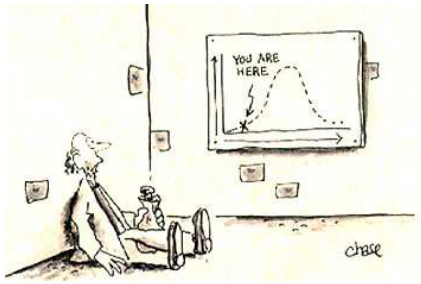
PREMIÈRE LEÇON

STATISTIQUES I



I - Lissage par les moyennes mobiles

Voici un tableau qui donne l'extension de la banquise au minimum de septembre de 1979 à 2007 :

Année	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
Extension ($\overline{\text{M}}\text{k}m^2$)	5,3	5,5	4,95	5,13	5,37	4,7	5	5,37	5,3	5,2	4,8	4,62	4,5	5,05	4,45
Année	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Extension ($\overline{\text{M}}\text{k}m^2$)	4,85	4,4	5,25	4,9	4,26	4,2	4,14	4,55	4,05	4,12	4,3	4,05	4,06	3,75	3,8*

FIGURE 1 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2

Représentons cette évolution par un nuage de points :

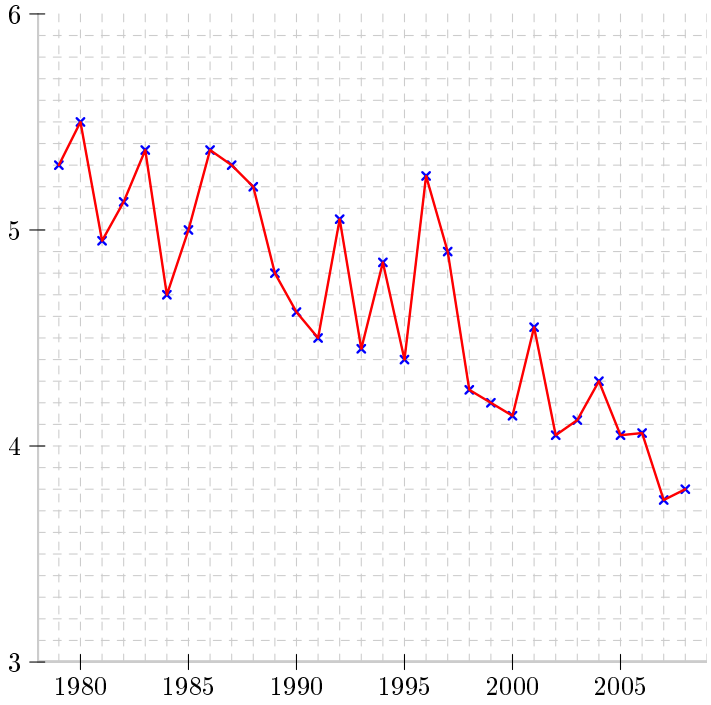


FIGURE 2 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2 - Graphe

On « sent » une tendance à la baisse mais des variations saisonnières la masque. Pour y remédier, on peut remplacer chaque terme par sa moyenne avec les termes voisins.

Définition 1 : moyenne mobile

On appelle moyenne mobile d'ordre k la moyenne arithmétique d'un terme avec les k termes voisins.

EN PRATIQUE

- si k est impair, c'est-à-dire s'écrit sous la forme $k = 2p+1$ alors on remplace chaque terme par la moyenne de ce terme, des p termes suivants et des p termes précédents ;
- Si k est pair, c'est-à-dire s'écrit sous la forme $k = 2p$, c'est moins évident : on remplace chaque terme par la moyenne de ce terme, des p termes suivants et des p termes précédents mais les termes extrêmes sont affectés d'un coefficient 1/2.

Définition 2 : Lissage par moyennes mobiles

Lisser une série chronologique c'est remplacer la série initiale par la série des moyennes mobiles.

