Statistique descriptive

Année académique 2020–2021

Carole.Baum@uliege.be

Chapitre 3 : Réduction des données

Exercice 1. (a) Calculer la moyenne arithmétique, la médiane, le mode et les premier et troisième quartiles des séries suivantes :

$$S_1 = \{3, 5, 2, 6, 5, 9, 5, 2, 8, 6\}$$

$$S_2 = \{84, 91, 72, 68, 87, 78, 52, 92, 71, 85, 62, 82, 99\}.$$

(b) Construire des séries possédant la même moyenne arithmétique, le même mode et la même médiane mais différant notamment par les effectifs et/ou les quartiles.

Solution :

(a) Pour la série S_1, on a

$$\bar{x}_1 = \frac{3+5+2+\ldots+6}{10} = 5.1 \qquad x_{M,1} = 5 \qquad \tilde{x}_1 = 5 \qquad Q_1 = 3 \qquad Q_3 = 6.$$

Pour la série S_1, on a

$$\bar{x}_1 = \frac{84+91+72+\ldots+99}{13} = 78.69 \qquad x_{M,2} = \nexists \qquad \tilde{x}_2 = 82 \qquad Q_1 = 71 \qquad Q_3 = 87.$$

(b) Par exemple, $S_1 = \{1, 2, 2, 2, 3\}$ et $S_2 = \{2, 2, 2\}$.

Exercice 2. Un grossiste dispose d'un stock important de pommes. Celles-ci sont réparties dans des caisses contenant chacune 12 pommes. La distribution du nombre de pommes de qualité supérieure par caisse est décrite dans le tableau suivant :

Nombre de pommes	0	1	2	3	4	5	6	7	8	9	10	11	12
Nombre de caisses	1	0	1	3	5	7	14	33	54	66	72	78	66

(a) En moyenne, combien y a-t-il de pommes de qualité supérieure par caisse ?

(b) Combien de pommes de qualité supérieure un client trouvera-t-il le plus fréquemment s'il achète, sans être suffisamment attentif, une caisse de pommes chez ce grossiste ?

(c) Si le grossiste décide de ne conserver que la moitié des caisses (évidemment celles contenant le plus possible de pommes de qualité supérieure), quel est le nombre minimum de pommes de qualité supérieure contenues dans ces caisses privilégiées ?

(d) Sachant que les amis du grossiste reçoivent en cadeau 10% des caisses (choisies parmi les meilleures), tandis que les 10% des caisses les moins bonnes sont gardées pour confectionner de la compote, déterminer le nombre moyen de pommes de qualité supérieure dans les caisses restantes.

Solution :

(a) Soit X le nombre de pommes de qualité supérieure, on a

$$\bar{x} = \frac{0 \cdot 1 + 1 \cdot 0 + 2 \cdot 1 + \ldots + 10 \cdot 66}{400} = 9.4425.$$

(b) Il s'agit du mode de la variable X, $x_M = 11$.

(c) En ne conservant que la moitié des caisses parmi les meilleures, ces caisses contiendront au minimum 10 pommes de qualité supérieure. En effet, on va garder les 66 caisses avec 12 pommes, les 78 caisses avec 11 pommes et 200-144=56 caisses parmi celles qui ont 10 pommes de qualité supérieure. Ce résultat est en réalité la médiane de la variable.

(d) Nous allons retirer (10% de 400) 40 caisses de chaque coté de la distribution. Parmi les caisses les meilleures, on retire 40 caisses avec 12 pommes de qualité supérieure. Parmi les moins bonnes, on retire toutes les caisses ayant au maximum 6 pommes de qualité supérieure (31 caisses) et 9 caisses avec 7 pommes de qualité supérieure.

La moyenne devient alors

$$\bar{x}_{0.1} = \frac{7 \cdot 24 + 8 \cdot 54 + \ldots + 12 \cdot 26}{320} = 9.6375.$$

Il s'agit en réalité d'une moyenne tronquée.

Exercice 3. Répondant à une offre d'emploi, une personne s'interroge sur le montant de ses rémunérations futures en cas d'embauche. Le directeur de l'entreprise de vente à domicile lui répond que le salaire moyen de la firme est supérieur à 1600 euros par mois, mais que, pendant la période de formation, l'employé ne gagnera que 500 euros chaque mois, puis sera augmenté dans la suite.

Avant de signer le contrat d'embauche, la personne a mené son enquête et a obtenu les renseignements suivants :

- Le directeur gagne 12500 euros par mois.
- Le sous-directeur gagne 6000 euros par mois.
- Chacun des 4 chefs de secteur gagne 1200 euros par mois.
- Chacun des 5 techniciens gagne 950 euros par mois.
- Chacun des 10 démarcheurs gagne 600 euros par mois.

(a) Le directeur a-t-il dit la vérité au candidat ?

(b) Quelle question le candidat aurait-il dû poser au patron pour avoir une estimation plus réaliste de son salaire futur ?

Solution :

(a) Le salaire moyen est

$$\bar{x} = \frac{12500 + 6000 + 4 \cdot 1200 + 5 \cdot 950 + 10 \cdot 600}{21} = \frac{34050}{21} = 1621.43€.$$

Le directeur a donc dit la vérité au candidat.

(b) Il aurait dû demander le salaire médian $\tilde{x} = x_{(11)} = 950€$, le salaire modal $x_M = 600€$ ou encore une moyenne pondérée où on retire la direction (directeur et sous-directeur) $\bar{x}_w = 818€$.

Exercice 4. Un village se compose de 4 quartiers. On connaît le nombre d'habitants par quartier et le nombre de véhicules par habitant.

