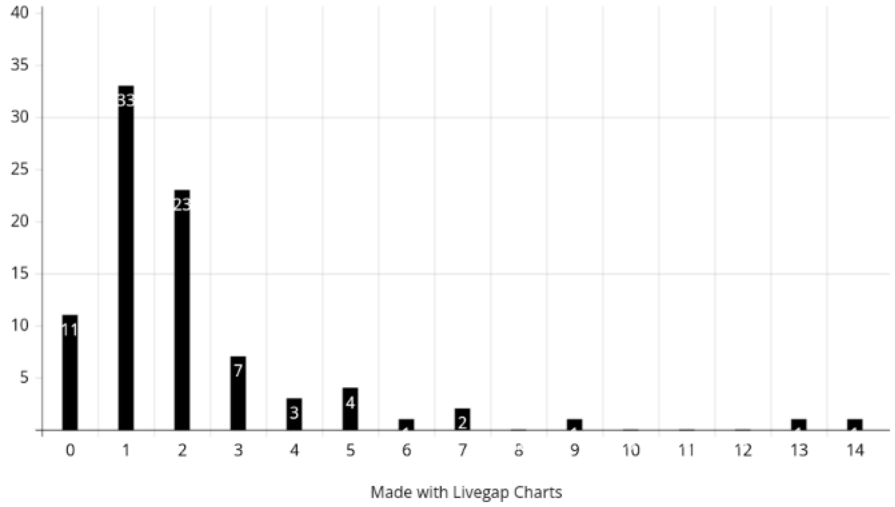
TD 6 - Statistiques (solutions)

Exercice 1. Les 87 étudiants d'une promotion ont répondu à la question « combien avez-vous de frère et sœurs ? » Voici la série brute obtenue :
2 2 5 2 3 1 1 1 1 2 1 1 0 2 2 5 1 4 2 2 1 0 2 2 1 2 1 1 2 1 3 1 2 1 0 0 1 3 13(treize) 1 0 2 1 1 1 1 6 1 3 1 0 1 5 7 2 1 2 2 3 0 3 2 2 7 14(quatorze) 9 5 4 1 0 1 2 1 3 2 1 2 2 0 4 1 0 1 0 1 1 1.

a) Présenter cette série statistique dans un tableau contenant les effectifs de chaque modalité ainsi que les effectifs cumulés, fréquences et fréquences cumulées.

	$Effectifs$	$Eff.\ cumulés$	$Fréq.\ (\%)$	$Fréq.\ cumulées\ (\%)$
0	11	11	$12,64$	$12,64$
1	33	44	$37,93$	$50,57$
2	23	67	$26,44$	$77,01$
3	7	74	$8,05$	$85,06$
4	3	77	$3,45$	$88,51$
5	4	81	$4,60$	$93,11$
6	1	82	$1,15$	$94,26$
7	2	84	$2,30$	$96,56$
9	1	85	$1,15$	$97,71$
13	1	86	$1,15$	$98,86$
14	1	87	$1,15$	100

b) Représenter graphiquement les effectifs.



c) Calculer la moyenne.

Solution : On a $\overline{x} = \dfrac{0 \times 11 + 1 \times 33 + \cdots + 14 \times 1}{87} = \dfrac{188}{87} \simeq 2,16.$

d) Déterminer la médiane. Quel est le pourcentage d'éléments de la série inférieurs ou égaux à la médiane ? Commenter.

Solution : L'effectif total est impair, la médiane est donc la $\frac{87+1}{2} = 44$e observation (après classement dans l'ordre croissant). D'après le tableau des effectifs cumulés, la médiane est $m = 1$. Le pourcentage d'observations inférieures ou égales à la médiane est, d'après le calcul des fréquences cumulées, égal à $50,57\%$ (très proche de 50%). Le terme « médiane » est bien adapté ici car sa fréquence cumulée est très proche de 50%, elle sépare donc bien les données en deux ensembles qui ont à peu près la même taille. Attention, il existe des séries statistiques dont la médiane a une fréquence cumulée très supérieure à 50% (par exemple si tout le monde avait répondu 1 au sondage), il convient donc de toujours préciser les fréquences cumulées lorsque l'on calcule une médiane.

e) Déterminer le premier quartile et le troisième quartile.

Solution : On a $87 \times \frac{1}{4} = 21,75$ et $87 \times \frac{3}{4} = 65,25$ donc le premier quartile est la 22e observation et le troisième quartile est la 66e observation. On a donc $q_1 = 1$ et $q_3 = 2$.

f) Calculer l'étendue de la série. Commenter l'utilité de cet indicateur dans le cas particulier de cette série.