Par exemple, reprenons les données du tableau ?? et effectuons un lissage par moyennes mobiles d'ordre 5. Il y a 30 termes au départ. L'ordre 5 étant impair, on va remplacer chaque terme par la moyenne de ce terme, des 2 précédents et des deux suivants : on est donc obligé de commencer par le troisième terme et de terminer par le vingt-huitième.
On remplacera alors t_3 par

$$t'_3 = \frac{t_1 + t_2 + t_3 + t_4 + t_5}{5} = \frac{5,3 + 5,5 + 4,95 + 5,13 + 5,37}{5} = 5,25$$

Remplissez le tableau suivant :

Année	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93
Extension ($\overline{\text{M}}km^2$)	/	/	5,25	5,13	5,03	...	...	5.114	5.134	5.058	4,884	4,834	...	4,694	4,65
Année	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08
Extension ($\overline{\text{M}}km^2$)	...	...	4,732	4,602	4,55	4,41	4,24	4,212	4,232	4,214	4,116	4,056	...	/	/

FIGURE 3 – Extension de la banquise au minimum de septembre de 1979 à 2007 en millions de km^2 lissé par moyennes mobiles d'ordre 5

Tracez alors sur le graphique ?? la courbe obtenue.

II - Mesures de tendance centrale

a. Le mode

i. Définition

Définition 3 : mode

Mesure qui correspond à la valeur ou à la modalité la plus fréquente. S'il y a plusieurs mode, on dit que la distribution est multimodale.

C'est la mesure la plus simple à évaluer.

ii. Distribution polymodale

Considérons le nouvel exemple suivant :

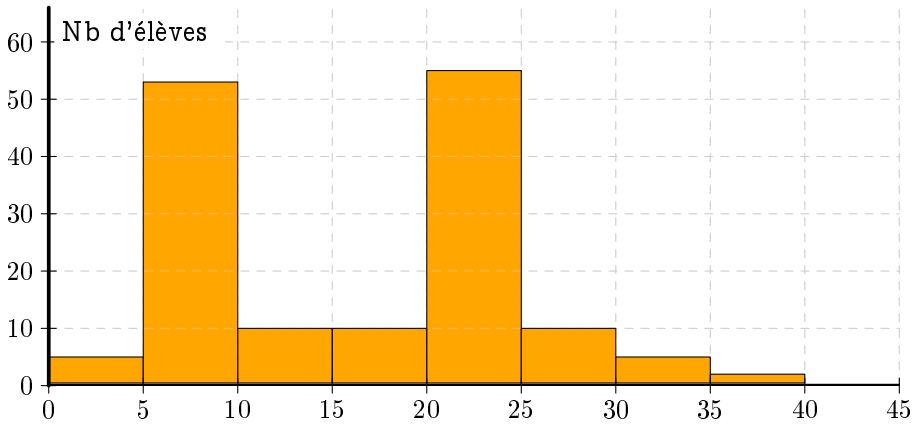


FIGURE 4 – Répartition des prisonniers syldaves selon la durée de leur peine (en années)

Ce graphique présente deux classes modales : [5 ; 10[et [20 ; 25[.

b. La médiane

i. Conventions

Dans toute la suite, on étudiera une population, notée E, et une variable statistique quantitative X définie sur E.

Exemple 1 :

Si on étudie par exemple le nombre de poupées Barbue que possèdent les élèves de la classe de BTS domotique
- E est l'ensemble des élèves de la classe
- X est la fonction qui, à un élément de E, associe le nombre de poupées Barbue qu'il ou elle possède.

Si E possède n éléments, on notera $V = \{x_1, x_2, \cdots, x_{n-1}, x_n\}$ l'ensemble **ordonné par valeurs croissantes** des valeurs prises par X.

Remarque 1 :

Notez bien que certains éléments de V peuvent être égaux : en effet, deux élèves différents peuvent avoir le même nombre de poupées Barbue.

ii. Définition

Définition 4 : médiane

La médiane M_e est un nombre tel que :
- **au moins** 50% des éléments de V sont inférieurs à M_e,
- **au moins** 50% des éléments de V sont supérieurs à M_e.

Exemple 2 : calcul de la médiane

- $V = \{1, 2, 2, 5, 5, 8, 8, 9, 37\}$
  - ▷ $M_e = 5$
- $V = \{1, 2, 2, 5, 8, 8, 9, 37\}$
  - ▷ $M_e = \frac{5+8}{2} = 6,5$

iii. Cas des données groupées par valeurs

Considérons par exemple la répartition des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges :

Nombre d'années après le Brevet	Nombre de ministres	Effectifs cumulés croissants
0	69	69
1	31	100
2	15	
3	6	
4	3	
5	1	

FIGURE 5 – Répartition des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges

Il y a un nombre impair de données (125). La médiane correspond donc à la donnée de rang 63 (62 après, 62 avant). Grâce à la troisième colonne, on trouve que la 63^e valeur vaut 0. Ainsi la médiane $M_e = 0$. On en déduit qu'au moins 50 % des ministres n'ont pas poursuivi d'étude après le Brevet.

