UNIVERSITÉ DE NANTES

Liens entre deux variables quantitatives (analyse de corrélation) Test de Bravais-Pearson

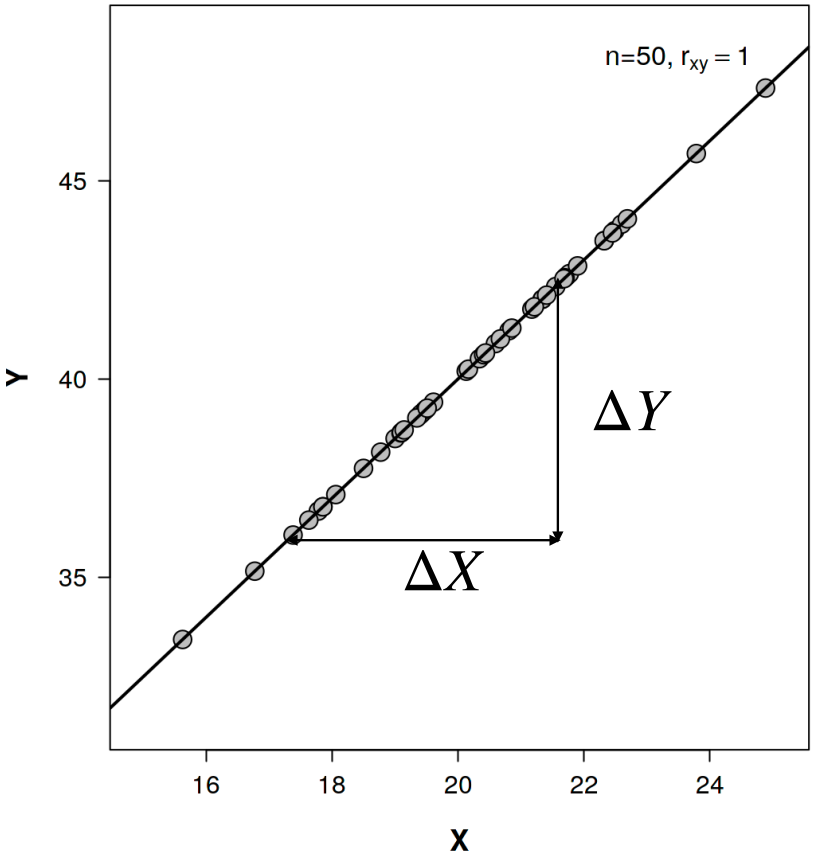
Module HPS3-32
Année 2021-22

GALHARRET J-M,
Laboratoire de Mathématiques Jean Leray
Faculté de Psychologie.

Test de Bravais-Pearson

Conditions d'application

- Soient X et Y deux variables quantitatives mesurées sur une même population.
- On suppose que le couple (X,Y) suit une loi normale (il ne suffit pas que X et Y suivent une loi normale!)
- On recherche un lien de type linéaire entre X et Y : c'est à dire que l'on cherche à prouver qu'une augmentation au niveau de Y est proportionnelle à une augmentation de X.



Variabilité commune

Covariance de (X,Y)

- **<u>Rappel :</u>** Soit X une variable quantitative et $(X_i)_{i=1,\dots,n}$ un échantillon de valeurs de X de moyenne $\bar{X}$ une estimation de la variance de X est donnée par la formule $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Une estimation de la covariance de (X,Y) est donnée par $s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$. Pour $X = Y$ on retrouve la variance de X et de Y.
- Formule de calcul : $s_{X,Y} = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i Y_i - n \times \bar{X}\bar{Y}\right)$

Variabilité commune

Coefficient de corrélation

- Le coefficient de corrélation de (X,Y) est donné par la formule $r_{X,Y} = \frac{s_{X,Y}}{s_X s_Y}$
- Il est égal à la covariance des variables $\tilde{X}$, $\tilde{Y}$ qui correspondent aux variables X,Y centrées réduites.
- Il donne une mesure de l'alignement des points du nuage (X_i, Y_i)
  - (https://shiny.rit.albany.edu/stat/rectangles/)
- On a les propriétés suivantes :
  1. $-1 \leq r \leq 1$
  2. Si $r = +1$ ou $r = -1$ alors les points du nuage (X_i, Y_i) sont parfaitement alignés

