Statistique descriptive
Solutions des exercices

Chapitre 3 : Réduction des données

1. $\bar{x}_1 = 5,1$; $x_{M,1} = 5$ $\tilde{x}_1 = 5$ $Q_1 = 3$ $Q_3 = 6$.
   $\bar{x}_2 = 78,69$; $x_{M,2} = \nexists$ $\tilde{x}_2 = 82$ $Q_1 = 71$ $Q_3 = 87$.
   Par exemple, $S_1 = \{1, 2, 2, 2, 3\}$ et $S_2 = \{2, 2, 2\}$.
2. (a) $\bar{x} = 9.4425$ (b) $x_M = 11$ (c) 10 (d) $\bar{x}_{0,1} = 9.6375$
3. (a) $\bar{x} = 1621.43$ (b) Salaire médian : $\tilde{x} = x_{(11)} = 950$, salaire modal : $x_M = 600$.
4. $\bar{x} = 0.4779$. Il s'agit d'une moyenne pondérée, où les poids correspondent aux nombres d'habitants dans les différents quartiers.
5. (a) L'estimation de la moyenne est obtenue en remplç chaque observation par le centre de la classe à laquelle elle appartient. On obtient $\bar{x} = 68.79$.
   L'estimation de la médiane correspond à l'abscisse associée à l'ordonnée 0.5 de l'ogive des fréquences cumulées. On obtient $\tilde{x} = 68.33$.
   La classe modale est donnée par $]70; 75]$ (attention : les amplitudes des classes n'étant pas constantes, ce sont les fréquences ajustées qu'il faut comparer et non les effectifs).
   Une estimation du mode est fournie par la relation de Yule et Kendall. On obtient $x_M \approx 69.72$.

   (b) Les valeurs estimées sont assez proches des vraies valeurs en ce qui concerne la moyenne et la médiane. La moyenne calculée sur la nouvelle série est donnée par

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

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

$$\bar{x}' - \bar{x} = \frac{1}{60}\left(12(c_1 - \bar{x}_1) + 14(c_4 - \bar{x}_4)\right) = -0.63.$$

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

   (c) L'approximation des quartiles sont données par les abscisses associées aux ordonnées 0.25 et 0.75 de l'ogive des fréquences cumulées. On obtient $Q_1 = 61.5$, $Q_3 = 73.93$. Avec ces approximations, il y a 12 observations inférieures à Q_1 et 18 observations supérieures à Q_3.
6. (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 $(8;6+n_2)$. De ceci, on déduit que $n_2 = 25$ et, par conséquent, $n_3 = 24$.

(b) En utilisant l'expression de la moyenne d'une série groupée, on obtient $e_4 = 16$.

7. (a) 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. On en tire que $x_2 \approx 50$.
Enfin, en utilisant l'expression de la moyenne d'une série groupée, on obtient $x_3 \approx 80$.

(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

Ogive des fréquence cumulées : voir Figure 7.

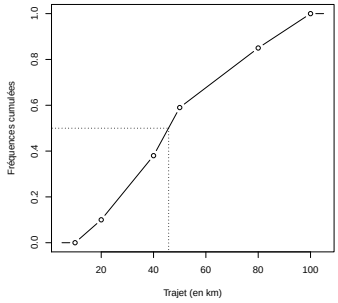

FIGURE 7 – Exercice 3.7(b)

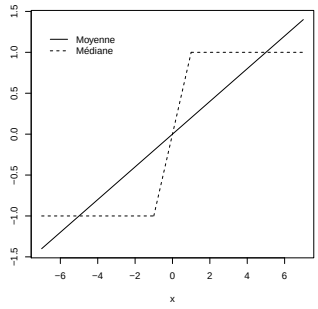

FIGURE 8 – Exercice 3.8

8. Voir Figure 8. 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 les valeurs extrêmes.
9. $\bar{z} = 0$ et $s_z^2 = 1$.
10. En utilisant le fait que l'effectif total vaut 20, l'expression de la moyenne d'une série groupée et l'expression de la variance d'une série groupée, on obtient trois équations à trois inconnues. Il suffit alors de résoudre le système pour obtenir les valeurs de n_1, n_4 et n_5. La classe médiane peut alors être déterminée et l'estimation de la médiane se fait par interpolation linéaire à partir de l'ogive. On obtient $\tilde{x} = 4.67$.

11. En utilisant l'expression de la variance d'une série groupée, on déduit que $n = 60$. Grâce aux fréquences cumulées, on peut calculer les fréquences de chaque classe, et en déduire les effectifs. On obtient alors $n_1 = 6$, $n_2 \approx 14$, $n_3 \approx 16$, $n_4 = 12$ et $n_5 = 12$.

12. (a) $\bar{x} = 2.4$; $\tilde{x} = x_M = 2$. La moyenne est moins appropriée dans le cas d'une variable discrète.

    (b) Au vu du tableau des effectifs, on a $Q_1 = 1$ et $Q_3 = 3$.
    Voir Figure 9 pour la boite à moustaches.