Il peut être plus pratique de travailler avec les fréquences :

Nombre d'années après le Brevet	Nombre de ministres	Pourcentage des ministres	Fréquence	Fréquences cumulées croissantes
0	69	69		
1	31	100		
2	15			
3	6			
4	3			
5	1			

FIGURE 6 – Répartition en pourcentage des ministres syldaves selon le nombre d'années qu'ils ont étudié après leur Brevet des Collèges

Il suffit alors de regarder à quelle valeur correspond la fréquence cumulée 50 %.

iv. Cas des données groupées par classes

Lorsque les données sont regroupées par classes, on peut, soit procéder graphiquement, soit effectuer un calcul. Dans les deux cas il s'agira d'une approximation.
Considérons l'exemple suivant :

Nombre d'heures	Nombre d'élèves	Pourcentage des élèves	Pourcentage cumulé des élèves
0-5	300		
5-10	420		
10-15	500		
15-20	330		
20-25	250		
25-30	160		
30-35	40		
Total	2000	100 %	

FIGURE 7 – Répartition et répartition cumulée des 2000 élèves des lycées Perrin et Goussier de Rezé selon le nombre d'heures de travail personnel par mois.

Méthode graphique On trace le *polygone des fréquences cumulées croissantes :*

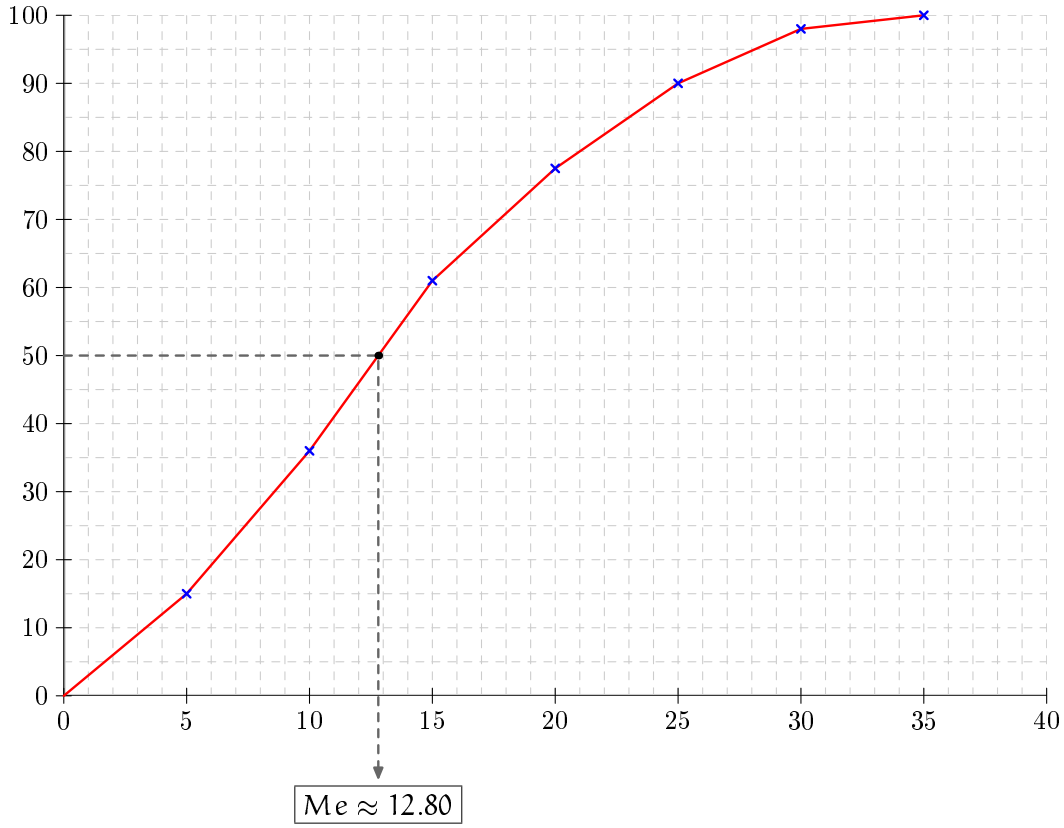


FIGURE 8 – Polygone des fréquences cumulées croissantes des 2000 élèves des lycées Perrin et Goussier de Rezé selon le nombre d'heures de travail personnel par mois.

