Séquence 4

Statistiques

Sommaire

Introduction

« *Etude méthodique des faits sociaux par des procédés numériques (classements, dénombrements, inventaires chiffrés, recensements) destinée à renseigner les gouvernements* » : ceci est la définition du mot « **statistique** » dans le dictionnaire Petit Robert.

Dès l'Antiquité (à Sumer, en Mésopotamie, en Egypte…), des gouvernements ont effectivement utilisé des « séries statistiques » pour être mieux renseignés sur leurs Etats et les gérer en conséquence.

Peu après 1750, on commence à faire des représentations graphiques, la moyenne et la médiane sont de plus en plus utilisées pour **résumer et décrire une série statistique**.

Les physiciens, et depuis longtemps les astronomes, doivent tenir compte de séries de mesures pour un même phénomène, ces variations étant en partie aléatoires. A partir des observations statistiques, les économistes tentent de faire des prévisions **en essayant de maîtriser l'incertitude**.

Un chapitre des mathématiques va répondre à ces besoins car les mathématiciens ont commencé (1650) à créer des outils pour étudier les phénomènes aléatoires : les **probabilités**.

Dans notre environnement quotidien (météo, sondages…), professionnel (cabinets d'assurance, de gestion, laboratoires d'analyses médicales, contrôles qualité dans l'industrie), universitaire (physique, chimie, biologie, psychologie, économie, archéologie…), dans tous ces domaines, les statistiques et les probabilités interviennent.

Il est indispensable au citoyen d'aujourd'hui de comprendre ce que sont les statistiques pour savoir ce que veulent réellement dire les informations qu'il reçoit.

Et il est souhaitable qu'un élève de la série ES connaisse et sache utiliser les notions de base des statistiques et de calcul des probabilités.

Dans cette séquence, il s'agit de **statistiques descriptives**. On va s'attacher à résumer des séries statistiques par des nombres significatifs pour permettre l'utilisation et la comparaison de ces séries.

On précisera et on complètera les notions étudiées les années précédentes, en particulier ce qui concerne la dispersion d'une série statistique.

Pour les explications, les exemples qui ont été choisis comportent peu de données. Dans la réalité du travail des statisticiens, il s'agit d'étudier des séries statistiques pour lesquelles les données sont beaucoup plus nombreuses et les outils informatiques permettent de le faire.

1 Pré-requis

1 Vocabulaire

Une série statistique porte sur un caractère (taille, poids, sport pratiqué…)

Nous étudierons ici uniquement des séries statistiques à **caractère quantitatif**, par exemple la taille des élèves d'une classe (mais pas le sport pratiqué qui est un caractère qualitatif).

On dit qu'une série statistique est à **caractère quantitatif discret** quand les valeurs prises par le caractère sont des valeurs numériques précises (par exemple le nombre de frères et sœurs).

Et on dit qu'une série statistique est à **caractère quantitatif continu** quand on connaît seulement les effectifs des termes de la série appartenant à des intervalles (par exemple la taille des élèves d'une classe).

2 Effectifs, fréquences, fréquences cumulées croissantes

Deux exemples vont rappeler ces notions.

▶ **Exemple 1**

Pour une classe de 30 élèves, on connaît le nombre de frères et sœurs de chaque élève.

Il s'agit d'une série statistique à caractère discret.

On obtient le tableau suivant :

Nombre de frères et sœurs x_i	0	1	2	3	4	5
Effectif n_i	4	12	8	3	2	1
Effectif cumulé croissant	4	**16**	**24**	27	29	30
Fréquence f_i (valeur approchée)	0,13	0,40	0,27	0,10	0,07	0,03
Fréquence cumulée croissante (valeur approchée)	0,13	0,53	0,80	0,90	0,97	1

Par exemple, l'effectif cumulé 24 obtenu pour $x_i = 2$ signifie que 24 élèves ont 2 frères et sœurs au maximum. Ce nombre 24 est obtenu en ajoutant les deux nombres écrit en bleu dans le tableau : 16 l'effectif cumulé précédent et 8 l'effectif correspondant à $x_i = 2$.

