IUT Dijon-Auxerre
GEA 1ère année TD3, 2016-2017, S1

Corrigé du Contrôle Continu n° 1

Exercice 1 :

1. La population étudiée est l'ensemble des employés de l'entreprise et la variable statistique étudiée est la distance X séparant leurs domiciles respectifs de leur lieu de travail. Il s'agit d'une variable quantitative.
2. On dresse le tableau statistique complet de cette série en notant $N = 130$ 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 10, de sorte que $a_i^r = \frac{a_i}{10}$,
   — 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	$[1;2[$	1,5	1	0,1	3	0,0231	30	0,2308	3	0,0231
2	$[2;5[$	3,5	3	0,3	24	0,1846	80	0,6154	27	0,2077
3	$[5;10[$	7,5	5	0,5	50	0,3846	100	0,7692	77	0,5923
4	$[10;30[$	20	20	2	51	0,3923	25,5	0,1962	128	0,9846
5	$[30;100[$	65	70	7	2	0,0154	0,2857	0,0022	130	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 d'employés habitant à plus de 10 Km de leur lieu de travail est :

$$\mathbf{P}[X > 10] = 1 - \mathbf{P}[X \leq 10] = 1 - F(10) \simeq 1 - 0,5923 = 0,4077 = 40,77\%.$$

   (b) Le pourcentage d'employés habitant à au moins 2 Km mais au plus à 20Km de leur lieu de travail est :

$$\mathbf{P}[2 < X \leq 20] = \mathbf{P}[X \leq 20] - \mathbf{P}[X \leq 2] = F(20) - F(2).$$

   La valeur de $F(2)$ se lit directement dans le tableau statistique de la variable X. On détermine $F(20)$ par interpolation linéaire. Clairement, 20 est dans la classe $[10; 30[$.

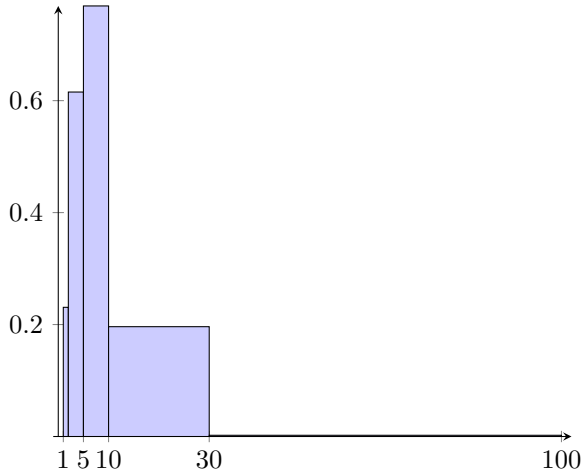

FIGURE 1 – Histogramme des fréquences relatives.

