Les statistiques descriptives

Table des matières

1 Vocabulaire en statistiques, et histogramme.

1.1 Définitions

Définition 1: Série

Une série statistique est un ensemble de nombres qui représentent un même phénomène.

Exemple 1

L'ensemble des tailles des élèves de seconde au lycée, le nombre de calories ingérées pendant trente jours d'une personne, le temps passé sur son téléphone chaque jour pendant un an. . . Tous ces exemples sont des séries que l'on peut étudier en statistiques.

Question 1

Citer différentes séries statistiques que vous pouvez mesurer devant vous.

Définition 2: Indicateur statistique

Un **indicateur statistique** est un calcul à partir d'une série, qui permet de **résumer** l'information contenue dans cette série.
Il existe plusieurs types d'indicateurs : les indicateurs de position, et les indicateurs de dispersion.

1. Un **indicateur de position** d'une série donne un nombre qui situe l'ensemble de la série.
2. Un **indicateur de dispersion** est sensible à la dispersion de la série, c'est-à-dire à quel point il y a un écart entre les nombres qui constituent la série.

Exemple 2

La moyenne fait partie des indicateurs !

Question 2

À votre avis, la moyenne est un indicateur de position ou de dispersion ?

Définition 3: Effectif

L'effectif de la série correspond au nombre de fois où une valeur est répétée dans une série.

Exemple 3

Dans série statistique suivante :

$$x = (1; 1; 30; 90; 100; 100; 100; 1000)$$

La valeur 1 est répétée deux fois, et la valeur 100 trois fois.
On résumera la série sous la forme d'un tableau :

Valeur	Effectifs
1	2
30	1
90	1
100	3
1000	1

Il y a bien $2 + 1 + 1 + 3 + 1 = 8$ valeurs dans notre série !

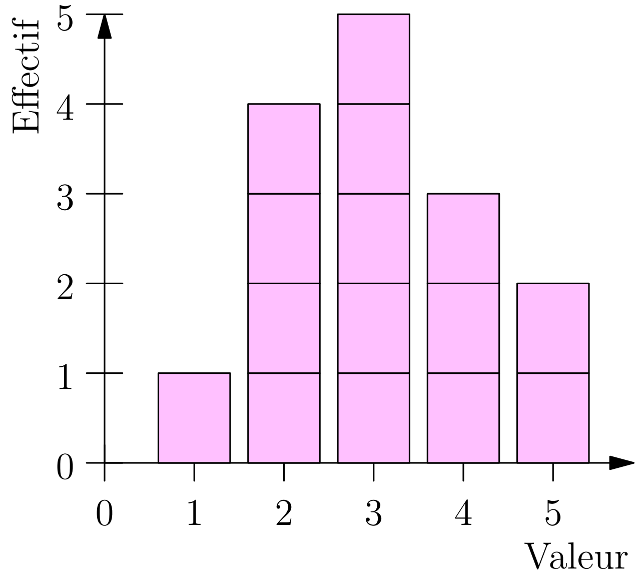
1.2 Histogramme

On regarde la série suivante.

$$x = (1; 2; 2; 2; 2; 3; 3; 3; 3; 3; 4; 4; 4; 5; 5)$$

On donne la représentation de cette série dans un tableau, puis dans un histogramme.

(x_i)	Effectifs
1	1
2	4
3	5
4	3
5	2



Exemple 4

Cette série pourrait représenter par exemple les notes d'un contrôle noté sur 5. Ici, 1 élève a eu $1/5$, 4 élèves qui ont eu $2/5$, 5 élèves ont eu $3/5$, 3 élèves ont eu $4/5$, et enfin 2 élèves ont eu $5/5$.

Question 3

Quel calcul permet de retrouver la taille de la série (x_i) de l'exemple précédent ?

Question 4

À quoi ressemblerait le **tableau** qui correspond à la série statistique suivante ?

$$y = (3; 3; 3; 3; 4; 4; 7; 7; 7; 7; 7; 7)$$

Question 5

Construisez l'**histogramme** de la série suivante :

$$y = (3; 3; 3; 3; 4; 4; 7; 7; 7; 7; 7; 7)$$

2 Indicateur de position

2.1 Moyenne d'une série

2.1.1 Définition et calculs d'une moyenne

Définition 4: Moyenne d'une série (sans regroupement par effectif)

Soit (x_i) une série statistique contenant n données. Alors, si on note $\overline{x}$ la moyenne de cette série, par définition :

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Définition 5: Moyenne d'une série (avec regroupement par effectif)

Soit (x_i) une série statistique, d'effectif (e_i) de taille n. Alors, si on note $\overline{x}$ la moyenne de cette série, on a :

$$\overline{x} = \frac{x_1 \times e_1 + x_2 \times e_2 + \ldots + x_n \times e_n}{e_1 + e_2 + \ldots e_n}$$

Exemple 5

Si on considère le tableau suivant, qui récapitule des notes d'élèves, regroupé par tranche de 5 points :

La série (x_i)	Valeur de la série retenue	Effectifs
$0 \leq x_i < 5$	2,5	12
$5 \leq x_i < 10$	7,5	23
$10 \leq x_i < 15$	12.5	8
$15 \leq x_i \leq 20$	17,5	10

