TP 4-Statistiques bivariées

Durée : 3h

1 Série statistique à deux variables

1.1 Généralités

Dans une population donnée, on peut être amené à étudier simultanément deux caractères X et Y. On peut alors étudier chacun de ces caractères séparément (statistiques univariées, voir TP2) ou étudier le lien entre les deux. Dans ce deuxième cas, on étudie alors le couple (X,Y) : c'est ce qu'on appelle des statistiques bivariées.

Définition 1 (Série statistique à deux variables)

Soient $\{\omega_1,\dots,\omega_n\}$ un échantillon d'une population Ω et $X=[x_1,\dots,x_n]$, $Y=[y_1,\dots,y_n]$ deux séries statistiques représentant les valeurs prises par deux caractères X et Y sur cet échantillon :

$$\forall i\in[\![1,n]\!]\quad x_i=X(\omega_i)\quad\text{et}\quad y_i=Y(\omega_i).$$

On appelle série statistique double la donnée de la liste

$$\big[(x_1,y_1),\dots,(x_n,y_n)\big]$$

Chaque couple est associé à un seul individu de l'échantillon : pour tout $i\in[\![1,n]\!]$, $(x_i,y_i)=(X(\omega_i),Y(\omega_i))$.

Lorsqu'on étudie un couple de caractères (X,Y), il est fréquent de se demander si l'un des deux caractères (par exemple X) est une cause de l'autre (par exemple Y). Dans ce cas, on dit que X est la **variable explicative** et que Y est la **variable à expliquer**.

► Un sociologue veut analyser s'il existe une relation entre le taux de criminalité dans des villes et la densité de population de ces villes. Entre la densité de population et le taux de criminalité, que choisiriez-vous comme variable explicative X et comme variable à expliquer Y ?

1.2 Nuage de points

Définition 2 (Nuage de points)

On se place dans un repère orthonormé du plan.
On appelle **nuage de points** associé à une série statistique double $\big[(x_1,y_1),\dots,(x_n,y_n)\big]$ l'ensemble des points M_i de coordonnées (x_i,y_i) pour $i=1,\dots,n$.

On appelle **point moyen** le point de coordonnées $(\bar{x},\bar{y})$ où $\bar{x}$ est la moyenne de la série $[x_1,\dots,x_n]$ et $\bar{y}$ la moyenne de la série $\big[y_1,\dots,y_n\big]$.

Méthode 1 (*Tracer un nuage de points en Scilab*)

1. *On construit d'abord les vecteurs* X *et* Y *(de même taille) qui contiennent respectivement la série* $[x_1,\dots,x_n]$ *et la série* $\big[y_1,\dots,y_n\big]$.
2. *On utilise ensuite la commande* `plot2d(X,Y,style=Z)` *où* Z *est une variable pouvant prendre les valeurs suivantes :*

-6	-5	-4	-3	-2	-1	0	1	2	3	4	5
Δ	$\Diamond$	♦	$\oplus$	$\times$	$+$	$\cdot$	*noir*	*bleu foncé*	*vert*	*bleu clair*	*rouge*

Le tableau suivant contient la liste de 11 pays d'Amérique du Nord et d'Amérique Centrale, dont la population dépassait le million d'habitants en 1985. Pour chaque pays, on mesure le taux de natalité Y (nombre de naissances annuel pour 1000 habitants) ainsi que le taux d'urbanisation X (pourcentage de la population vivant dans des villes de plus de 100000 habitants).

	Canada	Costa Rica	Cuba	USA	El Salvador	Guatemala	Haïti	Honduras	Jamaïque	Mexique	Nicaragua
X	55.0	27.3	33.3	56.5	11.5	14.2	13.9	19.0	33.1	43.2	28.5
Y	16.2	30.5	16.9	16.0	40.2	38.4	41.3	43.9	28.3	33.9	44.2

- Créer les vecteurs `X` et `Y` qui contiennent respectivement les taux d'urbanisation et les taux de natalité de ces 11 pays.
- Tracer le nuage de points associé en marquant les points par le symbole $\oplus$. Recopier vos lignes de commande ci-dessous.

- Déterminer les coordonnées du point moyen et placer le sur la même fenêtre graphique que le nuage de point, marqué par le symbole $\blacklozenge$. Recopier les coordonnées du point moyen.

