La corrélation linéaire

Tests visant à mettre en évidence une corrélation entre deux variables quantitatives

M. L. Delignette-Muller
VetAgro Sup

5 décembre 2019

Objectifs pédagogiques

- Connaître la définition et les conditions d'utilisation du coefficient de corrélation linéaire (de Pearson).
- Connaître le principe et les conditions d'utilisation des tests paramétriques et non paramétriques de corrélation ainsi que les limites de ces tests.
- Savoir identifier dans la pratique les cas où l'utilisation d'un test de corrélation n'est pas approprié et dans les autres cas savoir faire le choix entre le test paramétrique et le test non paramétrique, et réaliser ces tests.**
- Savoir interpréter les conclusions de ces tests.

** *savoir faire évalué uniquement en S6 après entraînement en TD*

Plan

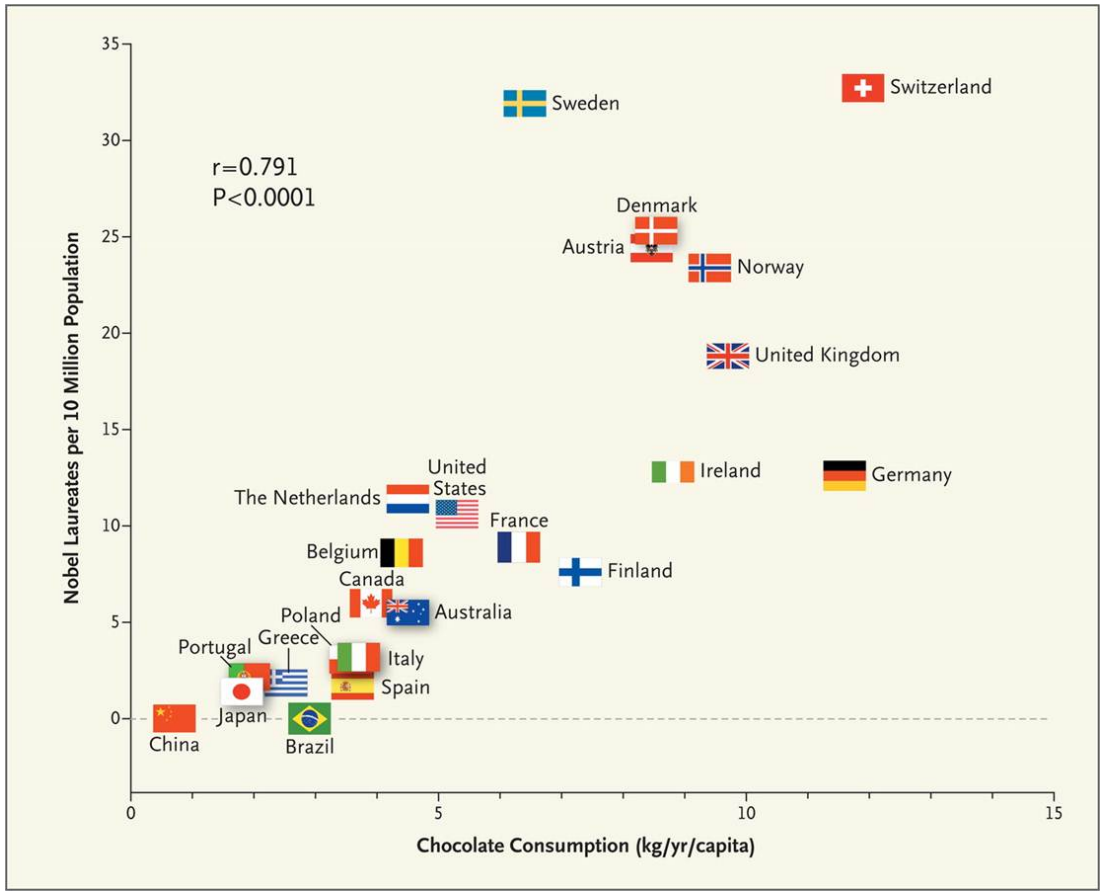
1 Le test de corrélation linéaire de Pearson

2 Le test de corrélation de rangs de Spearman

3 Les limites des tests de corrélation

Exemple tiré de la littérature

Figure extraite de Messerli (2012), Chocolate Consumption, Cognitive Function, and Nobel Laureates, *the New England Journal of Medicine*



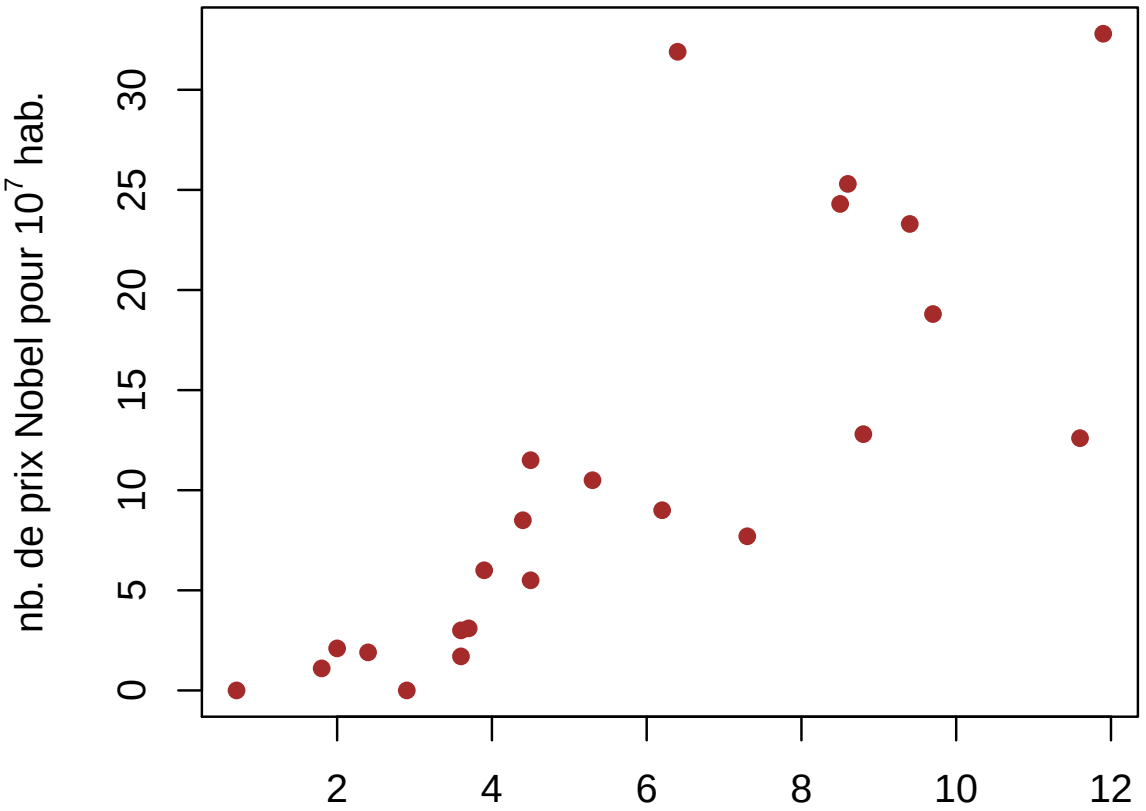
Les données observées sur un échantillon de 23 pays