Toutes les fréquences sont obtenues en divisant les effectifs par l'effectif total qui est égal à 30 ; on obtient toujours un nombre compris entre 0 et 1.

Les fréquences peuvent aussi être exprimées en pourcentage : par exemple 13% correspond à 0,13. Dans les activités et les exercices nous utiliserons les deux formes.

Ces fréquences sont souvent des valeurs approchées, sans que cela soit précisé.

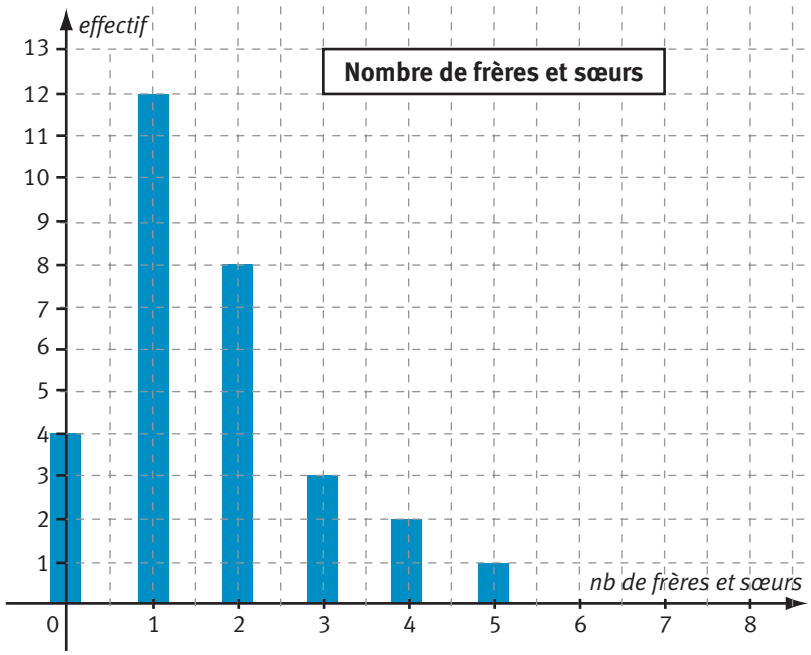
Le logiciel *sinequanon* (libre et gratuit) permet de travailler aisément sur les séries statistiques et notamment d'en faire des représentations graphiques.

Nous vous conseillons de réaliser les graphiques qui suivent avec ce logiciel.

Il suffit de cliquer sur « définir », « série statistique simple », valeurs isolées » et de rentrer les données dans le tableau proposé.

Cette série à caractère discret peut être représenté par un « *diagramme en bâtons* ».

Enfin, « définir » et « repère » permettent ensuite d'ajuster le graphique dans une fenêtre convenable.



▶ Exemple 2 On a relevé dans une entreprise de 125 employés le temps, en minutes, consacré à la pratique d'un sport par semaine.

Il s'agit d'une série statistique à caractère continu.

On obtient le tableau suivant :

Temps en minutes x_i	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 100[	[100 ; 140[	[140 ; 200]
Effectif n_i	35	41	**30**	12	5	2
Effectif cumulé croissant	35	**76**	**106**	118	123	125
Fréquence	0,28	0,32	0,24	0,10	0,04	0,02
Fréquence cumulée croissante	0,28	0,60	0,84	0,94	0,98	1

Le troisième effectif cumulé est 106 ; cela signifie que 106 employés de l'entreprise consacrent moins d'une heure par semaine à la pratique d'un sport.

Pour représenter cette série, utilisons encore le logiciel **sinequanon**.