Quartiers	Nbre d'habitants	Nbre de véhicules/habitant
A	1835	0,4583
B	1624	0,4883
C	729	0,5048
D	974	0,4774

Déterminer le nombre moyen de véhicules par habitant dans le village. De quelle moyenne s'agit-il ?

Solution :

Le nombre moyen de véhicules par habitant dans le village est donné par

$$\bar{x} = \frac{1835 \cdot 0.4583 + 1624 \cdot 0.4883 + 729 \cdot 0.5048 + 974 \cdot 0.4774}{1835 + 1624 + 729 + 974} = 0.4779$$

Il s'agit d'une moyenne pondérée.

Exercice 5. La série groupée des poids de 60 étudiants masculins de premier bachelier est décrite dans le tableau suivant :

Classes	Effectifs
$[50;60]$	12
$]60;65]$	10
$]65;70]$	12
$]70;75]$	14
$]75;80]$	5
$]80;100]$	7
Total	60

(a) Pour cette répartition en classes, estimer la moyenne arithmétique $\bar{x}$ et la médiane $\widetilde{x}$ de la série des poids. Déterminer la classe modale et calculer une valeur approchée du mode x_M.

(b) Comment ces valeurs se comparent-elles par rapport aux paramètres $\bar{x}, \widetilde{x}$ et x_M calculés directement à partir des données brutes de la série sachant que celles-ci valent $\bar{x} = 70,02$, $\widetilde{x} = 69,5$ et $x_M = 65$? Donner un ordre de grandeur de l'erreur commise dans le calcul de $\bar{x}$ en remplaçant les observations des classes c_1 et c_4 par les centres des classes. Pour rappel, dans la classe $]50;60]$, les observations distinctes sont 52, 52, 55, 56, 57, 58, 59, 60, 60, 60, 60 et 60 tandis que dans la classe $]70;75]$, les observations sont 71, 71, 72, 72, 72, 72, 72, 72, 75, 75, 75, 75, 75 et 75.

(c) Donner une approximation des quartiles Q_1 et Q_3. A peu près 15 observations de la série initiale doivent avoir une valeur inférieure à Q_1 et le même nombre, une valeur supérieure à Q_3. En pratique, est-ce le cas ?

Solution :

(a) Pour estimer la moyenne arithmétique, comme les données sont groupées, il faut utiliser les centres des classes. On a donc

$$\bar{x} = \frac{55 \cdot 12 + 62.5 \cdot 10 + \ldots + 90 \cdot 7}{60} = 68.79$$

L'estimation de la médiane correspond à l'abscisse associée à l'ordonnée 0.5 dans l'ogive des fréquences cumulées. Dans cette ogive les points $(65, 0.37)$ et $(70, 0.57)$ sont alignés sur une même droite d'équation

$$y - 0.37 = \frac{0.57 - 0.37}{70 - 65}(x - 65) \tag{1}$$

En remplaçant y par 0.5 dans l'équation (1), on trouve

$$\tilde{x} = (0.5 - 0.37)\frac{70 - 65}{0.57 - 0.37} + 65 = 68.33$$

La classe modale est la classe $]70;75]$. **Attention**, comme les amplitudes des classes ne sont pas constantes, ce sont les fréquences ajustées qu'il faut comparer et non les effectifs. Les fréquences ajustées sont calculées en divisant les fréquences par les amplitudes des classes.

Une estimation du mode est donnée par Yule et Kendall avec la formule $x_M \approx 3\tilde{x} - 2\bar{x} = 67.41$.

(b) Les valeurs estimées sont assez proches des vraies valeurs. La moyenne calculée sur la nouvelle série est données par

$$\bar{x}' = \frac{1}{60}(12c_1 + 10\bar{x}_2 + 12\bar{x}_3 + 14c_4 + 5\bar{x}_5 + 7\bar{x}_6),$$

où $\bar{x}_1, \ldots, \bar{x}_6$ sont les moyennes calculées sur les observations des classes correspondantes. Ainsi, l'erreur commise est donnée par

$$\bar{x}' - \bar{x} = \frac{1}{60}(12(c_1 - \bar{x}_1) + 14(c_4 - \bar{x}_4)) = \frac{1}{60}(12(55 - 57.42) + 14(72.5 - 73.14)) = -0.63$$

On sous-estime donc un tout petit peu la moyenne.

(c) Les approximations des quartiles sont données par les abscisses associées aux ordonnées 0.25 et 0.75 de l'ogive des fréquences cumulées.

L'ordonnée 0.25 se trouve sur le segment de droite reliant les points $(60, 0.2)$ et $(65, 0.37)$. On a donc

$$x = (0.25 - 0.2)\frac{65 - 60}{0.37 - 0.2} + 60 = 61.5$$

L'ordonnée 0.75 se trouve sur le segment de droite reliant les points $(70, 0.57)$ et $(75, 0.8)$. On a donc

$$x = (0.75 - 0.57)\frac{75 - 70}{0.8 - 0.57} + 70 = 73.91$$

Avec ces approximations, il y a 12 observations inférieures à Q_1 et 18 observations supérieures à Q_3.

Exercice 6. Un institut de sondage a enquêté sur les frais d'entretien déclarés par les ménages possédant une résidence secondaire. Les dépenses (exprimées en une certaine unité monétaire et groupées en 7 classes) ainsi que les effectifs sont repris dans le tableau suivant, mais certaines données fournies par l'enquêteur sont restées indéchiffrables.

Classes pour les frais déclarés	Effectifs
$[0, 4]$	6
$]4, 8]$	n_2
$]8, 12]$	n_3
$]12, e_4]$	17
$]e_4, 22]$	14
$]22, 30]$	11
$]30, 42]$	3
Total	100

(a) Retrouver les valeurs manquantes n_2 et n_3 sachant que le premier quartile vaut $Q_1 = 7$.

