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.

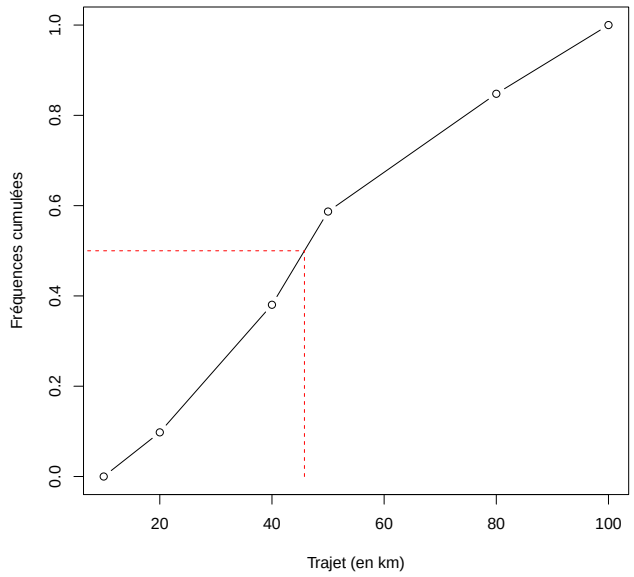

FIGURE 1 – Ogive des fréquences cumulées pour la variable nombre de kilomètres

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.

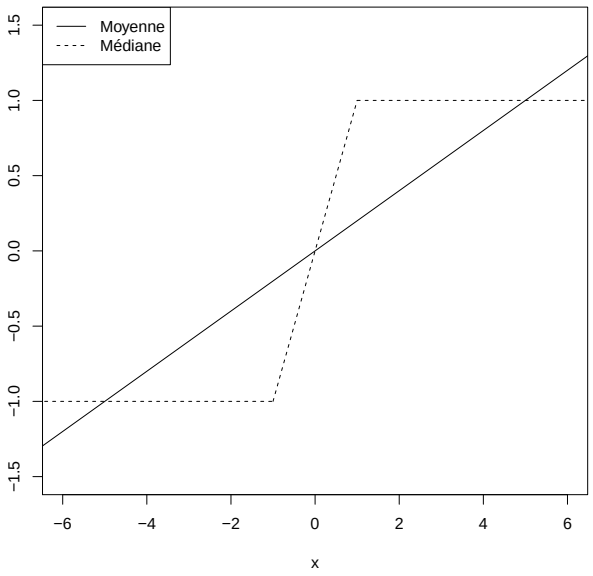

FIGURE 2 – Représentation de la moyenne et de la médiane en fonction de x