CHAPITRE 1

Contrôle par équipe du 30 mai 2022 : statistiques descriptives

Exercice 1.1
On considère deux groupes d'étudiants.
- Notes du groupe A : 8, 8, 9, 9, 10, 11.
- Notes du groupe B : 6, 6, 8, 8, 9, 9, 13, 14.

Comparer les deux groupes en calculant la moyenne, la médiane et l'écart type.

Correction

- Groupe A : $N = 6$ et $\mathscr{C} = \{8, 9, 10, 11\}$.

Valeur α_i	Effectif n_i	Fréquence f_i	Effectif cumulé	Fréquence cumulée
8	2	2/6	2	2/6
9	2	2/6	4	4/6
10	1	1/6	5	5/6
11	1	1/6	6	6/6
	$\sum_{i=1}^{p=4} n_i = 6$	$\sum_{i=1}^{p=4} f_i = 1$		

La moyenne est $\dfrac{8\times 2 + 9\times 2 + 10 + 11}{6} = \dfrac{55}{6} \approx 9.1667$.

La médiane est $\dfrac{9+9}{2} = 9$.

L'écart-type est $\sqrt{\dfrac{\sum_{i=1}^{n=6} x_i^2}{6} - \bar{\mathbf{x}}^2} = \sqrt{\dfrac{2\times 8^2 + 2\times 9^2 + 10^2 + 11^2}{6} - \bar{\mathbf{x}}^2} \approx 8.7178$.

- Groupe B : $N = 8$ et $\mathscr{C} = \{6, 8, 9, 13, 14\}$.

Valeur α_i	Effectif n_i	Fréquence f_i	Effectif cumulé	Fréquence cumulée
6	2	2/8	2	2/8
8	2	2/8	4	4/8
9	2	2/8	6	6/8
13	1	1/8	7	7/8
14	1	1/8	8	9/8
	$\sum_{i=1}^{p=6} n_i = 8$	$\sum_{i=1}^{p=6} f_i = 1$		

La moyenne est $\dfrac{6\times 2 + 8\times 2 + 9\times 2 + 13 + 14}{8} = \dfrac{73}{8} \approx 9.125$.

La médiane est $\dfrac{8+9}{2} = 8.5$.

L'écart-type est $\sqrt{\dfrac{\sum_{i=1}^{n=8} x_i^2}{6} - \bar{\mathbf{x}}^2} = \sqrt{\dfrac{2\times 6^2 + 2\times 8^2 + 2\times 9^2 + 13^2 + 14^2}{8} - \bar{\mathbf{x}}^2} \approx 9.0416$.

On remarque que les moyennes et les médianes sont très proches. Cependant on ne peut pas pour autant conclure que ces deux groupes ont des niveaux identiques. En effet, après le calcule des écarts type, on note que le groupe B est beaucoup plus dispersé que le groupe A (les étudiants de ce groupe ont des notes plus irréguliers ; on peut dire donc que le groupe B est moins homogènes que le groupe A).

Exercice 1.2 (Distribution statistique groupée)

Considérons le tableau suivant :

Budget	Fréquence	Fréquence cumulée
]6;8]	0.08	0.08
]8;12]	0.1	0.18
]12;14]	0.16	0.34
]14;z]	0.1	0.44
]z;22]	0.3	0.74
]22;38]	0.26	1.0

1. Estimer la borne z manquante dans les deux cas suivants :
   1.1. le budget moyen est égal à 18.64 euros,
   1.2. le budget médian est égal à 18.80 euros.
2. Considérons dorénavant que la borne manquante est égale à 20. Estimer le budget moyenne et médian.