Solution : L'étendue est $e = 14 - 0 = 14$. Cet indicateur ne donne pas beaucoup d'informations ici car valeurs sont concentrées autour des petits chiffres ($88,51\%$ des valeurs sont inférieures à 4). Il convient donc de compléter l'étendue avec d'autres indicateur de dispersion.

g) Calculer l'écart absolu moyen ($EAM = \frac{1}{n}\sum_{i=1}^{n} |x_i - \overline{x}|$).

Solution : On a $EAM \simeq 1,5117$.

h) Déterminer la variance et l'écart-type. Comparer l'écart-type et l'écart absolu moyen.

Solution : On a $Var(x) \simeq 5,8631$ et $\sigma = \sqrt{Var(x)} \simeq 2,42$. On a $\sigma \geqslant EAM$ (propriété vraie en toute généralité).

Exercice 2. Une étude des notes obtenues par deux classes d'une école à un test commun a fourni les résultats suivants :

	Classe 1	*Classe* 2
$Effectif$	20	30
$Moyenne$	12	10
$Écart-type$	4	6

Pour chacune des affirmations suivantes, dire si elle est vraie ou fausse, en justifiant la réponse :

a) La note moyenne des deux classe réunies est égale à 11.

Solution : La moyenne des deux classes réunies est $M = \frac{M_1 + M_2}{20+30}$ où $M_1 = 12 \times 20$ et $M_2 = 10 \times 30$. Donc $M = \frac{12 \times 20 + 10 \times 30}{50} = 10,8 \neq 11$. L'affirmation est fausse.

b) L'écart-type des notes des deux classe réunies est égal à 5.

Solution : La variance des deux classes réunies est $V = \frac{S}{50} - 10,8^2$ où $S = S_1 + S_2$ est la somme des carrés de toutes les notes (= somme des carrés des notes de la classe 1 + somme des carrés des notes de la classe 2). On peut calculer S_1 et S_2 grâce aux écart-types donnés. On a $\sigma_1^2 = V_1 = \frac{S_1}{20} - 12^2$, d'où $S_1 = (144 + 16) \times 20 = 3800$. De même pour la classe 2, on trouve $S_2 = (100 + 36) \times 30 = 4080$. Donc $V = \frac{7280}{50} - 10,8^2 = 28,96$ et $\sigma = \sqrt{V} \simeq 5,3814$. L'affirmation est donc fausse.

Exercice 3 (*)**.** On a relevé dans un magasin le montant des achats (en euros) un jour donné :

Prix d'achat (en euros)	*Nombre de clients*
$[10, 50[$	12
$[50, 70[$	12
$[70, 90[$	38
$[90, 110[$	31
$[110, 150[$	20
$[150, 170[$	14

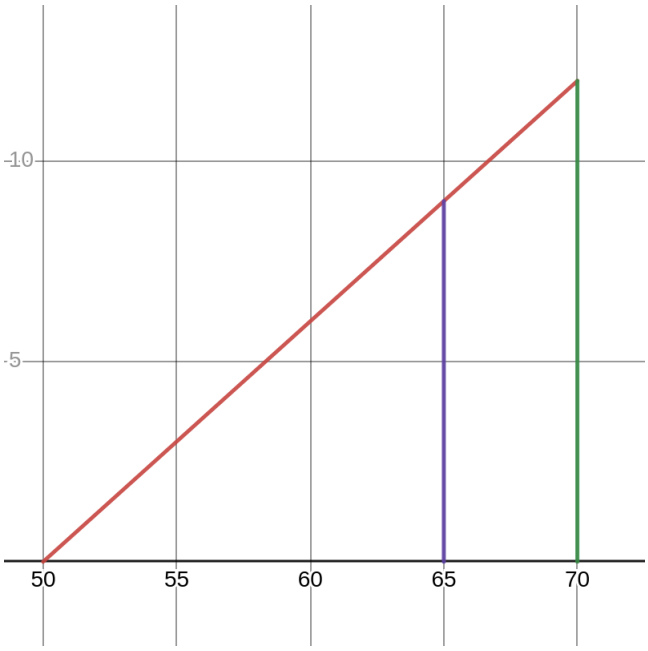
a) Tracer l'histogramme de cete série.

Solution : L'histogramme d'une série dont les valeurs sont regroupées en classe est un graphique où l'on dessine des rectangles dont la largeur correspond aux extrémités des classes (en abscisses) et la hauteur correspond aux densités d'effectifs (ou les densités de fréquences des classe) en ordonnées. Les densités d'effectifs (resp. de fréquence) sont les effectifs (resp. fréquences) des classes divisées par les longueurs (ou encore amplitudes) des classes. (Les fréquences cumulées seront utiles pour la dernière question.)

