TP 7

Statistiques descriptives bivariées

Compétences attendues.

- ✓ Représenter un nuage de points associé à une série statistique double.
- ✓ Représenter la droite des moindres carrés.
- ✓ Calculer le coefficient de corrélation linéaire et interpréter sa valeur.

Liste des commandes `Python` exigibles aux concours.

- Dans la librairie `numpy` : `np.mean`, `np.var`, `np.std`.
- Dans la librairie `matplotlib.pyplot` : `plt.plot`, `plt.show`.

Objectifs. Au TP 3, nous nous sommes intéressé aux séries statistiques univariées, c'est-à-dire l'étude d'un caractère sur une population. Cependant, les données statistiques ne vont pas toujours toutes seules, et pour un même individu, il est possible de s'intéresser à plusieurs caractères. Dans ce TP, nous nous limiterons à l'étude simultanée de deux caractères. Nous nous poserons alors la question suivante : peut-on exprimer l'un de ces caractères en fonction de l'autre ? Plus précisément, l'un est-il une fonction affine de l'autre ? De cette recherche de correspondances peuvent découler des analyses fines, explicatives voire prédictives, ou au contraire mettre en évidence des absences de corrélation entre ces caractères.

Anthony Mansuy
Professeur de Mathématiques en deuxième année de CPGE filière ECG au Lycée Clemenceau (Reims)
Page personnelle : `http://anthony-mansuy.fr`
E-mail : `mansuy.anthony@hotmail.fr`

1 Séries statistiques doubles ou bivariées

1.1 Définition

Soit Ω une population et $\{\omega_1, \dots, \omega_n\}$ un échantillon de taille n, sur lequel nous étudions deux caractères quantitatifs $X, Y : \Omega \to \mathbb{R}$ avec X supposé non constant. Pour tout $i \in [\![1, n]\!]$, on note :

- $x_i = X(\omega_i)$ la modalité de X prise par l'individu ω_i,
- $y_i = Y(\omega_i)$ la modalité de Y prise par l'individu ω_i.

Définition.

On appelle **série statistique double** ou **bivariée** de l'échantillon $\{\omega_1, \dots, \omega_n\}$ pour le couple de caractères (X, Y) la donnée du n-uplet $((x_i, y_i))_{1 \leq i \leq n}$ des modalités de (X, Y) sur Ω.

Exemple. En interrogeant un échantillon de la population mondiale Ω, on peut étudier l'âge X et l'acuité visuelle Y de chaque individu. On définit ainsi une série statistique bivariée.

Remarque. Les séries statistiques bivariées sont la version empirique des couples de variables aléatoires discrètes.

Représentation informatique. On représentera une série statistique double sur `Python` par deux vecteurs `x` et `y` de taille n, où `(x(i),y(i))` est la modalité $(X, Y)(\omega_i)$.

1.2 Représentation graphique

Nuage de points

On représente une série statistique double à l'aide d'un **nuage de points**. C'est l'ensemble des points M_i du plan de coordonnées (x_i, y_i) pour tout $1 \leq i \leq n$.

Définition.

Soient `x` et `y` deux vecteurs de même taille.
L'instruction `plt.plot(x,y,".")` trace le nuage de points dont les abscisses sont données par `x` et les ordonnées par `y`.

Remarque. L'option `"."` a pour effet de ne pas relier les points et de les marquer par des ronds. On peut également utiliser `"+"` ou `"*"` pour changer la forme des points.

Point moyen

On rappelle que la **moyenne** de la série statistique $x = (x_1, \dots, x_n)$ est donnée par :

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

On appelle **point moyen** de la série statistique double $((x_i, y_i))_{1 \leq i \leq n}$ le point de coordonnées $(\overline{x}, \overline{y})$.

Définition.

Soient `x` et `y` deux vecteurs de même taille.
L'instruction `plt.plot(np.mean(x),np.mean(y),"o")` trace le point moyen de la série statistique double $((x_i, y_i))_{1 \leq i \leq n}$.

Remarque. L'option `"o"` permet de différentier le point moyen des autres points du nuage.

1.3 Modèle de régression

Lorsqu'on étudie une série statistique bivariée, on peut penser que l'une des variables, par exemple X, est une cause de l'autre, par exemple Y. On dit alors que X est la **variable explicative** et que Y est la **variable à expliquer**.

Exemple. Reprenons l'exemple de l'âge et de l'acuité visuelle. A priori, la variable explicative est l'âge (caractère X) et la variable à expliquer est l'acuité visuelle (caractère Y).

