Corrélations de Pearson et de Spearman

Notion de corrélation

La chose la plus importante à se rappeler en [bio]statistiques au niveau des corrélations, c'est que **corrélation n'est pas causalité**, soit, en d'autres termes, ce n'est pas parce que x et y sont liés que y est la cause de x.

La corrélation au sens de Pearson calcule la force de la relation **linéaire** et linéaire seulement entre x et y alors que la corrélation au sens de Spearman calcule la force de la liaison **monotone** (au sens des fonctions monotones, donc croissantes ou décroissantes). Il est donc correct de dire que la corrélation de Pearson est une corrélation paramétrique et celle de Spearman une corrélation non paramétrique parce que la corrélation de Spearman est basée sur les rangs des valeurs. De façon plus précise, la corrélation de Spearman de x et y est exactement la corrélation de Pearson appliquée aux rangs de x et de y.

Exemple numérique

Pour bien illuster la différence entre ces deux types de corrélation, il suffit de regarder la corrélation entre x et $y = e^{x^2}$. Il s'agit d'une corrélation "exponentielle du carré" donc non linéaire mais strictement croissante donc monotone. Il n'est donc pas étonnant que pour les valeurs $x = 1, 2, 3..10$, on trouve une coefficient de corrélation de Pearson d'environ 0.522 et pour Spearman 1.000 exactement, avec comme p-values respectives 0.1215 et 0.0000.

La courbe suivante qui n'est pas facile à lire (regarder les valeurs sur l'axe Y) résume ce phénomène :