Statistiques descriptives bivariées

Cadre
- Tout le vocabulaire et notions de la fiche univariées statistiques

A) Cadre

Sur un échantillon $\mathscr{E}$ donné d'une population $\mathscr{P}$, on est amené à observer plusieurs caractères. Le cas qui nous intéresse ici est celui de deux caractères quantitatifs. Dans ce cas, on étudie des séries statistiques bivariées.

B) Données

On se donne un échantillon $\mathscr{E}$ de taille n d'une certaine population $\mathscr{P}$ sur lequel on a mesuré deux caractères quantitatifs x, y. On note :

a. $x_1 < x_2 < \cdots < x_p$ les différentes modalités observées de x sur les individus de $\mathscr{E}$ (quitte à les renuméroter).

b. $y_1 < y_2 < \cdots < y_q$ les différentes modalités observées de y sur les individus de $\mathscr{E}$ (idem).

c. On note $n_{i,j}$ l'effectif de l'observation (x_i, y_j) c'est-à-dire le nombre d'individus de $\mathscr{E}$ pour lesquelles l'observation conjointe de x et y donne les valeurs respectives x_i et y_j

d. On a donc $\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j} = n$.

C) Représentation des données

Dans un tableau à double entrée, on représente les modalités de x et de y. À l'intersection de la ligne x_i et de la colonne y_j, on place l'effectif conjoint $n_{i,j}$ de la modalité (x_i, y_j) (Comme pour les lois conjointes de variables aléatoires).

D) Indicateurs

[Point moyen] **Déf.1**

Pour une série statistique (x, y) bivariée, le point moyen est le point de coordonnées $(\overline{x}, \overline{y})$.

E) Covariance, coefficient de corrélation

[Covariance d'une série stat. biv.] **Déf.2**

Soit (x, y) une série statistique double, notons (x_i, y_j) ses modalités distinctes (*cf.* B) données), n la taille de l'échantillon. La covariance de la série est le nombre noté $s_{x,y}$ défini par :

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i - \overline{x})(y_j - \overline{y}).$$

[Formule de Kœnig pour la covariance] **Prop.1**

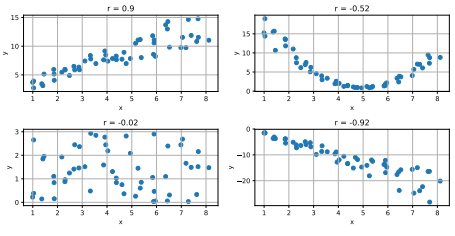
$$s_{x,y} = \overline{xy} - \overline{x}\cdot\overline{y}$$

[Coefficient de corrélation linéaire] **Déf.3**

C'est le nombre noté $r_{x,y}$ et défini par :

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

Rem.1 | Le coefficient de corrélation linéaire renseigne sur la tendance qu'a le nuage de points de la série à se concentrer le long d'une droite : il *ne détecte pas* d'éventuelles corrélations *non linéaires* ! (voir figure ci-dessous, le deuxième cas semble révéler une corrélation du style $y = a(x-b)^2$).



[Propriétés du coefficient de corrélation linéaire] **Prop.2**

a. $r_{x,y} \in [-1, 1]$.

b. $|r_{x,y}| = 1$ si et seulement si il existe deux réels a, b tels que :

$$y = ax + b.$$

(le nuage de points est aligné sur une droite).

Rem.2 | Ne pas confondre causalité et corrélation !

F) Ajustement affine - régression linéaire

[Données du problème] **Déf.4**

a. On se donne un nuage de n points du plan (x_i, y_i) $(1 \leq i \leq n)$, issu d'une statistique bivariée.

b. Deux réels a, b quelconques, et la droite d'équation $y = ax + b$.

c. On définit l'écart quadratique moyen EQM(a, b) mesurant la «distance» du nuage à la droite.