On cherche alors un **modèle de régression**, c'est-à-dire une expression de Y en fonction de X :

$$Y \simeq f(X)$$

où la fonction f est appelée **fonction de régression**. Pour déterminer un modèle de régression, on trace le nuage des points de Y en fonction de X afin de deviner une relation entre ces données.

Remarque. Obtenir un modèle de régression permet de généraliser les observations et de faire des prédictions. C'est un des principes de base de l'intelligence artificielle (voir cette vidéo pour plus de détails).

Exercice 1 (★)
Nous allons comparer deux sondages étudiant le lien entre la taille X et le poids Y d'un individu.

1. A priori, quelle est la variable explicative et quelle est la variable à expliquer ?
2. Sondage au Japon :

Individu	1	2	3	4	5	6	7	8
Taille x_i en cm	161	170	152	181	163	145	168	175
Poids y_i en kg	58	66	52	73	60	45	65	68

   (a) Tracer le nuage de points de cet exemple.
   (b) Placer le point moyen avec un symbole différent.
   (c) Peut-on déduire du nuage de points un modèle de régression entre X et Y ?
3. Sondage aux USA :

Individu	1	2	3	4	5	6	7	8
Taille x_i en cm	169	195	177	182	166	155	189	174
Poids y_i en kg	90	95	115	90	70	60	80	70

   (a) Sur une nouvelle figure, tracer le nuage de points de cet exemple.
   (b) Placer le point moyen avec un symbole différent.
   (c) Peut-on déduire du nuage de points un modèle de régression entre X et Y ?
4. Comment interpréter ces résultats ?

---

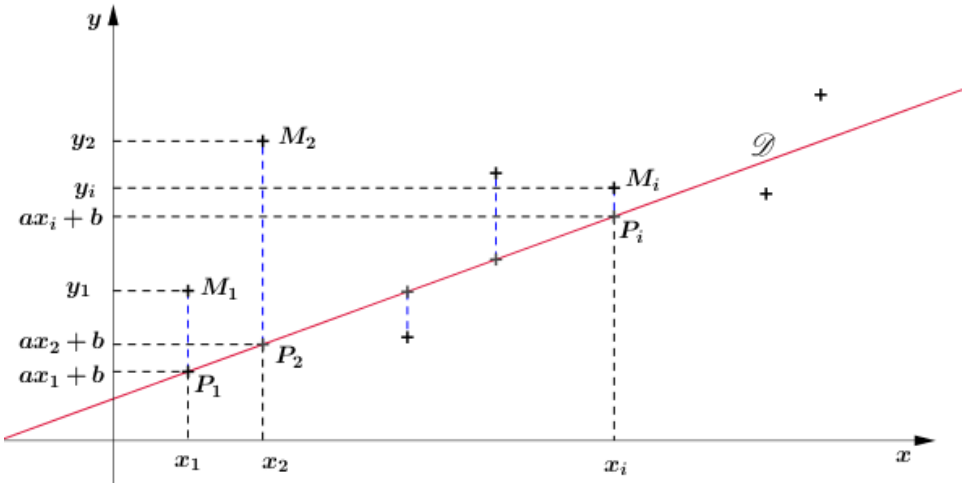
2 Régression linéaire

2.1 Méthode des moindres carrés

Dans cette section, on se place dans le cas particulier où l'on souhaite savoir si Y peut s'exprimer comme une fonction affine de X. Autrement dit, on cherche $(a, b) \in \mathbb{R}^2$ tel que :

$$Y \simeq aX + b.$$

Les points M_i du nuage associé à la série statistique (X, Y) n'étant très probablement pas alignés, on va chercher la « meilleure » droite $\mathscr{D}$ approchant ces points au sens suivant : pour tout $1 \leq i \leq n$, on mesure la distance M_iP_i entre M_i et le point $P_i \in \mathscr{D}$ d'abscisse x_i.



On cherche alors $(a,b) \in \mathbb{R}^2$ rendant minimale la quantité :

$$\sum_{i=1}^{n} M_i P_i^2 = \sum_{i=1}^{n} (y_i - ax_i - b)^2. \qquad (*)$$

On pourrait légitimement vouloir minimiser d'autres quantités, par exemple la somme des longueurs ou encore la plus grande des longueurs. Une des raisons pour lesquelles on s'intéresse à la somme des carrés est qu'on dispose alors d'un résultat garantissant l'existence et l'unicité d'une telle droite :

Théorème 1 (Problème des moindres carrés : Régression linéaire)

Considérons une série statistique double $((x_i, y_i))_{1 \leq i \leq n}$.
Il existe une et une seule droite minimisant la quantité $(*)$. On l'appelle la **droite de régression linéaire** associée à la série statistique double $((x_i, y_i))_{1 \leq i \leq n}$.

