

Daniel Borcard  
Département de sciences biologiques  
Université de Montréal

## La corrélation

- *A quoi sert-elle?*

A mesurer le degré de **liaison** entre deux variables.

- *Comment fonctionne-t-elle?*

Elle mesure la **dispersion conjointe** de deux **variables centrées-réduites** (démonstration: Scherrer, p.651) et se calcule comme le quotient de la covariance entre deux variables ( $s_{xy}$ ) par le produit de leurs écarts-types ( $s_x$  et  $s_y$ ):

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- *Pourquoi "centrées-réduites"?*

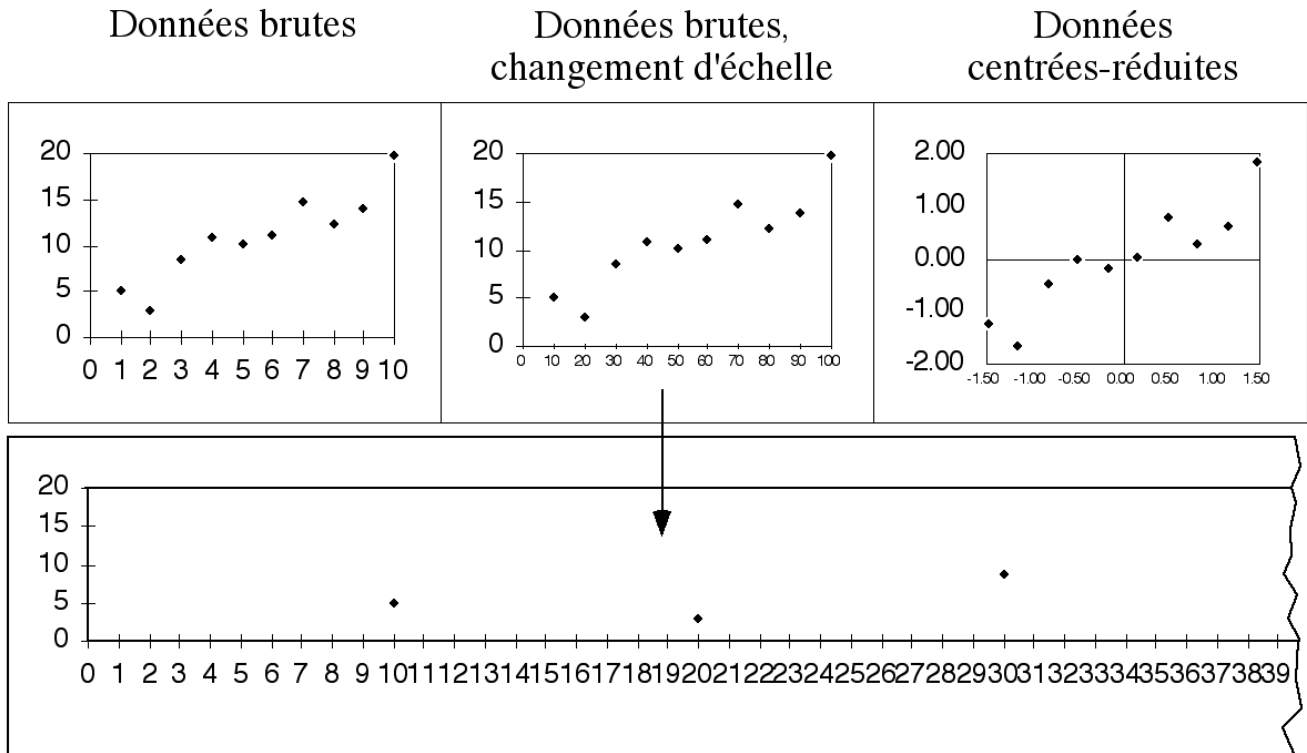
La mesure de dispersion conjointe est en fait la **covariance** (la manière dont deux variables varient ensemble, "co-varient").

Le **centrage** permet de comparer les dispersions par rapport à un **point de référence unique** (la moyenne, qui vaut zéro pour les deux variables après le centrage).

Si on ne réduit pas les variables, la covariance dépend du degré de dispersion de chacune des variables. Or, par exemple, cette dispersion dépend de choses aussi triviales que l'unité de mesure (si l'on change l'unité de mesure d'une des variables, la covariance change).

En **réduisant** les variables, on les exprime toutes deux en **unités d'écart-type**, et on leur donne à toutes deux une **variance égale à 1**. La variation de la mesure de covariance entre ces variables centrées-réduites ne dépendra donc plus que de la **liaison** entre elles.

Voir ci-dessous la démonstration graphique.