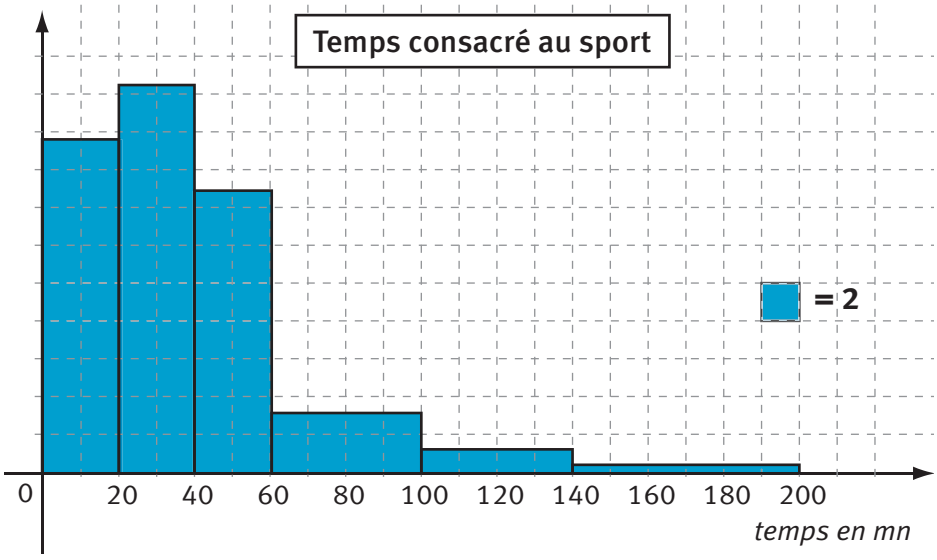
Il suffit de cliquer sur « définir », « série statistique simple », valeurs regroupées en classe » et de rentrer les données dans le tableau proposé.

Cette série à caractère continu peut être représenté par un « *histogramme* ».

Choisissons 1 petit carreau pour représenter un effectif de 2.

Définissons ensuite le repère en choisissant en abscisse 1 cm pour 20 minutes et en ordonnée 1 cm pour 1 par exemple.

Nous obtenons alors l'histogramme suivant :



Remarque

Lorsque les classes ont *même amplitude*, les rectangles de l'histogramme ont tous la même largeur. Leurs *aires* étant proportionnelles aux effectifs, leurs *hauteurs* le sont aussi. On peut alors « lire » les effectifs sur un « axe virtuel ». Mais lorsque les classes sont *d'amplitude différentes*, et c'est le cas pour notre exemple, les rectangles ont des largeurs différentes. Les aires des rectangles sont toujours proportionnelles aux effectifs, mais les hauteurs, elles, ne le sont plus.

Courbe des fréquences cumulées croissantes

Pour expliquer cette construction, utilisons l'exemple 2.

Dans ces graphiques, on indique en abscisse les valeurs du caractère : ici de 0 à 200. Et on indique les fréquences cumulées en ordonnée.

On place les points de coordonnées (20 ; 0,28), (40 ; 0,60), (60 ; 0,84)… (200 ; 1) qui correspondent aux informations suivantes : 28% des employés de l'entreprise consacre moins de 20 minutes par semaine au sport, 60% des employés moins de 40 minutes, etc.

On complète ces points par un premier point d'abscisse 0 (la plus petite valeur du caractère) et d'ordonnée 0 (0% des employés passent strictement moins de 0% de leur temps à la pratique d'un sport).

On joint alors les points par des segments de droite. La courbe obtenue est appelée **courbe des fréquences cumulées croissantes.**

On obtient toujours en utilisant le logiciel sinequanon le graphique ci-dessous.

Statistiques à une variable

Variable non numérique | Valeurs isolées | Valeurs regroupées en classes | Boîtes à moustaches multiples