Alors la moyenne de la série se calcule par :

$$\overline{x} = \frac{2{,}5 \times 12 + 7{,}5 \times 23 + 12{,}5 \times 8 + 17{,}5 \times 10}{12 + 23 + 8 + 10}$$

Question 6

Pour être sûr d'utiliser sa calculatrice correctement, montrez que la moyenne $\overline{x}$ de l'exemple précédent est :

$$\overline{x} = 9.00943396226415\ldots \approx 9$$

Question 7

Si j'ai eu trois notes (sur vingt) en dessous de 5, deux notes entre 5 et 10, et cinq notes entre 10 et 15, quelle est l'estimation de ma moyenne d'après les calculs précédents ?

2.1.2 Propriétés de la moyenne

Proposition 1: Linéarité de la moyenne

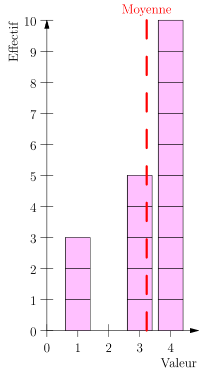
Quelque soit $k \in \mathbb{R}$, si on considère la série statistique $(x_i)_i$ et la série $(k \times x_i)_i$ alors :

$$\overline{k \times x} = k\overline{x}$$

Exemple 6

Si on considère la série suivante :

(x_i)	Effectifs
1	3
3	5
4	10

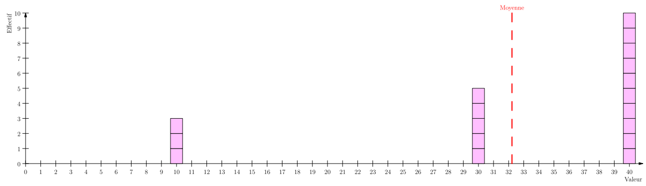


alors :

$$\overline{x} = \frac{1 \times 3 + 3 \times 5 + 4 \times 10}{3 + 5 + 10} = \frac{58}{18} = \frac{29}{9} \approx 3{,}2$$

Si maintenant on considère la série suivante :

(x_i)	Effectifs
10	3
30	5
40	10



Alors, on remarque que c'est la série précédente où chaque valeur est multipliée par 10. Donc

$$\overline{y} = 10\overline{x} = \frac{290}{9} \approx 32$$

Soit exactement la valeur trouvée pour la série précédente mais multipliée par 10.

Démonstration 1

Cette proposition se démontre avec l'égalité suivante, valable quelque soit $k \in \mathbb{R}$, et $(x_i)_{1 \leq i \leq n}$ une série statistique :

$$\begin{aligned}
\overline{k \times x} &= \frac{k \times x_1 + k \times x_2 + \ldots + k \times x_n}{n} \\
&= \frac{k \times (x_1 + x_2 + \ldots + x_n)}{n} \\
&= k \times \frac{x_1 + x_2 + \ldots + x_n}{n} \\
&= k \times \overline{x}
\end{aligned}$$

Proposition 2

Soit $a \in \mathbb{R}$ et $x = (x_i)$ une série statistique. Alors :

$$\overline{x+a} = \overline{x} + a$$

Et $x + a$ désigne la série $(x_i + a)_i$

Exemple 7

Si on considère la série :

$$x = (1; 2; 3; 4)$$

De moyenne $\overline{x} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2{,}5$, alors la série

$$y = (11; 12; 13; 14)$$

qui est la même que précédemment mais où **toutes les valeurs ont été augmentées de** 10 a comme moyenne

$$\overline{y} = \overline{x} + 10 = 12{,}5$$

Démonstration 2

Cette proposition se démontre à l'aide des égalités suivantes, vraies quelque soit la série statistique x et $a \in \mathbb{R}$:

$$\begin{aligned}
\overline{x+a} &= \frac{(x_1+a)+(x_2+a)+\ldots+(x_n+a)}{n} \\
&= \frac{(x_1+x_2+\ldots+x_n)+(a+a+\ldots+a)}{n} \\
&= \frac{x_1+x_2\ldots x_n+n\times a}{n} \\
&= \frac{x_1+x_2+\ldots x_n}{n} + \frac{n\times a}{n} \\
&= \overline{x} + a
\end{aligned}$$

Question 8

On considère la série suivante :

$$x = (1, 1, 3, 4, 8)$$

1. Calculez sa moyenne.
2. Calculer rapidement la moyenne de la série $(2, 2, 6, 8, 16)$
3. Calculer rapidement la moyenne de la série $(3, 3, 5, 6, 10)$

Question 9

Si on augmente de 10% toutes les valeurs d'une série, que se passe-t-il pour sa moyenne ?

Question 10

Si un professeur a oublié un point à tous ces élèves lors d'un contrôle, que peut-on dire de la moyenne des notes des élèves ?

2.2 Médiane d'une série

Définition 6: Médiane d'une série

La médiane m d'une série statistique (x_i) est un nombre tel qu'il y ait autant de valeurs x_i plus grande que m que de valeurs x_i plus petite. (Dit autrement « m coupe en deux la série statistique»).
Le mot «médiane» vient du latin *medianus* qui signifie «du milieu».