Données brutes (telles que saisies informatiquement) :

Country	Nobels	Chocolate
Australia	5.5	4.5
Austria	24.3	8.5
Belgium	8.5	4.4
Brazil	0.0	2.9
Canada	6.0	3.9
China	0.0	0.7
Denmark	25.3	8.6
Findland	7.7	7.3
France	9.0	6.2
Germany	12.6	11.6
Greece	1.9	2.4
Ireland	12.8	8.8
Italy	3.1	3.7
Japan	1.1	1.8
Norway	23.3	9.4
Poland	3.0	3.6
Portugal	2.1	2.0

...

La consommation de chocolat et le nombre de prix Nobel sont-ils corrélés ?



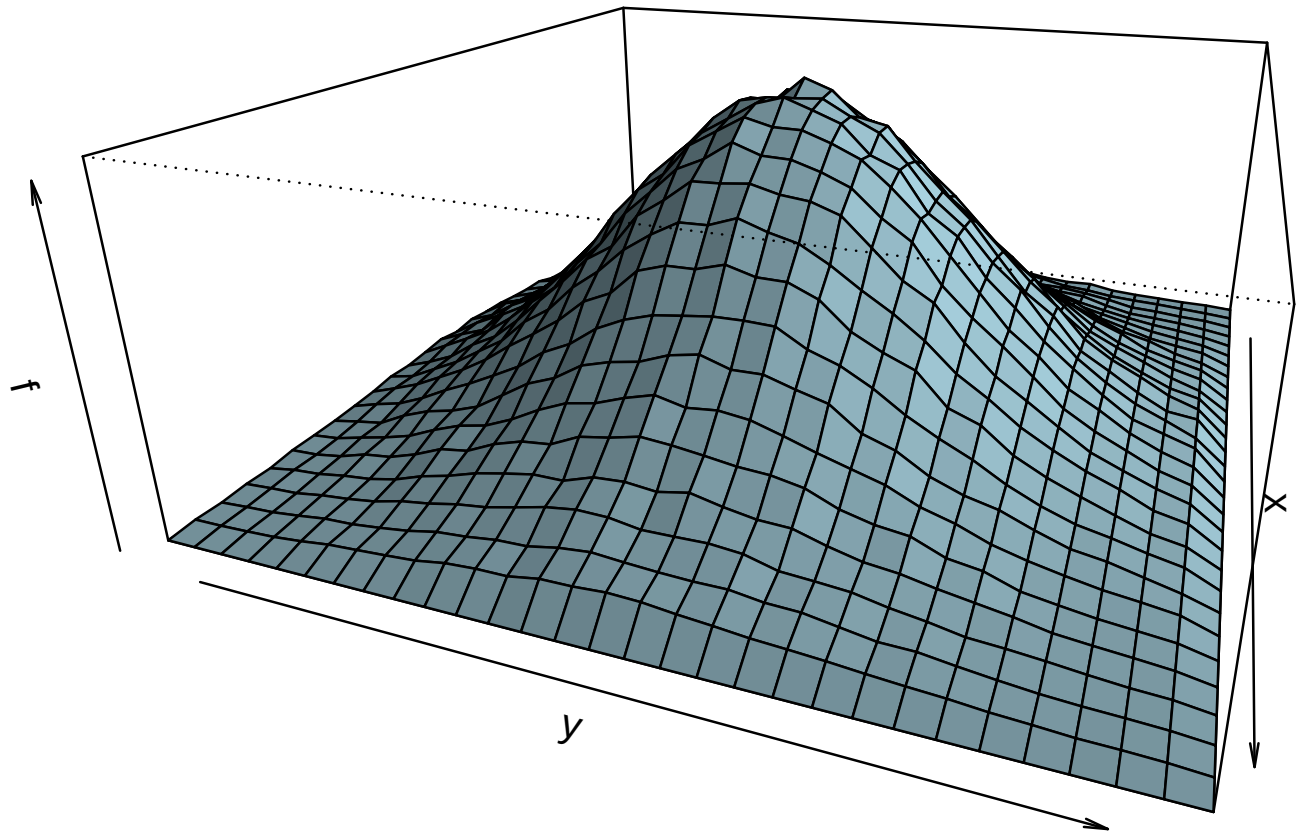
Le coefficient de corrélation linéaire de Pearson

- **Conditions d'utilisation**
  x et y 2 variables aléatoires observées sur un échantillon aléatoire simple et distribuées suivant une loi normale bivariée (ce qui induit un nuage de points à peu près elliptique)
- **Calcul et propriétés**