Procédure de test

Test de Bravais-Pearson (B-P)

- On teste $H_0 : r_{X,Y} = 0$ versus $r_{X,Y} \neq 0$.
- Ho s'exprime comme « il n'y a pas de lien (ou d'association) linéaire significatif entre les variations de X et les variations de Y »
- Le fait que le lien soit linéaire est souvent omis.
- Sous Ho, la distribution de $|r_{X,Y}|$ est donnée par la table de Bravais-Pearson à $(n-2)$ degrés de liberté.
- La taille d'effet du test de B-P est également basé sur $r_{X,Y}$:

$\|r\|$	[0.1,0.3[	[0.3,0.5[	[0.5,1]
Effet	Faible	Modéré	Fort

Exemple

JASP Database Big Five Personality Traits

This data set, "Big Five Personality Traits", provides scores of 500 participants on a Big Five personality questionnaire.

Variables:

Neuroticism - Score on the Neuroticism scale

Extraversion - Score on the Extraversion scale.

Openness - Score on the Openness to Experience scale.

Agreeableness - Score on the Agreeableness scale.

Conscientiousness - Score on the Conscientiousness scale.

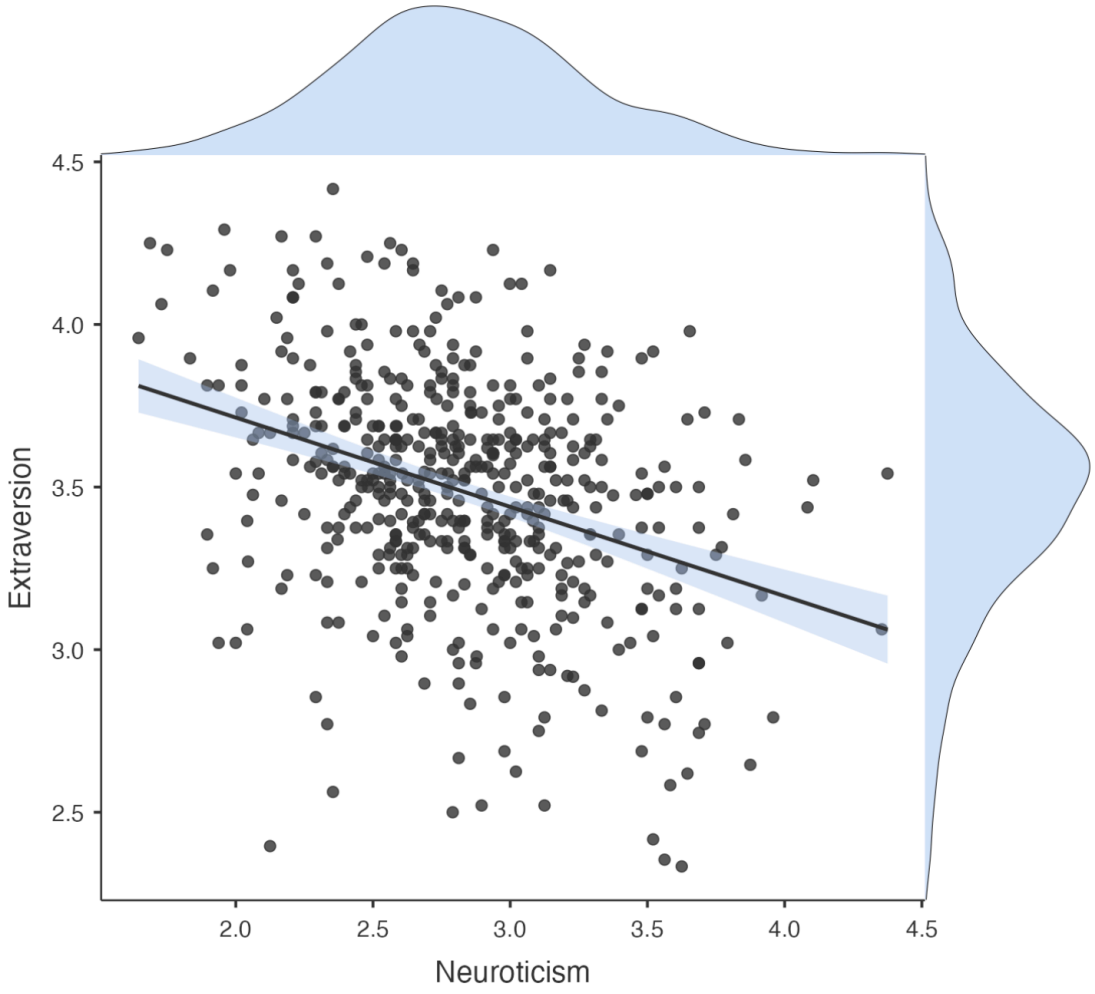
On va tester deux hypothèses :