Correction

1. Calcule de la borne z.
   1.1. Comme on ne dispose que du tableau des fréquences, alors on estime la moyenne par la formule

$$\bar{\mathbf{x}} \approx \sum_{i=1}^{p} f_i \frac{\alpha_i + \alpha_{i+1}}{2} = \frac{14}{25} + 1 + \frac{52}{25} + \frac{z+14}{20} + \frac{3(z+22)}{20} + \frac{39}{5},$$

   où $\frac{\alpha_i+\alpha_{i+1}}{2}$ est le centre de la i-ème classe et f_i sa fréquence. Si le budget moyen $\bar{\mathbf{x}}$ est égal à $18.64 = \frac{466}{25}$ euros alors on a

$$18.64 = \frac{466}{25} \approx \frac{z}{5} + \frac{386}{25}$$

   ce qui donne $z \approx 16$ euros.
   1.2. Si le budget médian est égal à $18.80 = \frac{94}{5}$ euros, on regarde le tableau des fréquences cumulées et on voit que la médiane est quelque part dans l'intervalle]z;22]. Par interpolation linéaire sur cet intervalle (on impose le passage par les points $(z, 0.44)$ et $(22, 0.74)$) on trouve :

$$y_{\text{fréq. cum.}} = \frac{0.74 - 0.44}{22 - z}(x_{\text{budget}} - z) + 0.44$$

   ainsi, si $y_{\text{fréq. cum.}} = 0.5$ et $x_{\text{budget}} = 18.80 = \frac{94}{5}$, on trouve $z = 18$ euros.
2. Considérons dorénavant que $z = 20$ euros.
   2.1. Le budget moyen est estimé par

$$\bar{\mathbf{x}} \approx \sum_{i=1}^{p} f_i \frac{\alpha_i + \alpha_{i+1}}{2} = \frac{14}{25} + 1 + \frac{52}{25} + \frac{17}{10} + \frac{63}{10} + \frac{39}{5} = \frac{486}{25} = 19.44.$$

   2.2. Le budget médian est quelque part dans l'intervalle]20;22]. Par interpolation linéaire sur cet intervalle (on impose le passage par les points $(20, 0.44)$ et $(22, 0.74)$) on trouve :

$$y_{\text{fréq. cum.}} = \frac{0.74 - 0.44}{22 - 20}(x_{\text{budget}} - 20) + 0.44$$

   ainsi, si $y_{\text{fréq. cum.}} = 0.5$, $x_{\text{budget}} = \frac{102}{5} = 20.4$ euros.

Exercice 1.3

Une série statistique est définie partiellement dans le tableau ci-dessous.

x_k :	2	3	4	5	6
y_k :	26	y_1	y_2	y_3	45

On sait aussi que l'équation de la droite de régression de y en fonction de x est

$$y = \frac{22}{5} + \frac{22}{5}x = 4.4 + 4.4x.$$

1. En déduire $\bar{\mathbf{x}}$ et $\bar{\mathbf{y}}$.
2. Si toutes les valeurs de x augmentent de 2, quelle est l'équation de la nouvelle droite de régression de y en fonction de x ?
3. Si toutes les valeurs de y augmentent de 2, quelle est l'équation de la nouvelle droite de régression de y en fonction de x ?

Correction

1. On a $n = 5$ ainsi

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n} x_k = 4.$$

On sait que le point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ appartient à la droite de régression qui a pour équation $y = \gamma_0 + \gamma_1 x$ avec $\gamma_0 = \frac{22}{5}$ et $\gamma_1 = \frac{22}{5}$, donc

$$\bar{\mathbf{y}} = \gamma_0 + \gamma_1 \bar{\mathbf{x}} = 22.$$

2. Si toutes les valeurs de x augmentent de 2, alors
   - la moyenne $\bar{\mathbf{x}}_{\text{new}} = \frac{1}{n}\sum_{k=1}^{n}(x_k + 2) = \bar{\mathbf{x}}_{\text{old}} + 2$ augmente de 2,
   - la variance $V(\mathbf{x}_{\text{new}}) = \frac{1}{n}\sum_{k=1}^{n}((x_k + 2) - \bar{\mathbf{x}}_{\text{new}})^2 = V(\mathbf{x}_{\text{old}})$ ne change pas,
   - la covariance $C(\mathbf{x}_{\text{new}}, \mathbf{y}) = \frac{1}{n}\sum_{k=1}^{n}((x_k + 2) - \bar{\mathbf{x}}_{\text{new}})(y_k - \bar{\mathbf{y}}) = C(\mathbf{x}_{\text{old}}, \mathbf{y})$ ne change pas,
   - la pente $(\gamma_1)_{\text{new}}$ de la nouvelle droite de régression ne change pas non plus car