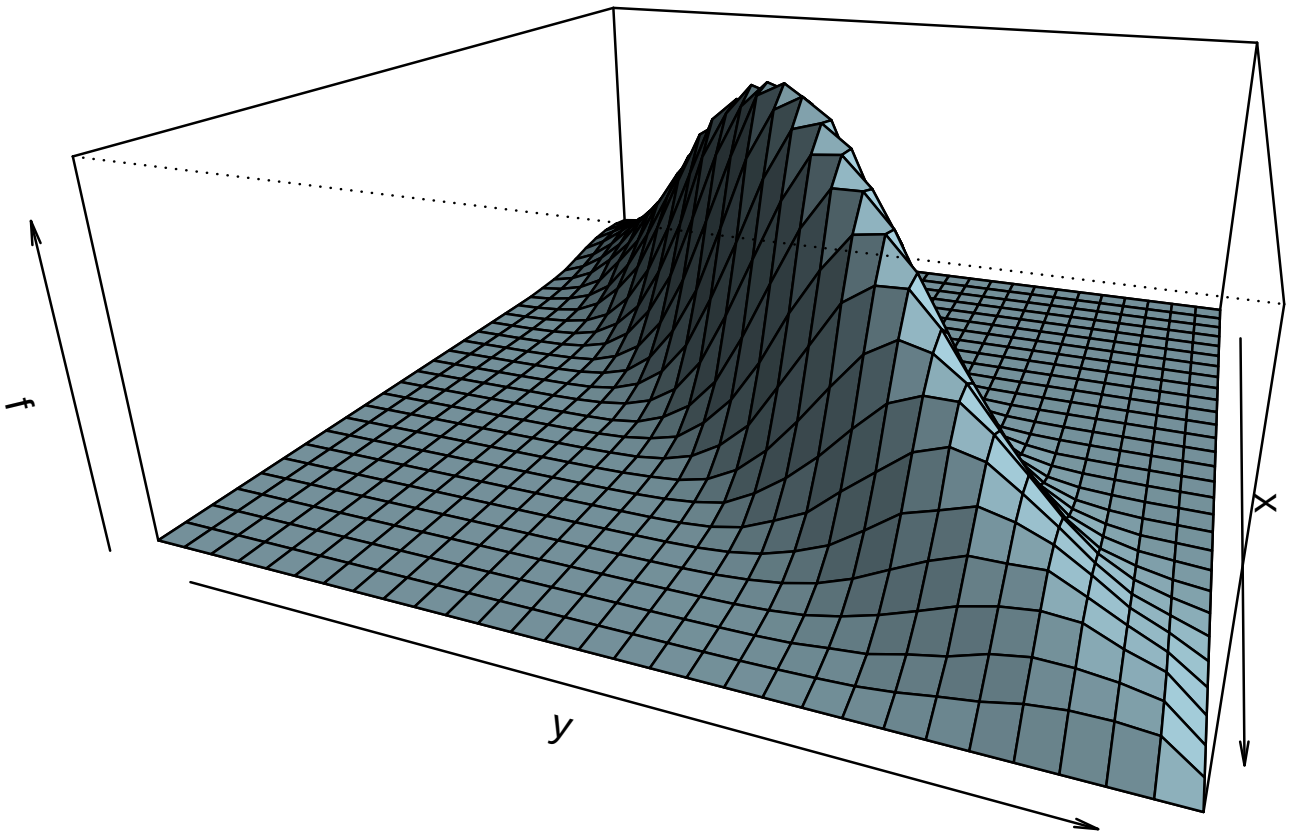
$$r = \frac{Cov(x,y)}{\sqrt{V(x)V(y)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\overline{x})^2 \times \frac{1}{N}\sum_{i=1}^{N}(y_i-\overline{y})^2}}$$

r est un indicateur unidirectionnel de l'allongement du nuage de points : $-1 \leq r \leq 1$ et plus les points sont alignés et plus $|r|$ est proche de 1.

Visualisation d'une loi normale bivariée sans corrélation entre x et y



Visualisation d'une loi normale bivariée avec une forte corrélation entre x et y : $r = 0.9$



Le test de corrélation linéaire de Pearson

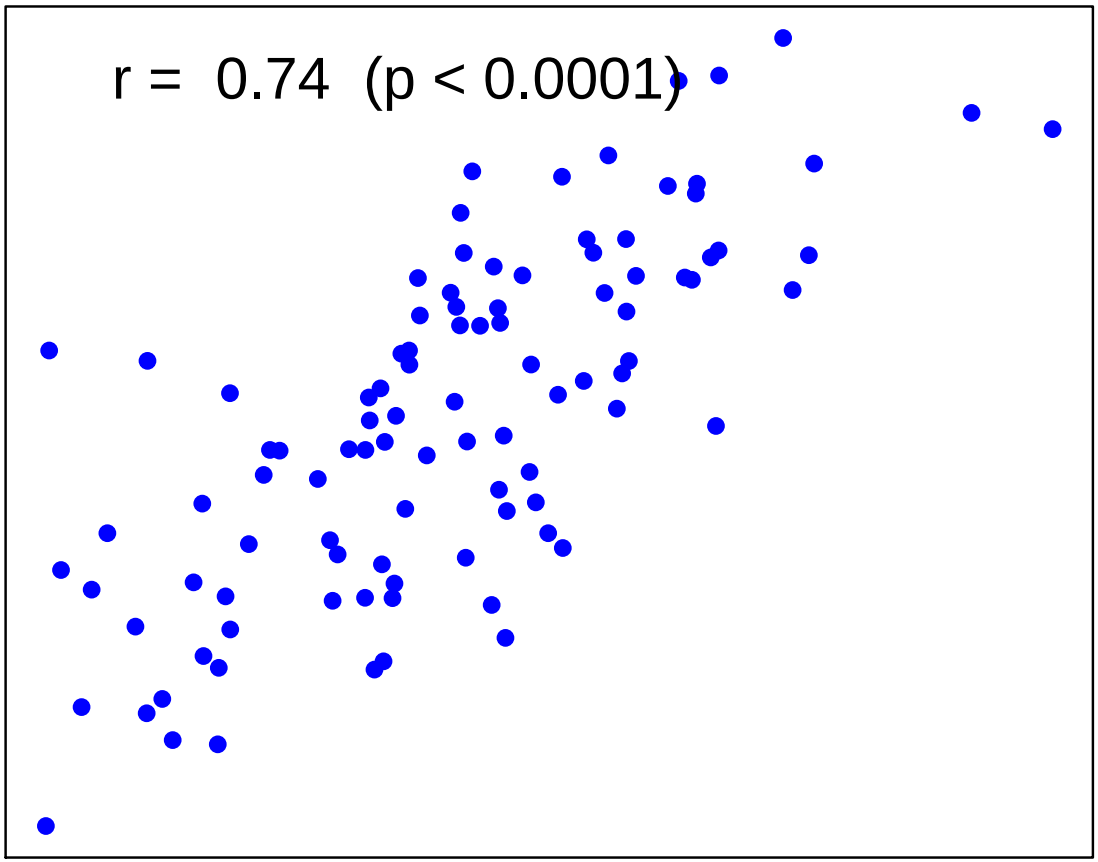
Test de la nullité du coefficient de corrélation linéaire de Pearson
H_0 : "$r = 0$"
c'est-à-dire absence de corrélation linéaire entre les 2 variables observées

Test basé sur une **statistique de Student**

$t = \sqrt{\frac{r^2(N-2)}{1-r^2}} \sim T(N-2)$ sous H_0
avec N le nombre de points observés

