ANNEXE 6
Note théorique sur la corrélation

Le coefficient de corrélation

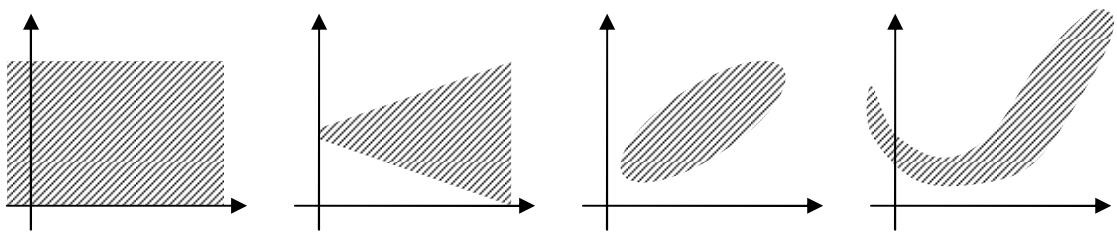
On dispose de deux ensembles de valeurs conjointes (des concentrations concomitantes) : $(x_i)_{1\le i\le n}$ et $(y_i)_{1\le i\le n}$, que l'on considère comme les observations de deux variables aléatoires *X* et *Y* dont les réalisations x_i et y_i sont les concentrations mesurées aux mêmes instants *i*.

Etudier la corrélation entre deux variables aléatoires (ou statistiques), c'est étudier l'intensité de la liaison qui peut exister entre ces variables. La liaison recherchée est une relation **affine**.

1. Etude graphique de la corrélation

Afin d'examiner s'il existe une liaison entre *X* et *Y* on représente chaque observation *i* comme un point de coordonnées (x_i,y_i) dans un repère cartésien. La forme du nuage de points ainsi tracé est fondamentale pour la suite : ainsi la figure 1 montre :

a) une absence de liaison ;
b) une absence de liaison en moyenne mais pas en dispersion ;
c) une corrélation linéaire positive ;
d) une corrélation non linéaire.



(a) (b) (c) (d)

Figure 1

On dit qu'il y a corrélation s'il y a dépendance en moyenne : à $X=x$ fixé, la moyenne $\overline{Y}$ est fonction de *x*. Si cette liaison est approximativement linéaire on se trouve dans le cas de la corrélation linéaire.

2. Le coefficient de corrélation linéaire

Ce coefficient mesure exclusivement le caractère plus ou moins **linéaire** du nuage de points, autrement dit la qualité de la relation linéaire ou le degré de dépendance linéaire entre les deux variables.

A. Définition

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

où s_x et s_y sont les écarts-types de *x* et *y* :

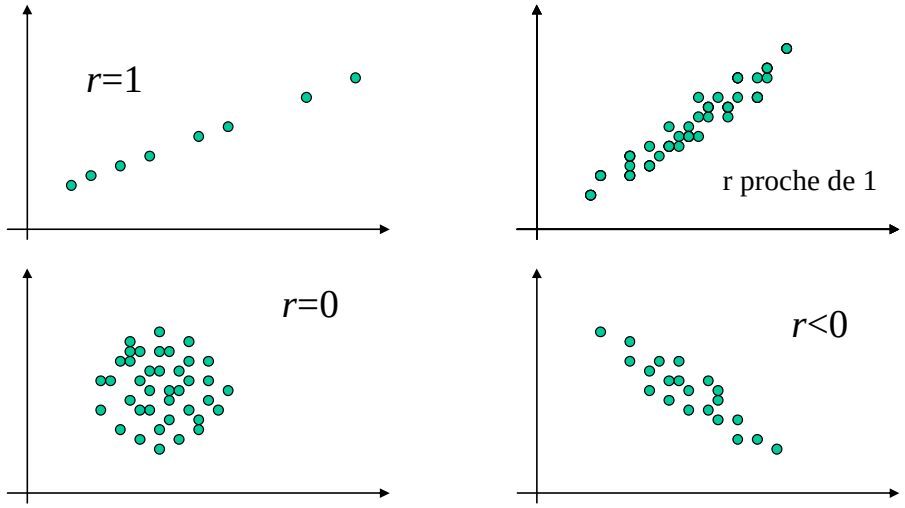
$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \qquad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

Remarque : le numérateur $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ est la covariance observée.