	Données brutes	Donn. brutes, changmt. éch.	Données centrées-réduites
Covar.	13.270	132.701	0.919
Corrél.	0.919	0.919	0.919

• *Ne pas oublier:*

- une des deux variables au moins doit être **aléatoire**.
- le coefficient de corrélation de Pearson ne mesure adéquatement que la liaison **linéaire** entre deux variables.
- la présence d'une corrélation n'implique **pas forcément** une relation de **causalité** entre les deux variables impliquées.

## Test de signification du $r$ de Pearson: rappel

- En général, on teste d'abord l'hypothèse que la corrélation entre deux variables est égale à zéro (dans la population statistique).
- On peut le faire par un test où la variable auxiliaire suivante suit une distribution  $t$  de Student à  $n-2$  degrés de liberté:

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

Règles de décision: Scherrer p. 653 pour  $z$  (voir plus bas).

Remarque: on peut aussi se servir de cette formule pour tester la pente d'une droite de régression linéaire. En effet, une corrélation linéaire significative entre deux variables  $x$  et  $y$  se traduit par une pente significative d'une droite de régression de  $y$  sur  $x$  (ou de  $x$  sur  $y$ ).

Conditions d'application du test du  $r$  de Pearson: il faut que:

- les deux variables soient quantitatives;
- la distribution conjointe des deux variables soit bi-normale;
- les observations soient indépendantes.

• Selon le contexte de l'étude, l'hypothèse contraire peut être:

- bilatérale ( $H_1: r_{xy} \neq 0$ )
- unilatérale à gauche ( $H_1: r_{xy} < 0$ )
- unilatérale à droite ( $H_1: r_{xy} > 0$ )

## La transformation de Fisher; intervalle de confiance du $r$ de Pearson et test de signification du $r$ si $H_0$ dit que $\rho =$ autre chose que zéro

Sokal et Rohlf (1981) : p. 583

- Le coefficient de corrélation de Pearson est borné de  $-1$  à  $+1$ . Sa distribution d'échantillonnage est complexe dès qu'on a affaire à une population dont  $\rho$  diffère de 0. C'est pourquoi on utilise une transformation pour calculer un intervalle de confiance et aussi dans le cas où on veut tester une hypothèse nulle où  $H_0: \rho = 0$ .
- Il arrive que l'hypothèse nulle habituelle ( $H_0: \rho = 0$ ) soit triviale et inintéressante. Par exemple, en morphométrie, on sait pertinemment qu'au cours de la croissance d'un animal la longueur du fémur et celle du premier métatarse sont corrélées. Si ces deux longueurs ne font que refléter la croissance de l'animal, la corrélation attendue vaut 1 ( $H_0: \rho = 1$ ). En revanche, si l'animal modifie sa manière de se déplacer en grandissant, sa morphologie peut changer et la corrélation risque de se modifier ( $H_1: \rho < 1$ ). D'autres exemples peuvent être trouvés, notamment en génétique.
- On ne peut pas tester l'hypothèse  $H_0$  d'une corrélation différente de zéro de la manière habituelle. En effet, lorsque la corrélation paramétrique ( $\rho$ ) (= la corrélation dans la population statistique) n'est pas nulle, la distribution de  $r$  n'est pas symétrique (puisque'elle est bornée à  $\pm 1$ ). Deux voies s'ouvrent donc: soit on invente un test qui tient compte de cette asymétrie, soit on restaure la symétrie de la distribution à l'aide d'une transformation.
- La deuxième solution est réalisable. La **transformation de Fisher**:

$$z = \frac{1}{2} \ln \frac{1+r}{1-r} = \text{tgh}^{-1}(r)$$

Scherrer (2007) p. 652

(arc-tangente hyperbolique!) restaure la **symétrie** de la distribution et **étire** l'étendue de variation de  $-1$  à  $+1$ .

• A partir de là, la logique est la suivante: toutes les opérations qu'on veut réaliser se font sur les corrélations **transformées en z**, puis, s'il y a lieu, on revient aux vraies valeurs par une transformation inverse, c'est-à-dire:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} = \operatorname{tgh}(z)$$

• *Test d'hypothèse*: en faisant subir la transformation de Fisher à la fois à la corrélation de l'échantillon  $r$  et à celle de l'hypothèse nulle [ $\xi_0 = \operatorname{tgh}^{-1}(r_0)$ ], on peut construire une statistique-test appelée  $t_{(\infty)}$  par référence au  $t$  de Student, mais qui se comporte à peu près comme une **distribution normale centrée-réduite**:

$$t_{(\infty)} = (z - \xi_0) \sqrt{n - 3}$$

• Comme d'habitude, la nature de l'hypothèse (uni- ou bilatérale) détermine les zones d'acceptation ou de rejet de l'hypothèse nulle.