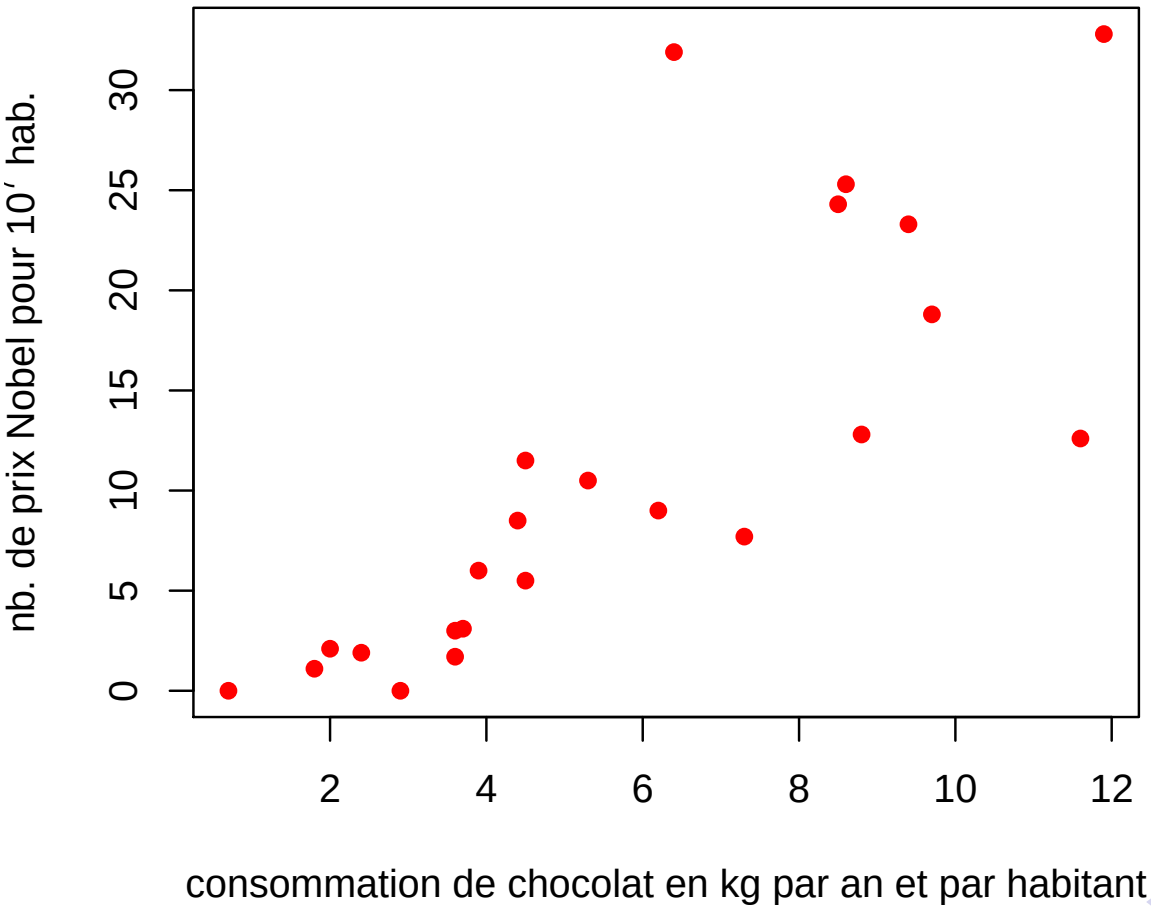
Test très peu robuste, d'où la nécessité de bien vérifier les conditions d'utilisation en examinant le nuage de points.

Coefficient de corrélation et valeur de p associée sur un nuage de points à peu près elliptique



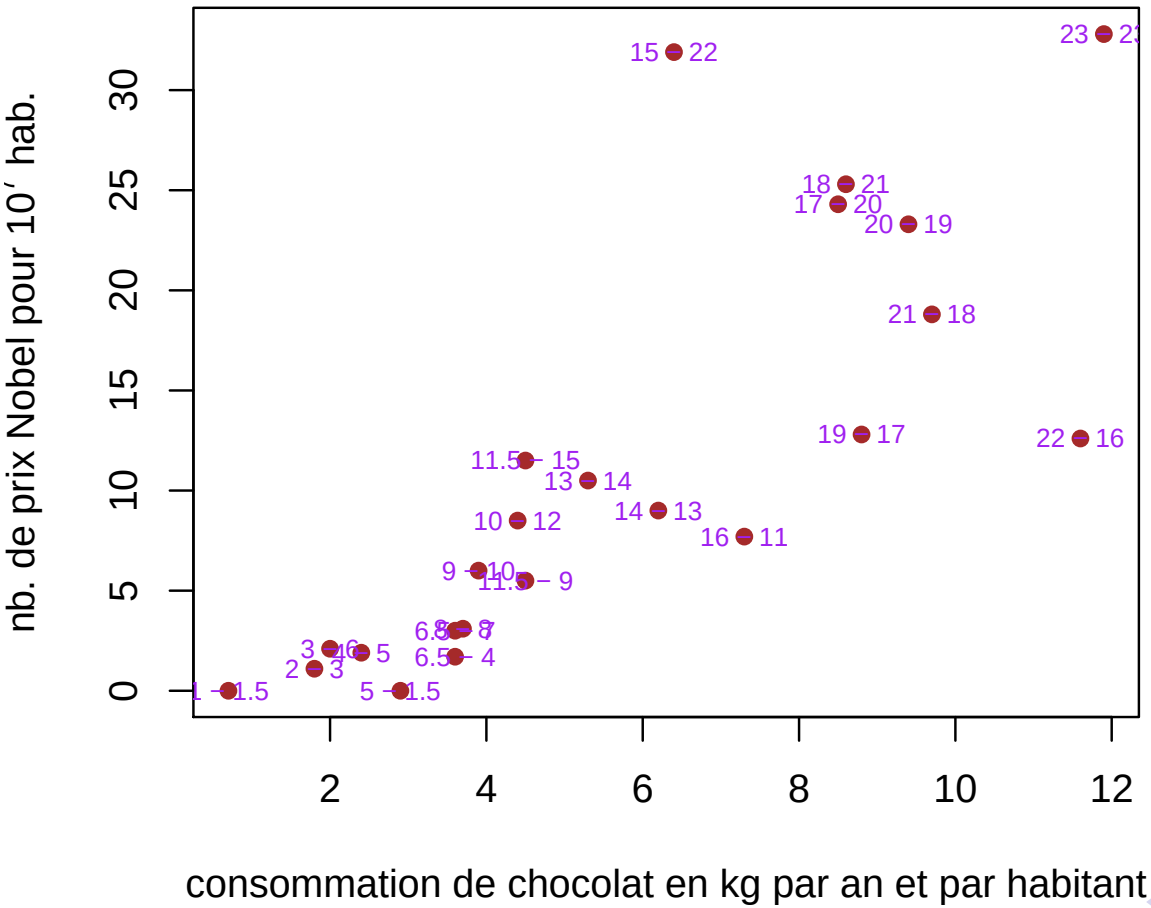
Revenons à notre exemple

Tendance linéaire mais nuage de points non elliptique : on préférera utiliser une méthode non paramétrique