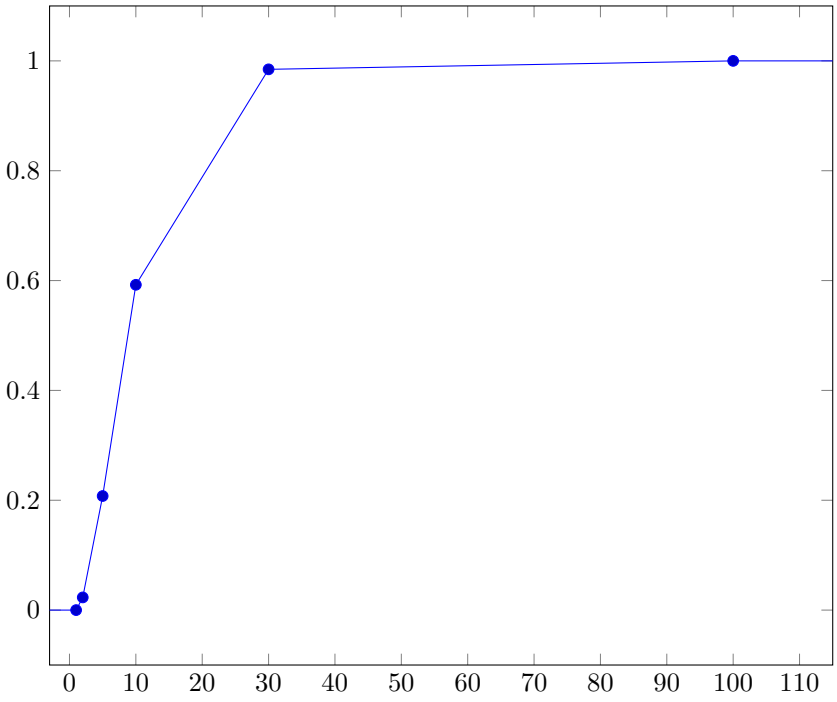

FIGURE 2 – Fonction de répartition de X.

On en déduit par interpolation linéaire que :

$$\frac{F(20) - F(10)}{F(30) - F(10)} = \frac{20 - 10}{30 - 10} = \frac{1}{2}$$

puis que

$$F(20) = \frac{1}{2}\left(F(30) - F(10)\right) + F(10) \simeq \frac{1}{2}\left(0,9846 - 0,5923\right) + 0,5923 \simeq 0,7885.$$

Finalement, le pourcentage d'employés habitant à au moins 2 Km mais au plus à 20Km

de leur lieu de travail est :

$$\mathbf{P}\left[2 < X \leq 20\right] = F(20) - F(2) \simeq 0,7885 - 0,0231 = 0,7654 = 76,54\%.$$

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 $C_3 = [5; 10[$.
   (b) La moyenne (arithmétique) est donnée par :

$$\overline{X} = \frac{n_1 c_1 + \cdots + n_5 c_5}{N} \simeq 12,4115.$$

   (c) La médianne me est la valeur telle que $F(me) = \mathbf{P}[X \leq me] = 0,5$. On a $F(5) = 0,2077$ et $F(10) = 0,5923$. Ainsi, cette valeur est dans la classe $C_3 = [5; 10[$. On détermine me par interpolation linéaire :

$$\frac{me - 5}{10 - 5} = \frac{F(me) - F(5)}{F(10) - F(5)}$$

   soit

$$me = 5 \times \frac{0,5 - 0,2077}{0,5923 - 0,2077} + 5 \simeq 8,8001$$

   (d) La variance est donnée par :

$$V[X] = \frac{n_1(c_1 - \overline{X})^2 + \cdots + n_5(c_5 - \overline{X})^2}{N} = \frac{n_1 c_1^2 + \cdots + n_5 c_5^2}{N} - \overline{X}^2 \simeq 91,8248.$$

   (e) L'écart-type est donné par :

$$\sigma = \sqrt{V[X]} \simeq 9,5825.$$

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

$$C_V = \frac{\sigma}{\overline{X}} \simeq 0,7721.$$

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

$$EAM = \frac{n_1|c_1 - \overline{X}| + \cdots + n_5|c_5 - \overline{X}|}{N} \simeq 7,5721.$$

6. La masse de la distribution est concentrée vers la gauche. La médianne est significativement inférieure à la moyenne. La distribution présente une queue étalée vers la droite.

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	de 9 mois à 2 ans	de 2 ans à 5 ans	plus de 5 ans	Eff. marg. $n_{i\cdot}$ en X
16-20 ans	37	2	0	39
20-25 ans	5	25	2	32
25-40 ans	2	15	20	37
plus de 40 ans	1	8	33	42
Eff. marg. $n_{\cdot j}$ en Y	45	50	55	150

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 $N = 150$ est l'effectif total. On obtient le tableau suivant pour les T_{ij} :

X \ Y	de 9 mois à 2 ans	de 2 ans à 5 ans	plus de 5 ans	Eff. marg. $n_{i\cdot}$ en X
16-20 ans	11,7	13	14,3	39
20-25 ans	9,6	10,6667	11,7333	32
25-40 ans	11,1	12,3333	13,5667	37
plus de 40 ans	12,6	14	15,4	42
Eff. marg. $n_{\cdot j}$ en Y	45	50	55	150

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

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{(n_{ij} - T_{ij})^2}{T_{ij}} \simeq 152,3078.$$

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

$$\chi^2_{\max} = N \times \min(l-1; r-1) = 150 \times \min(3;2) = 300$$

puis que le coefficient de Cramér de la série est

$$C = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} \simeq 0,7125.$$

Ce coefficient étant assez proche de 1, on déduit que la durée totale de location d'un appartement est plutôt liée à l'âge des occupants au début du contrat.

