TP 6 : Statistiques descriptives bivariées

Dans une population donnée, il peut arriver que l'on souhaite étudier simultanément deux caractères X et Y.
On peut alors s'intéresser aux propriétés de chacun des 2 caractères pris séparément (statistiques univariées), mais aussi au lien entre ces 2 caractères (statistiques bivariées) ; on étudie alors le couple de caractères $Z = (X, Y)$.
En particulier, on peut penser que l'une des variables, X par exemple, est une cause de l'autre, par exemple Y.
On dit alors que X est la **variable explicative** et Y est la **variable à expliquer**.
Dans ce cas, on tentera d'exprimer Y en fonction de X en commençant par tracer le nuage des points de Y en fonction de X pour deviner la relation entre ces données.

Remarque : En informatique, nous nous intéresserons à l'étude statistique du couple (X, Y) (étude d'une série statistique obtenue en effectuant plusieurs simulations de ce couple). Les notions introduites dans ce TP permettent de donner une définition pratique des notions probabilistes qui seront introduites dans le chapitre "couple de variables aléatoires" en mathématiques.

Exemple
En interrogeant un échantillon de la population mondiale Ω, on peut étudier le lien entre l'âge et l'acuité visuelle.
A priori, la variable explicative est alors l'âge (caractère X) et la variable à expliquer est l'acuité visuelle (caractère Y).
On interroge alors un échantillon d'individus ; pour l'individu $n^o\ i$ sondé au hasard, on note $M_i = (x_i, y_i)$ le couple de résultats obtenus et on trace alors le nuage comme expliqué ci-après :

1. Modèle de régression

a) Nuage de points et point moyen

Remarque : Prérequis Scilab : la fonction plot2d
Taper les instructions :

```
-->x=1:5, y=[2 8 1 5 4],  plot2d(x,y,style=-1,rect=[0 0 6 10])
```

Que fait exactement la fonction `plot2d` ? A quoi sert l'option `rect=[0 0 6 10]` ?
Relancer les instructions en changeant uniquement le style : on testera les styles -2, -3, 1,2,5, 0.

• On appelle **nuage de points** associé à la série des n couples (x_i, y_i) l'ensemble des points M_i de coordonnées (x_i, y_i) tracés dans un repère orthonormé du plan .
L'examen du nuage de points permet de faire des constatations qualitatives :
- est-il concentré ou dispersé ?
- relève-t-on une tendance ?
  (variations dans le même sens (covariance positive) ? suit une courbe particulière ? (courbe de régression) ?...)
- y-a-t-il des valeurs a priori aberrantes ?

• Si l'on note $\overline{x}$ la moyenne des x_i et $\overline{y}$ la moyenne des y_i, alors le point de coordonnées $(\overline{x}; \overline{y})$ est le **point moyen du nuage**.

• Si l'on considère X comme la variable explicative et Y comme la variable à expliquer, chercher un **modèle de régression** consiste à savoir si Y est à peu près fonction de X (à quelques erreurs près) ($Y \simeq f(X)$) ou plus précisément à mettre Y sous la forme :

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

où la fonction f est appelée fonction de régression et ε est une variable aléatoire appelée erreur d'ajustement (ou résidus).

On considère alors que la régression est correcte si ε est d'espérance nulle et que le nuage de points de (X, ε) ne se répartit pas autour de 0 selon une direction précise.

Si une telle formule existe, on dit que les variables sont **corrélées**. Sinon, on dit qu'elles sont **non corrélées**.
On dit que la corrélation est forte quand l'erreur est globalement très faible. Sinon, on dit que la corrélation est faible.

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 $n^o\, i$	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

   Tracer le nuage de points de cet exemple et placer le point moyen avec un symbole différent.
   Modifier les axes : on souhaite que l'axe des abscisses aille de 140 à 200 cm et l'axe des ordonnées de 0 à 140 kg.
   X et Y sont-elles corrélées ? fortement corrélées ?

3. **Sondage aux USA :**

individu $n^o\, i$	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

   Sur une nouvelle figure, tracer le nuage de points de cet exemple et placer le point moyen avec un symbole différent.
   Modifier les axes : on souhaite que l'axe des abscisses aille de 140 à 200 cm et l'axe des ordonnées de 0 à 140 kg.
   X et Y sont-elles corrélées ? fortement corrélées ?

4. Donner des explications à ces résultats.

b) Covariance

La **covariance théorique (probabiliste)** d'un couple (X, Y) de variables aléatoires est défini par :

$$cov(X, Y) = E\left((X - E(X))(Y - E(Y))\right)$$

mais le théorème de Koenig-Huygens permet également de la calculer par la formule

$$cov(X, Y) = E(XY) - E(X)E(Y)$$

Par analogie, La **covariance empirique (statistique)** d'un n-échantillon du couple (X, Y) est définie par

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

Le théorème de Koenig-Huygens donne :

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

Remarque
En Scilab, la covariance se calcule soit par l'instruction `corr(x,y,1)` (on ne l'utilisera pas), soit en utilisant la définition i.e. `mean((x-mean(x)).*(y-mean(y)))`,ou le théorème de Koenig-Huygens i.e. `mean(x.*y)-mean(x)*mean(y)`

La covariance détermine le lien de variations entre X et Y.
En particulier, si X et Y sont indépendantes alors $cov(X,Y)=0$. La réciproque est fausse !
En général, si $cov(X,Y)>0$ alors X et Y varient plutôt simultanément dans le même sens (par exemple si X et la taille et Y le poids des individus), si $cov(X,Y)<0$ 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).