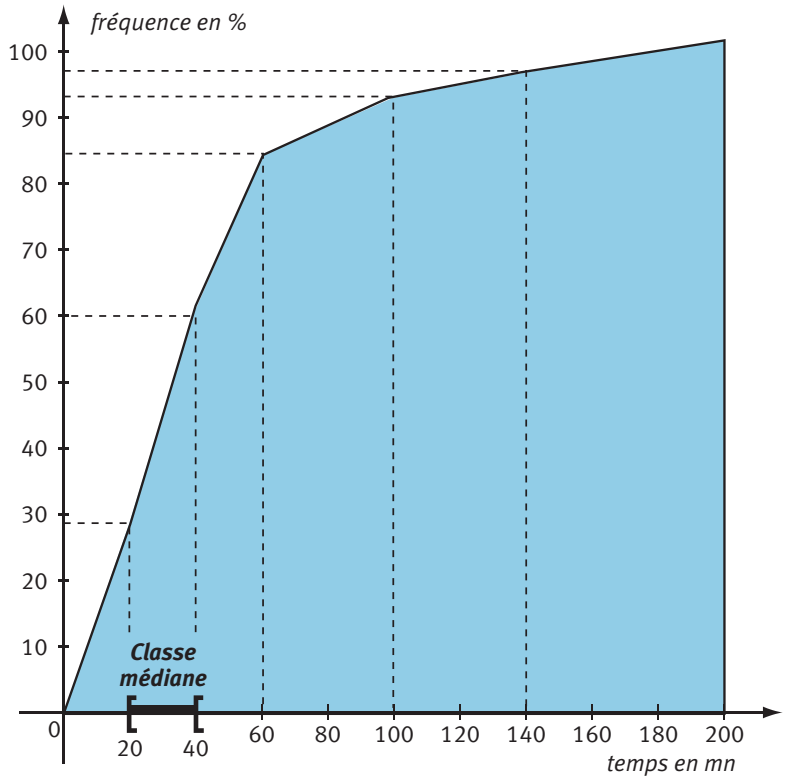
n°	Intervalles : borne inférieure	borne supérieure	Effectifs (entiers)
1	0	20	35
2	20	40	41
3	40	60	30
4	60	100	12
5	100	140	5
6	140	200	2

1 carreau(x) = Couleur

Dessin réalisé avec les pourcentages
Dessin réalisé avec les effectifs

Style du graphique
Histogramme — Afficher les effectifs
Diagramme en boîte (ou boîte à moustaches)
Diagramme en boîte avec déciles
Fréquences cumulées — Croissantes — Décroissantes
Droite de Henry

Pour définir le repère, on peut prendre par exemple, 1 cm pour 20 minutes en abscisses, et 1cm pour 10% en ordonnées.



Remarque

Ce choix (de relier les points par des segments de droite) revient à considérer que les valeurs du caractère sont régulièrement distribuées à l'intérieur de chaque classe, ce qui n'est pas forcément réel.
C'est pourquoi ces graphiques devront être utilisés avec précaution.

2 Paramètres numériques

Vous avez déjà utilisé quelques nombres qui permettent de résumer une série statistique.

a) Médiane d'une série statistique

Les valeurs du caractère d'une série statistique étant rangées par ordre croissant, on définit ***la médiane***. C'est un nombre tel qu'il y a autant de valeurs de la série qui lui sont inférieures que de valeurs qui lui sont supérieures. Plusieurs définitions plus précises sont possibles.
Celle qui sera utilisée dans ce cours, conformément au programme, est la suivante :

■ Définition

- si l'effectif N de la série est un nombre impair, $N = 2n+1$, la médiane de la série est la valeur centrale du caractère, celle qui est numérotée $n+1$.
- si l'effectif N de la série est un nombre pair, $N = 2n$, la médiane est le nombre égal à la demi somme des deux valeurs centrales, celles qui sont numérotées n et $n+1$.

Dans l'exemple des frères et sœurs des élèves, l'effectif total est égal à 30 ; la médiane est donc la demi somme des 15ème et 16ème valeurs, elle est donc égale à 1.

Remarque

Dans le cas où l'effectif de la série statistique est un nombre pair, la médiane n'est pas toujours une valeur de la série statistique.

Pour une série à caractère continu, on pourra seulement définir **la classe médiane**.

Dans l'exemple 2, l'effectif total est égal à 125 ; la médiane est donc la valeur du caractère du 63ème terme ; les effectifs cumulés croissants nous montrent que ce terme est dans la classe [20 ; 40[: c'est la classe médiane de la série statistique.

b) Moyenne d'une série statistique

Supposons donnée une série statistique à caractère quantitatif discret.

