IUT Dijon-Auxerre
GEA 1ère année TD4, 2016-2017, S1

# Corrigé du Contrôle Continu n° 1

**Exercice 1 :**

1. La population étudiée est l'ensemble des ménages ayant répondu à l'enquête et la variable statistique étudiée est la quantité $X$ d'électricité consommée par an par le ménage (en KWh). Il s'agit d'une variable quantitative.
2. On dresse le tableau statistique complet de cette série en notant $N = 310$ l'effectif total et pour $i \in \{1, \dots, 5\}$ :
   — $C_i = [x_i; x_{i+1}[$ la $i^{\text{ème}}$ classe,
   — $c_i = \frac{x_i + x_{i+1}}{2}$ le centre de la $i^{\text{ème}}$ classe,
   — $a_i = x_{i+1} - x_i$ l'amplitude de la $i^{\text{ème}}$ classe,
   — on choisit (arbitrairement) de normaliser par 1000, de sorte que $a_i^r = \frac{a_i}{1000}$,
   — $n_i$ l'effectif de la $i^{\text{ème}}$ classe,
   — $f_i = \frac{n_i}{N}$ la fréquence de la $i^{\text{ème}}$ classe,
   — $n_i^r$ l'effectif relatif de la $i^{\text{ème}}$ classe,
   — $f_i^r = \frac{n_i^r}{N}$ la fréquence relative de la $i^{\text{ème}}$ classe,
   — $N_i$ l'effectif cumulé croissant jusqu'à la $i^{\text{ème}}$ classe,
   — $F_i = \frac{N_i}{N}$ la fréquence cumulée croissante jusqu'à la $i^{\text{ème}}$ classe.

| $i$ | $C_i$ | $c_i$ | $a_i$ | $a_i^r$ | $n_i$ | $f_i$ | $n_i^r$ | $f_i^r$ | $N_i$ | $F_i$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | $[2000; 4000[$ | 3000 | 2000 | 2 | 50 | 0,1613 | 25 | 0,0806 | 50 | 0,1613 |
| 2 | $[4000; 5000[$ | 4500 | 1000 | 1 | 75 | 0,2419 | 75 | 0,2419 | 125 | 0,4032 |
| 3 | $[5000; 5500[$ | 5250 | 500 | 0,5 | 69 | 0,2226 | 138 | 0,4452 | 194 | 0,6258 |
| 4 | $[5500; 6000[$ | 5750 | 500 | 0,5 | 66 | 0,2129 | 132 | 0,4258 | 260 | 0,8387 |
| 5 | $[6000; 9000[$ | 7500 | 3000 | 3 | 50 | 0,1613 | 16,6667 | 0,0538 | 310 | 1 |

   L'histogramme des fréquences de cette série statistique est représenté dans la Figure 1.
3. La fonction de répartition $F$ de cette série statistique représente la proportion de valeurs inférieures ou égales à une valeur prescrite $x$ ou encore la probabilité empirique $F(x) = \mathbf{P}[X \leq x]$ d'avoir observé une valeur inférieure ou égale à $x$. Sa courbe représentative interpole linéairement les points $(x_1; 0), (x_2; F_1), \dots$. Elle est représentée dans la Figure 2.
4. (a) Le pourcentage de ménages consommant plus de 5000 KWh par an est :

$$\mathbf{P}[X > 5000] = 1 - \mathbf{P}[X \leq 5000] = 1 - F(5000) \simeq 1 - 0,4032 = 0,5968 = 59,68\%.$$

   (b) Le pourcentage de ménages consommant au moins 5500 KWh mais moins de 7000 KWh par an est :

$$\mathbf{P}[5500 < X \leq 7000] = \mathbf{P}[X \leq 7000] - \mathbf{P}[X \leq 5500] = F(7000) - F(5500).$$

   La valeur de $F(5500)$ se lit directement dans le tableau statistique de la variable $X$. On détermine $F(7000)$ par interpolation linéaire. Clairement, 7000 est dans la classe