(b) Sachant que $\bar{x} = 13$, déterminer la borne e_4 de la classe.

Solution :

(a) Puisque l'effectif total vaut 100, on sait que $n_2 + n_3 = 49$. Par ailleurs, puisque le premier quartile correspond à la 25ème plus petite observation, on sait que le segment de l'ogive des effectifs cumulés correspondant à la classe $]4; 8]$ passe par les points de coordonnées $(4, 6)$, $(7, 25)$ et $(8, 6 + n_2)$.

On a donc,

$$6 + n_2 - 6 = \frac{25 - 6}{7 - 4}(8 - 4)$$
$$\Leftrightarrow \quad n_2 = \frac{25 - 6}{7 - 4}(8 - 4) = 25.33$$

On en déduit donc que $n_2 = 25$ et que $n_3 = 49 - 25 = 24$.

(b) En utilisant l'expression de la moyenne d'une série groupée, on a

$$13 = \frac{6 \cdot 2 + 25 \cdot 6 + 24 \cdot 10 + 17 \cdot \left(\frac{12+e_4}{2}\right) + 14 \cdot \left(\frac{e_4+22}{2}\right) + 11 \cdot 26 + 3 \cdot 36}{100}$$
$$\Leftrightarrow \quad e_4 = 16$$

Exercice 7. On a interrogé 92 représentants de commerce sur le nombre de kilomètres qu'ils effectuaient par jour pour leur travail. Les résultats sont repris dans le tableau ci-dessous, duquel certaines données ont disparu.

Trajets en km	Nombres de représentants
$[10, 20]$	x_1
$]20, 40]$	26
$]40, x_2]$	19
$]x_2, x_3]$	24
$]x_3, 100]$	14

(a) Retrouver les valeurs manquantes x_1, x_2 et x_3 sachant que le trajet médian est égal à 45,79 km et que le trajet moyen est égal à 49,89 km.

(b) Construire l'ogive des fréquences cumulées et vérifier graphiquement la valeur de la médiane.

Solution :

(a) Pour commencer, puisque l'effectif total vaut 92, on a $x_1 = 9$.

La classe médiane est la classe $]40; x_2]$. On sait par ailleurs que le point de coordonnées $(45.79; 46)$ est situé sur le segment de droite correspondant à cette classe sur l'ogive des effectifs cumulés (c'est-à-dire passant par les points $(40, 35)$ et $(x_2, 54)$). Ce segment a pour équation :

$$y - 35 = \frac{46 - 35}{45.79 - 40}(x - 40)$$

Ce qui mène, en remplaçant y par 54 dans l'équation à

$$x_2 = (54 - 35)\frac{45.79 - 40}{46 - 35} + 40 \approx 50.$$

Enfin, en utilisant l'expression de la moyenne d'une série groupée, on obtient

$$49.89 = \frac{9 \cdot 15 + 26 \cdot 30 + 19 \cdot 45 + 24 \cdot \left(\frac{x_3+50}{2}\right) + 14 \cdot \left(\frac{x_3+100}{2}\right)}{92}$$
$$\Leftrightarrow \quad x_3 \approx 80$$

(b) Tableau statistique :

Classes	Effectifs	Effectifs cumulés	Fréquences	Fréquences cumulées
$[10, 20]$	9	9	$0,1$	$0,1$
$]20, 40]$	26	35	$0,28$	$0,38$
$]40, 50]$	19	54	$0,21$	$0,59$
$]50, 80]$	24	78	$0,26$	$0,85$
$]80, 100]$	14	92	$0,15$	1

L'ogive est reprise en Figure 1.

Exercice 8. Soit la série statistique $S = \{-2, 2, -1, 1, x\}$, où x désigne un nombre réel arbitraire.

Calculer la médiane de S lorsque le paramètre x varie. Représenter graphiquement la fonction qui, à tout réel x, associe la médiane de S.

Même problème en remplaçant la médiane par la moyenne arithmétique de S.

Commenter les résultats et comparer les deux situations.

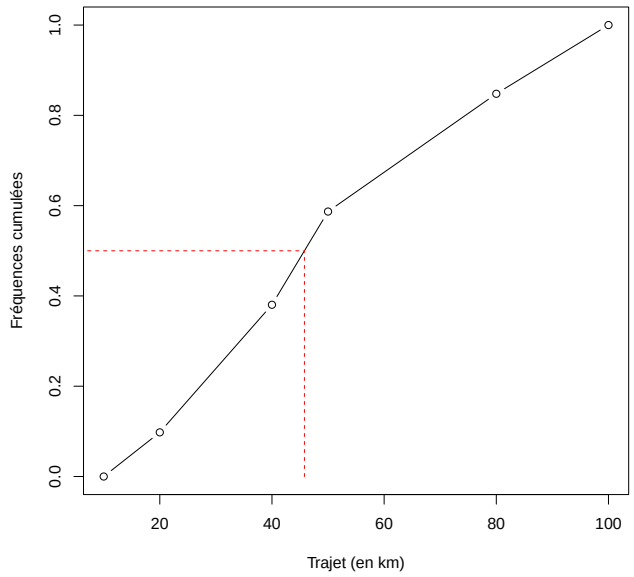

FIGURE 1 – Ogive des fréquences cumulées pour la variable nombre de kilomètres

Solution :

Pour $x < -1$, la médiane de la série est -1. Pour $x > 1$, la médiane de la série est 1. Pour $x \in [-1, 1]$, la médiane vaut x.