Voici un exemple :

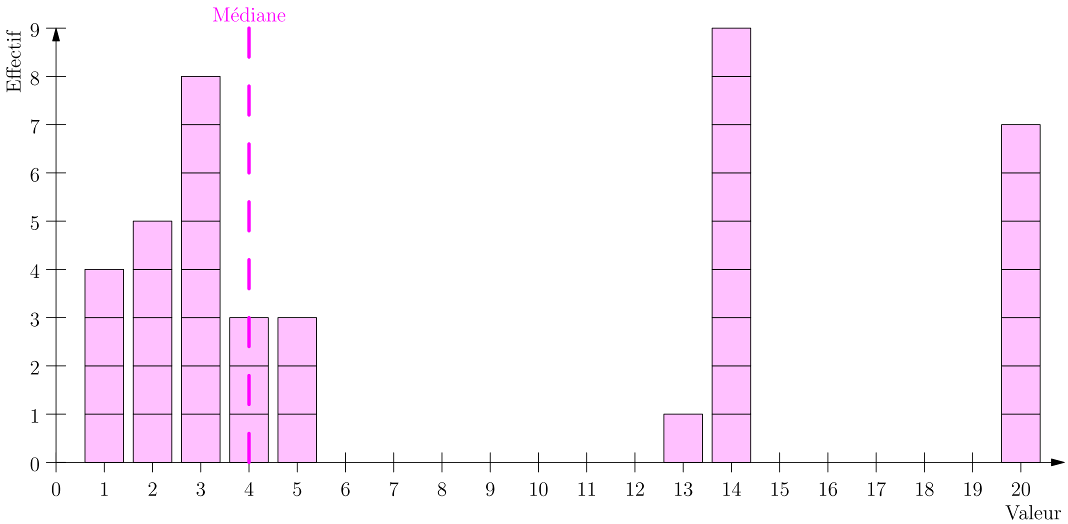


Figure 1 – Une série statistique pour comprendre la médiane.

Exemple 8

Dans la série illustrée à la figure 1 la médiane est 4.
C'est-à-dire qu'il y a autant de rectangles roses **avant** 4 que de rectangles roses **après**.

- $4+5+8+3=20$ valeurs sont *inférieures ou égales* à 4
- $3+1+9+7=20$ valeurs sont *supérieures* à 4

donc 4 est bien la valeur *médiane* de la série.

Question 11

D'où vient le calcul $4+5+8+3=20$ de l'exemple précédent ?

Question 12

Comment modifier la série statistique pour faire bouger la valeur de la moyenne **sans modifier la valeur de la médiane** ?

Question 13

Quelle est la valeur de la moyenne de la série illustrée à la figure 1 ?

Question 14

On change les dix 20 de la série 1 par dix 200.

— Comment change la médiane ?

— Comment change la moyenne ?

Que remarquez-vous ?

2.3 Quartile

Le quartile est le même concept que la médiane mais pour couper la série en **quart**.

2.3.1 Premier quartile

Définition 7: Premier quartile

Le premier quartile Q_1 est le nombre tel que 25% des valeurs de la série statistique soient plus petite que Q_1.

Exemple 9

Regardez la figure 2 pour un exemple !

2.3.2 Troisième quartile

Définition 8: Troisième quartile

Le troisième quartile Q_3 est le nombre tel que 75% des valeurs de la série statistique soient plus petite que Q_3.

Exemple 10

En prenant comme exemple la série de la figure 1 , et en plaçant cette fois-ci le premier quartile, la médiane, et le troisième quartile on obtient l'histogramme de la figure 2.