• *Intervalle de confiance* (Scherrer p. 652 sq.): l'exemple ci-dessous le définit pour  $\alpha = 0.05$  (Scherrer:  $z_{(1-\alpha/2)}$ : p. 750)

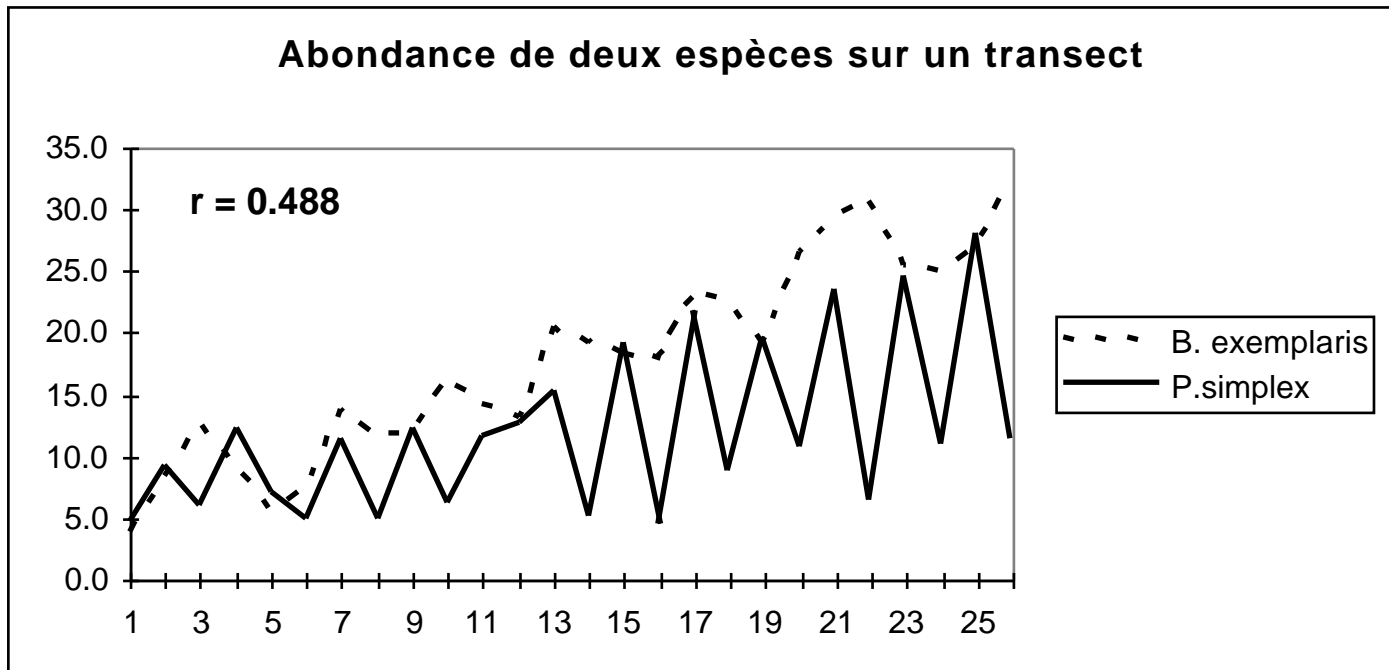
$$\xi_1 = z - \frac{t_{(0.05; \infty)}}{\sqrt{n-3}} = z - \frac{1.959964}{\sqrt{n-3}} \quad \xi_2 = z + \frac{t_{(0.05; \infty)}}{\sqrt{n-3}} = z + \frac{1.959964}{\sqrt{n-3}}$$

• Lorsqu'on a trouvé  $\xi_1$  et  $\xi_2$ , il faut encore les **retransformer** pour obtenir l'intervalle de confiance dans l'échelle du  $r$  d'origine:

$$r_1 = \operatorname{tgh}(\xi_1) \quad \text{et} \quad r_2 = \operatorname{tgh}(\xi_2)$$

## Exemple (fictif):

On a fait des prélèvements de sol sur un transect, et compté les abondances de deux espèces d'insectes:



Nombre de prélèvements: 26

### *Bidonia exemplaris:*

Somme 463.8

Moyenne 17.8

Ecart-type 8.1

### *Predator simplex:*

Somme 310.0

Moyenne 11.9

Ecart-type 6.9

*r* de Pearson **0.4878**

*t* de Student 2.7372

---

$r$ de Pearson	<b>0.4878</b>
$t$ de Student	2.7372
degrés de liberté	24

### Hypothèse nulle $H_0: \rho = 0$

#### *Hypothèse contraire $H_1: \rho \neq 0$ (bilatérale)*

$t$ crit. pour $\alpha = 0.05$ :	2.064
$r$ crit. pour $\alpha = 0.05$ :	0.388
=> corrélation significative au seuil 0.05	