$$(\gamma_1)_{\text{new}} = \frac{C(\mathbf{x}_{\text{new}}, \mathbf{y})}{V(\mathbf{x}_{\text{new}})} = \frac{C(\mathbf{x}_{\text{old}}, \mathbf{y})}{V(\mathbf{x}_{\text{old}})} = \gamma_1 = \frac{22}{5};$$

   - comme la moyenne $\bar{\mathbf{x}}_{\text{new}}$ a changé, alors la nouvelle ordonnée à l'origine est

$$(\gamma_0)_{\text{new}} = \bar{\mathbf{y}} - \gamma_1 \bar{\mathbf{x}}_{\text{new}} = \gamma_0 + \gamma_1 \bar{\mathbf{x}}_{\text{old}} - \gamma_1 \bar{\mathbf{x}}_{\text{new}} = \gamma_0 - 2\gamma_1.$$

   En conclusion, l'équation de la nouvelle droite de régression de y en fonction de x est

$$y = (\gamma_0)_{\text{new}} + (\gamma_1)_{\text{new}} x = -\frac{22}{5} + \frac{22}{5}x.$$

3. De la même manière, si toutes les valeurs de y augmentent de 2, alors
   - la moyenne $\bar{\mathbf{y}}_{\text{new}} = \frac{1}{n}\sum_{k=1}^{n}(y_k + 2) = \bar{\mathbf{y}}_{\text{old}} + 2$ augmente de 2,
   - la covariance $C(\mathbf{x}, \mathbf{y}_{\text{new}}) = \frac{1}{n}\sum_{k=1}^{n}(x_k - \bar{\mathbf{x}})((y_k + 2) - \bar{\mathbf{y}}_{\text{new}}) = C(\mathbf{x}, \mathbf{y}_{\text{old}})$ ne change pas,
   - la pente $(\gamma_1)_{\text{new}}$ de la nouvelle droite de régression ne change pas non plus car

$$(\gamma_1)_{\text{new}} = \frac{C(\mathbf{x}, \mathbf{y}_{\text{new}})}{V(\mathbf{x})} = \frac{C(\mathbf{x}, \mathbf{y}_{\text{old}})}{V(\mathbf{x})} = \gamma_1 = \frac{22}{5};$$

   - comme la moyenne $\bar{\mathbf{y}}_{\text{new}}$ a changé, alors la nouvelle ordonnée à l'origine est

$$(\gamma_0)_{\text{new}} = \bar{\mathbf{y}}_{\text{new}} - \gamma_1 \bar{\mathbf{x}} = (\bar{\mathbf{y}}_{\text{old}} + 2) - (\bar{\mathbf{y}}_{\text{old}} - \gamma_0) = \gamma_0 + 2.$$

   En conclusion, l'équation de la nouvelle droite de régression de y en fonction de x est

$$y = (\gamma_0)_{\text{new}} + (\gamma_1)_{\text{new}} x = \frac{32}{5} + \frac{22}{5}x.$$

Exercice 1.4 (Distribution statistique bivariée)

On considère les données suivantes :

$$\{(x_i, y_i)\} = [(1,2), (2,2), (2,2), (2,3), (2,3), (1,3)].$$

Compléter les tableaux/valeurs/graphique suivants directement sur cette feuille.