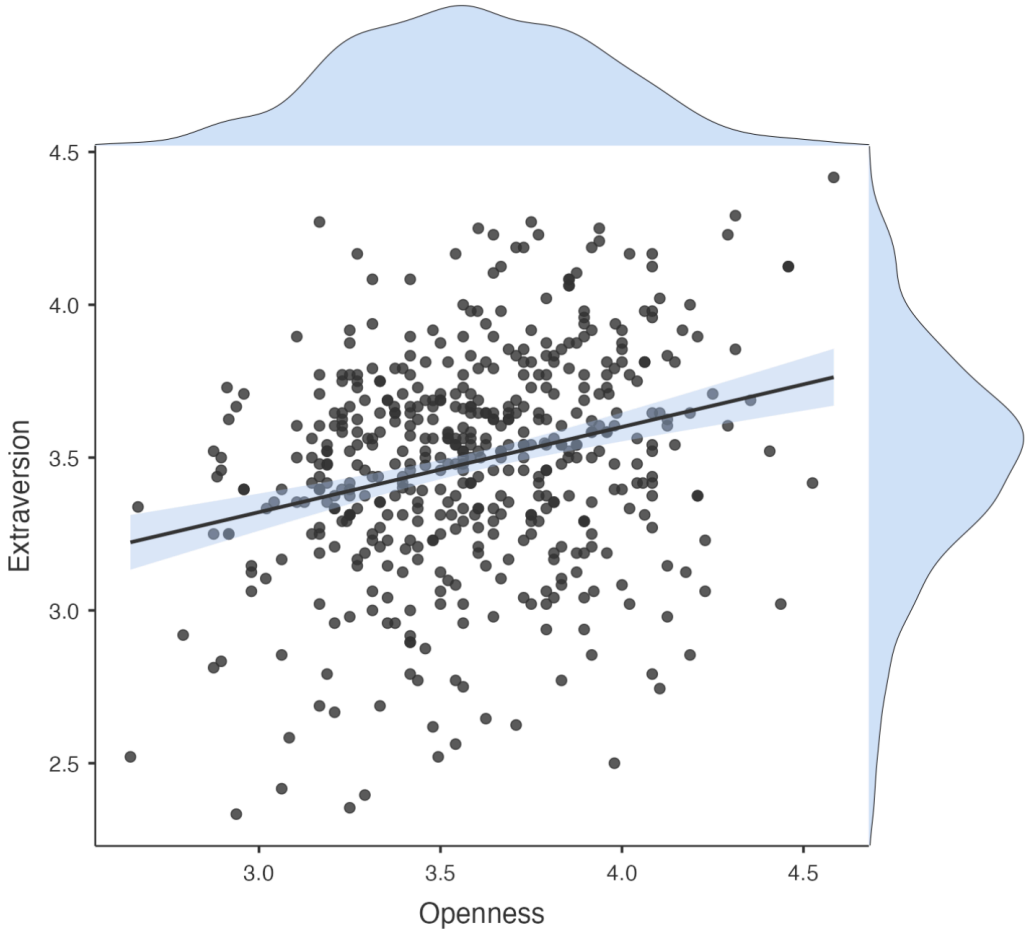
- Une diminution du score Neuroticism est associé à une augmentation du score d'Extraversion.
- Une augmentation du score d'Extraversion est associé à une augmentation du score d'Openness.