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 la pointure X d'un individu et sa taille Y.
2. Déterminons le coefficient de corrélation linéaire. On a ici :

$$\overline{X} = \frac{X_1 + \cdots + X_{10}}{10} = 41,3$$

$$\overline{Y} = \frac{Y_1 + \cdots + Y_{10}}{10} = 168,5$$

$$V[X] = \frac{(X_1 - \overline{X})^2 + \cdots + (X_{10} - \overline{X})^2}{10} = 7,81$$

$$V[Y] = \frac{(Y_1 - \overline{X})^2 + \cdots + (Y_{10} - \overline{Y})^2}{10} = 65,05$$

$$Cov(X,Y) = \frac{(X_1 - \overline{X})(Y_1 - \overline{Y}) + \cdots + (X_{10} - \overline{X})(Y_{10} - \overline{Y})}{10} = 20,15$$

et donc

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{V[X]V[Y]}} \simeq 0,8940.$$

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

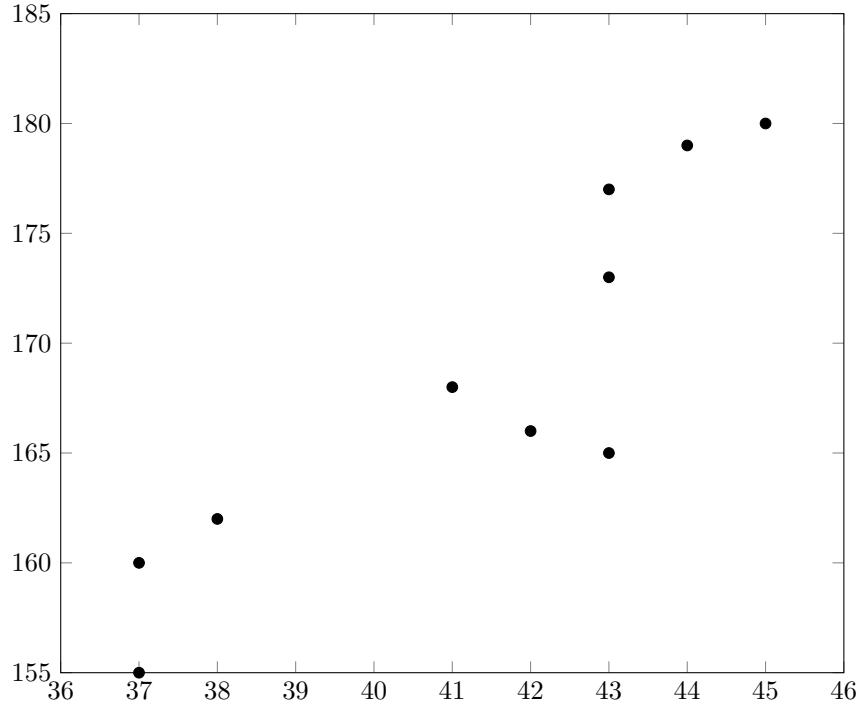

FIGURE 3 – Nuage de points représentant la série bivariée de l'Exercice 3.

3. La droite de régression expliquant la pointure d'un individu en fonction de sa taille est la droite de régression de X en Y. Elle admet pour équation :

$$x = ay + b,$$

avec

$$a = \frac{Cov(X,Y)}{V[Y]} \simeq 0,3098 \quad \text{et} \quad b = \overline{X} - a\overline{Y} \simeq -10,8949.$$

4. La droite de régression, déterminée dans la question précédente, permet de prédire pour un individu dont la taille est de $y^* = 164$ cm une pointure d'environ :

$$x^* = ay^* + b \simeq 0,3098 \times 164 - 10,8949 \simeq 40.$$