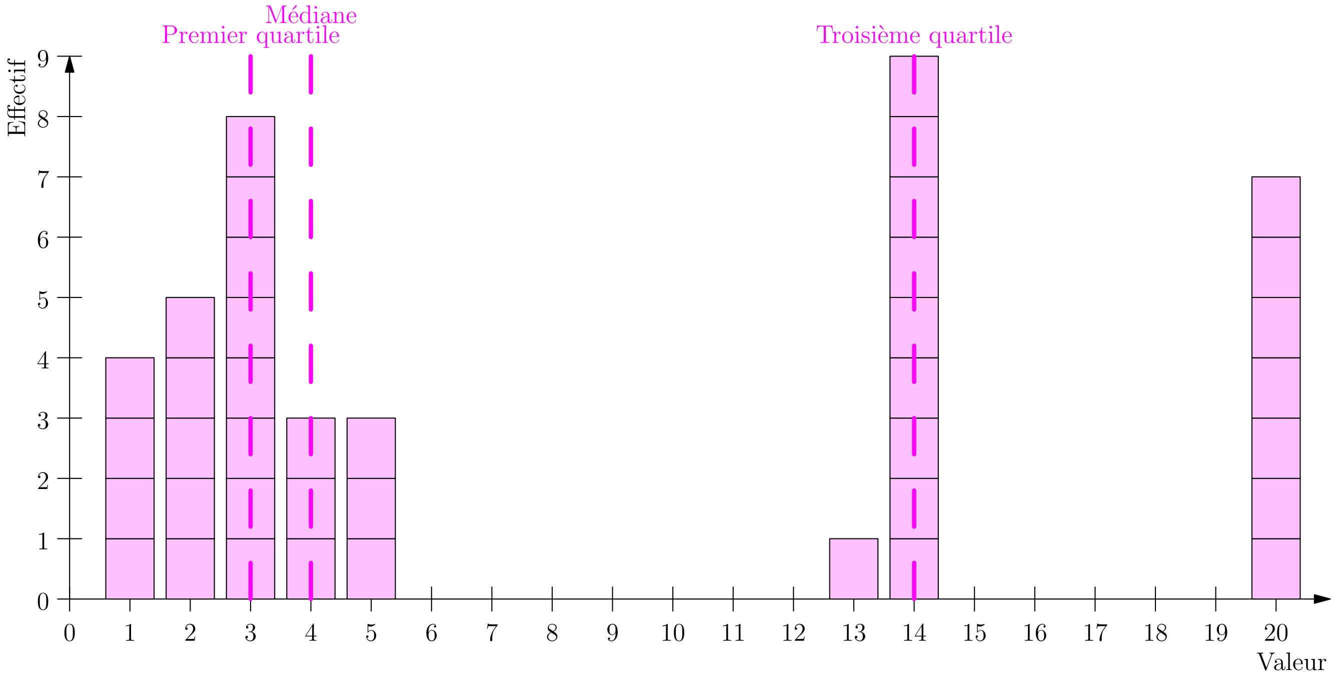


Figure 2 – Les trois quartiles représentés pour une série statistique

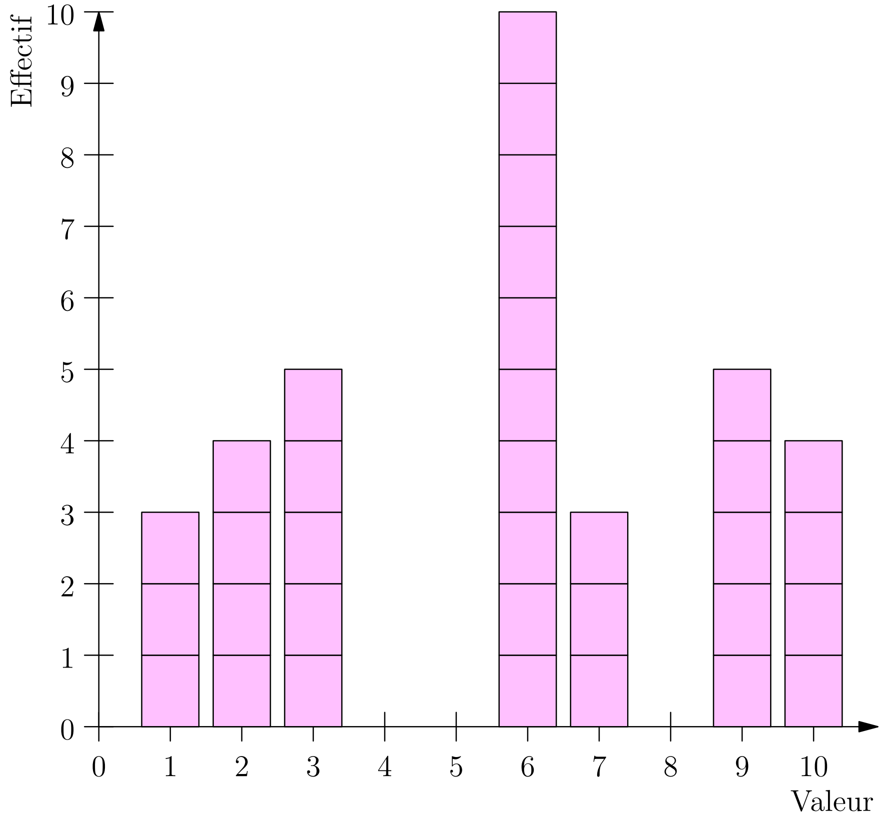
Exemple 11

Ici, sur la figure 2, on peut donc lire les informations suivantes :

- 25% des valeurs de la série sont inférieures ou égale à 3 (le premier quartile),
- 50% des valeurs de la série sont inférieures ou égale à 4 (la médiane, qui correspondrait au «deuxième quartile»),
- 75% des valeurs de la série sont inférieures ou égale à 14 (le troisième quartile),

Question 15

Voici la représentation d'une série. Déterminer le premier quartile, la médiane, ainsi que le troisième quartile.