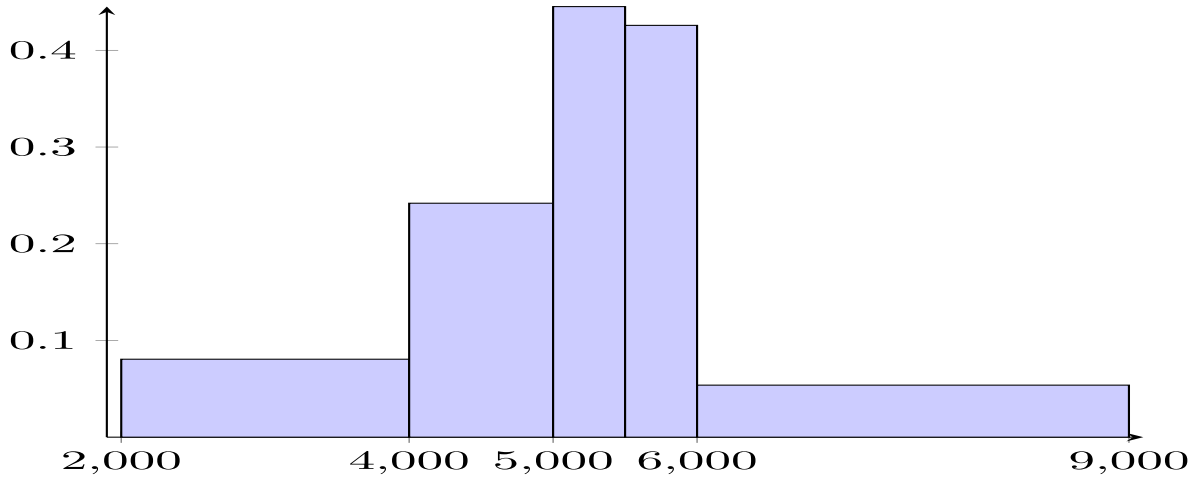


FIGURE 1 – Histogramme des fréquences relatives.

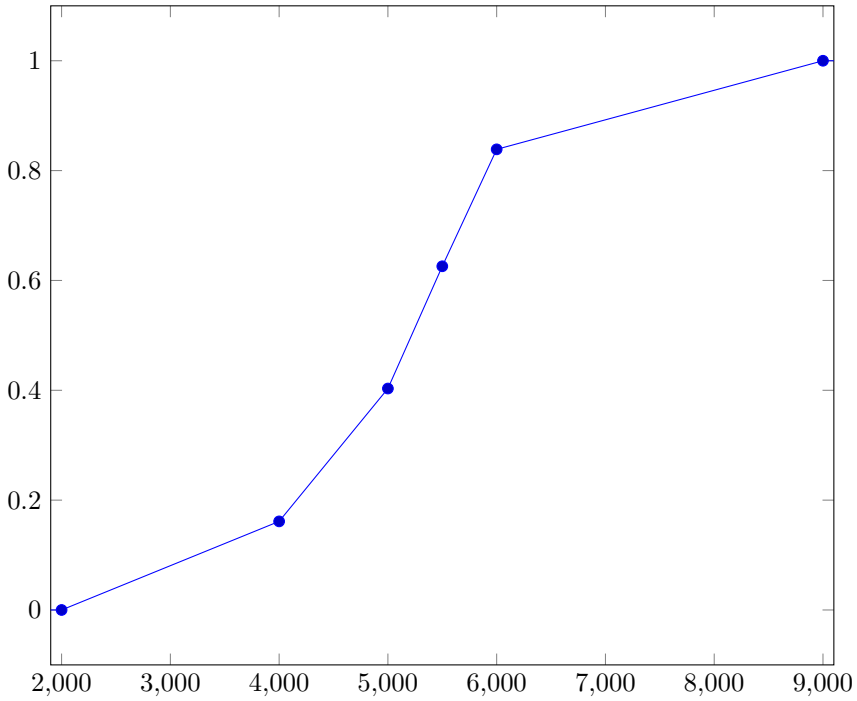


FIGURE 2 – Fonction de répartition de $X$.