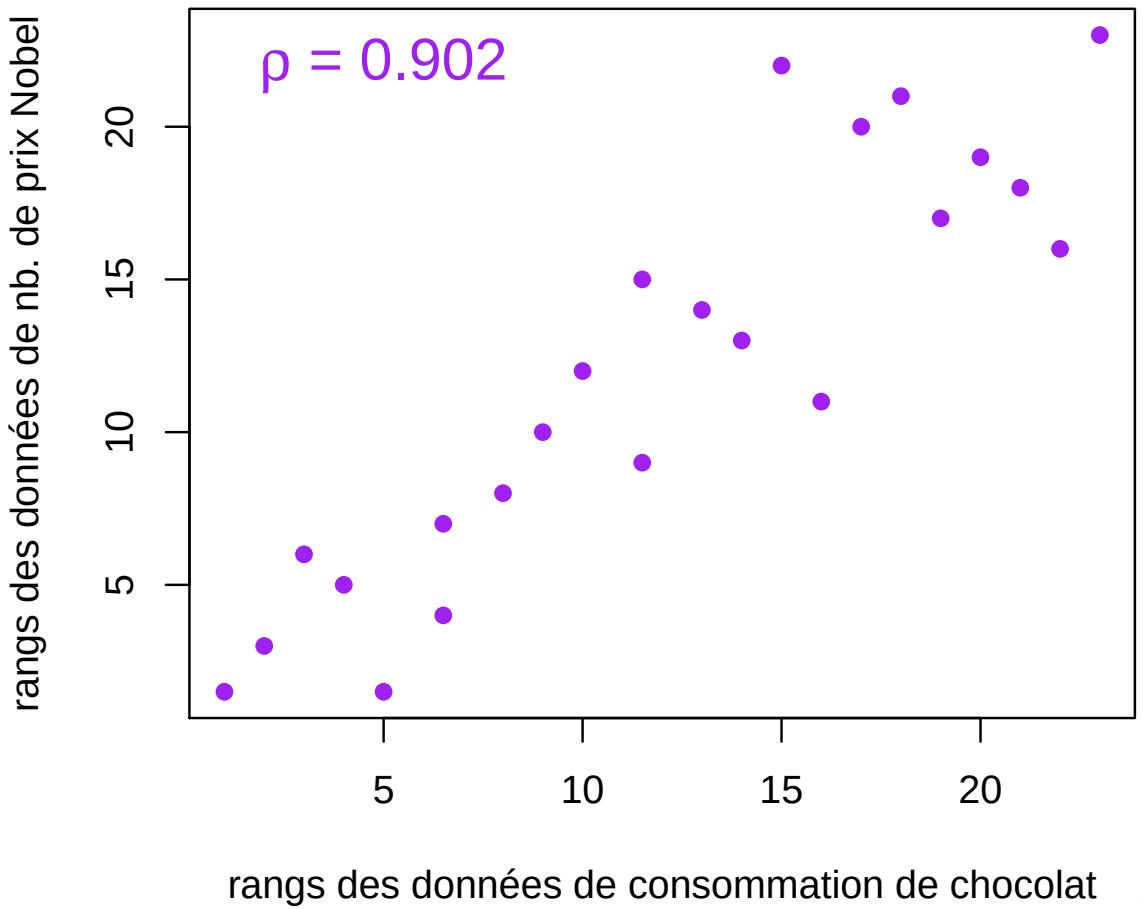
Coefficient de corrélation de rangs de Spearman (1)

On classe les valeurs de x d'un côté, et celle de y de l'autre et on associe à chaque point du nuage le rang de x et le rang de y



Coefficient de corrélation de rangs de Spearman (2)

On calcule le coefficient de corrélation linéaire sur les rangs des x et les rangs des y



Test de corrélation de rangs de Spearman sur l'exemple

$\rho = 0.902$ associé à $p < 0.0001$
On observe une corrélation significative entre la consommation individuelle de chocolat dans les états et le nombre de prix Nobel pour 10 millions d'habitants.
ATTENTION ! On n'en déduira bien entendu pas de lien de causalité.
Une corrélation entre 2 variables observées n'implique pas forcément un lien de causalité

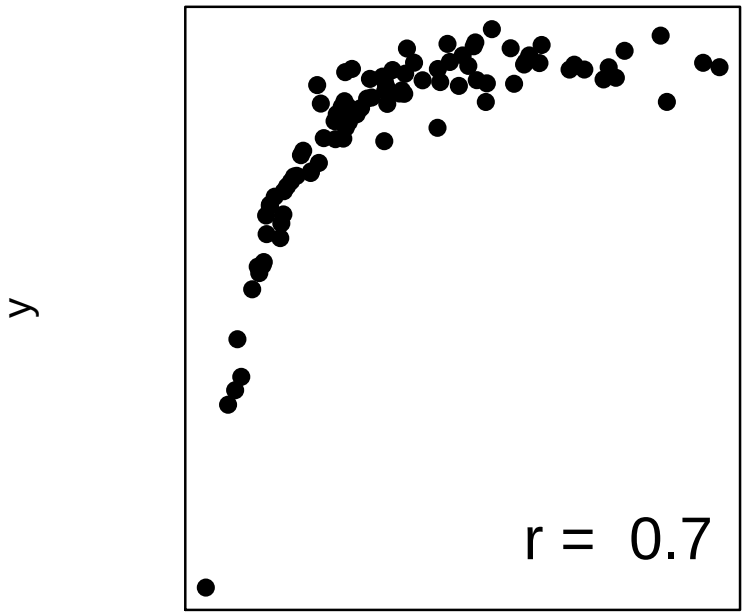
Grande prudence nécessaire avant utilisation d'un test de corrélation