1. Tableau de la distribution conjointe de deux variables quantitatives **x** et **y** :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$
$\alpha_1 = 1$	$n_{1,1} =$	$n_{1,2} =$
$\alpha_2 = 2$	$n_{2,1} =$	$n_{2,2} =$

2. Tableau des fréquences :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$	Fréquence marginale de α_i
$\alpha_1 = 1$	$f_{1,1} =$	$f_{1,2} =$	$f_{1,\cdot} =$
$\alpha_2 = 2$	$f_{2,1} =$	$f_{2,2} =$	$f_{2,\cdot} =$
Fréquence marginale de β_j	$f_{\cdot,1} =$	$f_{\cdot,2} =$	1

3. Tableau des profiles en colonne $f_{i|j}$:

Profiles en colonne $f_{i\|j}$		
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$
$\alpha_1 = 1$	$f_{1\|1} =$	$f_{1\|2} =$
$\alpha_2 = 2$	$f_{2\|1} =$	$f_{2\|2} =$
	1	1

4. Tableau des profiles en ligne $f_{j|i}$:

Profiles en ligne $f_{j\|i}$			
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$	
$\alpha_1 = 1$	$f_{1\|1} =$	$f_{1\|2} =$	1
$\alpha_2 = 2$	$f_{2\|1} =$	$f_{2\|2} =$	1

5. Calculer les moyennes, variances et covariances indiquées :

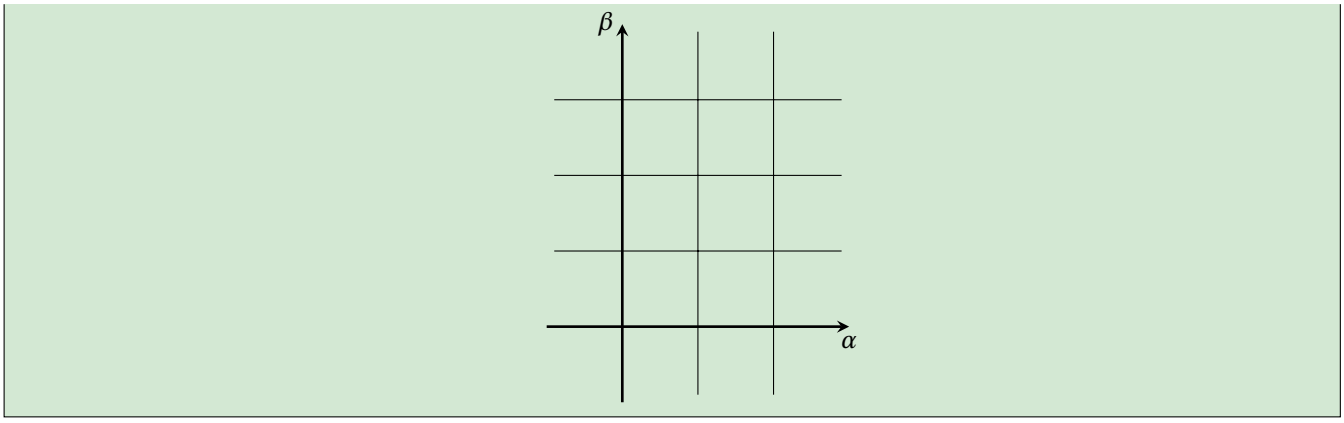
$\bar{\mathbf{x}} =$ $V(\mathbf{x}) =$ $C(\mathbf{x},\mathbf{y}) =$

$\bar{\mathbf{y}} =$ $V(\mathbf{y}) =$ $C(\mathbf{y},\mathbf{x}) =$

6. Calculer la droite de régression de $\mathbf{y} = \gamma_0 + \gamma_1 x$ et le coefficient de corrélation :

$\gamma_1 =$ $\gamma_0 =$ $r(\mathbf{x},\mathbf{y}) =$

7. Représenter la distribution conjointe sur un plan comme un nuage de points (chaque point avec son poids). Représenter aussi le point $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ et la droite de régression.