À chaque borne supérieure des classes, on fait correspondre la fréquence cumulée et on relie les points par des segments (cela constitue une approximation mais le fait de regrouper les élèves en classe en était déjà une).

Comme la médiane correspond à 50 % des effectifs, on lit sur le graphique l'abscisse du point du polygone d'ordonnée 50 : environ 13.

Méthode analytique P est le point de la droite (M_2M_3) d'ordonnée 50. On détermine donc une équation de la droite (M_2M_3) et on calcule l'abscisse cherchée.
Ici $M_2(10\,;36)$ et $M_3(15\,;61)$. L'équation réduite de (M_2M_3) est de la forme $y = ax + b$ avec a le coefficient directeur.

$$a = \frac{y_{M_3} - y_{M_2}}{x_{M_3} - x_{M_2}} = \frac{61-36}{15-10} = 5$$

L'équation devient donc $y = 5x + b$. Pour déterminer b, il suffit d'utiliser le fait que M_2 appartient à (M_2M_3) donc vérifie son équation :

$$36 = 5 \times 10 + b \iff b = 36 - 50 = -14$$

Finalement, l'équation réduite cherchée est $y = 5x - 14$. Il reste à chercher le point de cette droite d'ordonnée 50 :

$$50 = 5x_p - 14 \iff x_p = \frac{50+14}{5} = 12,8$$

On trouve donc $M_e \approx 12,8$.

c. Moyenne

i. Données non groupées

Soit $x_1, x_2, \ldots, x_n$ une série statistique. La moyenne $\overline{x}$ est l'unique valeur que devrait prendre chacune des données pour que la somme des données soit préservée.
En d'autres termes, on cherche $\overline{x}$ tel que

$$x_1 + x_2 + \cdots + x_n = \underbrace{\overline{x} + \overline{x} + \cdots + \overline{x}}_{n \text{ termes}}$$

On en déduit que :

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

On utilise souvent le symbole Σ pour représenter une somme (sigma est la lettre grecque correspondant à notre S).
Ainsi $\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$.
La formule de la moyenne devient donc :

$$\overline{x} = \frac{\sum_{i=1}^{n} x_1}{1} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Faut-il vraiment un exemple ?...
Bon, voici le relevé des pourcentages de réussite lors du dernier exercice de tir des généraux de l'armée syldave :

12 22 32 2 24 2 5 33

Quel est le pourcentage moyen de réussite ?

ii. Données groupées par valeurs

Reprenons le tableau **??**. Si nous voulons calculer le nombre moyen d'années d'étude après le brevet des ministres syldaves il faudrait additionner 69 zéros, 31 un, 15 deux, 6 trois, 3 quatre, 1 cinq et diviser le tout par le nombre total de ministres à savoir 125 d'après ce que nous venons de voir.
Nous pouvons simplifier les choses en notant e_i l'effectif de la valeur x_i :

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} e_i \cdot x_i$$

Pour les ministres, $\overline{x} = \frac{1}{125}(69 \times 0 + 31 \times 1 + 15 \times 2 + 6 \times 3 + 3 \times 4 + 1 \times 5) = \cdots$

iii. Données groupées par classes

Regrouper des données par classe constitue déjà une approximation mais a le désavantage de nous empêcher d'utiliser nos moyens de calculer la moyenne. Pour y remédier, nous allons effectuer une approximation supplémentaire en prenant comme représentant d'une classe son **milieu**.

Âge	Milieu de classe (m_i)	Nombre de femmes
15-20		202
20-25		204
25-35		359
35-45		338
45-65		304
65-75		10
Total		1414

FIGURE 9 – Répartition des femmes du harem du Grand Protecteur de la Syldavie selon l'âge avec le milieu des classes

Nous pouvons alors trouver une approximation de l'âge moyen des femmes du Harem :

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{n} m_i e_i = \frac{202 \times 17,5 + \cdots}{1414} = \cdots$$

III - Quartiles et diagrammes en boîte

a. L'idée

Pour avoir une idée un peu plus précise de la série statistique étudiée, on voudrait séparer notre population en 4 groupes au lieu de 2 comme cela a été fait avec la médiane.
On a donc envie de calculer les médianes des parties basses et hautes.
On a également comme cahier des charges d'avoir au moins 25% des valeurs prises par x inférieures au premier quartile Q_1 et au moins 75% des valeurs prises par x inférieures au troisième quartile Q_3.