La moyenne de la série vaut quand à elle

$$\bar{x} = \frac{-2 + 2 - 1 + 1 + x}{5} = \frac{x}{5}.$$

Les deux statistiques sont représentées à la Figure 2. La médiane est bornée (comprise entre -1 et 1, quelle que soit la valeur de x), alors que la moyenne arithmétique ne l'est pas : lorsque $x \to \pm\infty$, on a $\bar{x} \to \pm\infty$. Elle est donc fortement influencée par des valeurs extrêmes.

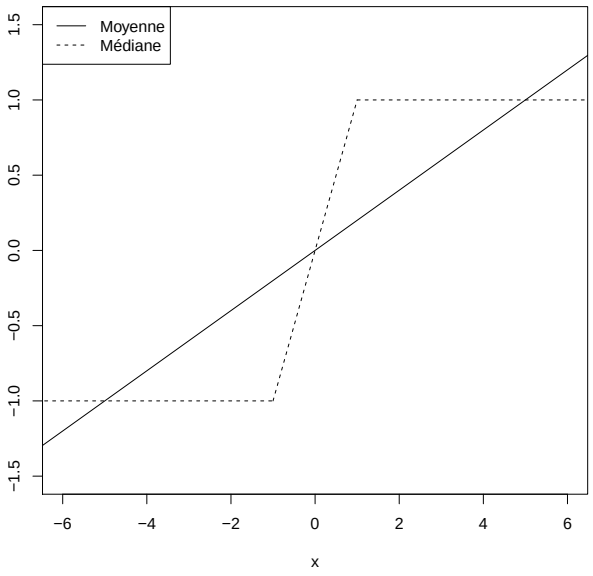

FIGURE 2 – Représentation de la moyenne et de la médiane en fonction de x

Exercice 9. Soit $S = \{x_1, \ldots, x_n\}$ une série statistique univariée de moyenne $\bar{x}$ et de variance s_x^2. Calculer la moyenne et la variance de la série des valeurs centrées et réduites $Z = \{z_1, \ldots, z_n\}$ où $z_i = \frac{x_i - \bar{x}}{s_x}$.

Solution :

On a

$$\begin{aligned}\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i &= \frac{1}{n}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \\ = \frac{1}{ns_x}\sum_{i=1}^{n}(x_i - \bar{x}) &= \frac{1}{ns_x}\left(\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}\right) \\ = \frac{1}{ns_x}(n\bar{x} - n\bar{x}) &= 0\end{aligned}$$

et

$$\begin{aligned}s_z^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2 \\ = \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{s_x^2}{s_x^2} = 1\end{aligned}$$

Exercice 10. On dispose de la série suivante groupée en 5 classes dont seuls les effectifs des deuxième et troisième classes sont connus.

Classes	Effectifs n_i
$[0, 2]$	n_1
$]2, 4]$	5
$]4, 6]$	6
$]6, 8]$	n_4
$]8, 10]$	n_5
Total	20

Calculer la médiane de cette série sachant que la moyenne vaut 4,7 et la variance 5,71.

Solution :

On sait, comme l'effectif total vaut 20 que, $n_1 + 5 + 6 + n_4 + n_5 = 20$. De plus, par le calcul de la moyenne à l'aide des centres des classes, on a

$$\frac{1 \cdot n_1 + 3 \cdot 5 + 5 \cdot 6 + 7 \cdot n_4 + 9 \cdot n_5}{20} = 4.7$$

et grâce au calcul de la variance, on a

$$\frac{(1 - 4.7)^2 \cdot n_1 + (3 - 4.7)^2 \cdot 5 + (5 - 4.7)^2 \cdot 6 + (7 - 4.7)^2 \cdot n_4 + (9 - 4.7)^2 \cdot n_5}{20} = 5.71$$

Nous avons donc trois équations à trois inconnues. Il suffit de résoudre le système pour trouver les valeurs de $n_1 = 3$, $n_4 = 4$ et n_5=2. La médiane se trouve ensuite par interpolation linéaire dans l'ogive des effectifs cumulés. Les points $(4; 8)$, $(6; 14)$ et $(\tilde{x}; 10.5)$ sont alignés sur une même droite :

$$\begin{aligned}10.5 - 8 &= \frac{14 - 8}{6 - 4}(\tilde{x} - 4) \\ \Leftrightarrow \tilde{x} &= (10,5 - 8)\frac{6 - 4}{14 - 8} + 4 = 4.83\end{aligned}$$

Exercice 11. Le tableau suivant donne la répartition en 5 classes des salaires (en unités monétaires) des employés d'une entreprise.

Classes de salaire c_i	Effectifs n_i	Fréquences cumulées F_i
$[2, 6]$	n_1	0,1
$]6, 10]$	n_2	0,33
$]10, 12]$	n_3	0,6
$]12, 14]$	n_4	0,8
$]14, 18]$	n_5	1

Sachant que $s^2 = 8$, $\sum_{i=1}^{5} f_i \widetilde{x}_i^2 = 133,8$ et que $\sum_{i=1}^{5} n_i \widetilde{x}_i = 673$, où $\widetilde{x}_i$ est le centre et f_i la fréquence de la i-ème classe, calculer les effectifs n_i de chaque classe et l'effectif total n.

Solution :

Il suffit de trouver l'effectif total. Par définition, la variance est

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Cependant, ici, les données sont groupées, on a donc

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{5} \tilde{x}_i n_i$$

et

$$\begin{aligned}
s^2 &\approx \frac{1}{n}\sum_{i=1}^{5} n_i \tilde{x}_i^2 - \bar{x}^2 \\
8 &\approx \sum_{i=1}^{5} \frac{n_i}{n} \tilde{x}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{5} \tilde{x}_i n_i\right)^2 \\
8 &\approx \sum_{i=1}^{5} f_i \tilde{x}_i^2 - \left(\frac{1}{n}\sum_{i=1}^{5} \tilde{x}_i n_i\right)^2 \\
8 &\approx 133.8 - \left(\frac{673}{n}\right)^2
\end{aligned}$$

Ce qui permet de conclure que $n \approx 60$. Grâce aux fréquences cumulées, on obtient finalement les effectifs $n_1 = 0.1 \cdot 60 = 6$, $n_2 = 0.23 \cdot 60 = 14$, $n_3 = 0.27 \cdot 60 = 16$, $n_4 = 0.2 \cdot 60 = 12$ et $n_5 = 0.2 \cdot 60 = 12$.