1.3 Covariance et corrélation

Définition 3 (Covariance et corrélation)

Étant donné une série statistique double $[(x_1, y_1), \ldots, (x_n, y_n)]$, on appelle :

- **covariance** de la série double le réel :

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{k=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})$$

- **coefficient de corrélation linéaire** de la série double le réel :

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

où σ_X est l'écart-type de la série X et σ_Y est l'écart-type de la série Y.

Méthode 2 (*Calcul de la covariance et du coefficient de corrélation linéaire*)

1. *La fonction* `corr(X,Y,1)` *calcule la covariance de* X *et de* Y.
2. *Pour obtenir le coefficient de corrélation linéaire il suffit de calculer la covariance, les écart-types et d'utiliser la formule de la définition.*

- Compléter la fonction ci-dessous pour qu'elle renvoie la variance de `x`.

```
    function res=Variance(x)
    n=length(x)
    s=0
    for i=1:n
        s=......
    end
    res=s/n
endfunction
```

On reprend l'exemple où X est le taux d'urbanisation et Y le taux de natalité de 11 pays d'Amérique.

► Déterminer la covariance de la série double :

► Déterminer les écart-types σ_X et σ_Y. On recopiera les lignes de commande permettant le calcul de σ_Y ainsi que les valeurs de σ_X et σ_Y ci-dessous :

► Donner la valeur du coefficient de corrélation linéaire :

2 Régression

2.1 Généralités

On considère un couple de variables (X, Y) où X est la variable explicative et Y la variable à expliquer.

Déterminer un **modèle de régression** consiste à savoir si Y est une fonction de X :

$$Y = f(X)$$

où f est une fonction. En pratique, il est très rare que Y soit une fonction de X mais on peut souvent approcher Y par une variable aléatoire qui est une fonction de X :

$$Y = f(X) + \varepsilon$$

où la fonction f est appelée **la fonction de régression** et la variable aléatoire ε est appelée **l'erreur d'ajustement**.
Dans la suite, on s'intéressera à des modèles de **régression linéaire** c'est-à-dire modèle où la fonction f est **affine** :

$$f(X) = aX + b.$$

Plus précisément, on va chercher la droite qui « colle au mieux » au nuage de points.

2.2 Moindres carrés

Si le nuage de points associé à une série statistique double possède une forme étirée, on peut chercher une droite qui approche le mieux possible les points de ce nuage. Le problème consiste donc à déterminer une droite $y = ax + b$ qui ajuste bien le nuage de points.
Si on approche notre nuage par une droite $y = ax + b$, l'erreur que l'on commet en utilisant la droite de régression pour prédire y_i à partir de x_i est $y_i - (ax_i + b)$.

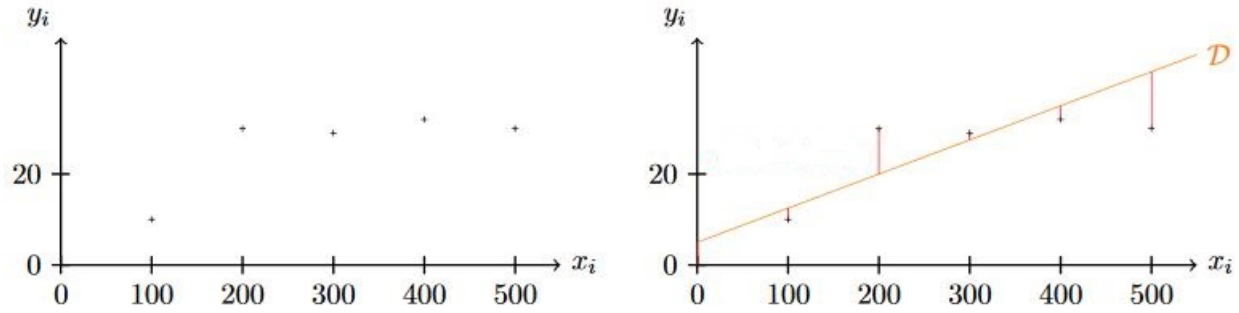

FIGURE 1 – A gauche : nuage de points ; à droite : le même nuage avec une droite d'ajustement et les erreurs (en rouge)

La **méthode des moindres carrés** consiste à trouver (s'ils existent) les coefficients a et b qui minimisent **l'erreur quadratique** :

$$\sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 .$$

Théorème (Droite de régression linéaire)

Étant donné une série statistique double $\left[(x_1, y_1), \ldots, (x_n, y_n)\right]$, l'unique droite minimisant l'erreur quadratique est la droite d'équation :

$$y = \frac{\text{cov}(X, Y)}{\sigma_X^2}(x - \bar{X}) + \bar{Y}.$$

Cette droite est appelé **la droite de régression linéaire de** Y **en** X.