$[6000; 9000[$. On en déduit par interpolation linéaire que :

$$\frac{F(7000) - F(6000)}{F(9000) - F(6000)} = \frac{7000 - 6000}{9000 - 6000} = \frac{1}{3}$$

puis que

$$F(7000) = \frac{1}{3}\left(F(9000) - F(6000)\right) + F(6000) \simeq \frac{1}{3}\left(1 - 0,8387\right) + 0,8387 \simeq 0,8925.$$

Finalement, le pourcentage de ménages consommant au moins 5500 KWh mais moins

de 7000 KWh par an est :

$$\mathbf{P}\left[5500 < X \leq 7000\right] = F(7000) - F(5500) \simeq 0,8925 - 0,6258 = 0,2667 = 26,67\%.$$

5. Les paramètres de cette série statistique sont les suivants.
   (a) La classe modale est la classe ayant la fréquence relative la plus élevée ; il s'agit ici de la classe $C_3 = [5000; 5500[$.
   (b) La moyenne (arithmétique) est donnée par :

$$\overline{X} = \frac{n_1 c_1 + \cdots + n_5 c_5}{N} \simeq 5175.$$

   (c) La médianne $me$ est la valeur telle que $F(me) = \mathbf{P}[X \leq me] = 0,5$. On a $F(5000) = 0,4032$ et $F(5500) = 0,6258$. Ainsi, cette valeur est dans la classe $C_3 = [5000; 5500[$. On détermine $me$ par interpolation linéaire :

$$\frac{me - 5000}{5500 - 5000} = \frac{F(me) - F(5000)}{F(5500) - F(5000)}$$

   soit

$$me = 500 \times \frac{0,5 - 0,4032}{0,6258 - 0,4032} + 5000 \simeq 5217,4304.$$

   (d) La variance est donnée par :

$$V[X] = \frac{n_1(c_1 - \overline{X})^2 + \cdots + n_5(c_5 - \overline{X})^2}{N} = \frac{n_1 c_1^2 + \cdots + n_5 c_5^2}{N} - \overline{X}^2 \simeq 1816754,032.$$

   (e) L'écart-type est donné par :

$$\sigma = \sqrt{V[X]} \simeq 1347,8702.$$

   (f) L'écart-type relatif (ou coefficient de variation) est donné par :

$$C_V = \frac{\sigma}{\overline{X}} \simeq 0,2605.$$

   (g) L'écart-absolu moyen est donné par :

$$EAM = \frac{n_1|c_1 - \overline{X}| + \cdots + n_5|c_5 - \overline{X}|}{N} \simeq 1028,2258.$$

6. La distribution est assez symétrique. La médianne et la moyenne sont du même ordre de grandeur.

**Exercice 2 :** Pour déterminer le coefficient de Cramér de la série, on commence par compléter le tableau de contingence avec les effectifs marginaux observés $n_{i\cdot}$ en $X$ et $n_{\cdot j}$ en $Y$.

| $X$ \ $Y$ | 3 | 5 | 6 | Eff. marg. $n_{i\cdot}$ en $X$ |
|---|---|---|---|---|
| blanche | 12 | 34 | 22 | 68 |
| jaune | 6 | 19 | 10 | 35 |
| rose | 4 | 13 | 8 | 25 |
| violette | 3 | 7 | 5 | 15 |
| Eff. marg. $n_{\cdot j}$ en $Y$ | 25 | 73 | 45 | 143 |

On calcul ensuite les effectifs théoriques $T_{ij}$ que l'on obtiendrait si les variables $X$ et $Y$ étaient indépendantes et avec les mêmes marginales. Ceci se fait à l'aide de la formule :

$$T_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

où $N = 143$ est l'effectif total. On obtient le tableau suivant pour les $T_{ij}$ :

| $X$ \ $Y$ | 3 | 5 | 6 | Eff. marg. $n_{i\cdot}$ en $X$ |
|---|---|---|---|---|
| blanche | 11,8881 | 34,7133 | 21,3986 | 68 |
| jaune | 6,1189 | 17,8671 | 11,0140 | 35 |
| rose | 4,3706 | 12,7622 | 7,8671 | 25 |
| violette | 2,6224 | 7,6573 | 4,7203 | 15 |
| Eff. marg. $n_{\cdot j}$ en $Y$ | 25 | 73 | 45 | 143 |

On détermine, ensuite la distance du Chi-2 entre les deux tableaux par la formule :

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \simeq 0,3656.$$

En utilisant que le nombre de lignes est $l = 4$ et le nombre de colonnes est $r = 3$, on obtient que :

$$\chi^2_{\max} = N \times \min(l-1; r-1) = 143 \times \min(3; 2) = 286.$$

puis que le coefficient de Cramér de la série est

$$C = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} \simeq 0,0358.$$

Ce coefficient étant très proche de 0, on en déduit que la couleur d'une fleur de lys et son nombre de pétales sont deux caractères indépendants.