Exercice 12. Une enquête auprès de 100 travailleurs a fourni une série de données relative au nombre d'emplois antérieurs occupés au cours des dix dernières années. Les deux premières colonnes du tableau statistique sont reprises dans le tableau suivant :

Valeurs x_i	Effectifs n_i
0	8
1	19
2	32
3	17
4	14
5	10
Total	100

(a) Calculer la moyenne arithmétique, la médiane et le mode de la série. Interpréter les résultats obtenus. Quel est le paramètre le plus indiqué pour cet ensemble de données ?

(b) Déterminer les quartiles Q_1 et Q_3. Représenter ensuite la boite à moustaches correspondante.

(c) Calculer la variance et l'écart-type.

(d) Le diagramme en bâtons associé à ces données discrètes semble-t-il symétrique ? Calculer les coefficients de dissymétrie de Fisher et Pearson.

Solution :

(a) La moyenne arithmétique est donnée par

$$\frac{0 \cdot 8 + 1 \cdot 19 + 2 \cdot 32 + 3 \cdot 17 + 4 \cdot 14 + 5 \cdot 10}{100} = 2.4$$

Comme l'effectif total est pair, la médiane est donnée par

$$\frac{x_{(50)} + x_{(51)}}{2} = \frac{2+2}{2} = 2$$

Le mode de la série est 2.

Ces trois paramètres sont presque tous égaux. Cependant, la moyenne est moins appropriée dans le cas discret puisqu'elle donne une valeur qui ne fait pas partie de nos valeurs initiales.

(b) On a $Q_1 = \frac{x_{(25)} + x_{(26)}}{2} = 1$ et $Q_3 = \frac{x_{(75)} + x_{(76)}}{2} = 3$
La boite à moustache est reprise en Figure 3.

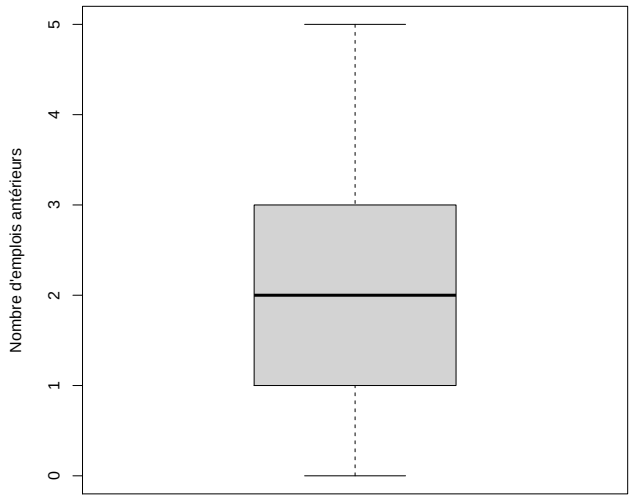

FIGURE 3 – Boite à moustache de la variable nombre d'emplois antérieurs occupés au cours des dix dernières années

(c) La variance est donnée par

$$s^2 = \frac{(0-2.4)^2 \cdot 8 + (1-2.4)^2 \cdot 19 + (2-2.4)^2 \cdot 32 + (3-2.4)^2 \cdot 17 + (4-2.4)^2 \cdot 14 + (5-2.4)^2 \cdot 10}{100} = 1.98$$

et l'écart-type vaut $s = \sqrt{1.98} = 1.407$.

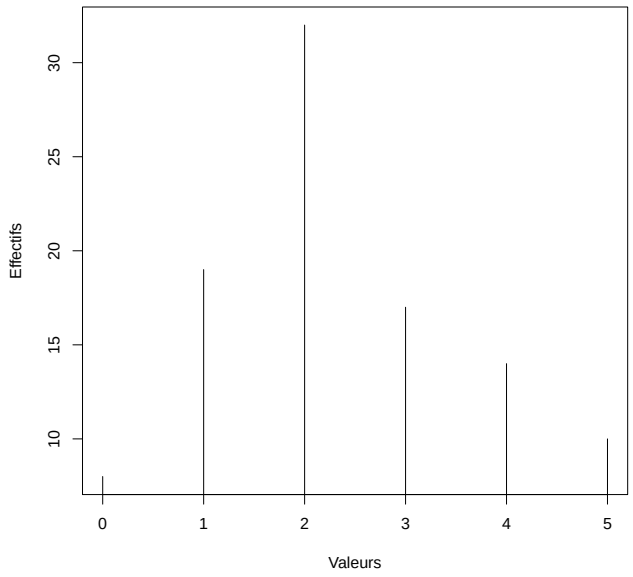

FIGURE 4 – Diagramme en bâtons de la variable nombre d'emplois antérieurs occupés au cours des dix dernières années

(d) Le diagramme en bâtons se trouve à la Figure 4. On peut constater un léger étalement à droite, c'est à dire un légère dissymétrie à gauche. Cela est notamment du au fait qu'il est impossible d'avoir des valeurs inférieurs 0 car notre variable est un comptage. Cependant, il est en théorie possible d'avoir des valeurs dans l'ensemble des naturels $\mathbb{N}$.

