'est un bon exemple de corrélation positive. Lorsqu'il fait plus chaud (la chaleur augmente), les ventes de crème glacée augmentent car les gens vont plus en acheter pour se refroidir. Inversement, s'il fait plus froid (la chaleur diminue), les ventes de crème glacée diminuent.On parle de **corrélation négative** lorsqu'une des deux variables diminue et que l'autre augmente. Par exemple, le degré d'adhésion aux valeurs écologistes et le niveau de pollution d'une personne. Logiquement, plus quelqu'un adhère aux valeurs écologistes, moins il pollue.

Pour calculer la corrélation, il faut d'abord transformer vos données en **score Z** (ou scores standardisés). Un score Z indique la position d'une observation par rapport aux autres observations de la distribution. En gros, on change la moyenne pour 0 et l'écart-type devient 1. Pour calculer les score Z, vous avez besoins de la moyenne et de l'écart-type de vos échantillon, si vous n'êtes pas familier avec ces notions, je vous conseille les capsules vidéo sur les mesures de tendances et les mesures de variabilité, de vous référez à votre manuel ou votre professeur de statistique ou de prendre rendez-vous avec un tuteur de soutien à l'apprentissage et à la réussite du CSIPU.

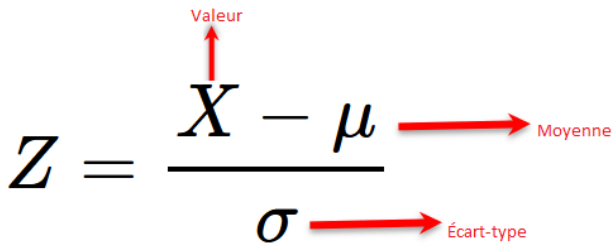
Ensuite, il faut calculer le **coefficient de Pearson (*r*)** qui indique la force de la corrélation. Le coefficient de Pearson peut avoir une valeur entre -1 et 1. Si votre coefficient de Pearson est égal à 0, votre corrélation est dite nulle, il n'y a aucun lien entre les deux variables. Plus la valeur de votre coefficient s'éloigne de 0 (dans les positifs ou dans les négatifs), plus la corrélation est forte. Un coefficient de Pearson positif indique une corrélation positive et un coefficient de Pearson négatif indique une corrélation négative. Pour calculer le

coefficient de Pearson, il faut multiplier les deux scores Z de chaque participant, additionner les résultats et diviser la somme obtenue par le nombre de participants-1. Vous pouvez par la suite élever votre coefficient de Pearson au carré pour calculer votre **coefficient de détermination (r^2)**. Aussi appelé pourcentage de variance expliquée, le coefficient de détermination vous informe sur à quel point les changements dans la variable y sont dûs à la variable x.

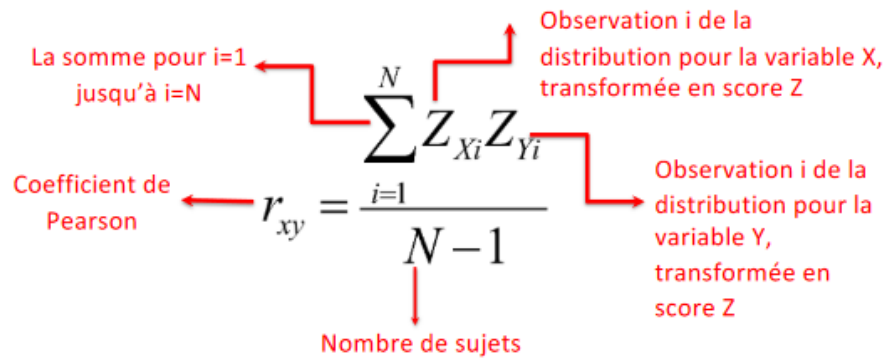
Finalement, on calcule le **score *t***. Le score *t* permet de faire des inférences à partir des résultats obtenus à l'aide de notre échantillon pour les généraliser à la population. Il permet de voir si nos résultats sont dûs au hasard ou s'il existe bien un lien dans la population.

Formules

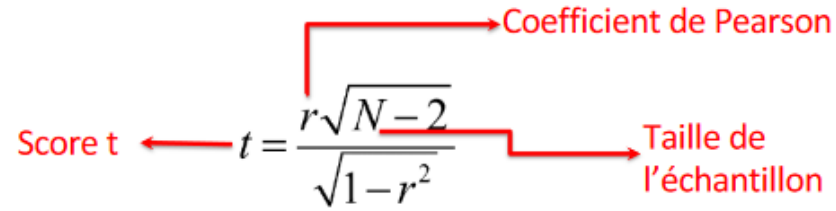
Score Z de chaque variable



Coefficient de Pearson



Score *t*



Mise en situation

Les chercheurs S. Buck et Van Houtte se demandent si la consommation de café a un lien avec le niveau de stress. Ils recrutent 5 participants et leur demandent combien de tasses de café est-ce qu'ils boivent dans une journée et leur font passer un test de stress.