	Neuroticism	Extraversion
N	500	500
Missing	0	0
Mean	2.83	3.49
Median	2.80	3.52
Standard deviation	0.453	0.355
Minimum	1.65	2.33
Maximum	4.38	4.42

On a $s_{X,Y} = -0.056$

Ce qui donne $r = \frac{s_{X,Y}}{s_X s_Y} = \frac{-0.056}{0.453 \times 0.355} = -.350$





$r = .267$

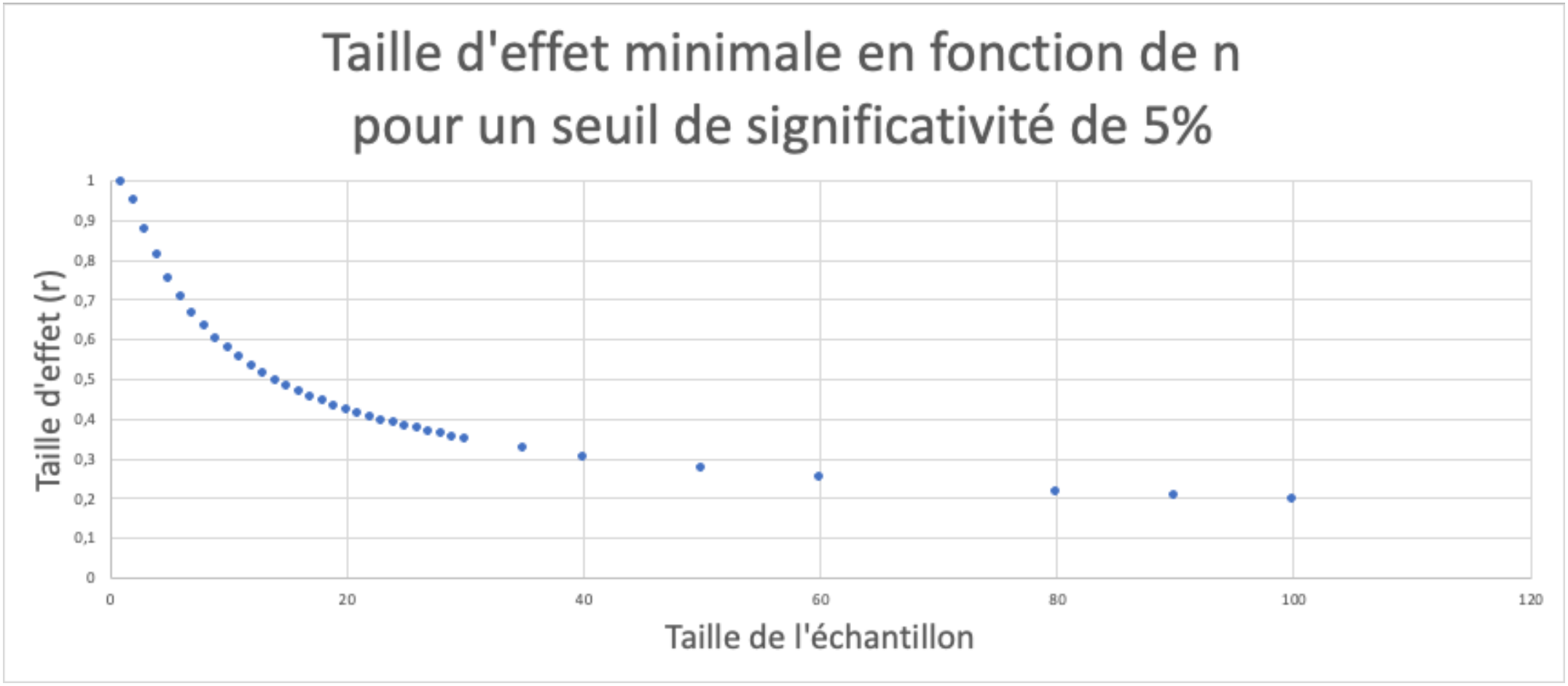
Table de Bravais-Pearson

	p_value						
ddl	**0,2**	**0,1**	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
1	0,951	0,988	0,997	1,000	1,000	1,000	1,000
2	0,800	0,900	0,950	0,980	0,990	0,998	0,999
3	0,687	0,805	0,878	0,934	0,959	0,986	0,991
4	0,608	0,729	0,811	0,882	0,917	0,963	0,974
5	0,551	0,669	0,754	0,833	0,875	0,935	0,951
6	0,507	0,621	0,707	0,789	0,834	0,905	0,925
7	0,472	0,582	0,666	0,750	0,798	0,875	0,898
8	0,443	0,549	0,632	0,715	0,765	0,847	0,872
9	0,419	0,521	0,602	0,685	0,735	0,820	0,847
10	0,398	0,497	0,576	0,658	0,708	0,795	0,823
30	0,233	0,296	0,349	0,409	0,449	0,526	0,554
40	0,202	0,257	0,304	0,358	0,393	0,463	0,490
50	0,181	0,231	0,273	0,322	0,354	0,419	0,443
60	0,165	0,211	0,250	0,295	0,325	0,385	0,408
80	0,143	0,183	0,217	0,257	0,283	0,336	0,357
90	0,135	0,173	0,205	0,242	0,267	0,318	0,338
100	0,128	0,164	0,195	0,230	0,254	0,303	0,321
200	0,091	0,116	0,138	0,164	0,181	0,216	0,230
1000	0,041	0,052	0,062	0,073	0,081	0,098	0,104