Le coefficient de dissymétrie de Fisher est donné par

$$\gamma_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3} = \frac{0.72}{2.79} = 0.258$$

Le coefficient de dissymétrie de Pearson est donné par

$$S_k = \frac{\bar{x} - x_M}{s} = \frac{2.4 - 2}{1.407} = 0.284$$

Les deux coefficients sont positifs, ce qui confirme que la distribution est dissymétrique à gauche.

Exercice 13. Les poids de 22 étudiantes de premier bachelier sont donnés par la série ordonnée suivante :

$$S = \{47, 48, 49, 50, 53, 55, 55, 55, 56, 56,$$

$$58, 59, 61, 62, 62, 63, 63, 63, 64, 65, 65, 66\}.$$

(a) Calculer la moyenne arithmétique, la médiane et le mode de cette série et interpréter ces paramètres.

(b) Calculer l'écart-type pour la série des poids des filles. Estimer le nombre et le pourcentage d'étudiantes dont le poids se trouve entre $\bar{x} - 2s$ et $\bar{x} + 2s$ et comparer avec les nombres atteints réellement.

(c) Calculer les quartiles Q_1 et Q_3. Sachant que la série des poids des étudiants masculins de la même section correspond aux paramètres suivants :

$$Q_1 = 65; \widetilde{x} = 69, 5; Q_3 = 75,$$

tandis que les observations individuelles sont reprises ci-dessous, comparer les deux distributions de poids à l'aide de boîtes à moustaches.

68	70	67	75	72	71	67	65	60	60	65	65
77	95	85	70	70	72	66	75	90	65	62	70
52	60	59	65	68	71	97	65	57	75	77	75
85	56	77	67	62	52	67	72	79	60	72	69
58	55	75	75	78	65	95	65	90	72	72	60

Quelle distribution donne la boîte la plus symétrique ? Calculer le coefficient de Yule et Kendall pour les deux boîtes. Pour la série des poids des filles, comparer ce coefficient avec le coefficient de dissymétrie de Fisher.

(d) Dans une autre section comportant 25 étudiantes, les poids ont été aussi mesurés et en voici un résumé succinct : $\bar{x} = 62$, $s = 3,334$. Pour une répétition, les étudiantes des deux sections sont regroupées. A partir des paramètres individuels calculés sur les deux séries, calculer la moyenne arithmétique et la variance des poids de ce groupe de 47 étudiantes. Commenter la décomposition de la variance.

Solution :

(a) La moyenne de la série est obtenue en sommant toutes les observations et en divisant par l'effectif total qui est 22 ; cela donne $\bar{x} = 57.95$kg. La médiane de la série est obtenue par la formule $\tilde{x} = \frac{x_{(11)}+x_{(12)}}{2} = \frac{58+59}{2} = 58.5$kg. Dans ce cas, le mode n'existe pas ; il y a 3 valeurs 55 et 3 valeurs 63. Comme $\bar{x} < \tilde{x}$, on devrait avoir une légère dissymétrie à droite.

(b) L'écart-type pour cette série peut être obtenu par la formule

$$s = \sqrt{\frac{1}{22}\sum_{i=1}^{22} x_i^2 - \bar{x}^2} = 5.8\text{kg}$$

Par la propriété de Tchebychev, au moins 3/4 des observations sont dans l'intervalle

$$[\bar{x} - 2s; \bar{x} + 2s] = [46.35; 69.56].$$

En réalité, elles y sont toutes.

(c) Chez les filles, le premier quartile vaut $Q_1 = x_{(6)} = 55$ et le troisième quartile vaut $Q_3 = x_{(17)} = 63$. Les boites à moustaches sont reprises en Figure 5. La distribution des garçons semble plus symétrique.

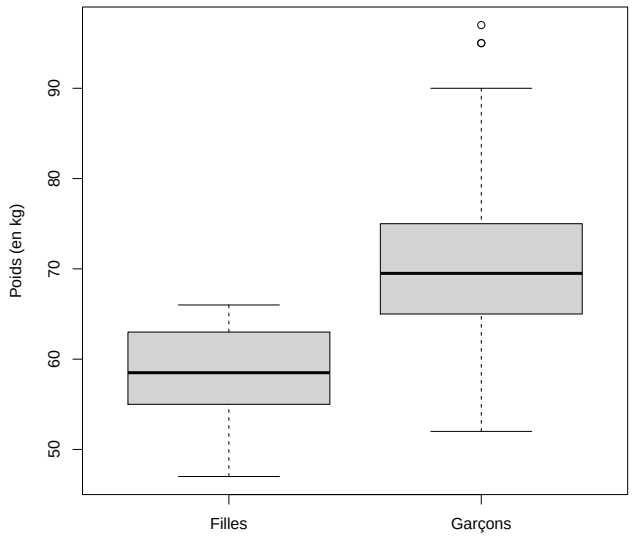

FIGURE 5 – Boites à moustaches de la variable poids pour la série des filles (à gauche) et la série des garçons (à droite)

On a d'ailleurs $Y_k = \dfrac{Q1 + Q3 - 2\tilde{x}}{Q3 - Q1} = \dfrac{55 + 63 - 2 \cdot 58.5}{63 - 55} = 0.125$ pour les filles et $Y_k = \dfrac{65 + 75 - 2 \cdot 69.5}{75 - 65} = 0.1$ pour les garçons.

Par contre, pour les filles, le coefficient de dissymétrie de Fisher vaut $\gamma_1 = -0.41$. Cela est du au fait que Y_k s'intéresse juste aux observations centrales.