On note N l'effectif total, x_i les valeurs du caractère, n_i les effectifs et f_i les fréquences correspondantes.

■ Définition

La moyenne de la série, est le nombre $\bar{x}$ défini par :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N} = f_1x_1 + f_2x_2 + \ldots + f_px_p$$

Dans l'exemple des frères et sœurs des élèves, on a :

$$\bar{x} = \frac{1}{30}(4\times 0 + 12\times 1 + \ldots + 1\times 5) = \frac{50}{30} = \frac{5}{3} \approx 1{,}7$$

On trouve que la moyenne vaut $\frac{5}{3}$, donc environ 1,7. En moyenne, un élève de la classe a donc 1,7 frères et sœurs. Il ne faut pas s'étonner de ce résultat bizarre ; en effet, la moyenne n'est pas nécessairement une valeur du caractère de la série statistique (ici 0, 1, 2…).

Dans l'exemple 2, qui est celui d'une série continue, on fait des calculs analogues en utilisant les centres des classes et on trouve que la moyenne vaut 39,84 min.

c) Le symbole ∑

Les calculs effectués en statistiques nécessitent d'ajouter de nombreux termes. Le symbole $\sum$ permet d'éviter d'écrire la liste de ces termes.

Par exemple, si $x_1, x_2, x_3, ..., x_{12}$ désignent 12 nombres réels, leur somme $x_1 + x_2 + x_3 + ... + x_{12}$ sera notée $\sum_{i=1}^{i=12} x_i$.

▶ Exemple Si on considère les 2 listes de nombre :

$x_1 = 3$; $x_2 = 5$; $x_3 = 8$; $x_4 = 4$; $x_5 = 6$;

et $y_1 = 21$; $y_2 = 20$; $y_3 = 18$; $x_4 = 22$; $y_5 = 21$, on a alors :

$$\sum_{i=1}^{i=5} x_i = 3+5+8+4+6 = 26 \ ; \ \sum_{i=1}^{i=5} y_i = 102 \ ;$$

$$\sum_{i=1}^{i=5} x_i y_i = 3\times 21 + 5\times 20 + ... + 6\times 21 = 521.$$

Retour à la moyenne

La moyenne d'une série statistique peut être écrite à l'aide du symbole $\sum$.

On a : $\overline{x} = \frac{n_1 x_1 + n_2 x_2 + ... + n_p x_p}{N} = f_1 x_1 + f_2 x_2 + ... + f_p x_p = \frac{\sum_{i=1}^{i=p} n_i x_i}{N} = \sum_{i=1}^{i=p} f_i x_i.$

2 Médiane, quartiles, diagramme en boîte

A Activités

Activité 1

Médiane, quartiles, déciles d'une série à caractère discret.

On a demandé à 50 personnes prenant l'autobus, le nombre de fois où chacune de ces personnes a utilisé ce type de transport pendant la semaine écoulée.

Voici les résultats :

Nombre de voyages en autobus : x_i	1	2	3	4	5	6	7	8	9	10
Effectif	3	3	5	7	6	9	5	4	5	3
Effectif cumulé croissant : n_i										
Fréquence en %										
Fréquence cumulée croissante en %										

1. Compléter les lignes du tableau.
2. Déterminer la médiane.
3. Quelle est la plus petite valeur q du caractère pour laquelle au moins 25% ont une valeur inférieure à q ? Même question avec 75%.
4. Mêmes questions avec 10% et 90%.

Activité 2

Avec deux séries à caractère continu.

On reprend l'entreprise de l'exemple 2 du chapitre 1, on l'appelle l'entreprise A.

On rappelle les données :

Temps en minutes x_i	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 100[	[100 ; 140[	[140 ; 200]
Effectif n_i	35	41	30	12	5	2
Effectif cumulé croissant	35	76	106	118	123	125
Fréquence	0,28	0,32	0,24	0,10	0,04	0,02
Fréquence cumulée croissante	0,28	0,60	0,84	0,94	0,98	1

On a vu que la classe médiane est la classe [20 ; 40].