Prix	*Nb clients*	*Longueur*	*Densité*	*Fréq.*	*Fréq. cumulées*
$[10, 50[$	12	40	$0,3$	$9,45\%$	$9,45\%$
$[50, 70[$	12	20	$0,6$	$9,45\%$	$18,9\%$
$[70, 90[$	38	20	$1,9$	$29,9\%$	$48,8\%$
$[90, 110[$	31	20	$1,55$	$24,4\%$	$73,2\%$
$[110, 150[$	20	40	$0,5$	$15,7\%$	$88,9\%$
$[150, 170[$	14	20	$0,7$	$11,0\%$	$99,9\%(100\%)$

b) Donner (par interpolation linéaire) une valeur approchée du nombre de clients qui ont fait un achat d'un montant inférieur à 65 euros.

Solution : 12 clients ont dépensé entre 50 et 70 euros. Le nombre de clients ayant parmi eux qui ont dépensé moins de 65 euros est d'environ $\frac{15 \times 12}{20} = 9$ (produit en croix ou même théorème de Thalès comme dans le dessin ci-dessous). Il y a donc environ $12 + 9 = 21$ clients qui ont fait un achat d'un montant inférieur à 65 euros.



c) Calculer la médiane, le premier quartile et le troisième quartile.
Solution : Grâce au fréquences cumulées, on sait que la médiane se trouve dans la classe $[90, 110[$ (de longueur 20), le premier quartile dans la classe $[70, 90[$ (de longueur 20) et le troisième dans la classe $[110, 150[$ (de longueur 40). Par interpolation linéaire comme dans la question précédente, la médiane des prix d'achat est $m \simeq 90 + \dfrac{1,2 \times 20}{24,4} \simeq 91$ euros. De même, $q_1 \simeq 70 + \dfrac{6,1 \times 20}{29,9} \simeq 74,1$ euros et $q_3 \simeq 110 + \dfrac{1,8 \times 40}{15,7} \simeq 114,6$ euros

Exercice 4 (*)**.** La distribution des salaires dans une entreprise de 200 salariés est la suivante :

Salaire	*Eff.*	*Fréq.*	*Longueur*	*Densité fréq.*	*Fréq. cumulées.*
$[1000, 1600[$	25	$12,5\%$	600	$0,02$	$12,5\%$
$[1600, 2000[$	65	$32,5\%$	400	$0,08$	45%
$[2000, 2500[$	60	30%	500	$0,06$	75%
$[2500, 3000[$	30	15%	500	$0,03$	90%
$[3000, 4000[$	20	10%	1000	$0,01$	100%

a) Tracer l'histogramme de cette série.
On peut tracer des rectangles de largeur correspondant aux extrémités des classes et de hauteur la densité de fréquence (c'est pareil que la densité d'effectifs mais tout divisé par l'effectif total).

b) Déterminer la médiane, le premier quartile et le troisième quartile.
Solution : On a déjà q_3 grâce au tableau des fréquences cumulées, $q_3 = 2500$ (75% des gens ont un salaire inférieur à 2500 euros). La médiane est comprise entre 2000 et 2500 et q_1 est entre 1600 et 2000, comme pour l'exercice précédent, on trouve par interpolation linéaire $m \simeq 2083,3$ euros et $q_1 \simeq 1753,8$ euros.

c) Calculer une valeur approchée du salaire moyen.
Solution : On calcule le salaire moyen en prenant le centre de chaque classe,

$$\overline{x} \simeq \frac{1300 \times 25 + 1800 \times 65 + \cdots + 3500 \times 20}{200} \simeq 2185.$$

d) Calculer une valeur approchée de la variance et de l'écart-type.
Solution : On prend encore le centre des classes pour le calcul de la variance,

$$Var(x) \simeq \frac{1300^2 \times 25 + \cdots + 3500^2 \times 20}{200} - 2185^2 \simeq 368150,$$

et

$$\sigma = \sqrt{Var(x)} \simeq 606,75.$$

e) Pour chaque classe, calculer une valeur approchée de la somme des salaires de la classe.
Solution : La somme des salaires d'une classe peut être approchée par le centre de la classe $\times$ l'effectif de cette classe.

f) Dans un repère orthogonal, on considère le point qui a
— pour abscisse : la proportion des individus dont le salaire est inférieur à un nombre réel positif s (notons-la p) ;
— pour ordonnée : le rapport de la somme des salaires de la série inférieurs à s et de la somme de tous les salaires (notons-le Σ).

L'ensemble des points obtenus lorsque s varie de 0 à $+\infty$ est appelé la *courbe de concentration*. Construire à partir des données de l'exercice une courbe de concentration approchée.