(d) La moyenne globale dans le groupe des 47 étudiantes est la moyenne pondérée des deux groupes :

$$\bar{x}_T = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2} = \frac{22 \cdot 57.95 + 25 \cdot 62}{47} = 60.1\text{kg}.$$

La variance totale est la somme de la variance dans les groupes et la variance entre les groupes :

$$\begin{aligned} s_T^2 &= \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x}_T)^2 + n_2(\bar{x}_2 - \bar{x}_T)^2}{n_1 + n_2} \\ &= \frac{22 \cdot 5.8^2 + 25 \cdot 3.334^2}{47} + \frac{22(57.95 - 60.1)^2 + 25(62 - 60.1)^2}{47} \\ &= 21.66 + 4.08 = 25.74 \end{aligned}$$

La variance entre les groupes représente $\dfrac{4.08}{25.74} \cdot 100 = 16\%$ de la variance totale ; alors que la variance dans les groupes représente 84% de la variance totale. La variation vient donc principalement d'une variation de la variable et non d'une variation entre les groupes.

Exercice 14. Une firme F comprend deux filiales A et B. Les salaires moyens (en euros) par catégorie socio-professionnelle sont donnés dans le tableau suivant :

	A		B	
Catégories	Sal. moyens	Eff.	Sal. moyens	Eff.
Ouvriers	1500	50	1200	10
Employés	1750	85	1600	20
Cadres	3000	5	2500	30

Comparer la dispersion des salaires moyens dans les deux établissements en calculant les variances s_1^2 et s_2^2. Calculer la variance globale des salaires de l'entreprise complète. Discuter la décomposition de la variance.

Solution :

Commençons par calculer les moyennes dans les deux établissements ainsi que la moyenne globale :

$$\bar{x}_1 = \frac{1500 \cdot 50 + 1750 \cdot 85 + 3000 \cdot 5}{50 + 85 + 5} = 1705.36$$

$$\bar{x}_2 = \frac{1200 \cdot 10 + 1600 \cdot 20 + 2500 \cdot 30}{10 + 20 + 30} = 1983.33$$

$$\bar{x} = \frac{140\bar{x}_1 + 60\bar{x}_2}{200} = 1788.75$$

Calculons ensuite les variances dans les deux établissements ainsi que la variance globale :

$$s_1^2 = \frac{1500^2 \cdot 50 + 1750^2 \cdot 85 + 3000^2 \cdot 5}{50 + 85 + 5} - 1705.36^2 = 76132.02$$

$$s_2^2 = \frac{1200^2 \cdot 10 + 1600^2 \cdot 20 + 2500^2 \cdot 30}{10 + 20 + 30} - 1983.33^2 = 284722.2$$

$$s^2 = \frac{140s_1^2 + 60s_2^2}{200} + \frac{140(\bar{x}_1 - \bar{x})^2 + 60\bar{x}_2 - \bar{x})^2}{200} = 138709.1 + 16226.86 = 154935.9$$

La variance entre les groupes représente 10% de la variance totale, la variance dans les groupes représente 90% de la variance totale. La variance globale est donc due à une grande variabilité dans les groupes (donc due aux différentes catégories socio-professionnelles) et non à une grande différence entre les salaires moyens des deux filiales.

Exercice 15. Pendant 14 semaines, on a relevé la recette, en milliers d'euros, d'un supermarché le lundi et le samedi. Les résultats sont repris dans le tableau ci-dessous.

Sem.	Recette lundi (X)	Recette samedi (Y)
1	57	83
2	60	93
3	52	77
4	49	69
5	56	81
6	46	70
7	51	71
8	63	91
9	49	70
10	57	82
11	40	55
12	45	65
13	65	105
14	55	80

Décrire les deux séries univariées correspondant respectivement aux recettes du lundi et à celles du samedi par des boîtes à moustaches (la construction des boîtes, c'est-à-dire le calcul des quartiles, des valeurs pivots,..., doit être expliquée).

Solution :

Pour calculer les quartiles et la médiane, il faut tout d'abord trier les deux séries de recettes :

$$X = \{40, 45, 46, 49, 49, 51, 52, 55, 56, 57, 57, 60, 63, 65\}$$

$$Y = \{55, 65, 69, 70, 70, 71, 77, 80, 81, 82, 83, 91, 93, 105\}$$

Comme il y a 14 observations dans les deux séries, les deux médianes seront respectivement

$$\tilde{x} = \frac{x_{(7)} + x_{(8)}}{2} = \frac{52 + 55}{2} = 53.5 \qquad \text{et} \qquad \tilde{y} = \frac{y_{(7)} + y_{(8)}}{2} = \frac{77 + 80}{2} = 78.5$$

Les premiers et troisièmes quartiles sont ensuite, pour la série du lundi :

$$Q_1 = x_{(4)} = 49 \quad \text{et} \quad Q_3 = x_{(11)} = 57$$

L'écart inter-quartile de cette série vaut donc $EIQ = 57 - 49 = 8$, les valeurs pivots sont

$$a_1 = 49 - 1.5 \cdot 8 = 37 \quad \text{et} \quad a_2 = 57 + 1.5 \cdot 8 = 69$$

ce qui donne les valeurs adjacentes $x_{(g)} = 65$ et $x_{(d)} = 40$. Pour cette série, il n'y a donc pas de valeurs extérieures.

Pour la série du samedi, on a :

$$Q_1 = y_{(4)} = 70 \quad \text{et} \quad Q_3 = y_{(11)} = 83 \quad \text{et} \quad EIQ = 83 - 70 = 13$$

$$a_1 = 70 - 1.5 \cdot 13 = 50.5 \quad \text{et} \quad a_2 = 83 + 1.5 \cdot 13 = 102.5$$

ce qui donne les valeurs adjacentes $x_{(g)} = 55$ et $x_{(d)} = 93$. Pour cette série, l'observation 105 est une valeur extérieure.