On considère une deuxième entreprise, l'entreprise B, où on a relevé aussi le temps consacré au sport par semaine par ses 160 employés.

1. Compléter le tableau suivant pour l'entreprise B.

Temps en minutes x_i	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 100[	[100 ; 140[	[140 ; 200]
Effectif n_i	29	43	47	12	5	2
Effectif cumulé croissant						
Fréquence						
Fréquence cumulée croissante						

2. Quelle est la classe médiane pour l'entreprise B ?

3. Construire, sur un même graphique, les deux courbes des fréquences cumulées croissantes.

4. En utilisant les points des deux courbes d'ordonnée 0,5, d'ordonnée 0,25, et d'ordonnée 0,75, comparer les deux séries statistiques.

B Cours

1 Quartiles, écart interquartile

On cherche ici à déterminer des nombres qui partagent la série statistique (dont les valeurs sont rangées par ordre croissant) en quatre groupes de même effectif environ.

On utilise la médiane et deux nombres appelés le premier et le troisième quartile.

Pour ne pas avoir à distinguer encore plus de cas que pour la médiane, on choisit les deux définitions suivantes.

Elles semblent d'abord un peu désagréables, mais la pratique permet de se familiariser avec leur utilisation. D'ailleurs l'essentiel est de retenir l'idée de base et de savoir déterminer ces quartiles avec une calculatrice ou un tableur.

Définitions

Premier quartile Q_1 : c'est la plus petite valeur de la série telle qu'au moins 25% des données soient inférieures à Q_1.

Troisième quartile Q_3 : c'est la plus petite des valeurs de la série telle qu'au moins 75% des données soient inférieures à Q_3.

(Rappel : « inférieur » correspond à $\leq$)
Dans certains cas, on peut trouver facilement ces deux valeurs.
Et un moyen toujours efficace de les trouver est d'utiliser les fréquences cumulées croissantes.

On verra plus loin comment utiliser une calculatrice ou un tableur.

▶ **Exemple** Dans l'activité 1 sur le nombre des trajets en autobus, on a obtenu :

Nombre de voyages en autobus	1	2	3	4	5	6	7	8	9	10
Effectif cumulé croissant	3	6	11	18	24	33	38	42	47	50
Fréquence cumulée croissante en %	6%	12%	22%	36%	48%	66%	76%	84%	94%	100%

La médiane est égale à la demi somme des vingt-cinquième et vingt-sixième terme, ces termes sont égaux à 6, la médiane est donc égale à 6.

La ligne des fréquences cumulées croissantes nous montre que le premier quartile est égal à 4 et le troisième quartile est égal à 7.

Remarque

Et le deuxième quartile ? Une définition analogue avec 50% donne le deuxième quartile.

On retrouve la médiane si l'effectif N de la série est impair.

Mais on ne retrouve nécessairement pas la médiane si l'effectif N de la série est pair. En effet, si $N = 2n$, d'après la définition qui est choisie ici pour la médiane, la médiane est la demi-somme des termes de la série de rang n et de rang $n+1$. Si ces termes ont des valeurs différentes, le résultat n'est pas une valeur de la série contrairement au deuxième quartile.

Au lycée le choix a été fait d'utiliser la médiane, définie comme cela a été rappelé dans les prérequis, et de ne pas utiliser le deuxième quartile.

Les premier et troisième quartiles permettent de mieux savoir comment est répartie la série statistique autour de la médiane.

On définit alors un nouveau nombre pour caractériser la série.

■ Définition

L'intervalle $\mathbf{[Q_1 ; Q_3]}$ est appelé **l'intervalle interquartile** de la série statistique.

Le nombre $\mathbf{Q_3 - Q_1}$ est appelé **l'écart interquartile** de la série statistique.

▶ **Exemple** Dans l'activité 1, l'intervalle interquartile est l'intervalle [4 ; 7], l'écart interquartile est égale à 3.