Démonstration : On considère la fonction F de deux variables a et b :

$$\begin{aligned} F : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ (a, b) &\longmapsto \sum_{i=1}^{n} \left(y_i - (ax_i + b)\right)^2. \end{aligned}$$

Le but est de montrer que cette fonction possède un minimum et de le trouver. Nous verrons comment faire au second semestre. ■

Remarque 1

On remarque que le point moyen est toujours sur la droite de régression linéaire.

On reprend l'exemple où X est le taux d'urbanisation et Y le taux de natalité de 11 pays d'Amérique.

- Déterminer les coefficients a et b tels que $y = ax + b$ est la droite de régression linéaire de Y en X (on recopiera les lignes de commande permettant le calcul des coefficients ainsi que leur valeur) :

- Tracer la droite de régression linéaire en rouge avec la commande `plot2d` et recopier les lignes de commande :

2.3 Lien avec le coefficient de corrélation linéaire

Proposition 1

Étant donné une série statistique double $\left[(x_1, y_1), \ldots, (x_n, y_n)\right]$, on note $\rho_{X,Y}$ le coefficient de régression linéaire.

- Plus $|\rho_{X,Y}|$ est proche de 1 plus les points du nuage sont proches de l'alignement et plus les prévisions données par les droites de régression sont pertinentes.
- $|\rho_{X,Y}|$ vaut 1 si et seulement si les points du nuage sont alignés.
- Si $\rho_{X,Y} > 0$ la pente de la droite de régression est positive ; si $\rho_{X,Y} < 0$ la pente de la droite de régression est négative.

3 Exercices

Exercice 1

Dans le cadre d'une enquête visant à comparer les différents aliments vendus dans les fast-food, on a obtenu le tableau suivant :

Aliment	*Poids (en kg)*	*Prix (en euros)*
Sandwich végétarien	150	3.90
Sandwich parisien	92	3.40
Big Mac	193	5.70
Hamburger	90	2.90
Hot Dog	135	3.80
Fallafel	241	6.00
Quiche	169	3.30
Pizza	165	3.30

1. *Représenter le nuage de points (Poids, Prix).*
2. *Déterminer le point moyen et placer le sur le même graphique que le nuage de points mais dans un style différent.*
3. *Déterminer la droite de régression du Prix en Poids et afficher la sur la même fenêtre graphique que le nuage de points mais dans un style différent.*
4. *Utiliser la droite de régression pour estimer le prix d'un aliment pesant* 200 *grammes.*

Exercice 2

Récupérer les données du fichier .sce sur le site. On considère la série statistique double `(L1,L2)`.

1. *Déterminer le point moyen de la série double.*
2. *Tracer le nuage de points et le point moyen sur la même fenêtre graphique dans des styles différents.*
3. (a) *Déterminer le coefficient de corrélation linaire de* `(L1,L2)`.
   (b) *Déterminer les coefficients de la droite de régression linéaire de* `L2` *en* `L1`.
   (c) *Afficher la droite de régression sur le même graphique que précédemment en rouge. Une approximation affine est-elle pertinente?*
4. *On définit* `U=log(L1)` *et* `V=log(L2)`.
   (a) *Avec la commande* `clf()` *nettoyer la fenêtre graphique. Tracer le nuage de point associé au couple* (U,V).
   (b) *Déterminer les coefficients de la droite de régression linéaire de* `V` *en* `U`.
   (c) *Afficher la droite de régression sur le même graphique que le nuage.*
   (d) *Quelle relation cela induit entre* `L1` *et* `L2`*?*
   (e) *Tracer la courbe obtenue sur le même graphique que le nuage de points de la question 2.*