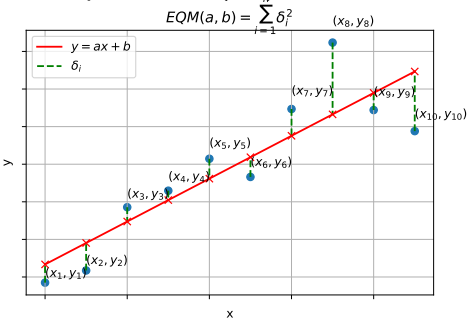
$$\text{EQM}(a,b) := \sum_{i=1}^{n} \underbrace{\left(y_i - \overbrace{(ax_i + b)}^{\hat{y}_i}\right)^2}_{\delta_i^2}.$$

Rem.3 | Ce choix de EQM est motivé par les considérations suivantes :

a. Les points (x_i, y_i) sont alignés sur la droite d'équation $y = ax + b$ si, et seulement si EQM(a, b) est nul.

b. Comme : $\forall (a, b) \in \mathbf{R}^2$, EQM(a,b)$\geq$ 0, plus EQM(a, b) est petit, mieux la droite d'équation $y = ax + b$ devrait épouser le nuage de points.

c. La «meilleure» droite épousant le nuage de points devrait donc être celle qui rend EQM(a, b) **minimal**.



légende : $\hat{y}_i = ax_i + b$ (×), et $\delta_i = |y_i - \hat{y}_i|$ (- - -).

[Ajustement affine] **Déf.5**

On appelle ajustement affine du nuage de points $((x_i, y_i))_{1\leq i\leq n}$ la recherche des paramètres a, b de sorte à rendre EQM(a, b) **minimal**.

La résolution du problème de l'ajustement affine se comprend mieux dans un cadre abstrait de projection orthogonale dans $\mathbf{R}^n$.

[Vecteurs abstraits associés] **Déf.6**

- [Vecteurs des] observations :
$$\mathbf{x} := (x_1, \dots, x_n) \in \mathbf{R}^n$$
$$\mathbf{y} := (y_1, \dots, y_n) \in \mathbf{R}^n$$
- Série certaine :
$$\mathbf{1} = (1, \dots, 1) \in \mathbf{R}^n$$
- Prévisions : $\hat{\mathbf{y}} = a\mathbf{x} + b\mathbf{1} \in \mathbf{R}^n$.
- Erreur de prévision du modèle (a, b) :
EQM$(a, b) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 \in \mathbf{R}_+$.

Rem.4

a. On appelle les $\hat{y}_i = ax_i + b$ les «prévisions du modèle linéaire (a, b)», puisque si x et y étaient effectivement liés par la relation $y = ax + b$, alors les valeurs de y_i seraient parfaitement prédites à partir de celles de x. En effet, dans ce cas le modèle dit :
$$\underbrace{\hat{y}_i}_{\text{prévisions}} = ax_i + b = \underbrace{y_i}_{\text{observations}}$$

b. Résoudre le problème de l'ajustement affine, c'est donc déterminer les couples (a, b) pour lesquels on a : EQM$(a, b) = \min_{(a,b)\in\mathbf{R}^2} \|\mathbf{y} - \hat{\mathbf{y}}\|^2$, c'est-à-dire le meilleur modèle linéaire collant aux observations.

c. L'ensemble des modèles possibles, c'est-à-dire l'ensemble des sets de prévisions possibles $\hat{\mathbf{y}}$ à partir d'observations $\mathbf{x}$, est donc l'ensemble des vecteurs de la forme $\hat{\mathbf{y}} = a\mathbf{x} + b\mathbf{1}$ quand (a, b) décrit $\mathbf{R}^2$. Cet ensemble est donc le sev de F de $\mathbf{R}^n$ donné par : $F = \text{Vect}(\mathbf{x}, \mathbf{1})$. D'après **b.**, on cherche donc $\hat{\mathbf{y}}$ tel que $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \min_{\mathbf{v}\in F} \|\mathbf{y} - \mathbf{v}\|^2$. De **thm.4 c.** de la fiche `Spé 13 Produit scalaire`, on tire le **thm.1** qui suit :