b. Expérimentons

Exemple 3 : huit éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8\}$

On peut séparer l'effectif en quatre groupes de même effectif égal à 25% de l'effectif total donc ça « colle »

$$V = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

On peut prendre
– $Q_1 = \frac{2+3}{2} = 2,5$ et 25% des effectifs ont une valeur inférieure à Q_1
– $M_e = \frac{4+5}{2} = 4,5$ et 50% des effectifs ont une valeur inférieure à M_e
– $Q_3 = \frac{6+7}{2} = 6,5$ et 75% des effectifs ont une valeur inférieure à Q_3
et Q_1 et Q_3 sont bien les médianes respectives des parties basses et hautes.
On peut séparer l'effectif en parties hautes et basses

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$$

On peut prendre

Exemple 4 : dix éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

- $Q_1 = 3$ et $3/_{10} = 30\% (\geqslant 25\%)$des effectifs ont une valeur inférieure à Q_1
- $M_e = \frac{5+6}{2} = 5,5$ et 50% des effectifs ont une valeur inférieure à M_e
- $Q_3 = 8$ et $8/_{10} = 80\% (\geqslant 75\%)$ des effectifs ont une valeur inférieure à Q_3

et Q_1 et Q_3 sont bien les médianes respectives des parties basses et hautes.

Exemple 5 : onze éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$

On peut séparer l'effectif en parties haute et basse

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$$

On peut prendre

- $Q_1 = 3$ et $3/_{11} \simeq 27\% (\geqslant 25\%)$des effectifs ont une valeur inférieure à Q_1
- $M_e = 6$ et 50% des effectifs ont une valeur inférieure à M_e
- $Q_3 = 9$ et $9/_{11} = 82\% (\geqslant 75\%)$ des effectifs ont une valeur inférieure à Q_3

et Q_1 et Q_3 sont bien les médianes respectives des parties basses et hautes.

Exemple 6 : neuf éléments

$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$

On peut séparer l'effectif en parties haute et basse avec la médiane au milieu :

$$V = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

et prendre $Q_1 = \frac{2+3}{2} = 2,5$, mais...
... seulement $2/_9 \simeq 22\%$ des effectifs ont une valeur inférieure à Q_1, ce qui est contraire à notre cahier des charges.
Dans ce cas particulier, on va inclure la médiane dans les parties basses et hautes pour éviter cet écueil.
Cela donne :

- Partie Basse = $\{1, 2, 3, 4, 5\}$ donc $Q_1 = 3$ et $3/_9 \simeq 33\% \geqslant 25\%$ des effectifs ont une valeur inférieure à Q_1, ce qui convient.
- Partie Haute = $\{5, 6, 7, 8, 9\}$ donc $Q_3 = 7$ et $7/_9 \simeq 78\% \geqslant 75\%$ des effectifs ont une valeur inférieure à Q_1, ce qui convient.

Remarque 2 : cycle

Il est aisé de constater que ces observations vont se répéter par cycle de longueur 4.
C'est quand V a un nombre d'éléments égal à un multiple de 4 plus 1 que nous devons être prudents.
D'ailleurs, toutes les machines à calculer ne s'accordent pas sur le calcul des quartiles.
La méthode que je vous propose est la plus cohérente et sera en accord avec la détermination graphique des quartiles que nous verrons bientôt.
Elle permet également d'avoir une définition rigoureuse comme nous allons le voir.

C'est quand V a un nombre d'éléments égal à un multiple de 4 plus 1 que nous devons être prudents.

c. Lecture graphique dans le cas des données groupées en classe

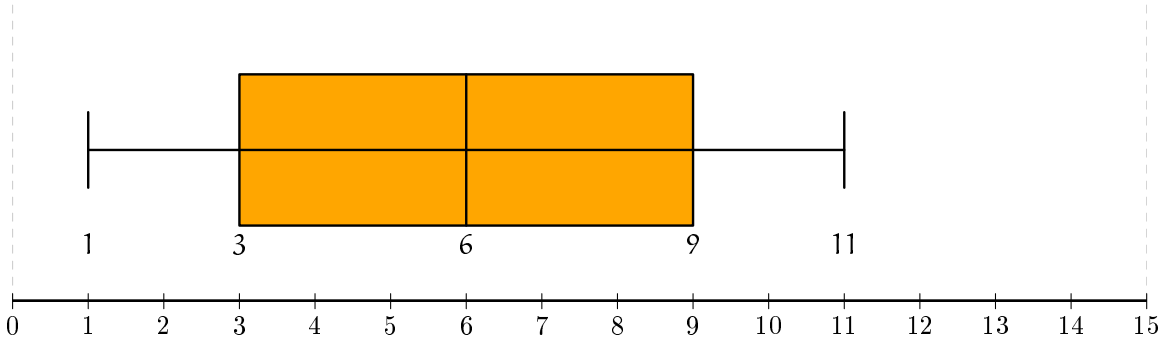
Complétez le graphique **??** qui avait été obtenu grâce au tableau **??**.