Vous devriez trouver les éléments suivants :

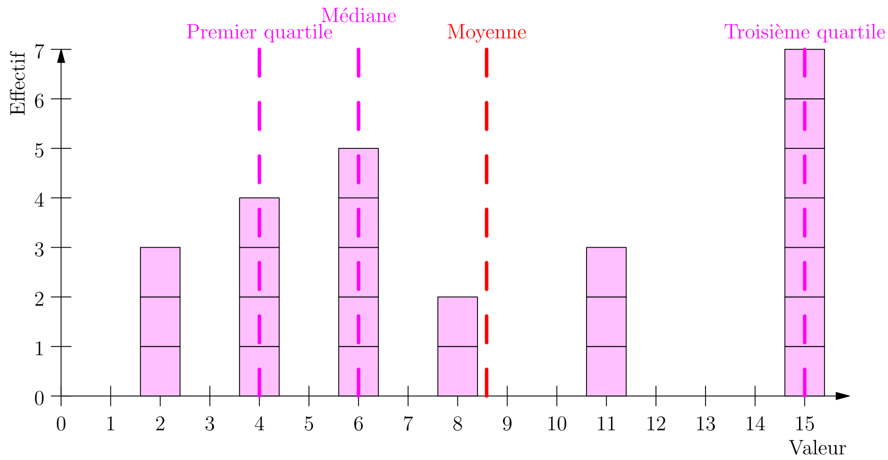
— $Q1 = 3$

— $M = 6$

— $Q3 = 9$

Question 16

Retrouver tous les indicateurs de la série suivante. Vérifiez bien que vous obtenez le même résultat que ceux affichés !



1. Combien de valeurs sont inférieures à la moyenne en proportion ?
2. Combien de valeurs sont inférieures à 11 en proportion ?

3 Indicateurs de dispersion

3.1 Écart-type, l'indicateur de dispersion lié à la moyenne

Définition 9: Écart-type

Soit (x_i) une série statistique, d'effectifs (n_i) de taille p. On note $\overline{x}$ la moyenne de la série. On calcule l'écart-type de cette série par :

$$\sigma = \sqrt{\frac{n_1(x_1 - \overline{x})^2 + \ldots + n_p(x_p - \overline{x})^2}{n_1 + \ldots + n_p}}$$

σ désigne la lettre «sigma» dans l'alphabet grec. Ce qui correspond à notre «s» minuscule.
L'écart-type mesure à quel point les valeurs sont dispersées **autour de la moyenne**.

Exemple 12

Voici une série :

$$(1\ ;\ 1\ ;\ 4\ ;\ 7\ ;\ 7\ ;\ 9\ ;\ 9\ ;\ 9)$$

Pour calculer son écart-type, on calcule d'abord la moyenne de la série :

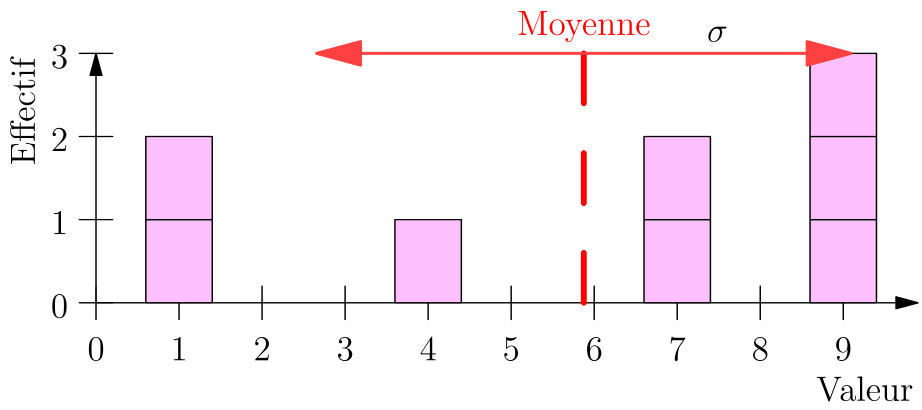
$$\overline{x} = \frac{2\times 1 + 4 + 2\times 7 + 3\times 9}{2+1+2+3} = \frac{47}{8} = 5{,}875$$

Puis, on utilise la formule vue plus haut :

$$\sigma = \sqrt{\frac{2(1-5{,}875)^2 + (4-5{,}875)^2 + 2(7-5{,}875)^2 + 3(9-5{,}875)^2}{8}} = 3{,}218\ldots$$

Sur la calculatrice, on peut d'abord calculer «sans la racine», et appliquer la racine carré à la toute fin.

On peut vérifier effectivement sur le graphique ci-dessous les calculs de l'exemple précédent :



Question 17

Pourquoi y'a-t-il un carré dans la formule de l'écart-type ? Et pourquoi une racine carrée ?

Reponse 1

On souhaite avoir un nombre σ qui soit grand dès qu'il y a des valeurs loin de la moyenne de la série. Dans l'exemple précédent, si on se concentre sur le premier terme, qui est $(1-5{,}875)^2$, on voit qu'on calcule la différence entre la valeur de la série 1, et la moyenne, $5{,}875$. Cette différence est élevée au carrée pour marquer cette différence d'autant plus, mais aussi pour que les différence s'ajoutent (un carré est toujours positif). La racine carrée sert à contrebalancer les carrés dans l'expression (attention, la racine n'annule pas les carrés ici, puisque en générale $\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$.

3.2 Écart interquartile, l'indicateur de dispersion lié à la médiane

Définition 10: Écart interquartile

Si Q_1 et Q_3 sont les deux quartiles d'une série, on définit $E = Q_3 - Q_1$ l'écart interquartile de la série.