Remarque. La droite des moindres carrés est la droite qui passe « la plus près » de tous les points du nuage de points au sens des moindres carrés (c'est-à-dire au sens où elle minimise la quantité $(*)$).

2.2 Équation de la droite de régression linéaire

Définition.

On définit :

- la **variance** (empirique) de la série statistique x par : $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.
- l'**écart-type** (empirique) de la série statistique x par s_x.
- la **covariance** (empirique) de la série statistique double $((x_i, y_i))$ par :

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Remarques.

1. La covariance détermine le lien de variations entre X et Y :
   - Si $s_{x,y} > 0$, alors X et Y varient plutôt simultanément dans le même sens (par exemple si X est la taille et Y le poids des individus).
   - Si $s_{x,y} < 0$, alors X et Y varient plutôt simultanément dans des sens opposés (par exemple si X est le poids et Y l'espérance de vie des individus).
2. Comme en probabilités, la formule de Koenig-Huygens est valable :

$$\begin{aligned} s_{x,y} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i \cdot y_i - x_i \cdot \overline{y} - \overline{x} \cdot y_i + \overline{x} \cdot \overline{y}) \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i y_i}_{=\overline{xy}} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{=\overline{x}} \cdot \overline{y} - \overline{x} \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{=\overline{y}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \overline{x} \cdot \overline{y}}_{=\overline{x}\cdot\overline{y}} = \overline{xy} - \overline{x} \cdot \overline{y}. \end{aligned}$$

Propriété 2 (Équation de la droite de régression linéaire)

Soit $((x_i, y_i))$ une série statistique double. La droite de régression linéaire a pour équation $y = ax + b$ où :

$$a = \frac{s_{x,y}}{s_x^2} \qquad \text{et} \qquad b = \overline{y} - a\overline{x}.$$

En particulier, cette droite passe toujours par le point moyen $(\overline{x}, \overline{y})$.

Définition.

Soient `x` et `y` deux vecteurs de même taille, `x` ayant au moins deux coefficients distincts.

- `np.var(x)` donne la variance du vecteur `x`.
- `np.std(x)` donne l'écart-type du vecteur `x`.
- `np.mean(x*y)-np.mean(x)*np.mean(y)` donne la covariance de la série statistique double $((x_i, y_i))$.
- On détermine les réels a et b tels que $y = ax + b$ est l'équation de la droite de régression linéaire pour la série statistique double $((x_i, y_i))$ à l'aide de la propriété précédente :

```
s = np.mean(x*y)-np.mean(x)*np.mean(y)
a = s/np.var(x)
b = np.mean(y)-a*np.mean(x)
```

Remarque. Ajoutons deux commandes utiles mais hors-programme :

- `np.corrcoef(x,y)` renvoie la matrice $\begin{pmatrix} 1 & r_{x,y} \\ r_{x,y} & 1 \end{pmatrix}$.

  Donc `np.corrcoef(x,y)[0,1]` renvoie le coefficient de corrélation linéaire $r_{x,y}$.
- `a,b = np.polyfit(x,y,1)` renvoie les coefficients a et b de la droite de régression linéaire.

  Plus généralement, `np.polyfit(x,y,k)` donne les coefficients de la courbe polynomiale de degré k qui approche le mieux le nuage de points.

2.3 Coefficient de corrélation linéaire

Définition.

On appelle **coefficient de corrélation linéaire** de la série statistique double $((x_i, y_i))$ le réel définit par :

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

Propriété 3 (du coefficient de corrélation linéaire)

Le coefficient de corrélation linéaire vérifie les propriétés suivantes :

- $|r_{x,y}| \leq 1$;
- $r_{x,y} = \pm 1$ si et seulement s'il existe a et b tels que $Y = aX + b$. Dans ce cas, le signe de a est le même que celui de $r_{x,y}$.

Remarques.

1. Un coefficient de corrélation linéaire proche de ± 1 indique que la droite des moindres carrés approche bien le nuage de points. En général, on estime que la corrélation linéaire entre les séries X et Y est forte quand $|r_{x,y}| \geq 0,9$.
2. Si le coefficient de corrélation linéaire est positif, c'est que x_i « a tendance » à augmenter lorsque y_i augmente, alors que s'il est négatif, x_i diminue lorsque y_i augmente.
3. Une corrélation linéaire proche de 0 indique une absence de relation de dépendance **linéaire** entre x et y.

