1 Objectifs

Dans ce T.D., nous nous intéressons à l'étude simultanée de deux variables statistiques se rapportant à une même population, comme par exemple la taille et le poids dans une population humaine. De la recherche de correspondances entre deux variables statistiques peuvent découler des analyses fines, explicatives voire prédicitives (dis-moi combien tu mesures, je te dirais à peu près combien tu pèses ...), ou au contraire mettre en évidence des absences de corrélation entre variables.

Nous commençons par donner les notions de base des statistiques bivariées, puis nous les appliquerons sur des données démogaphiques issues du T.D. précédent.

2 Corrélation linéaire et régression linéaire

2.1 Série statistique double, ou bivariée

Une série statistique bivariée est la donnée simultanée de deux variables. Dans le tableau suivant (1) on a reporté pour 24 offres de ventes d'appartements dans les $5^{\text{ème}}$ et $6^{\text{ème}}$ arrondissements de Paris la surface en m^2 (série $(x_i)_{1\leqslant i\leqslant 24}$) et le prix en milliers de francs (série $(y_i)_{1\leqslant i\leqslant 24}$) de chaque appartement :

x_i	28	50	196	55	190	110	60	48	90	35	86	65
y_i	130	280	800	268	790	500	320	250	378	25	350	300
x_i	32	52	40	70	28	30	105	52	80	60	20	100
y_i	155	245	200	325	85	78	375	200	270	295	85	495

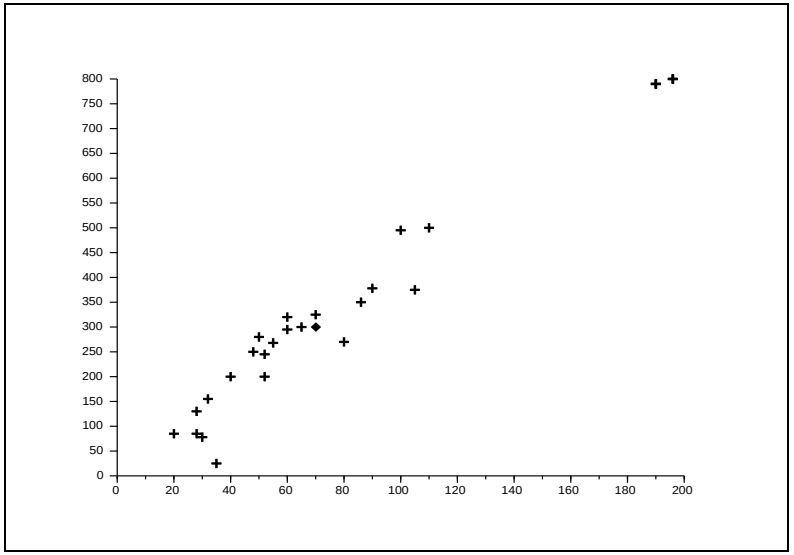
2.2 Nuage de points

On souhaite étudier s'il existe une corrélation entre prix et surface. On commence par tracer le **nuage de points** de la série, ensemble des points de coordonnées (x_i, y_i).

Le script suivant

```
xi=[28 50 196 55 190 110 60 48 90 35 86 65 32 52 40 70 28 30 105 52 80 60 20 100];
yi=[130 280 800 268 790 500 320 250 378 25 350 300 155 245 200 325 85 78 375 200 270 295 85 495];
mx=mean(xi);
my=mean(yi);
plot2d(xi,yi,-1);
plot2d(mx,my,-4);
```

fournit le graphique :



Le point de coordonnées $(\overline{x}, \overline{y})$ s'appelle le point moyen du nuage.

2.3 Droite de régression linéaire

À première vue, les points s'alignent assez bien le long d'une droite. On peut se demander s'il n'y aurait une relation affine $y = ax + b$ liant approximativement surface et prix.

Une telle droite s'appelle **droite de régression linéaire**.

Si je souhaite acheter un appartement de 150 m^2, la connaissance des réels a et b me permettrait d'estimer le prix de vente d'un tel appartement.

(1). Données de 1975 citées dans Probabilités, analyse des données et tatistique de G. SAPORTA

2.3.1 Notations statisticiennes

Complétons les notations déjà introduites au T.D. précédent.

Voici les principales notations, ainsi qu'une valeur approchée de ces grandeurs pour l'exemple que nous étudions.