Propriétés :

- $-1 \le r \le 1$ (ou bien $|r| \le 1$)
- r est un indicateur sans dimension ; il n'est pas sensible aux unités de chacune des variables. Ainsi par exemple, le coefficient de corrélation linéaire entre 2 mesures sera identique que l'une des mesures soit exprimée en ppb ou en $\mu g/m^3$.
- $|r| = 1$ est équivalent à l'existence d'une relation linéaire exacte (tous les points sont alignés sur une droite) : $ax_i + by_i + c = 0 \quad \forall i$.
- X et Y indépendantes $\Rightarrow r = 0$, mais la réciproque n'est pas vraie en général : **la non corrélation n'est pas nécessairement l'indépendance**, alors que l'indépendance entraîne forcément une non corrélation.
- Notons pour finir que la corrélation n'est pas transitive : X très corrélée avec Y, Y très corrélée avec Z, n'implique nullement que X soit corrélée avec Z.

Exemples :



B. Du bon usage du coefficient *r*

Attention, il est toujours possible de calculer un coefficient de corrélation (sauf cas très particulier) mais un tel coefficient de corrélation n'arrive pas toujours à rendre compte de la relation qui existe en réalité entre les variables étudiées. En effet, il suppose que l'on essaye de juger de l'existence d'une relation linéaire entre nos variables. Il n'est donc pas adapté pour juger de corrélations qui ne seraient pas linéaires et non linéarisables. Il perd également de son intérêt lorsque les données étudiées sont très hétérogènes puisqu'il représente une relation moyenne et que l'on sait que la moyenne n'a pas toujours un sens, notamment si la distribution des données est multi modale ou très asymétrique (c'est typiquement le cas pour la répartition du SO_2 par exemple).

Le coefficient *r* ne mesure que le caractère linéaire d'une liaison et son usage doit être réservé à des nuages où les points sont répartis de part et d'autre d'une tendance linéaire (figure 1.c).

La figure 2 montre les risques d'un usage inconsidéré du coefficient de corrélation linéaire *r*. On notera en particulier que *r* est très sensible à la présence de valeurs extrêmes et/ou aberrantes (valeurs très éloignées de la majorité des autres, pouvant être considérées comme des exceptions), et n'est donc pas « robuste ».

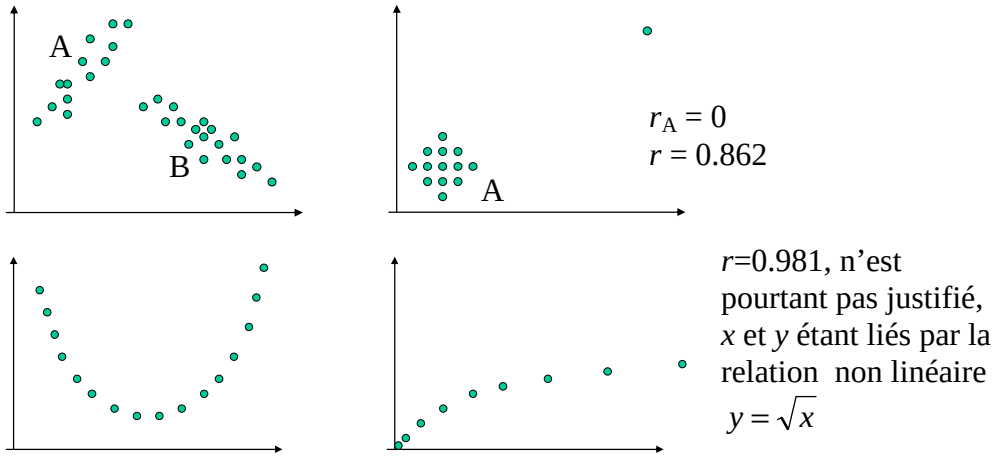


Figure 2

Dans l'exemple suivant, les 4 nuages de la figure 3 ont mêmes moyennes, mêmes variances et même coefficient de corrélation :

$$\bar{x} = 9 \qquad \bar{y} = 7.5$$
$$s_x^2 = 10.0 \qquad s_y^2 = 3.75$$
$$r = 0.82$$