Définition 5 : quartiles

Le premier quartile est obtenu en prenant la médiane de la sous-série contenant les observations dont le rang est strictement inférieur à celui de la médiane (*la partie basse*) pour autant qu'au moins 25% des observations soient inférieures ou égales à cette valeur.
Sinon, il faut inclure la médiane dans la partie basse.
Le troisième quartile est obtenu en prenant la médiane de la sous-série contenant les observations dont le rang est strictement supérieur à celui de la médiane (*la partie haute*) pour autant qu'au moins 75% des observations soient inférieures ou égales à cette valeur.
Sinon, il faut inclure la médiane dans la partie haute.

d. Boîte à moustache

Après avoir calculé les quartiles, on peut les regrouper dans un tableau. Il est toutefois plus parlant de dresser un diagramme en boîte, ou diagramme de Tukey ou encore boîte à moustache. Reprenons pour cela l'exemple **??**



IV - Mesures de dispersion

a. Écart interquartile

Continuons à exploiter nos quartiles. Environ 50 % de la population a une modalité entre Q_1 et Q_3 : en observant la boîte à moustache décrivant une série statistique, on peut affiner sa description grâce à **l'écart interquartile**.

Définition 6 : écart interquartile

La distance entre Q_1 et Q_3 est appelé écart interquartile.

Observons deux exemples :

Exemple 7 : comparaison des écarts interquartiles

Voici un tableau donnant les notes au Bac de deux classes :

Notes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Effectifs classe A	0	0	0	1	1	0	2	3	2	9	5	9	5	1	2	1	0	2	1	1	0
Effectifs classe B	0	5	5	5	5	0	0	0	0	0	3	0	5	5	0	2	0	5	0	5	0

Voici les boîtes à moustaches correspondant :