☞ La *taille* ou *effectif total* de l'échantillon se note N :

$$\mathrm{N} = 24$$

☞ La *moyenne des* x_i (respectivement y_i) se note $\overline{x}$ (resp. $\overline{y}$) :

$$\overline{x} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i \simeq 70,083 \text{ (resp. } \overline{y} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} y_i \simeq 309,333)$$

☞ La *moyenne des* x_i^2 (respectivement y_i^2) se note $\overline{x^2}$ (resp. $\overline{y^2}$) :

$$\overline{x^2} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_i^2 \simeq 6\,909 \text{ (resp. } \overline{y^2} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} y_i^2 \simeq 129\,158)$$

☞ La *moyenne des* x_iy_i se note $\overline{xy}$:

$$\overline{xy} = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} x_iy_i \simeq 29\,637$$

☞ La *covariance de x et de y* se note σ_{xy} :

$$\sigma_{xy} = \overline{(x-\overline{x})(y-\overline{y})} = \overline{xy} - \overline{x}.\overline{y} \simeq 7\,958,18$$

☞ L'*écart-type des* x_i (respectivement y_i) se note σ_x (resp. σ_y) :

$$\sigma_x = \sqrt{\overline{x^2} - \overline{x}^2} \simeq 44,69 \text{ (resp. } \sigma_y = \sqrt{\overline{y^2} - \overline{y}^2} \simeq 182,95)$$

☞ Le *coefficient de corrélation linéaire de x et de y* se note ρ_{xy} :

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \simeq 0,973$$

Ajoutons que si les données n'ont pas toutes la même fréquence (ou effectif) d'apparition, il convient de remplacer le facteur $\frac{1}{\mathrm{N}}$ par la fréquence $f_i = \frac{n_i}{\mathrm{N}}$ du couple (x_i, y_i) dans la série.

Par exemple, $\overline{xy}$ est alors définie par

$$\overline{xy} = \sum_i f_ix_iy_i = \frac{1}{\mathrm{N}}\sum_i n_ix_iy_i.$$

La **covariance** et le **coefficient de corrélation linéaire** sont des grandeurs qui permettent de quantifier les liens linéaires entre x et y.

2.3.2 Méthode de GAUSS des moindres carrés (en y).

Afin de déterminer une droite (G) : $y = ax + b$ optimale, on mesure l'écart e_i entre le point du nuage (x_i, y_i) et (G) par la différence e_i d'ordonnées à l'abscisse x_i : pour tout i de $[\![1\,;\,\mathrm{N}]\!]$, soit $e_i = y_i - (ax_i + b)$.

On cherche alors a et b réalisant le minimum de la fonction f définie par :

$$\text{pour tout } a \text{ et } b \text{ réels, } f(a,b) = \frac{1}{\mathrm{N}}\sum_{i=1}^{\mathrm{N}} e_i^2 = \frac{1}{\mathrm{N}}\sum_{k=i}^{\mathrm{N}} (y_i - (ax_i+b))^2.$$

Ainsi la *méthode des moindres carrés en y* consiste à rechercher **la droite rendant minimale la somme des carrés des écarts mesurés en ordonnées.**

Nous démontrerons que le minimum de cette fonction est atteint pour

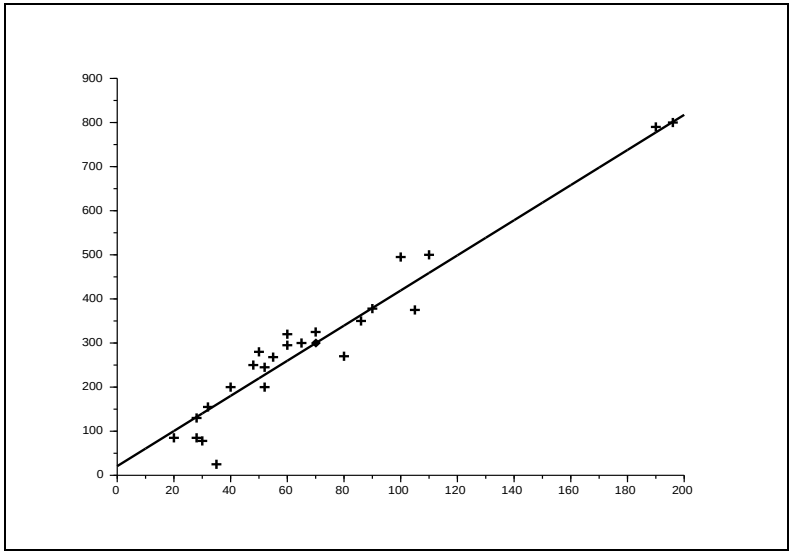
$$a = \rho_{xy}\frac{\sigma_y}{\sigma_x} \text{ et } b = \overline{y} - \rho_{xy}\frac{\sigma_y}{\sigma_x}\overline{x}.$$