    (c) On a $s^2 = 1.98$ et $s = 1.407$.

    (d) Le diagramme en bâtons est donné à la Figure 10. On constate un léger étalement à droite.
    On a par ailleurs $S_k = 0.284$ et $\gamma_1 = 0.258$, ce qui confirme la constatation précédente.

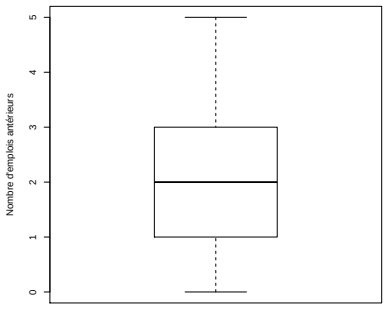

FIGURE 9 – Exercice 3.12(b)

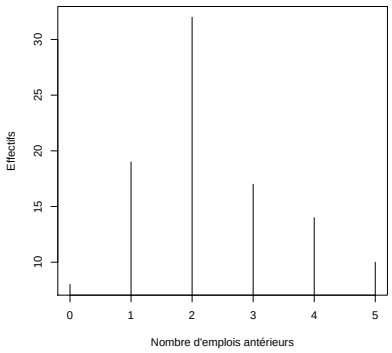

FIGURE 10 – Exercice 3.12(d)

13. (a) $\bar{x} = 57.95$; $\tilde{x} = 58.5$ et le mode n'existe pas (il y a 3 valeurs 55 et 3 valeurs 63).

    (b) $s = 5.80$.
    Par la propriété de Tchebychev, au moins $\frac{3}{4}$ des observations sont dans l'intervalle $[\bar{x} - 2s, \bar{x} + 2s] = [46,35; 69,56]$. En réalité, elles y sont toutes.

    (c) A l'aide de l'ogive des fréquences cumulées, on obtient $Q_1 = 55$ et $Q_3 = 63$. Voir Figure 11 pour les deux boites à moustaches. La distribution des garçons semble plus symétrique.
    On a d'ailleurs $Y_k = 0,125$ pour les filles et $Y_k = 0,1$ pour les garçons. Par contre, pour la série des filles, le paramètre de dissymétrie de Fisher vaut $\gamma_1 = -0,41$. Cela est dû au fait que Y_k s'intéresse juste aux observations centrales.

    (d) On a $\bar{x}_T = 60.10$ et $s_T^2 = 21.93+4.08 = 26.01$, où le premier terme correspond à la variance dans les groupe et le seconde à la variance entre les groupes.

La variance entre les groupes représente 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.

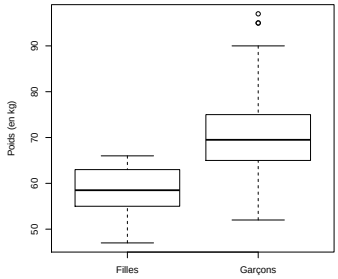

FIGURE 11 – Exercice 3.13(b)

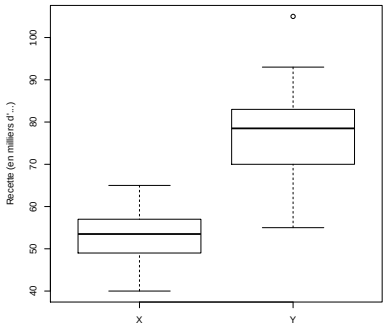

FIGURE 12 – Exercice 3-15

14. On a $s_1^2 = 75986$ et $s_2^2 = 284854$. De plus, $\bar{x}_1 = 1705.4$ et $\bar{x}_2 = 1983.3$, dont on déduit que $s^2 = 154864$.
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 catétogires socio-professionnelles) et non à une grande différence entre les salaires moyens des deux filiales.

15.(X) $Q_1 = 49$; $\tilde{x} = 53,5$; $Q_3 = 57$; $a_1 = 37$, $a_2 = 69$, $x_{(g)} = 40$ et $x_{(d)} = 65$.
(Y) $Q_1 = 70$; $\tilde{x} = 78,5$; $Q_3 = 83$; $a_1 = 50,5$, $a_2 = 102,5$, $x_{(g)} = 55$ et $x_{(d)} = 93$.
Voir Figure 12 pour les boites à moustaches.

16. (a) Notons d'abord que $n' = n + 1$ et $\bar{x}_{S'} = \dfrac{n\bar{x} + y}{n+1}$. De là, on déduite que $(s')^2 = \dfrac{n}{(n+1)^2}((n+1)s^2 + (\bar{x} - y)^2)$.

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

(c) $\lim_{n\to+\infty}(s')^2 = \lim_{n\to+\infty}\left(\dfrac{1}{1+\frac{1}{n}}s^2 + \dfrac{1}{n(1+\frac{2}{n}+\frac{1}{n^2}}(\bar{x} - y)^2\right) = s^2$. Ainsi, l'impact sur la variance d'une observation supplémentaire y dans un échantillon de taille infinie est nul.