TD 1 : Séries statistiques et séries statistiques doubles

Exercice 1. *Neuf étudiants émettent un avis pédagogique vis-à-vis d'un professeur selon une échelle d'appréciation de* 0 *à* 20*. On relève par ailleurs la note, sur* 20*, obtenue par ces étudiants dans la matière dispensée par ce professeur :*

X=Note	8	11	10	13	9	17	7	15	16
Y=Avis	5	7	16	6	12	14	10	9	8

1. *Représenter graphiquement le nuage de points de la série statistique double* (X, Y).
2. *Calculer le coefficient de corrélation linéaire* $r(X, Y)$.
3. *Déterminer la droite des moindres carrés de* Y *en fonction de* X *et compléter le graphique précèdent.*
4. *Prédire l'avis moyen d'un étudiant ayant obtenu la note* $12/20$.

Exercice 2. *Le fichier « tips.txt »contient, en particulier, les pourboires et additions recueillis par un serveur américain auprès de 244 groupes de consommateurs. On souhaite modéliser le pourboire perçu en fonction de l'addition.*

1. *En utilisant le logiciel R, charger le jeu de données « tips.txt »via l'onglet « import dataset ».*
2. *On pose* $(X, Y) = \{(x_1, y_1), \ldots, (x_{244}, y_{244})\}$ *la série statistique double où* X *représente l'addition et* Y *représente le pourboire. Le code R suivant permet de calculer la droite des moindres carrés de cette série :*

```
X=tips$TOTBILL
Y=tips$TIP
covXY=mean(X*Y)-mean(X)*mean(Y)
varX=mean(X^2)-mean(X)^2
a=covXY/varX
b=mean(Y)-a*mean(X)
plot(X,Y)
curve(a*x+b,add=TRUE,col="red")
```

3. *Prédire le pourboire moyen du serveur lorsque l'addition est de* 30$.
4. *On considère la série statistique double pondérée* $(X, Y) = \{(x_1, y_1, p_1), \ldots, (x_{244}, y_{244}, p_{244})\}$ *où le poids est la fraction du nombre de convives d'un groupe par le nombre total de clients servi. Le code R suivant donne les poids :*

```
N=tips$SIZE
Poids=N/sum(N)
```

Modifier le code pour prendre en compte la pondération.

Exercice 3. *Soit* $X = \{(x_1, p_1), \ldots, (x_n, p_n)\}$ *une série statistique pondérée et* $a \in \mathbb{R}$.

1. *Quelle est la formule pour la moyenne quand on change d'unité la série* X*, c'est-à-dire lorsque* X *devient* aX *? Donner et démontrer la formule.*
2. *Quelle est la formule pour la moyenne quand on translate la série* X*, c'est-à-dire lorsque* X *devient* $a + X$ *? Donner et démontrer la formule.*

3. *Quelle est la formule de la variance pour aX ? Donner et démontrer la formule.*

4. *Quelle est la formule de la variance pour $a + X$? Donner et démontrer la formule.*

Soit $(X, Y) = \{((x_1, y_1), p_1), \ldots, ((x_n, y_n), p_n)\}$ une série statistique double pondérée et $b \in \mathbb{R}$.

5. *Quelle est la formule du coefficient de corrélation linéaire entre aX et bY (avec $a \neq 0$ et $b \neq 0$) ? Donner et démontrer la formule.*

6. *Quelle est la formule du coefficient de corrélation linéaire entre $a + X$ et $b + Y$? Donner et démontrer la formule.*

Exercice 4. *n individus sont classés de 1 à n selon deux critères. Pour les deux classements, on suppose qu'il n'y a pas d'ex aequo. On note $(X, Y) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ la série statistique double où, x_i, y_i représentent les rangs de l'individu i pour le premier critère et second critère.*

1. *Montrer l'égalité suivante*

$$cov(X, Y) = v(X) - \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2.$$

2. *En déduire que le coefficient de corrélation linéaire des deux variables X et Y est donné par :*

$$r(X, Y) = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2 - 1)}.$$

Pour trouver cette identité on pourra également utiliser les relations suivantes

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \text{ et } \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Le coefficient $r(X, Y)$ est appelé coefficient des rangs de Spearman.

Exercice 5. *On rappelle que la droite des moindres carrés a pour coefficients :*

$$a^* = \frac{cov(X, Y)}{v(X)} \text{ pour la pente et } b^* = \overline{Y} - a^*\overline{X} \text{ pour l'ordonnée à l'origine .}$$

Redémontrer ces formules via du calcul différentiel.

Exercice 6. *(Ecrits CAPES 2013)*

Dans cette partie, n désigne un entier naturel non nul et $(x_1, ..., x_n)$, un n-uplet de réels. On définit sur $\mathbb{R}$ les deux fonctions G et L par :

$$G(x) = \sum_{i=1}^{n}(x - x_i)^2 \quad L(x) = \sum_{i=1}^{n}|x - x_i|.$$

1. **Minimisation de G**

   (a) *En écrivant $G(x)$ sous la forme d'un trinôme du second degré, démontrer que la fonction G admet un minimum sur $\mathbb{R}$ et indiquer pour quelle valeur de x il est atteint.*

   (b) *Que représente d'un point de vue statistique la valeur de x trouvée à la question précédente ?*

2. **Minimisation de L** *On supposera dans cette question que la série est ordonnée, c'est-à-dire que : $x_1 \leq x_2 \leq \cdots \leq x_n$.*

   (a) *Représenter graphiquement la fonction L dans le cas où : $n = 3$, $x_1 = -2$, $x_2 = 3$, $x_3 = 4$.*

(b) *Représenter graphiquement la fonction L dans le cas où : $n = 4$, $x_1 = -2$, $x_2 = 2$, $x_3 = 4$, $x_4 = 7$.*

(c) *Démontrer que la fonction L admet un minimum m sur $\mathbb{R}$ et indiquer pour quelle(s) valeur(s) de x il est atteint.On distinguera les cas n pair et n impair.*

(d) *Que représentent d'un point de vue statistique les valeurs de x trouvées à la question précédente ?*