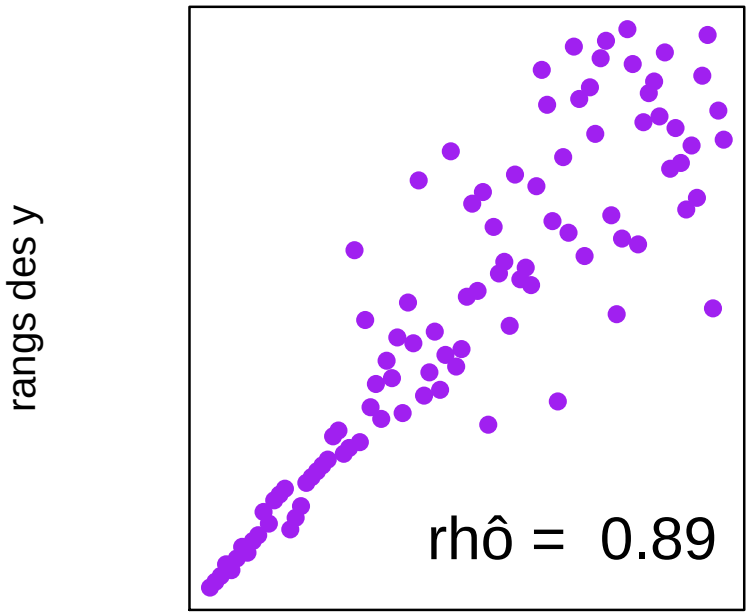
- Le test de corrélation de Pearson n'est pas adapté en cas de corrélation non linéaire
- Le test de corrélation de Pearson n'est pas du tout robuste (très influencé notamment par les valeurs extrêmes)
- Les tests de corrélation (Pearson et Spearman) ne sont pas adaptés en cas de corrélation non monotone et plus généralement de nuage de points non elliptique
- Les tests de corrélation (Pearson et Spearman) ne sont pas adaptés en cas de nuage de points formé de sous-nuages (sous-groupes se distinguant)

On ne devrait jamais reporter une valeur de r ou de ρ non assortie du nuage de points

Nuage de points avec corrélation monotone mais non linéaire

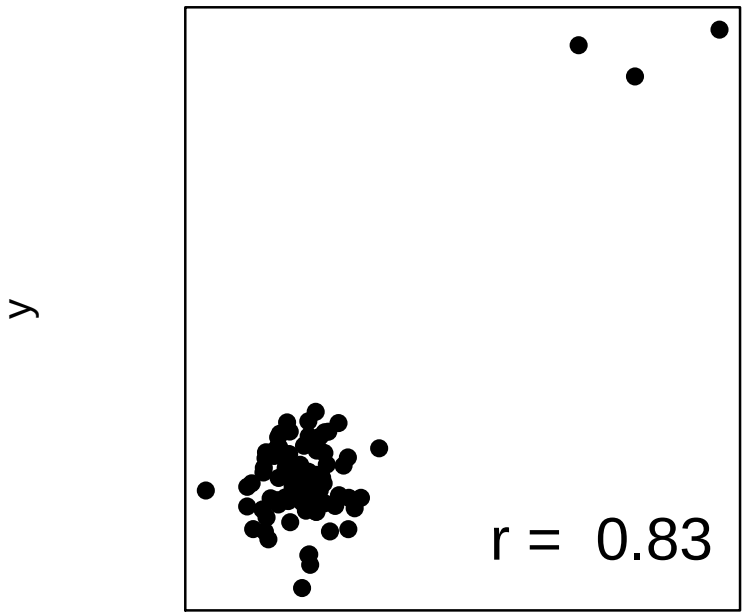
Ex. de nuage de points où ρ est plus adapté que r