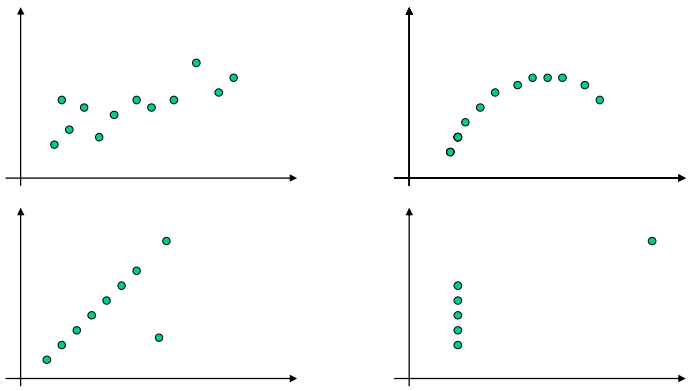


Figure 3

Seul le premier nuage justifie l'usage de *r*.

D'une manière générale, l'étude de la relation entre des variables, quelles qu'elles soient, doit s'accompagner de graphiques descriptifs, exhaustifs ou non dans l'appréhension des données à notre disposition, pour éviter de subir les limites purement techniques des calculs que nous utilisons.

3. Caractère significatif d'un coefficient de corrélation

En admettant que l'on se trouve dans le cas où l'usage de *r* est justifié, à partir de quelle valeur la liaison est-elle significative ?

On se fixe un risque acceptable α (en général 5 %) et on calcule la quantité r_0 suivante :

$$r_0 = \frac{t}{\sqrt{n-2+t^2}}$$

où *n* est le nombre d'observations et *t* le fractile d'ordre $\alpha/2$ de la loi de Student à $n-2$ degrés de liberté (donné par la table *ad-hoc*).

On rejette l'hypothèse que le coefficient de corrélation est nul, au risque α de se tromper, dès que $|r| \geq r_0$. On pourra alors dire, avec une probabilité α de se tromper, que le coefficient de corrélation est significativement différent de 0.

Par exemple, au risque de 5 % on déclarera qu'une liaison est significative sur un échantillon de 30 observations si $|r| > 0{,}36$ ($t = 2{,}048$).

Attention : on remarquera que le seuil de signification (r_0) décroît quand *n* croît ; **pour les grands échantillons, le fait de trouver que *r* diffère significativement de 0 ne garantit nullement que la liaison soit forte**. De plus, rappelons que *le fait de trouver que r n'est pas significativement différent de 0 n'entraîne pas nécessairement l'indépendance.*

On peut aussi déterminer un intervalle de confiance sur le coefficient de corrélation (afin de quantifier la qualité de la régression) grâce à l'introduction de la transformation suivante :

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad \text{et} \quad r = \frac{e^{2z_r}-1}{e^{2z_r}+1}$$

L'intervalle de confiance est défini par :

$$P\left(z_r - z_{\alpha/2}\sqrt{\frac{1}{n-3}} < Z < z_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}}\right) = 1-\alpha$$

où $z_{\alpha/2}$ est tel que $P(Y < z_{\alpha/2}) = 1 - \alpha/2$ avec *Y* une variable aléatoire normale centrée-réduite.