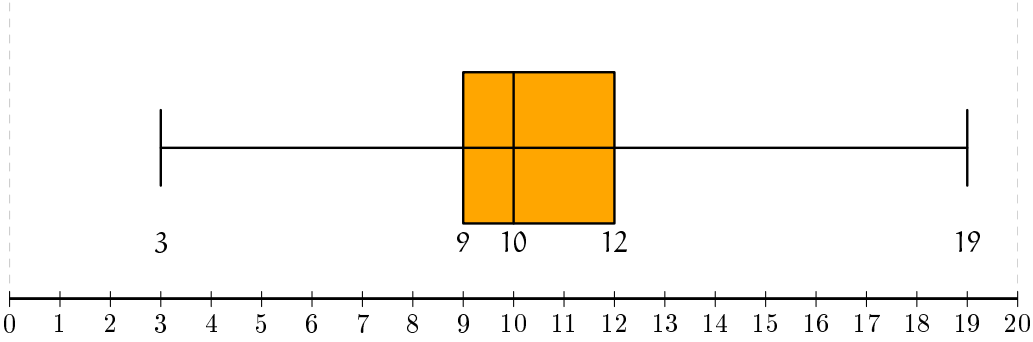


FIGURE 10 – Notes au Bac de la classe A

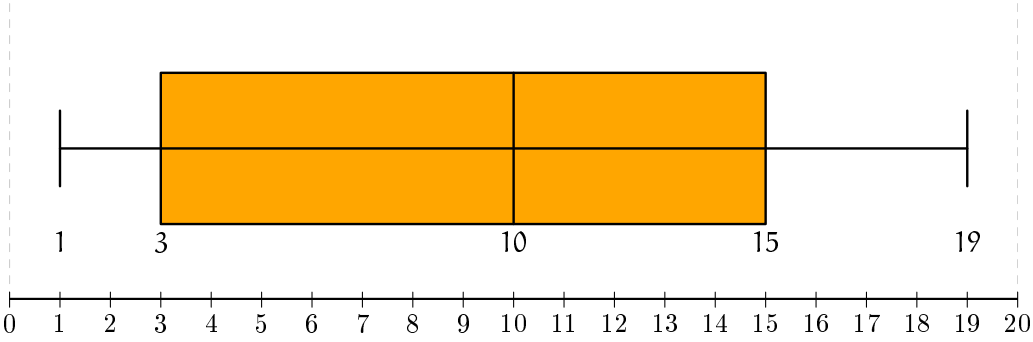


FIGURE 11 – Notes au Bac de la classe B

Dans les deux cas la médiane vaut 10 : cela signifie qu'au moins 50 % des élèves de chaque classe a eu moins de la moyenne. Cependant, les écarts interquartiles sont nettement différents : $Q_3 - Q_1 = 3$ dans la classe A mais $Q_3 - Q_1 = 12$ dans la classe B. La classe B est nettement plus hétérogène en terme de résultats. En effet, dans la classe A, 50 % environ des élèves ont eu entre 9 et 12 alors que dans la classe B la moitié des notes se situent entre 3 et 15.

b. Variance et écart-type

La dispersion peut également se mesurer autour de la moyenne.
Considérons une série statistique quelconque :

Valeurs x_i	0	1	2	3	4	7
Effectifs n_i	1	2	2	4	3	2

FIGURE 12 – Série statistique quelconque

À partir de ces données, calculez la moyenne $\overline{x}$ puis remplissez le tableau suivant proposant deux façons de « mesurer »pour chaque valeur « l'éloignement »par rapport à $\overline{x}$.

$x_i - \bar{x}$						
$(x_i - \bar{x})^2$						

FIGURE 13 – Essais de mesures de dispersion par rapport à la moyenne

Calculez dans chacun des trois cas l'éloignement moyen, c'est-à-dire la moyenne des écarts...

On peut visualiser les écarts sur le schéma suivant :

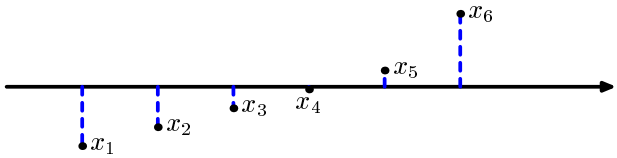


FIGURE 14 – Visualisation de l'écart par rapport à la moyenne

On peut prouver (à titre d'exercice...) que l'écart correspondant à la première ligne du tableau ?? est toujours nul. On préfère donc utiliser la moyenne des écarts de la deuxième ligne qu'on appelle **variance**.

Définition 7 : variance

On appelle variance d'une série quelconque à caractère quantitatif discret le nombre :

$$\mathrm{V} = \frac{1}{\mathrm{N}} \sum_{i=1}^{p} n_i (x_i - \overline{x})^2 = \sum_{i=1}^{p} f_i (x_i - \overline{x})^2$$

en notant n_i les effectifs et f_i les fréquences.

Remarque 3 : cas d'un regroupement en classe

dans le cas d'un regroupement en classe, on considère, comme pour le calcul de la moyenne (cf ??), le milieu des classes.

La variance est homogène au carré des valeurs x_i : on préfère donc en prendre la racine carrée pour revenir à une grandeur homogène aux valeurs mesurées. L'ÉCART-TYPE A DONC LA MÊME UNITÉ QUE LA POPULATION !

Définition 8 : écart-type

L'écart-type d'une série est la racine carrée de la variance. On le note souvant σ.

Reprenez les séries de l'exemple ?? et calculez les moyennes et écarts-type. On peut visualiser cette dispersion en traçant un diagramme à bâton et en mettant en évidence l'intervalle $[\overline{x} - \sigma\,;\, \overline{x} + \sigma]$.