H: La consommation de café a un lien avec le niveau de stress

H_0: La consommation de café n'a pas de lien avec le niveau de stress

Voici leur résultats:

Participant	**Café**	**Stress**
A	5	18
B	2	14
C	7	12
D	0	11
E	3	8
***Moyenne**	3.40	12.60
***Écart-type**	2.70	3.71

Calcul

Étape 1: Transformer les valeurs en score Z

$$Z = \frac{X - \mu}{\sigma}$$

a) pour la variable « café »

A $Z = \frac{5-3.4}{2.7} = \frac{1.6}{2.7} = 0.59$

B $Z = \frac{2-3.4}{2.7} = \frac{-1.4}{2.7} = -0.52$

C $Z = \frac{7-3.4}{2.7} = \frac{3.6}{2.7} = 1.33$

D $Z = \frac{0-3.4}{2.7} = \frac{-3.4}{2.7} = -1.26$

E $Z = \frac{3-3.4}{2.7} = \frac{-0.4}{2.7} = -0.15$

b) pour la variable « stress »

$$Z = \frac{18 - 12.6}{3.71} = \frac{5.4}{3.71} = 1.46$$

$$Z = \frac{14 - 12.6}{3.71} = \frac{1.4}{3.71} = 0.38$$

$$Z = \frac{12 - 12.6}{3.71} = \frac{-0.6}{3.71} = -0.16$$

$$Z = \frac{11 - 12.6}{3.71} = \frac{-1.6}{3.71} = -0.43$$

$$Z = \frac{8 - 12.6}{3.71} = \frac{-4.6}{3.71} = -1.24$$

Étape 2: Calculer le coefficient de Pearson et coefficient de détermination

Participant	Café	Stress	Café (score *Z*)	Stress (Score *Z*)
A	5	18	0.59	1.46
B	2	14	-0.52	0.38
C	7	12	1.33	-0.16
D	0	11	-1.26	-0.43
E	3	8	0.15	-1.24

$$r_{xy} = \frac{\sum_{i=1}^{N} Z_{Xi} Z_{Yi}}{N - 1}$$

$$r = \frac{0.59 * 1.46 + -0.52 * 0.38 + 1.33 * -0.16 + -1.26 * -0.43 + 0.15 * -1.24}{5 - 1}$$

$$r = \frac{0.86 + -0.20 + -0.21 + 0.54 + -0.19}{5 - 1}$$

$$r = \frac{0.80}{5 - 1}$$

$$r = \frac{0.80}{4}$$

$$r = 0.20$$

Vous voyez ci-contre un guide approximatif de comment on interprète généralement la force de lien avec le coefficient de Pearson. Dans notre cas, nous aurions donc une faible corrélation positive.

Pour calculer votre coefficient de détermination:

$$r^2 = 0.2^2 = 0.04$$

Cela signifie que 4% du niveau de stress est expliqué par la consommation de café.

Force de la corrélation	**Valeur de *r***
Parfait	(-)1
Fort	(-)0.9
	(-)0.8
	(-)0.7
Moyen	(-)0.6
	(-)0.5
	(-)0.4
Faible	(-)0.3
	(-)0.2
	(-)0.1
Nulle	0

Étape 3: Calculer le score *t*

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

$$t = \frac{0.2\sqrt{5-2}}{\sqrt{1-0.04}} = \frac{0.2\sqrt{3}}{\sqrt{0.96}} = \frac{0.2 * 1.73}{0.98} = \frac{0.35}{0.98} = 0.36$$

Étape 4: Le degré de liberté

Votre degré de liberté est égal à votre nombre de participants-2.

dl-N-2

dl=5-2

dl=3

Étape 5: trouver la valeur critique de *t*

Vous utilisez la table t (il s'agit de la même table que pour les tests t). Vous voyez ci-contre celle fournie par Tabachnick et Fidell dans leur manuel (Tabachnick et Fidell, 2021). Vous cherchez votre valeur critique à partir de votre degré de liberté et de votre seuil de rejet. Dans notre cas, nous avons un degré de liberté de 3 et utilisons un seuil de rejet 0.05. Notre valeur critique est donc de 3.182.

Si votre score *t* est **supérieur** à votre valeur critique, vous pouvez rejeter l'hypothèse nulle.

$0.6 < 3.182$

Notre score *t* n'est pas supérieur à la valeur critique. Nous ne pouvons donc pas rejeter l'hypothèse nulle et nous prononcer sur l'existence d'un lien entre la consommation de café et le niveau de stress dans la population.

TABLE C.2 Critical Values of the *t* Distribution for $\alpha = .05$ and .01, Two-Tailed Test

Degrees of Freedom	.05	.01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
∞	1.960	2.576

Source: Adapted from Table 9 in *Biometrika Tables for Statisticians*, vol. 1, 3d ed., edited by E. S. Pearson and H. O. Hartley (New York: Cambridge University Press, 1958).

Signification des symbole

X : Valeur

N : Nombre d'observations total

n : Nombre d'observations d'un groupe

- n_1 : nombre d'observations du groupe 1 ;
- n_2 : nombre d'observations du groupe 2 ;
- etc.

K : Nombre de groupes

Σ : Somme

$\bar{X}$: Moyenne d'un échantillon

μ : Moyenne d'une population

s : Écart-type d'un échantillon

σ :

Écart-type d'une population

s^2 : Variance d'un échantillon

σ^2 : Variance d'une population

χ^2 : Khi-carré

A: Fréquence attendue

O: Fréquence observé

L: Ligne

C: Colonne

t : Statistique *t* (ou score *t* dans le cas d'une corrélation/régression linéaire)

r : Coefficient de Pearson

Z : Score Z

b : Coefficient de régression

a : Ordonnée à l'origine

$\hat{Y}$: Valeur de la VD qu'on veut prédire l'aide la VI

F : Statistique F

Ouvrages de référence

- Field, A. (2017). *Discovering Statistics Using IBM SPSS Statistics: North American Edition* (5th ed.). Sage Edge
- Howell. D.C. (2008). *Méthodes statistiques en sciences humaines* (M. Rogier, V. Yzerbyt, & Y. Bestgen, Trans.). (6th ed.). De boeck. (Original work published 2008)
- Tabachnick, B. G., & Fidell, L. S. (2021). *Using Multivariate Statistics* (7th ed.). Pearson