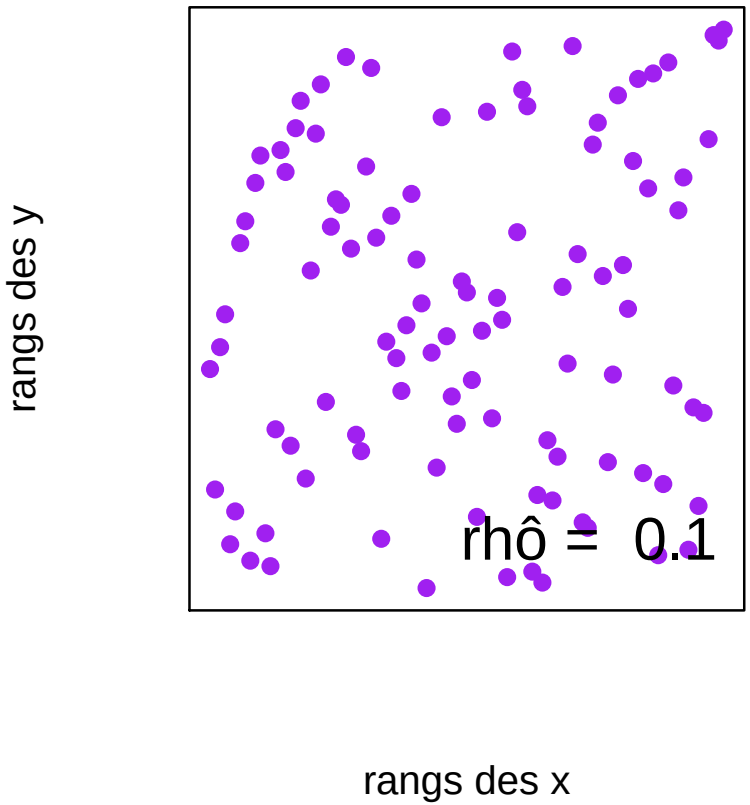




Nuage de points avec valeurs extrêmes

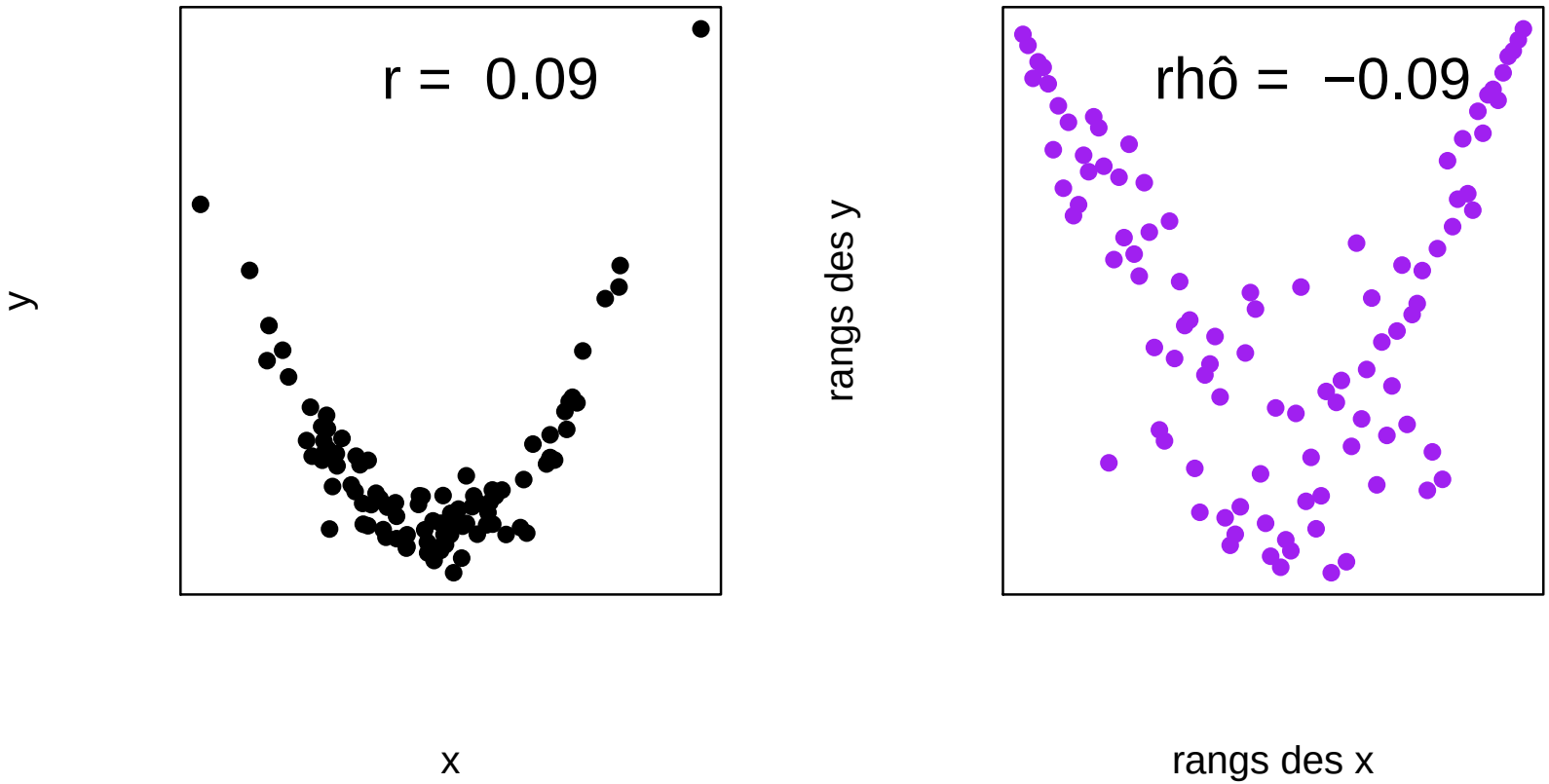
Ex. de nuage de points où ρ peut être calculé mais pas r qui serait artificiellement trop élevé





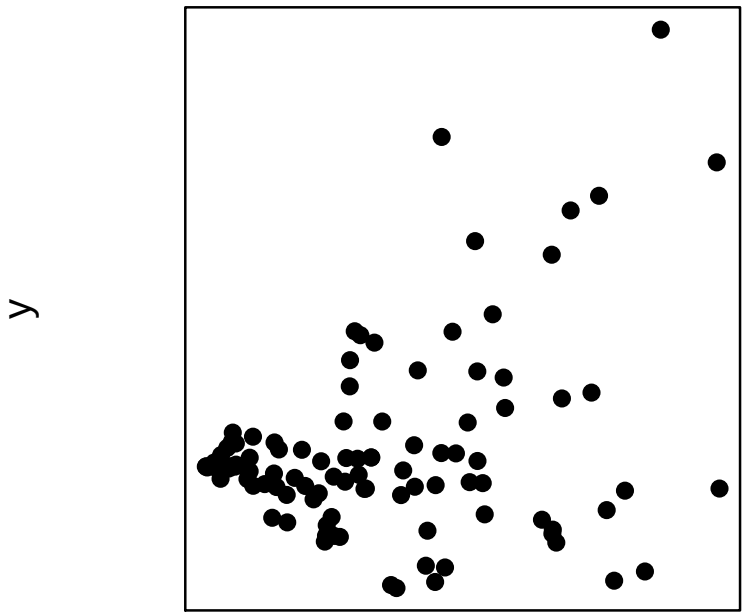
Nuage de points avec corrélation non monotone

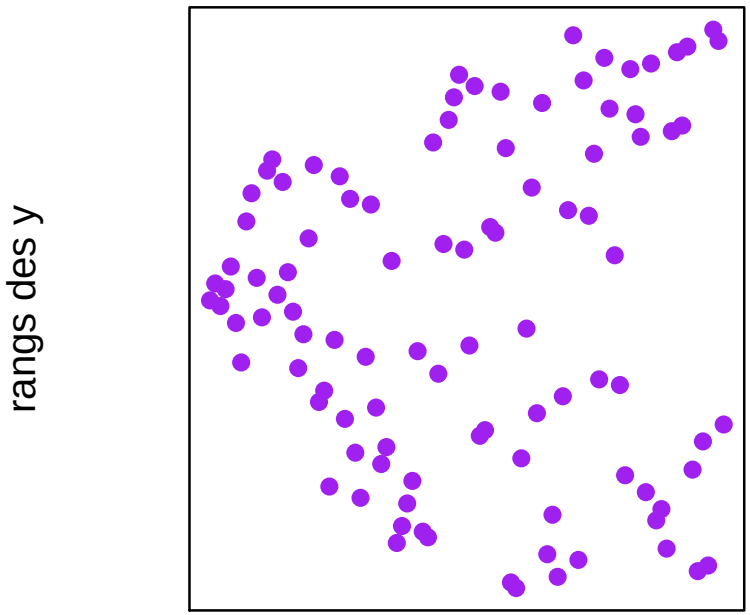
Ex. de nuage de points sur lequel il ne faut calculer ni r ni ρ (proche de 0)