**Exercice 3 :**

1. Le nuage de points est représenté dans la Figure 3. On observe que les points semblent se répartir le long d'une droite. L'utilisation d'un modèle de régression linéaire est donc approprié pour relier le nombre de jours de pluie $X$ dans l'année et la hauteur cumulée de précipitation $Y$ relevée la même année.
2. Déterminons le coefficient de corrélation linéaire. On a ici :

$$\overline{X} = \frac{X_1 + \cdots + X_{10}}{10} = 113,1$$

$$\overline{Y} = \frac{Y_1 + \cdots + Y_{10}}{10} = 835,33$$

$$V[X] = \frac{(X_1 - \overline{X})^2 + \cdots + (X_{10} - \overline{X})^2}{10} = 576,09$$

$$V[Y] = \frac{(Y_1 - \overline{X})^2 + \cdots + (Y_{10} - \overline{Y})^2}{10} = 6470,3061$$

$$Cov(X,Y) = \frac{(X_1 - \overline{X})(Y_1 - \overline{Y}) + \cdots + (X_{10} - \overline{X})(Y_{10} - \overline{Y})}{10} = 1758,597$$

et donc

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{V[X]V[Y]}} \simeq 0,9109.$$

On a $|Cor(X,Y)| \geq 0,8$ ; le modèle de régression linéaire est donc légitime ici.

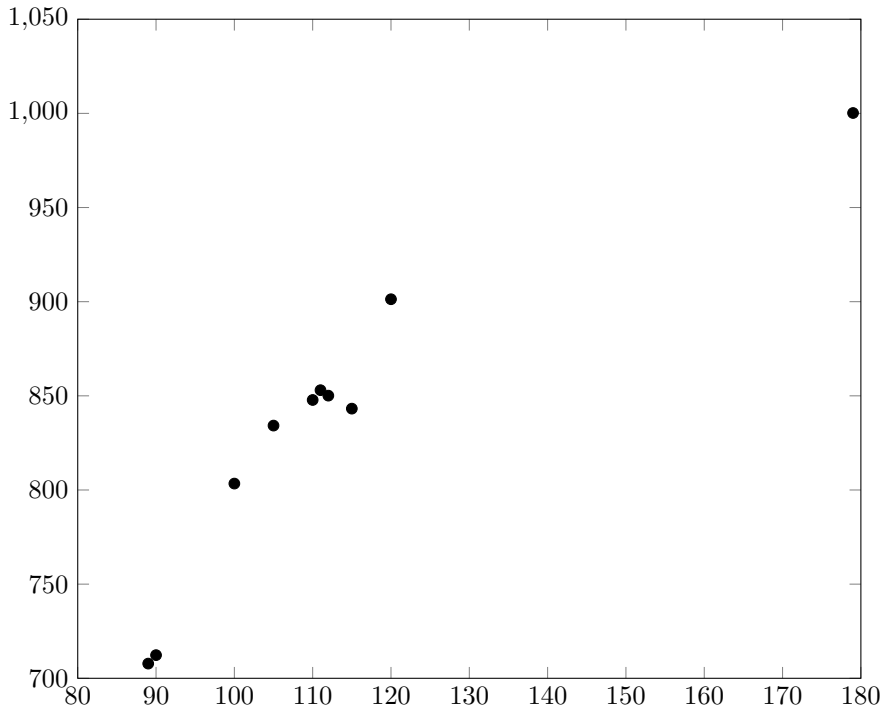


FIGURE 3 – Nuage de points représentant la série bivariée de l'Exercice 3.

3. La droite de régression expliquant le cumul annuel des précipitations en fonction du nombre de jours de pluie est la droite de régression de $Y$ en $X$. Elle admet pour équation :

$$y = ax + b,$$

avec

$$a = \frac{Cov(X,Y)}{V[X]} \simeq 3,0526 \quad \text{et} \quad b = \overline{Y} - a\overline{X} \simeq 490,0761.$$

4. Puisqu'en 2005 on a observé $x_{2005} = 130$ jours de pluie, la droite de régression déterminée dans la question précédente permet d'estimer que la hauteur d'eau mesurée en 2005 a été d'environ :

$$y_{2005} = ax_{2005} + b \simeq 3,0526 \times 130 + 490,0761 \simeq 886,9197\text{mm}.$$