Exercice 2 (★)
Reprenons les exemples de l'exercice 1.

1. Dans les deux cas (Japon et USA) :
   (a) Déterminer les coefficients a et b de la droite de régression.
   (b) Déterminer le coefficient de corrélation linéaire.
   (c) Représenter la droite de régression linéaire sur le même graphique que le nuage de points.
   (d) Estimer le poids d'un individu mesurant 1m80. Que pensez-vous de la pertinence de cette estimation ?
2. (a) Quel est le signe du coefficient directeur de la droite de régression ? Comment l'interprétez vous ?
   (b) Comparez les valeurs des coefficients de corrélation linéaire dans les deux cas. Sont-ils conformes aux représentations graphiques obtenues ?

---

3 Exercices

Exercice 3 (★)
1. (a) Créer deux vecteurs `x` et `y` contenant chacun 1000 nombres tirés au hasard dans $[0, 1]$ (on utilisera la commande `rd.random`).
   (b) Représenter le nuage de points ainsi que le point moyen et la droite des moindres carrés. Qu'en pensez vous ?
   (c) Calculer le coefficient de corrélation linéaire. Comment expliquer le résultat obtenu ?
2. On pose à présent :

```
>>> x = rd.uniform(-1, 1, 1000)
>>> y = x**2
```

   Ainsi défini, `x` est un vecteur de taille 1000 dont les coefficients sont des réels choisis aléatoirement dans l'intervalle $[-1, 1]$.
   (a) Représenter le nuage de points associés à ces séries statistiques ainsi que la droite des moindres carrés. Qu'en pensez vous ?
   (b) Calculer le coefficient de corrélation linéaire. Les variables `x` et `y` sont-elles indépendantes ?

À retenir. Si deux séries statistiques proviennent de caractères indépendants, alors le coefficient de corrélation linéaire est proche de zéro. À l'inverse, un coefficient de corrélation linéaire proche de zéro n'assure en rien l'indépendance des deux caractères étudiés : il indique seulement une indépendance **linéaire** entre ceux-ci.

Exercice 4 (★)
Considérons les séries statistiques `x` et `y` suivantes :

```
>>> x = np.arange(1,41)
>>> y = np.log(x)+rd.uniform(-1, 1, 40)
```

1. Représenter le nuage de points associé.
2. (a) Calculer le coefficient de corrélation linéaire. Semble-t-il bon ?
   (b) Calculer les coefficients a et b de l'équation de la droite de régression de `y` par rapport à `x`.
   (c) Superposer la droite de régression linéaire au nuage de points.
3. Vérifier que le nuage de points se superpose bien avec la courbe représentative de la fonction $f : x \mapsto \ln(x)$.
4. Étudier la corrélation linéaire de `y` par rapport à `z = np.log(x)`.

À retenir. Les relations entre les caractères X et Y ne sont pas nécessairement linéaires, elles peuvent être logarithmiques, exponentielles, ... L'étude faite dans ce TP peut cependant nous aider pour des régressions non linéaires, en étudiant la corrélation linéaire entre Y et $\ln(X)$, $\exp(X)$, ...

Exercice 5 (★)

Une bonne corrélation entre deux séries de données ne signifie pas pour autant qu'il existe un lien de cause à effet entre les deux. À titre d'exemple, considérons la série statistique suivante :

Année	1996	1997	1998	1999	2000
Morts	15.85	15.7	15.39	15.32	14.85
Importations de citrons	230	280	360	410	525

Ce tableau donne le nombre de morts (pour un million d'habitants) sur les autoroutes américaines, ainsi que le nombre de tonnes de citrons mexicains importés aux États-Unis de 1996 à 2000.

1. Calculer le coefficient de corrélation linéaire pour cette série double.
2. En déduisez vous une information pertinente ?

À retenir. Attention donc à l'erreur courante, notamment dans les médias, qui est de croire qu'un coefficient de corrélation linéaire élevé (en valeur absolue) induit une relation de causalité entre les deux phénomènes mesurés. Voir à ce sujet cette page des Décodeurs du `monde.fr` présentant un outil de corrélation géographique sur la base de données sans rapport, de manière à générer « vos propres cartes pour ne rien démontrer du tout ». Vous y apprendrez par exemple que la consommation de fromage est fortement corrélée au nombre de licences de football. Mais ce n'est pas pour autant que les footballeurs mangent plus de fromage.