[Existence et unicité de la droite des moindres carrés] **Thm.1**

a. Il existe une unique droite d'équation $y = ax + b$ minimisant la quantité EQM(a,b) (**Déf. 4**).

b. Les nombres a, b sont les coordonnées sur la base $(\mathbf{x}, \mathbf{1})$ du projeté orthogonal de $\mathbf{y}$ sur le plan F =Vect $(\mathbf{x}, \mathbf{1})$.
$$a = r_{x,y}\frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2}$$

c. Cette droite est appelée **droite de régression**, ou **droite des moindres carrés**, puisque toute autre droite a un EQM qui lui est *strictement* supérieur.

d. Elle passe *toujours* par le point moyen du nuage $G(\overline{x}, \overline{y})$ (**Déf.1**), donc $b = \overline{y} - a\overline{x}$.

e. On a donc l'équation :
$$y = r_{x,y}\frac{s_y}{s_x}(x - \overline{x}) + \overline{y}.$$

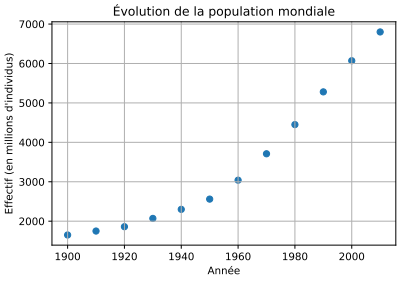
G) Recherche de lois approchées par changement d'échelle ...

Cas pour lesquels un ajustement affine est envisageable après changement d'échelle :

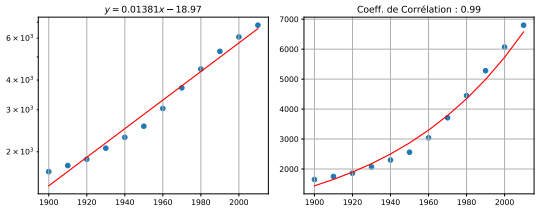
a. Modèles linéaires du premier ordre. (qui sont quasiment universels en première approche), la relation est du type : $y = Ce^{\lambda x}$, qui équivaut à : $\begin{cases} Y = AX + B \\ X = x, Y = \ln y, A = \lambda, B = \ln C \end{cases}$

Exple.1

Évolution de la population mondiale depuis 1900 :



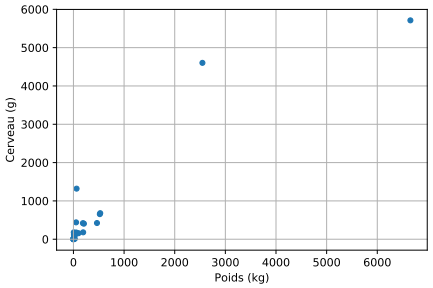
On peut ajuster ces données à un modèle malthusien (*cf.* fiche `Dynamique des populations`) :



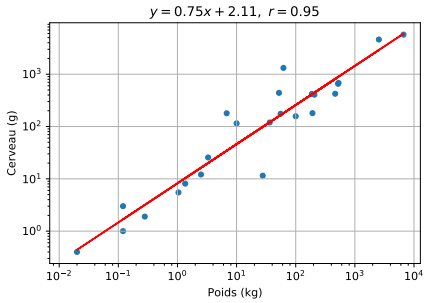
b. Relations allométriques. (pour des grandeurs dimensionnées) : On soupçonne un lien du type : $y = Cx^{\alpha}$, ce qui équivaut à
$\begin{cases} Y = AX + B \\ X = \ln x, Y = \ln y, A = \alpha, B = \ln C \end{cases}$

Exple.2

Corrélation entre la masse corporelle M (en kg) d'espèces de mammifères et celle m_c de leur cerveau (en g).



Les données sont très dispersées (recensées sur 23 espèces allant de la souris (0,02 kg pour un cerveau de 0,4g) à l'éléphant d'Afrique (7 tonnes environ pour un cerveau de 6kg)). En passant aux logarithmes, on voit que l'on peut tenter un ajustement affine :



duquel on tirerait : $m_c \simeq 8.25 \times M^{\frac{3}{4}}$