Exercice 2.
Nous allons vérifier que la loi normale centrée est une bonne modélisation d'une erreur due à plusieurs phénomènes aléatoires (utiliser Scinote).

1. Compléter et taper l'instruction suivante qui crée une série statistique $x=[x_1,x_2,...,x_{20}]$ pour laquelle $X_i=i+\mathcal{U}_i$ où $\mathcal{U}_i$ est un réel choisi au hasard dans le segment $[-\frac{1}{2};\frac{1}{2}]$:

   ```
   x= ************+grand(1,20,'unf',-1/2,1/2)
   ```

2. Compléter et taper l'instruction suivante qui crée une série statistique $z=[z_1,z_2,...z_{20}]$ pour laquelle les Z_i sont indépendants, indépendants des X_i, et suivent une loi normale centrée de variance 0.25 :

   ```
   z=grand(****,****,****,****,****)
   ```

3. On définit $Y=2-X+Z$.
   Dessiner le nuage de points associé à cet échantillon du couple (X,Y) et son point moyen.
   Quelle fonction de régression permet le mieux d'exprimer Y "à peu près" en fonction de X ?
   Le confirmer en tapant les instructions :

   ```
    plot2d(x,2-x,style=5)
   scf(1),plot2d(x, y-(2-x),style=-1)
   ```

   et en interprétant les résultats.
   Calculer et afficher la covariance empirique du couple (X,Y) sur cet échantillon et vérifier la cohérence du résultat.

4. Faire le même travail avec $Y=X^2+Z$ et $Y=exp(-X)+Z$.
   On pourra tracer respectivement les courbes d'équation $y=x^2$ et $y=\exp(x)$.
   Enoncez ce que vous constatez.

2. Régression linéaire

a) La méthode des moindres carrés

On se place dans le cas particulier où l'on souhaite savoir si la courbe de Y en fonction de X peut être approximée par une droite.
On cherche donc a et b deux constantes telles que $Y\simeq aX+b$ (ou $Y=aX+b+\varepsilon$).
Dans ce cas, on peut utiliser la méthode des moindres carrés qui nous donne l'équation de la droite la plus proche des points en terme de distance :

Théorème 1(admis)

La droite la plus proche du nuage des points est la droite d'équation $y=ax+b$ où

$$\boxed{a=\frac{cov(x,y)}{V(x)} \text{ et } b=\overline{y}-a\overline{x}}$$

Cette droite passe toujours par le point moyen, d'où l'équation $\overline{y}=a\overline{x}+b \Leftrightarrow b=\overline{y}-a\overline{x}$.

b) Coefficient de corrélation linéaire

• On appelle **coefficient de corrélation linéaire** le nombre ρ (rho) défini par

$$\rho(X,Y)=\frac{cov(x,y)}{\sigma(x)\sigma(y)} \text{ (où } \sigma \text{ désigne l'écart-type)}$$

• **Théorème :** $\rho\in[-1;1]$. Et $\rho=\pm1$ si et seulement si la régression $Y=aX+b$ est exacte.

• **Conséquence :** Si ρ est proche de 1 (en sciences humaines et en sciences économiques où beaucoup de

facteurs interviennent, une valeur de $|\rho(x,y)|$ de l'ordre 0,85 est souvent considérée comme bonne), cela peut nous permettre de confirmer qu'il y a bien une corrélation linéaire entre X et Y. Mais à condition d'avoir visualisé une relation linéaire entre les données car ce n'est pas toujours vrai : **Exercice 3.** contre-

exemple
Un bon coefficient de corrélation n'assure pas une relation affine entre X et Y :

1. Créer la série statistique $x = [1, 2, 3, ..., 20]$ et $y = [y_1, y_2, ..., y_{20}]$ tel que $Y_i = i^2 + U_i$ où U_i est un réel choisi au hasard dans le segment $[0; 1]$.
   Tracer le nuage de points (x, y), le point moyen et la courbe d'équation $y = x^2$ sur la même figure. Commenter.
2. Mémoriser dans une variable `c` la covariance empirique du couple (x, y) et dans des variables `v1` et `v2` les variances empiriques des séries statistiques x et y.
3. Calculez et affichez les coefficients a et b de la droite de régression, faites tracer la droite de régression puis calculez et affichez le coefficient de corrélation linéaire. Commentez.
4. Nous allons confirmer que l'approximation n'est pas bonne en traçant les résidus (erreur d'ajustement).
   En effet, si l'on fait une régression (ici, $Y = aX + b + \varepsilon$), l'approximation de Y par $aX + b$ est satisfaisante si les résidus de ε sont disposés symétriquement et sans structure particulière. Si ce n'est pas le cas, c'est que la relation entre X et Y n'est pas affine, il faut alors trouver une autre formule. Tapez :