Conclusion aux normes APA

Test de B-P

- Un test de B-P a permis de mettre en évidence une corrélation négative, modérée et significative entre le score de Neuroticism et le score d'Extraversion, $r(498) = -.350$, $p < .001$, c'est à dire qu'une augmentation du score d'Extraversion est en moyenne associée à une diminution du score de Neuroticism.
- Un test de B-P a permis de mettre en évidence une corrélation positive, modérée et significative entre le score d'Openness et le score d'Extraversion, $r(498) = .267$, $p < .001$, c'est à dire qu'une augmentation du score d'Extraversion est en moyenne associée à un haut score d'Openness.



Comme toujours plus le lien entre les deux variables est faible plus il faudra des échantillons grands pour prouver que ce lien est significatif.

Corrélation et causalité

Lien ou effet ?

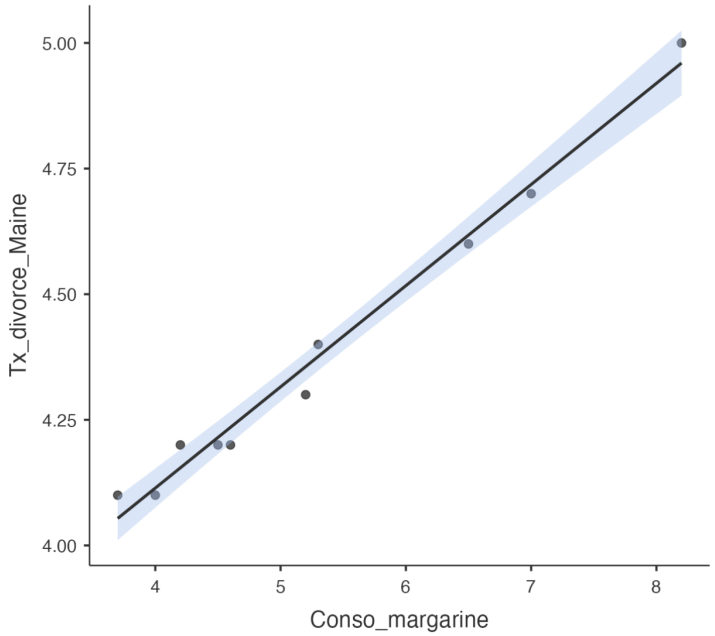
Variable de confusion

- Pour trouver des liens fallacieux voir :

  https://www.tylervigen.com/spurious-correlations

- Prenons un des exemples qui est traité sur ce site : le lien entre le taux de divorce dans le Maine et la consommation de margarine.
- On obtient une corrélation de $r = .993$. Cette corrélation est significative ($p<.001$)

		1	2	3
1-Année	Pearson's r	—		
	p-value	—		
2-Tx_divorce_Maine	Pearson's r	-0.888	—	
	p-value	< .001	—	
3-Conso_margarine	Pearson's r	-0.919	0.993	—
	p-value	< .001	< .001	—



Corrélation semi-partielle

Semipartial Correlation

		Tx_divorce_Maine	Conso_margarine
Tx_divorce_Maine	Pearson's r	—	0.446
	p-value	—	0.228
Conso_margarine	Pearson's r	0.384	—
	p-value	0.308	—

Note. controlling for 'Année'

Note. variation from the control variables is only removed from the variables in the columns

On efface la variabilité due à l'année dans le Tx_divorce et dans la consommation de margarine. Ainsi on rend compte que le lien qu'on avait trouvé entre ces deux variables était uniquement associé à la variable année (les deux p-values sont non significatives).