Application : en clair, pour un coefficient de corrélation calculé *r* :

- on se fixe un risque acceptable, en général $\alpha = 0.05$ (5 %) ;
- on détermine $z_{\alpha/2}$ avec la table de la loi normale centrée-réduite, en général $z_{\alpha/2} = 1{,}96$;
- on calcule z_r à partir de la valeur *r* :

$$z_r = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

- on calcule les bornes de l'intervalle de confiance sur *Z*, soit :

$$z_{\text{inf}} = z_r - z_{\alpha/2}\sqrt{\frac{1}{n-3}} \text{ et } z_{\text{sup}} = z_r + z_{\alpha/2}\sqrt{\frac{1}{n-3}},$$

- puis celles de l'intervalle de confiance sur *R*, à savoir :

$$r_{\text{inf}} = \frac{e^{2z_{\text{inf}}}-1}{e^{2z_{\text{inf}}}+1} \text{ et } r_{\text{sup}} = \frac{e^{2z_{\text{sup}}}-1}{e^{2z_{\text{sup}}}+1}.$$

$\Rightarrow$ Le coefficient de corrélation, estimé par *r*, appartient à l'intervalle $[r_{\text{inf}} ; r_{\text{sup}}]$ avec une probabilité de $1 - \alpha$.

Exemple : Soit r = 0,54 obtenu sur un échantillon de taille n = 69. On souhaite construire l'intervalle de confiance à 99 % autour de cette valeur. On obtient successivement z_r = 0,604. Dans la table de la loi normale, on lit $z_{0.005}$ = 2,575 et donc $P(0{,}293 < Z < 0{,}927) = 0{,}99$. Par inversion, on obtient l'intervalle de confiance sur l'estimation du coefficient de corrélation : $P(0{,}285 < R < 0{,}729) = 0{,}99$.

4. **Corrélation partielle** *(pour information)*

Une erreur courante est de croire qu'un coefficient de corrélation élevé induit une relation de cause à effet entre les deux phénomènes mesurés. En réalité, les deux phénomènes peuvent être corrélés à un même « phénomène-source » : une troisième variable non mesurée, et dont dépendent les deux autres. Autrement dit, il arrive fréquemment que la dépendance apparente entre deux variables soit due en réalité aux variations d'une 3ème variable.

La littérature statistique abonde en exemples de fausses corrélations surprenantes entre phénomènes variés qui disparaissent lorsque l'on fixe une 3ème variable (souvent non aléatoire comme le temps) ; ainsi de la corrélation entre le nombre de maladies mentales déclarées chaque année et le nombre de postes de radio installés, ou encore le nombre de coups de soleil observés dans une station balnéaire, qui est fortement corrélé au nombre de lunettes de soleil vendues ; pourtant aucun des deux phénomènes n'est bien sûr la cause de l'autre… On cite souvent aussi l'exemple plus morbide de la consommation de pétrole et de la mortalité des personnes âgées.

Les coefficients de corrélation partielle constituent un moyen d'éliminer l'influence d'une ou plusieurs variables.

Le coefficient de corrélation partielle entre X et Y conditionnellement à une 3ème variable Z est donné par :

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{\left(1 - r_{XZ}^2\right)\left(1 - r_{YZ}^2\right)}}$$

où r_{AB} est le coefficient de corrélation entre les variables A et B.

5. **Problèmes d'interprétation** : corrélation entre deux stations urbaines pour le NO_2

Un coefficient qui peut sembler plutôt élevé (ici 83%) peut conduire à des défauts d'interprétation…

En fait, il y a un assez grand nombre de points éloignés de la bissectrice, et notamment dans les concentrations les plus fortes, il n'y a aucun point sur la bissectrice. En fait, si on s'intéresse aux valeurs supérieures à 40, le coefficient de corrélation tombe à 0,56 et si on ne garde que les valeurs > 80, il tombe à 0,066…
Notons aussi que dans de nombreux cas, la différence de concentrations peut aller « du simple au double » : quand PQV enregistre 60 µg/m^3, JUS enregistre dans le même temps une valeur pouvant aller de 20 à 100 µg/m^3…