La moitié au moins des personnes interrogées ont donc fait un nombre de voyages compris entre 4 et 7.

La médiane est au « centre » de la série, les valeurs sont réparties de part et d'autre de la médiane.

La moitié de ces valeurs se trouve dans l'intervalle interquartile : l'amplitude de cet intervalle (c'est-à-dire l'écart interquartile) indique la dispersion plus ou moins grande des valeurs autour de la médiane.

La **médiane** est un **indicateur de position**, **l'écart interquartile** est un **indicateur de dispersion**.

Résumé d'une série statistique

On peut alors ainsi **résume**r une série statistique par le couple **(médiane ; écart interquartile).**

▶ **Exemple** Dans l'activité 1, on résume la série statistique en donnant sa médiane qui vaut 6 et l'écart interquartile qui vaut 3.

▶ **Commentaire** Quand on résume une série statistique par le couple (médiane ; écart interquartile), la médiane et les quartiles ne dépendent pas des valeurs des termes extrêmes. En effet, les valeurs des termes extrêmes peuvent changer un peu sans modifier la médiane et les quartiles.

Pour exprimer cela on dit que la **médiane** est un indicateur **« robuste ».**

Pour étudier l'évolution des salaires, on peut choisir de regarder comment progresse le salaire médian et le salaire correspondant au premier quartile, car ces renseignements ne sont pas dépendants des cas particuliers extrêmes.

De même, dans une classe, on peut observer l'évolution des résultats des élèves en regardant la progression de la médiane et du premier quartile des séries statistiques formées par les notes. On utilise ainsi des indicateurs qui ne sont pas influencés par les valeurs des notes les meilleures et les plus basses.

2 Déciles, écart interdécile d'une série statistique

De façon analogue à ce qui précède, on peut chercher à déterminer des nombres qui partagent la série statistique (dont les valeurs sont rangées par ordre croissant) en dix groupes de même effectif environ.

Ces nombres sont appelés les **déciles** de la série statistique.

Nous utiliserons seulement le premier et le dernier.

■ Définition

Premier décile D_1 : c'est le plus petit élément des valeurs de la série tel qu 'au moins 10% des données soient inférieures à D_1.

Neuvième décile D_9: c'est le plus petit élément des valeurs de la série tel qu'au moins 90% des données soient inférieures à D_9.

L'intervalle **$[D_1 ; D_9]$** est appelé **l'intervalle inter-décile** de la série statistique.

Le nombre **$D_9 - D_1$** est appelé **l'écart inter décile** de la série statistique.

▶ **Exemple** Dans l'activité 1, la ligne des fréquences cumulées croissantes nous permet de lire les déciles.

Nombre de voyages en autobus	1	2	3	4	5	6	7	8	9	10
Fréquence cumulée croissante en %	6%	12%	22%	36%	48%	66%	76%	84%	94%	100%

Le premier décile est égal à 2, le neuvième décile est égal à 9, l'intervalle interdécile est l'intervalle [2 ; 9] et l'écart interdécile est égal à 9 – 2, c'est-à-dire 7.

3 Diagrammes en boîte

Il est très utile de représenter graphiquement une série statistique.

Un seul coup d'œil permet de recueillir beaucoup d'informations, ce qui est en particulier très commode quand on compare des séries statistiques.

On a dit plus haut que l'on peut résumer une série statistique par le couple **(médiane ; écart interquartile).**

On visualise cela par un **diagramme en boîte**, appelé parfois « **boîte à moustaches** » ou « **boîte à pattes** ».

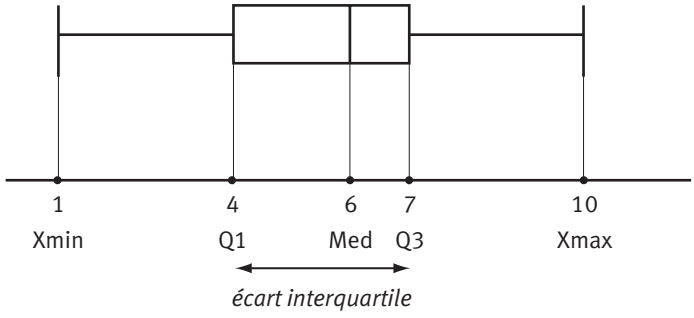
Les diagrammes suivants illustrent les constructions les plus fréquentes pour ce type de graphique.