Pour notre T.D., nous retiendrons ceci :

La droite (G_y) de régression linéaire par la méthode des moindres carrés en y a pour équation :

$$(\mathrm{G}_y) : \quad y - \overline{y} = \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - \overline{x}).$$

Numériquement, on obtient $(\mathrm{G}_y) \quad y = 3,9833x + 30,0921.$



1. Le point moyen est-il sur la droite de régression obtenue ?

2. Observer le graphique final obtenu et estimer graphiquement le prix de vente d'un appartement de 150 m^2 à l'aide de la droite (G).

2.3.3 Interprétation du coefficient de régression linéaire

Nous avons montré en probabilité que le coefficient de régression linéaire $\rho_{x,y}$ est toujours compris entre -1 et 1.
D'autre part, l'équation de la droite montre [(2)] que $\rho_{x,y}$ est positif lorsque x et y varient dans le même sens et négatif lorsque x et y varient en sens contraire.
Plus $\rho_{x,y}$ est proche de ± 1, plus les points sont alignés, à tel point que lorsque $\rho_{x,y} \pm 1$, les points sont complètement alignés.
En général, on admet que, lorsque ce coefficient atteint $0,8$ (en valeur absolue), il y a une certaine corrélation linéaire entre les variables x et y étudiées, au-delà de $0,9$, la corrélation est forte.

3 Autres régressions et recherche de modèles théoriques

Bien évidemment, les relations entre variables ne sont pas nécessairement linéaires, elles peuvent être logarithmique, exponentielle, ect.... Mais l'étude précédente peut nous aider même pour des régressions non linéaires.

4 T.D. : espérance de vie, revenu et mortalité

4.1 Chargement des données

Au cours de ce T.D., nous allons étudier certaines correspondances entre variables issues des données du précédent T.D.. J'ai ajouté une donnée : le revenu national brut par habitant en 2012 en dollars US. Cette donnée est fournit dans le document de l'INED, sauf pour 30 pays que j'ai donc retiré de la base de donnée. Notre étude portera donc sur 178 pays.

(2). ... puisque σ_x et σ_y sont positifs.

Le fichier `2014_TD_2.sce` est à télécharger à la rubrique informatique du site http ://ecs2poincare.free.fr et à ouvrir. Normalement, il s'ouvre automatiquement dans SciNotes, en provoquant l'ouverture de Scilab. Il n'y a plus qu'à taper simultanément les touches [CTRL] et[L] pour le faire lire par Scilab.
Vous retrouverez

- `pays` : les 178 noms de pays.
- `population` : nombre d'habitants (en millions) ;
- `mort` : taux de mortalité (nombre de décès sur 1000 habitants) ;
- `homme` : espérance de vie des hommes à la naissance (en années) ;
- `femme` : espérance de vie des femmes à la naissance (en années) ;
- `revenu` : revenu national brut par habitant en 2012 (en dollars US).

De plus, j'ai ajouté une fonction `nindex` donnant le nouvel index des pays, qui a fatalement changé avec la disparition de 30 d'entre eux.
`nindex(164)` donne `143` ce qui signifie que le nouvel index de la France (anciennement 164) est 143 dans cette base de donnée, ce que l'on peut vérifier en tapant `pays(143)`.
Petit rappel : l'**effacement de la figure précédente** se fait

- soit en fermant la fenêtre que Scilab a ouverte pour tracer la figure ;
- soit en tapant `clf()` dans la console (`clf()`= « clear figure »).

Bien évidemment, pour compléter un graphique, il peut être intéressant de ne pas effacer la figure courante.