Solution : On remplit le tableau suivant pour calculer p et Σ :

$centre$	$effectif$	$somme\ salaires$	$\sum sal.cumulés$	Σ	p
1300	25	32500	32500	$0,074$	$0,125$
1800	65	117000	149500	$0,343$	$0,45$
2250	60	135000	284000	$0,651$	$0,75$
2750	30	82500	367000	$0,839$	$0,9$
3500	20	70000	437000	1	1

On peut dessiner la courbe de concentration en reliant les points obtenus par des segments.

g) On appelle *indice de concentration* le nombre, compris entre 0 et 1, égal au double de l'aire comprise entre la courbe de concentration et la première bissectrice. Calculer, à partir des données de l'exercice, une valeur approchée de cet indice.

Solution : On peut calculer l'aire entre la courbe obtenue et la bissectrice $\mathcal{A}$ en calculant l'aire des trapèzes sous la courbe. L'indice de concentration est donc $I = 2\mathcal{A} = 2 \times \left(\frac{1}{2} - \Sigma \text{ trapèzes}\right) \simeq 0,137$. L'indicateur I est compris entre 0 et 1, il vaut 0 si la courbe suit exactement la bissectrice. Cet indice mesure à quel point les salaires sont bien répartis (plus l'indice est grand, plus la répartition des salaires est « inégalitaire »).

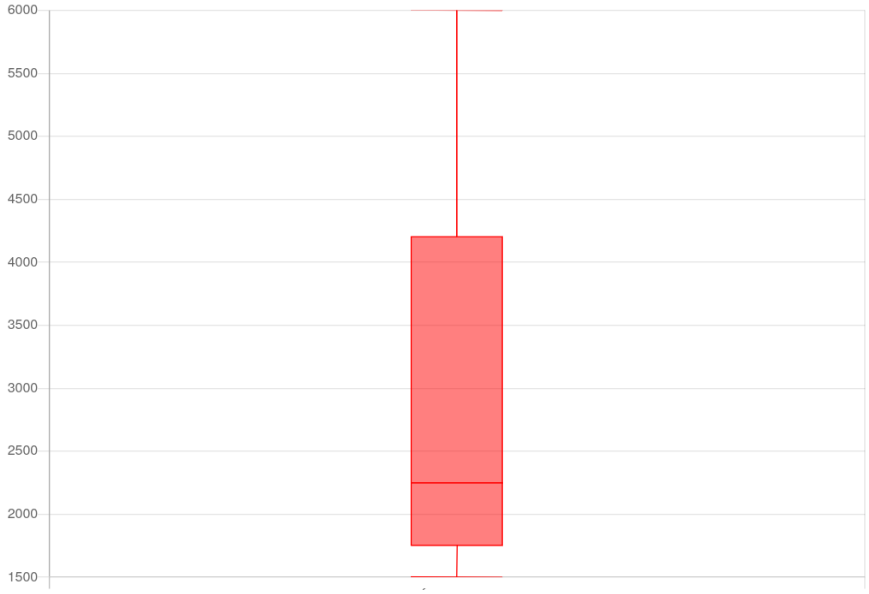
Exercice 5. La distribution des salaires dans une entreprise de 13 salariés est la suivante :
1500, 1750, 4000, 5000, 6000, 2250, 1600, 3500, 4200, 5500, 1700, 1800, 1900.

a) Calculer le salaire moyen.

Solution : On a $\overline{x} = \dfrac{1500 + 1750 + \cdots + 1900}{13} \simeq 3130,77$.

b) Résumer cette série à l'aide d'une boîte à moustaches (box-plot).

Solution : Pour dessiner la boîte à moustaches, on a besoin de la médiane et des quartiles. On trouve $m = 2250$ (la 7e observation), $q_1 = 1750$ (4e observation) et $q_3 = 4200$ (10e observation).



Les extrémités de la boîte sont en q_1 et q_3, les extrémités des moustaches sont en 1500 et 6000, on marque également la médiane (on peut aussi dessiner la boîte horizontalement).

Remarque : On peut décider si des observations sont aberrantes ou non à l'aide de l'écart interquartile, $EIQ = q_3 - q_1 = 2450$. Une observation est considérée comme aberrante si elle est supérieure à $q_3 + 1,5 \times EIQ$ ou inférieure à $q_1 - 1,5 \times EIQ$.

Mais ici, aucune observation ne semble aberrante car

— $q_3 + 1,5 \times EIQ = 4200 + 1,5 \times 2450 = 7875$, mais l'observation la plus élevée est 6000. La moustache va donc jusqu'à 6000.

— $q_1 - 1,5 \times EIQ = 1750 - 1,5 \times 2450 = -1925$, mais l'observation la plus basse est 1500. La moustache va donc jusqu'à 1500.