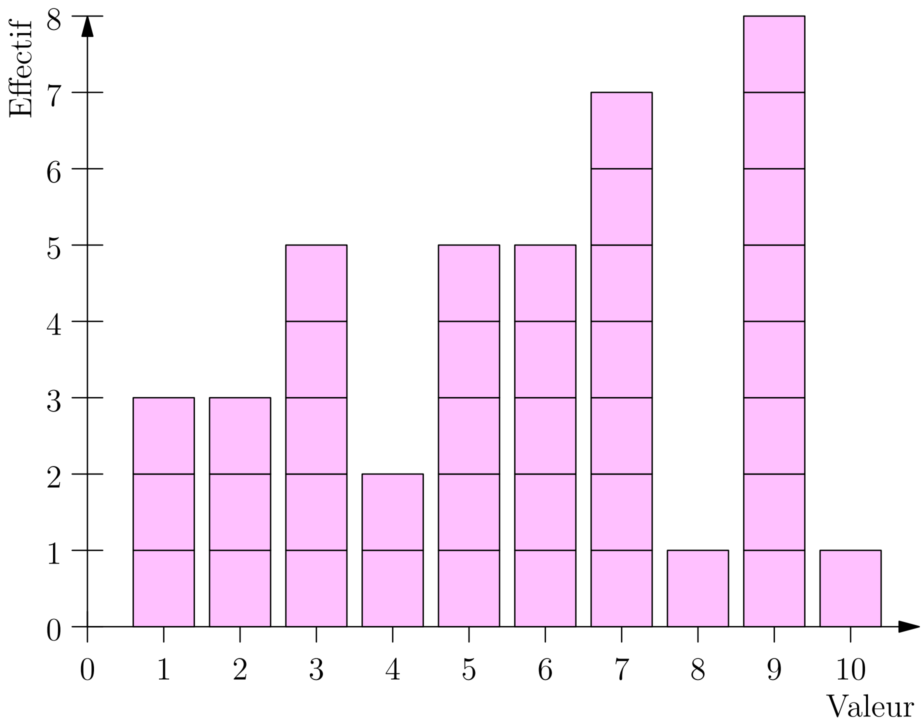
3.3 Visualiser les indicateurs de dispersion

3.3.1 Visualiser directement sur les diagrammes

Proposition 3

- L'écart-type est l'indicateur de dispersion associé à la moyenne. Plus l'écart-type est élevé, moins la moyenne est représentative de la série.
- De même, l'écart interquartile est un indicateur de dispersion associé à la médiane. Plus il est élevé, moins la mediane est représentative de la série.

À titre d'exemple, prenons la série (générée aléatoirement) suivante :



Puis, montrer la moyenne, et l'écart-type :

Exemple 13

Dans la figure 3, on voit la moyenne, et l'écart type qui est représenté par deux flèches rouges. La valeur de l'écart type correspond à la distance entre la moyenne et le bout d'une flèche (donc ici, $\sigma \approx 3$). On voit qu'effectivement les bouts de chaque flèche est proche d'une forte concentration des effectifs.

On va maintenant montrer l'écart-inter quartile, et la médiane :

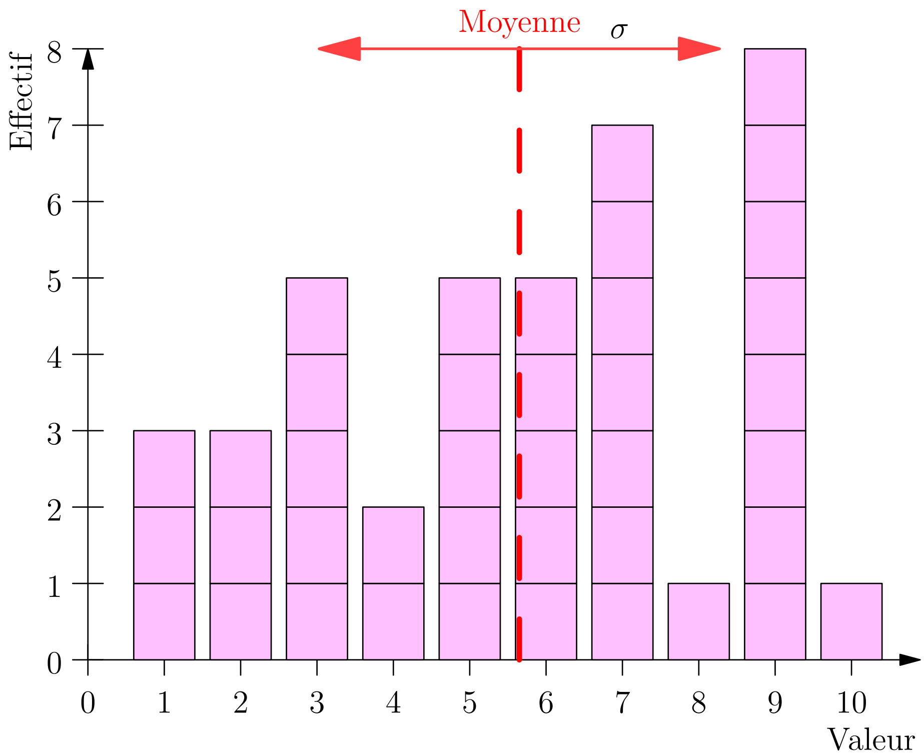


Figure 3 – Série avec la moyenne et l'écart-type.