4.2 Corrélation entre l'espérance de vie des hommes et celle des femmes

1. Taper `plot2d(homme,femme)`, puis `plot2d(homme,femme,-1)`.
   L'argument « -1 » demande à Scilab de ne pas relier les points.
2. Les fonctions `meanf` et `stdevf` (« standard deviation » signifie « écart-type ») de Scilab permettent de tenir compte des fréquences (ou des effectifs), à la différence des fonctions `mean` et `stdev`.
   Comparer les valeurs données par `mean(homme)` et `meanf(homme,population)` et préciser ce que représentent ces valeurs.
3. Faire de même avec `stdev` et `stdevf`.

4. La fonction `correl` donne le coefficient de corrélation en tenant compte des effectifs. La syntaxe à respecter est

   `correl(x,y,diag(effectif))`

   pour obtenir le coefficient de corrélation de `x` et `y` pondérés des effectifs [3] contenus dans `effectif`.

   Compléter la fonction suivante pour qu'elle fournissent le coefficient de corrélation ρ ainsi que les coefficients a et b de la droite de régression, et trace sur un même graphique le nuage de points et la droite de régression.

   Remarque : le « `2` » de la seconde fonction `plot2d` permet que la droite soit tracée dans une autre couleur que le nuage de points.

```
function [rho,a,b]=droite(x,y,eff)
  rho=.....................
  a=.....................
  b=.....................
  plot2d(x,y,-1)
  abscisse=[min(x) max(x)]
  plot2d(abscisse,.......................,2)
endfunction
```

5. Déterminer le coefficient de corrélation linéaire entre les espérances de vie des hommes et des femmes, ainsi que l'équation de la droite de régression.

   Ce coefficient de corrélation révèle-t-il une forte corrélation ?

6. Sur un même graphique, tracer le nuage de points et la droite de régression.

7. Placer aussi le point moyen. Appartient-il à la droite de régression ? Était-ce prévisible ?

8. Observer dans quelles zones géographiques sont les pays où les écarts d'espérances de vie entre femmes et hommes sont les plus extrêmes (maximum, puis minimum).

(3). Le dernier argument de la fonction `correl` est une matrice de taille *longueur de x* × *longueur de y* qui permet de donner l'effectif ou la fréquence de chaque (x_i, y_j). Comme nous utilisons uniquement les effectifs de (x_i, y_i), nous plaçons les effectifs sur la diagonale.

4.3 Corrélation espérance de vie et revenu par habitant

1. Tracer le nuage de points et la droite de régression linaire en prenant le revenu par habitant en abscisse et l'espérance de vie des femmes en ordonnées.
2. Que vaut le coefficient de corrélation linéaire ? Comment l'interpréter ?
3. Comme pour les hauteurs de la mousse de bière, nous allons chercher une régression autre que linéaire. Comme les revenus sont fortement dispersés nous allons rabougrir l'échelle grâce à la fonction logarithme.

   Tracer le nuage de points `log(revenu)` en abscisse et `femme` en ordonnées.
4. Les points semblent-ils plus alignés ? Que vaut le coefficient de corrélation linéaire ? Comment l'interpréter ?
5. Écrire l'équation de la droite de régression linéaire liant l'espérance de vie des femmes F au logarithme du revenu ln(R).
6. Sur un même graphique, tracer à nouveau le nuage de points `revenu-femme` et superposer la courbe de régression $\mathrm{F} = a\ln(\mathrm{R}) + b$. Commentaire ?

   Lorsqu'on pense qu'une variable peut expliquer, même partiellement, une autre variable, on la qualifie de **variable explicative**. Ici, le revenu est peut-être [4] une des variables explicatives de l'espérance de vie, **variable expliquée**.

4.4 Correlation mortalité et revenu par habitant

1. Tracer le nuage de points en prenant le revenu par habitant en abscisse et la mortalité en ordonnée.
2. Que vaut le coefficient de corrélation linéaire ? Comment l'interpréter ?
3. Recommencer en prenant `log(revenu)` et `mort`, puis `revenu` et `log(mort)`, puis enfin `log(revenu)` et `log(mort)`.
4. Comment peut-on interpréter toutes ces tentatives ?

(4). Notons que, dans le paragraphe précédent, on peut douter que l'espérance de vie des femmes explique celle des hommes (ou vice-versa), mais elles sont certainement corrélées parce que découlant de mêmes variables explicatives.