   ```
   scf(1), plot2d(x, y-(a*x+b),style=-1), scf(0)
   ```

   Conclure.
5. Pour ajuster au moyen d'une formule qui n'est pas affine, il faut avoir une idée de cette formule.
   Dans cet exemple, la formule exprimant Y en fonction de X est $Y = X^2 + N$, il est donc raisonnable de poser le changement de variable $Z = X^2$ et chercher un ajustement affine de la forme $Y = aZ + b$.
   Modifier le programme précédent pour répondre aux mêmes questions mais avec le couple (Z, Y) au lieu du couple (X, Y) (ps : fermer les fenêtres graphiques avant d'exécuter). Commenter les résultats obtenus.

Exercice 4.
Cet exercice a pour but de montrer que lorsqu'on manipule des données *réelles*, les choses sont moins simples.
Les tableaux suivants fournissent des données arrondies sur la France métropolitaine :

A	1950	1956	1960	1965	1969	1973	1979	1985	1990	2000	2005	2006	2008	2009	2010	2011	201
W	63	85	90	115	145	180	193	202	229	269	277	276	273	261	263	265	259
G	30	41	50	66	80	100	120	132	154	188	203	208	213	206	210	214	214

où A est l'année, W est la consommation d'énergie primaire en TEP et G est le PIB en volume.

1. Tracez le nuage de points associé à (W, G).
   Une partie des points est bien alignée et on a un "petit tas" de points.
   Rajouter 'instruction `plot2d(W,G,style=5)` : elle permet d'observer l'évolution temporelle. On constate qu'à partir de l'an 2000, la contrainte énergétique semble avoir précédé l'arrêt de la croissance.
2. En ne considérant que les données entre 1950 et 2000, tracez la droite de régression associée au couple (W, G) et affichez la valeur du coefficient de corrélation, faites tracer cette droite. Interprétez.
3. Etudiez sur une nouvelle figure le nuage de points associé à la série chronologique $(A, G/W)$.
   On voit que l'efficacité énergétique G/W a fortement augmenté entre 1973 et 1985 et depuis l'an 2000.

3. Tableau de contingence

Exercice 5.
On considère X une variable de Poisson de paramètre 2 et Y une variable aléatoire dont la loi conditionnelle sachant $[X = k]$ est une loi de Poisson de paramètre k.

1. Dans le programme suivant, compléter la fonction `simul_couple` qui effectue une simulation du couple (X, Y) et renvoit les valeurs x et y obtenues :

```
function [x,y]=simul_couple()
    x=grand(****,****,****,****)
    y=grand(****,****,****,****)
endfunction

x=zeros(1,20)
y=zeros(1,20)
for k=1:20 do
    [x(k),y(k)]=simul_couple()
end

disp([x;y])
```

   Expliquer les différentes instructions du programme puis l'exécuter.

2. Sans le taper, trouver ce que fait le programme suivant et donner la matrice F obtenue (on fera tourner le programme à la main en utilisant le tableau affiché ci-avant) :

```
l=max(x)
m=max(y)
N=zeros(l+1,m+1)
for i=1:20 do
   N(x(i)+1, y(i)+1)=N(x(i)+1,y(i)+1)+1
end
F=N/20
disp(F)
```

   Vérifier le résultat en tapant les instructions.

   On a alors obtenu le tableau de la **loi empirique** du couple (X, Y). si on effectuait un grand nombre de simulations (10000 au lieu de 20), cette loi empirique serait très proche de la loi théorique (loi de probabilité) du couple (X, Y).

Bilan : Comme pour les statistiques univariées, il y a deux façons de donner des données statistiques :
- Soit sous forme de données brutes :
Exemple

x	0	1	1	2	2	1	2	2	1	1
y	4	6	8	8	8	4	6	6	8	6

- Soit sous forme groupée : sous forme d'un tableau de contingence qui donne uniquement les valeurs (modalités) <u>distinctes</u> prises par X, les valeurs (modalités) <u>distinctes</u> prises par Y et l'effectif (ou la fréquence) de chacun des couples de valeurs :
Exemple
Les données ci-dessus sont regroupées ainsi :

$x \setminus y$	4	6	8
0	1	0	0
1	1	2	2
2	0	2	2

Il faut être capable de reconnaître un programme qui range des données brutes dans un tableau de contingence (cf exo 5).
On peut alors créer le tableau de contingence des fréquences en divisant les effectifs par l'effectif total `F= N/sum(N)` :

$x \setminus y$	4	6	8
0	0.1	0	0
1	0.1	0.2	0.2
2	0	0.2	0.2