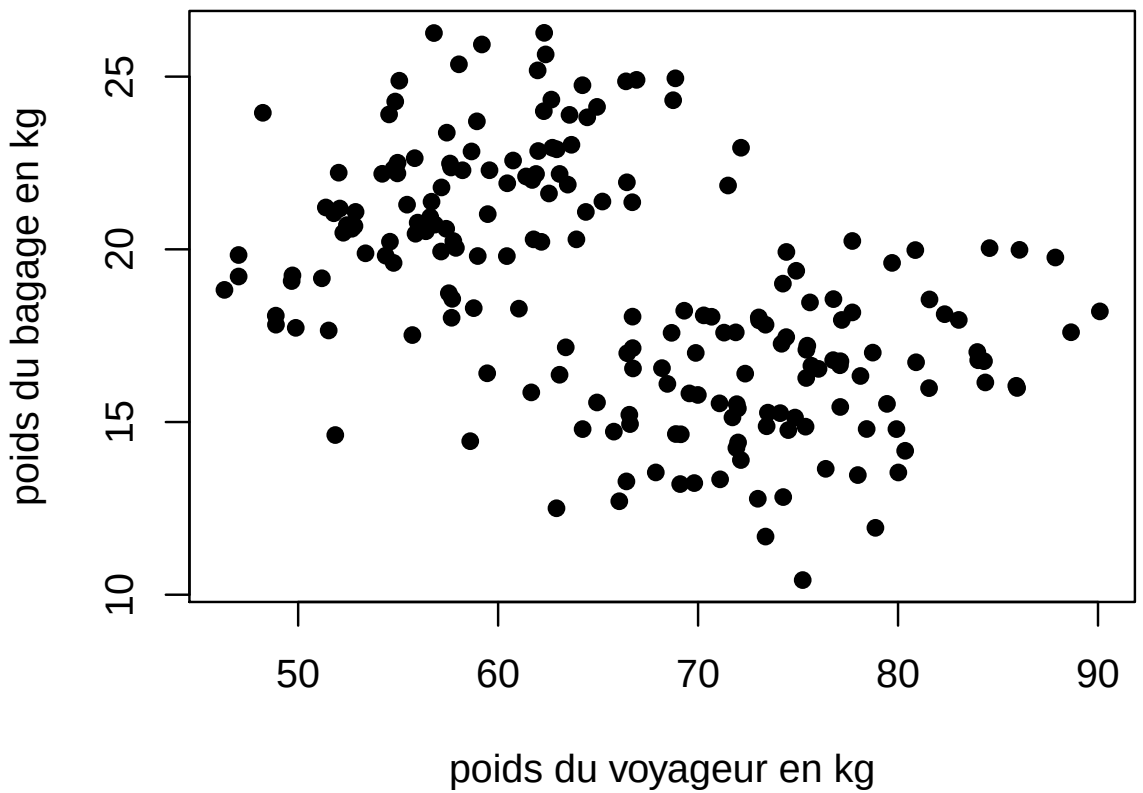
Autre nuage de points non elliptique

Ici la corrélation entre y et x n'est pas directe : plus x est grand, plus y est variable (pas décrit ni par r ni par ρ)





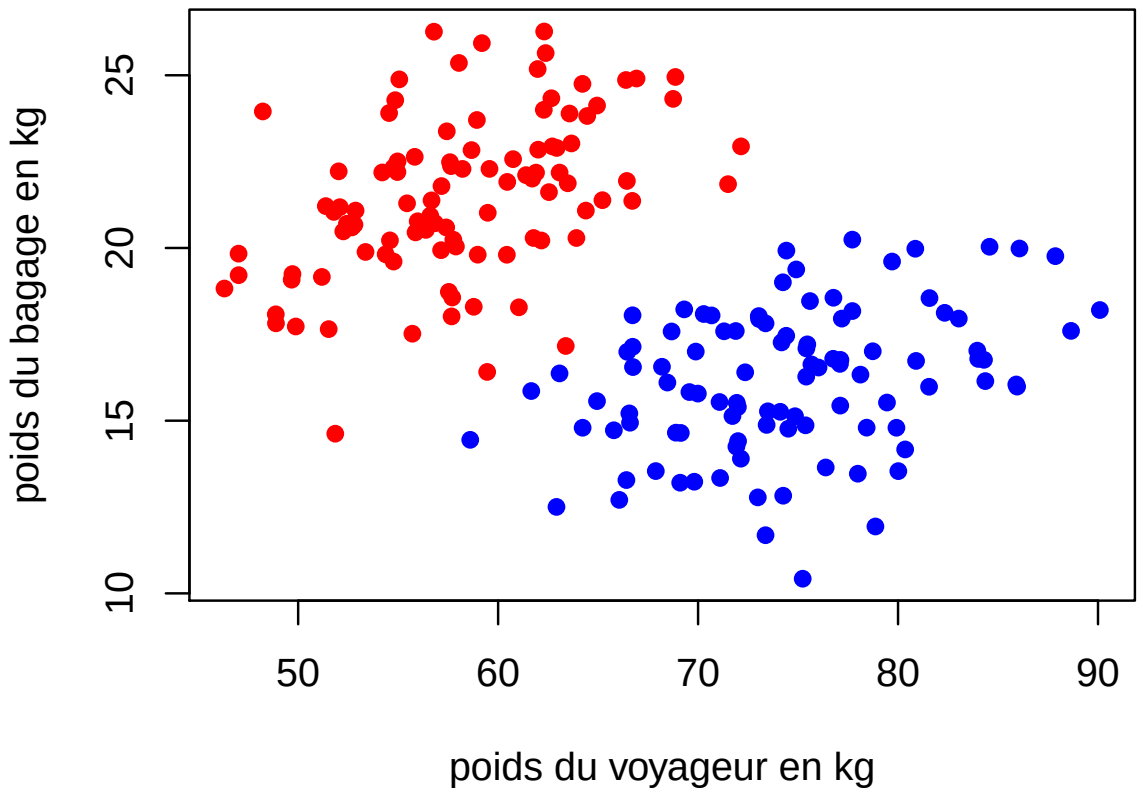
Un dernier exemple fictif mais réaliste



$r = -0.47$ ($p < 0.0001$), $\rho = -0.52$ ($p < 0.0001$)

Mais peut-on pour autant conclure à une corrélation négative ?

Deux sous-groupes distincts



$r = 0.48$ $(p < 0.0001)$ pour les femmes

$r = 0.41$ $(p < 0.0001)$ pour les hommes

En conclusion

Ne jamais calculer et/ou interpréter un coefficient de corrélation sans avoir vu le diagramme de dispersion correspondant !

Ouverture sur le concept de causalité

Corrélation n'implique pas forcément causalité.

Une corrélation peut être :

- due à un facteur de causalité commun (ex. : corrélation entre vente de glaces et noyades)
- due à une causalité dans le sens opposé à celui présenté (ex. : myopie des enfants et veilleuse)
- complètement fortuite (cf. site qui montre des corrélations fortuites entre deux variables suivies au cours du temps : http://www.tylervigen.com/spurious-correlations)

En réponse à l'étude de Messerli (chocolat / prix Nobel), des chercheurs ont montré une corrélation entre la consommation de chocolat et le nombre de tueurs en séries que le pays engendre.

Un exemple de corrélation fortuite ($r = 0.992$)

Extrait de

`http://www.tylervigen.com/spurious-correlations`