Les boites à moustaches sont reprises à la Figure 6.

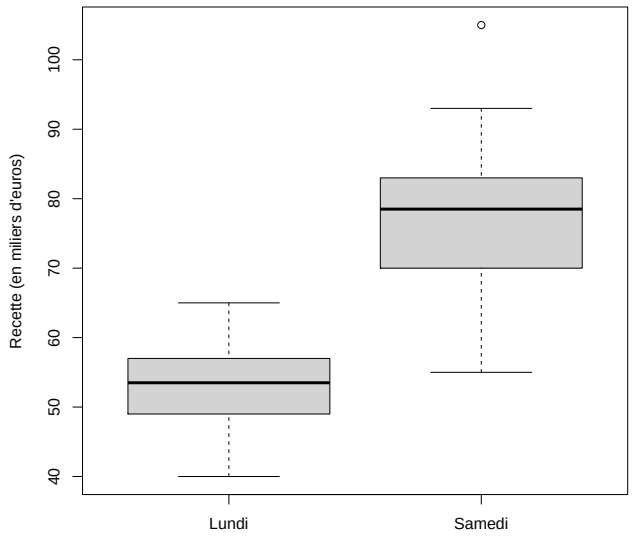

FIGURE 6 – Boites à moustaches de la variable recette, pour les séries du lundi (à gauche) et du samedi (à droite)

Exercice 16. Pour une série statistique $S = \{x_1, x_2, ..., x_n\}$, on connaît la moyenne arithmétique $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et la variance $s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$.

On ajoute à S un élément y de manière à former la nouvelle série $S' = S \cup \{y\}$.

(a) Calculer la variance $(s')^2$ de S' en fonction de y et des paramètres connus de S.

(b) Étudier le comportement de $(s')^2$ lorsque y varie.

(c) Étudier le comportement de $(s')^2$ lorsque l'effectif n de S grandit indéfiniment.

Solution :

(a) Remarquons tout d'abord que $n' = n + 1$ et

$$\bar{x}' = \frac{1}{n+1}\left(\sum_{i=1}^{n} x_i + y\right) = \frac{n}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) + \frac{y}{n+1} = \frac{n\bar{x} + y}{n+1}$$

On a alors

$$\begin{aligned}
(s')^2 &= \frac{1}{n+1}\left[\sum_{i=1}^{n}(x_i-\bar{x}')^2+(y-\bar{x}')^2\right]\\
&= \frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x}')^2+\frac{1}{n+1}(y-\bar{x}')^2\\
&= \frac{1}{n+1}\sum_{i=1}^{n}\left(x_i-\frac{n\bar{x}+y}{n+1}\right)^2+\frac{1}{n+1}\left(y-\frac{n\bar{x}+y}{n+1}\right)^2\\
&= \frac{1}{n+1}\sum_{i=1}^{n}\left(x_i-\bar{x}+\bar{x}-\frac{n\bar{x}+y}{n+1}\right)^2+\frac{1}{n+1}\left(\frac{(n+1)y-n\bar{x}-y}{n+1}\right)^2\\
&= \frac{1}{n+1}\sum_{i=1}^{n}\left(x_i-\bar{x}+\frac{\bar{x}-y}{n+1}\right)^2+\frac{1}{n+1}\left(\frac{n\cdot y-n\bar{x}}{n+1}\right)^2\\
&= \frac{1}{n+1}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2+2(x_i-\bar{x})\frac{\bar{x}-y}{n+1}+\left(\frac{\bar{x}-y}{n+1}\right)^2\right]+\frac{1}{n+1}\left(\frac{n}{n+1}(y-\bar{x})\right)^2\\
&= \frac{1}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{2}{n+1}\sum_{i=1}^{n}(x_i-\bar{x})\frac{\bar{x}-y}{n+1}+\frac{1}{n+1}\sum_{i=1}^{n}\left(\frac{\bar{x}-y}{n+1}\right)^2+\frac{n^2}{(n+1)^3}(y-\bar{x})^2\\
&= \frac{n}{n+1}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2+0+\frac{n}{(n+1)^3}(\bar{x}-y)^2+\frac{n^2}{(n+1)^3}(y-\bar{x})^2\\
&= \frac{n}{n+1}s^2+\frac{n(1+n)}{(n+1)^3}(\bar{x}-y)^2=\frac{n}{n+1}s^2+\frac{n}{(n+1)^2}(\bar{x}-y)^2
\end{aligned}$$

(b) Le comportement de $(s')^2$ en fonction de y est celui de $(\bar{x}-y)^2$. Or $(\bar{x}-y)^2=\bar{x}^2-2\bar{x}y+y^2$ est une parabole convexe qui prend son minimum en $y=\bar{x}$.

(c)

$$\begin{aligned}
\lim_{n\to+\infty}(s')^2 &= \lim_{n\to+\infty}\left(\frac{n}{n+1}s^2+\frac{n}{(n+1)^2}(\bar{x}-y)^2\right)\\
&= \lim_{n\to+\infty}\left(\frac{n}{n+1}s^2+\frac{n}{n^2+2n+1}(\bar{x}-y)^2\right)\\
&= \lim_{n\to+\infty}\left(\frac{1}{1+\frac{1}{n}}s^2+\frac{1}{n\left(1+\frac{2}{n}+\frac{1}{n^2}\right)}(\bar{x}-y)^2\right)=s^2
\end{aligned}$$

Le premier terme tend vers s^2 puisque $\lim_{n\to+\infty}\frac{1}{n}=0$ et le deuxième terme tend vers 0 pour la même raison. Ainsi, l'impact sur la variance d'une observation supplémentaire y dans un échantillon de taille infinie est nul.