Correction

1. Tableau de la distribution conjointe de deux variables quantitatives **x** et **y** :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$
$\alpha_1 = 1$	$n_{1,1} = 1$	$n_{1,2} = 1$
$\alpha_2 = 2$	$n_{2,1} = 2$	$n_{2,2} = 2$

2. Tableau des fréquences :

$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$	Fréquence marginale de α_i
$\alpha_1 = 1$	$f_{1,1} = \frac{1}{6}$	$f_{1,2} = \frac{1}{6}$	$f_{1,\cdot} = \frac{1}{3}$
$\alpha_2 = 2$	$f_{2,1} = \frac{1}{3}$	$f_{2,2} = \frac{1}{3}$	$f_{2,\cdot} = \frac{2}{3}$
Fréquence marginale de β_j	$f_{\cdot,1} = \frac{1}{2}$	$f_{\cdot,2} = \frac{1}{2}$	1

3. Tableau des profiles en colonne $f_{i|j}$:

Profiles en colonne $f_{i\|j}$		
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$
$\alpha_1 = 1$	$f_{1\|1} = \frac{1}{3}$	$f_{1\|2} = \frac{1}{3}$
$\alpha_2 = 2$	$f_{2\|1} = \frac{2}{3}$	$f_{2\|2} = \frac{2}{3}$
	1	1

4. Tableau des profiles en ligne $f_{j|i}$:

Profiles en ligne $f_{j\|i}$			
$\mathscr{A}$ \ $\mathscr{B}$	$\beta_1 = 2$	$\beta_2 = 3$	
$\alpha_1 = 1$	$f_{1\|1} = \frac{1}{2}$	$f_{1\|2} = \frac{1}{2}$	1
$\alpha_2 = 2$	$f_{2\|1} = \frac{1}{2}$	$f_{2\|2} = \frac{1}{2}$	1

5. Calcule des moyennes, variances et covariances :

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{p} n_{i,.}\alpha_i = \frac{5}{3}$$
$$\bar{\mathbf{y}} = \frac{1}{n}\sum_{j=1}^{q} n_{.,j}\beta_j = \frac{5}{2}$$
$$V(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{p} n_{i,.}\alpha_i^2 - \bar{\mathbf{x}}^2 = \frac{2}{9}$$
$$V(\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{q} n_{.,j}\beta_j^2 - \bar{\mathbf{y}}^2 = \frac{1}{4}$$
$$C(\mathbf{x},\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{i,j}\alpha_i\beta_j - \bar{\mathbf{x}}\bar{\mathbf{y}} = 0$$
$$C(\mathbf{y},\mathbf{x}) = C(\mathbf{x},\mathbf{y})$$

6. Calcule de la droite de régression de $\mathbf{y}$ par rapport à $\mathbf{x}$ et du coefficient de corrélation r :

$$\gamma_1 = \frac{C(\mathbf{x},\mathbf{y})}{V(\mathbf{x})} = 0$$
$$\gamma_0 = \bar{\mathbf{y}} - \gamma_1\bar{\mathbf{x}} = \frac{5}{2}$$
$$r(\mathbf{x},\mathbf{y}) = \frac{C(\mathbf{x},\mathbf{y})}{\sqrt{V(\mathbf{x})V(\mathbf{y})}} = 0$$

La droite cherchée a donc pour équation $y = \gamma_0 + \gamma_1 x$ et le coefficient de corrélation est r.

7. On peut représenter la distribution conjointe sur un plan comme un nuage de points : chaque point correspond à un couple (α_i, β_j) affecté de son poids $n_{i,j}$, autrement dit chaque point correspond à une observation (x_k, y_k) et à coté on indique combien de fois cette observation apparaît. Il y aura donc $p \times q$ points (autant que de cases que dans le tableau de contingence), chaque point se trouvant sur un coin de la grille de coordonnées (α_i, β_j).