$t$ crit. pour $\alpha = 0.01$ :	2.797
$r$ crit. pour $\alpha = 0.01$ :	0.496
=> corrélation non significative au seuil 0.01	

#### *Hypothèse contraire $H_1: \rho > 0$ (unilatérale)*

$t$ crit. pour $\alpha = 0.05$ :	1.711
$r$ crit. pour $\alpha = 0.05$ :	0.330
=> corrélation significative au seuil 0.05	

$t$ crit. pour $\alpha = 0.01$ :	2.492
$r$ crit. pour $\alpha = 0.01$ :	0.453
=> corrélation significative au seuil 0.01	

Dans le contexte de cet exemple, l'hypothèse unilatérale pourrait être celle d'une corrélation positive entre prédateur et proie (les prédateurs tendent à se trouver là où ils rencontrent le plus de proies). Cette hypothèse unilatérale, mieux cernée *a priori*, permet d'augmenter la puissance du test.

$$r \text{ de Pearson} = \mathbf{0.4878} \quad t \text{ de Student} = 2.7372$$

### Hypothèse nulle $H_0: \rho = +0.5$

Imaginons que la population de prédateurs est en fait constituée de deux sous-espèces distribuées en mosaïque sur le transect. Une des sous-espèces se nourrit de *B. exemplaris*, l'autre pas. Si le pas de l'échantillonnage correspond au diamètre des taches de distribution de chaque sous-espèce, alors un prélèvement sur deux seulement touchera la sous-espèce prédatrice de *B. exemplaris*. On pourrait donc s'attendre *a priori* à une corrélation de 0.5 entre les deux variables.

Pour tester cette hypothèse, il faut transformer le  $r$  de l'échantillon en  $z$  et le  $\rho_0$  de la population en  $\xi_0$  (transformation de Fisher):

$$r = 0.4878 \quad \text{donc } z = 0.5331$$

$$\rho_0 = 0.5000 \quad \text{donc } \xi_0 = 0.5493$$

$$\text{On calcule ensuite le } t_{(n-3)} : t_{(n-3)} = (z - \xi_0) \sqrt{n-3} = -0.0776$$

### Hypothèse contraire $H_1: \rho \neq 0.5$ (bilatérale)

On compare la valeur de  $t_{(n-3)}$  ci-dessus avec l'aire de la courbe normale centrée-réduite pour  $\alpha = 0.05$ . L'hypothèse étant bilatérale, on cherchera l'aire pour  $\alpha/2 = 0.025$ ; on cherche donc pour une probabilité cumulée  $1 - \alpha/2 = 0.975$ .

La valeur peut être trouvée dans une table de l'aire de la courbe normale, et aussi à la dernière ligne de la table du  $t$  de Student:  $z$  (ou  $t_{(n-3)}$ ) critique = 1.959964  $\approx$  1.96

Comme le  $t_{(n-3)}$  observé est plus petit **en valeur absolue** (test bilatéral ici) que le  $t_{(n-3)}$  critique ( $|-0.0776| < 1.96$ ), on ne peut pas rejeter l'hypothèse nulle  $\rho = +0.5$

*Intervalle de confiance:*

Pour cet exemple, à 95%,  $\zeta_1$  et  $\zeta_2$  valent respectivement 0.1244 et 0.9418. En retransformant ces valeurs (tangente hyperbolique) pour les ramener dans l'échelle de  $r$ , on obtient:

$$\zeta_1 = 0.1238$$

$$\zeta_2 = 0.7360$$

Il est intéressant de noter que la valeur de  $r$  (0.4878) n'est pas située au centre de l'intervalle de confiance, une conséquence logique de l'asymétrie de la distribution de  $r$ .

**Remarque:** la transformation de Fisher est valable (c'est à dire qu'elle fournit un  $z$  distribué approximativement normalement) pour autant que  $n$  soit plus grand que 50 ou, à la rigueur, 25. Lorsque  $n$  est petit (entre 10 et 25), Hotelling propose une correction à la formule de Fisher, dont une variante est donnée par Sokal & Rohlf (1995):

$$z^* = z - \frac{3z + r}{4(n-1)} \text{ et, pour l'hypothèse nulle: } \zeta_0^* = \zeta_0 - \frac{3\zeta_0 + \rho_0}{4n}$$

qui se teste par 
$$t_{(1-\alpha)} = z^* - \zeta_0^* \sqrt{n-1}$$

Les intervalles de confiance s'estiment par:

$$\zeta_1^* = z^* - \frac{t_{(0.05, n-1)}}{\sqrt{n-1}} \quad \text{et} \quad \zeta_2^* = z^* + \frac{t_{(0.05, n-1)}}{\sqrt{n-1}}$$