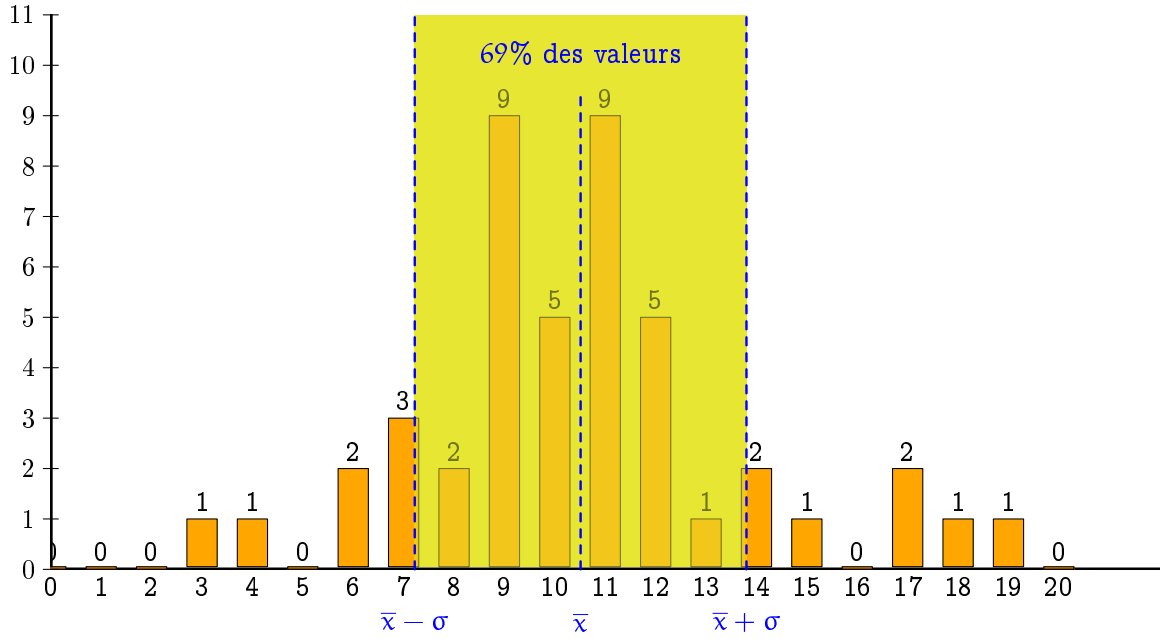


FIGURE 15 – Notes au Bac de la classe A

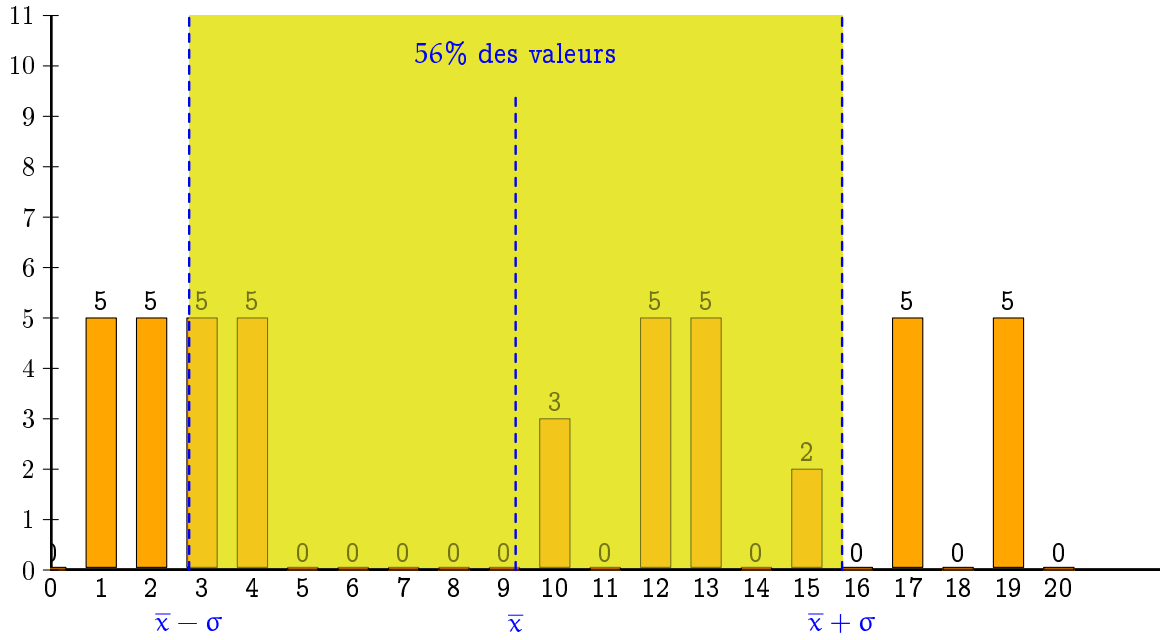


FIGURE 16 – Notes au Bac de la classe B