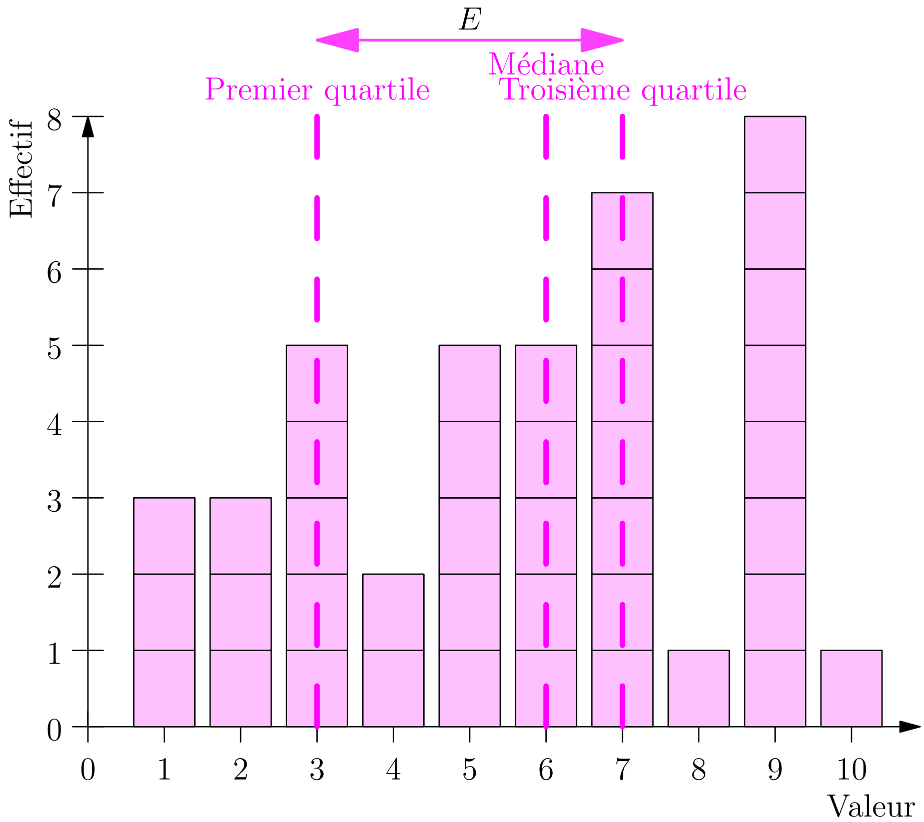


Figure 4 – Même série, mais en montrant la médiane et l'écart interquartile

Exemple 14

Cette fois, dans l'image , on a montré l'écart interquartile (qui vaut 4, puisque $7 - 3 = 4$) par rapport à la médiane. L'écart interquartile n'est pas centré autour de la médiane, puisqu'il correspond à la distance entre le premier quartile, et le troisième quartile. L'écart interquartile correspond est à rapprocher de l'étendue d'une série, sauf que l'on prend le premier et le troisième quartile comme extrémité.

Question 18

Dans cette situations, les valeurs de la médiane et de la moyenne sont proches, et leur indicateur respectif aussi.
Pourriez-vous imaginez une situation ou cela n'est pas le cas ?

3.3.2 À l'aide d'un diagramme en boite

Définition 11: Diagramme en boite

Un diagramme en boite (ou appelé aussi **moustache de chat**, allez savoir pourquoi) permet de visualiser la médiane, le premier quartile, et le troisième quartile. La boite commence au premier quartile, et se termine au troisième quartile. Un trait est ajouté pour montré la médiane. Les extrémités des « moustaches » correspondent au minimum et au maximum de la série.

Exemple 15

La figure 5 vous donnera un exemple avec la série vue précédemment, que vous pouvez retrouver sur la table 1.
La figure a été obtenue avec Python. Remarquez que l'on peut aussi tracer les diagrammes en boite verticalement.

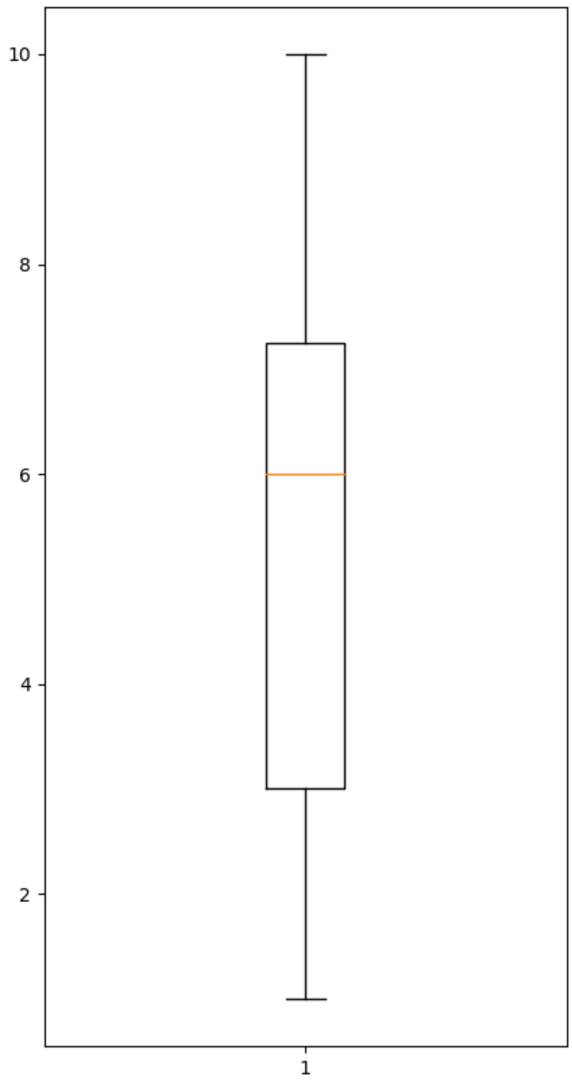


Figure 5 – Un diagramme en boite obtenu avec Python

Table 1 – Le diagramme en boite provient de cette série.

x	Effectif
1	3
2	3
3	5
4	2
5	5
6	5
7	7
8	1
9	8
10	1

Question 19

Retrouver la valeur de la médiane, et des quartiles sur le diagramme en boite de la figure 5, et comparer avec les valeurs obtenues à la figure 4 !

4 Étude de cas : pourquoi la moyenne n’est-elle pas un indicateur fiable ?

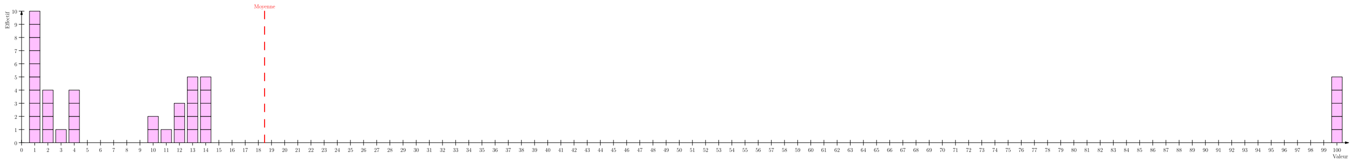


Figure 6 – Nous allons regarder cette série plus en détail.

Regardez la série 6 (oui, il faut zoomer). J’ai affiché la moyenne.

Question 20

Pourquoi la moyenne n’est pas un indicateur adapté à cette série ? Combien de valeurs sont autour de la moyenne ?

On devrait avoir une confirmation avec l’affichage de l’ecart-type :

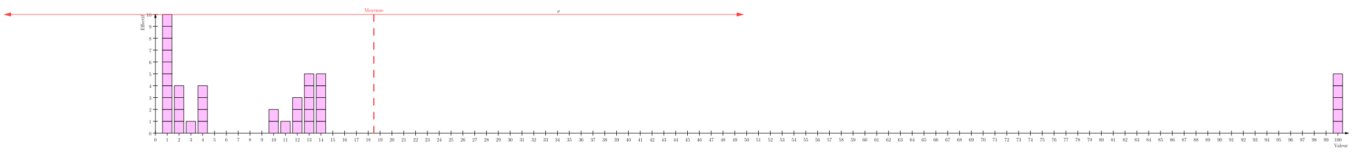


Figure 7 – Avec l’écart-type

Exemple 16

L’écart type est gigantesque ! C’est-à-dire que si l’on se contente de résumer la série à sa moyenne, on passe à coté de **beaucoup** d’information.

Question 21

Pourquoi l’écart-type est-il gigantesque dans la figure ?

On peut faire le même travail avec la médiane

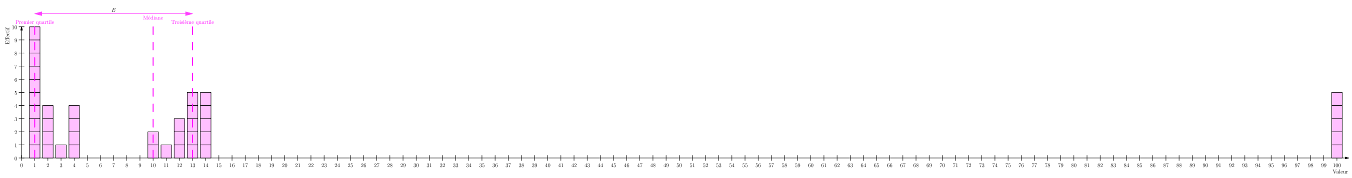


Figure 8 – Avec l’écart-type

Proposition 4

La moyenne est très sensible aux valeurs extrême, alors que la médiane est plus robuste.

Exemple 17

Si un milliardaire rentre dans une salle de classe, alors la moyenne de la richesse des personnes en classe sera de centaine de milliers d'euro. Cette somme ne représente la richesse d'absolument personne dans cette classe (le milliardaire est bien au dessus, et je vous promet que je ne possède pas des centaines de milliers d'euros, et je doute que vous non plus).