Ils correspondent à l'exemple des trajets d'autobus de l'activité 1.

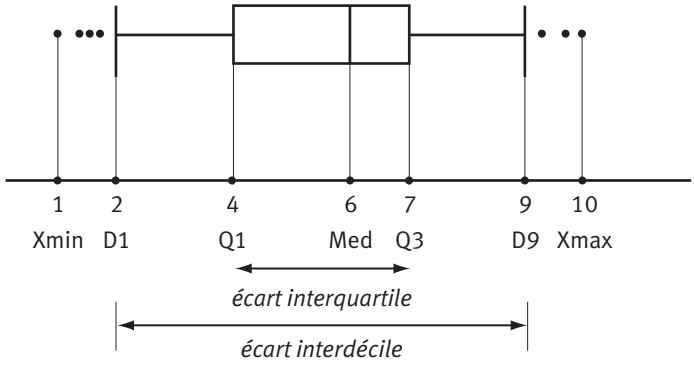
On utilise un axe gradué (ici, il est horizontal, il peut être vertical).

On dessine un rectangle (la boîte) limité par les quartiles, on indique la médiane.

A partir du rectangle, vers l'extérieur, on construit deux segments (les moustaches, les pattes) dont les autres extrémités correspondent aux valeurs extrêmes de la série.

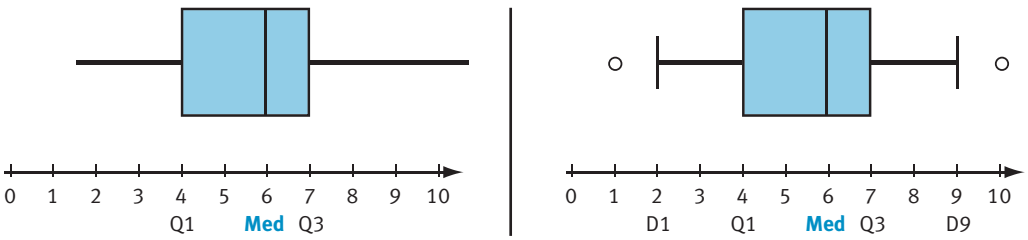


On peut aussi indiquer le premier et le neuvième décile :



Remarque

Sur ce deuxième graphique, on peut lire beaucoup d'informations : 7 paramètres de la série statistique sont lisibles ainsi que l'écart interquartile et l'écart interdécile.
Le logiciel sinequanon construit directement le diagramme en boîtes ou le diagramme en boîtes avec déciles.



Remarque

La hauteur de la boîte n'a pas de signification et peut être choisie selon son bon vouloir.

4 Cas des séries à caractère continu

Pour ce type de série statistique il est délicat d'utiliser les notions de médiane et de quartiles car on n'a pas d'information sur la répartition des valeurs à l'intérieur de chaque classe.

a) En utilisant les fréquences cumulées croissantes

Les fréquences cumulées croissantes permettent de repérer dans quelle classe se situe la médiane, c'est-à-dire dans quelle classe on franchit la fréquence cumulée égale à 50%.

■ Définition

La première classe pour laquelle la fréquence cumulée croissante dépasse 50% s'appelle la **classe médiane.**

▶ Exemple Dans l'entreprise A, on a vu que la médiane appartient à l'intervalle [20 ; 40], cet intervalle forme donc la classe médiane.

b) En utilisant la courbe des fréquences cumulées

Dans les cas où peut supposer que la répartition dans la classe médiane est régulière, homogène, on peut trouver graphiquement un nombre qui pourra être considéré comme une valeur approchée de la médiane.

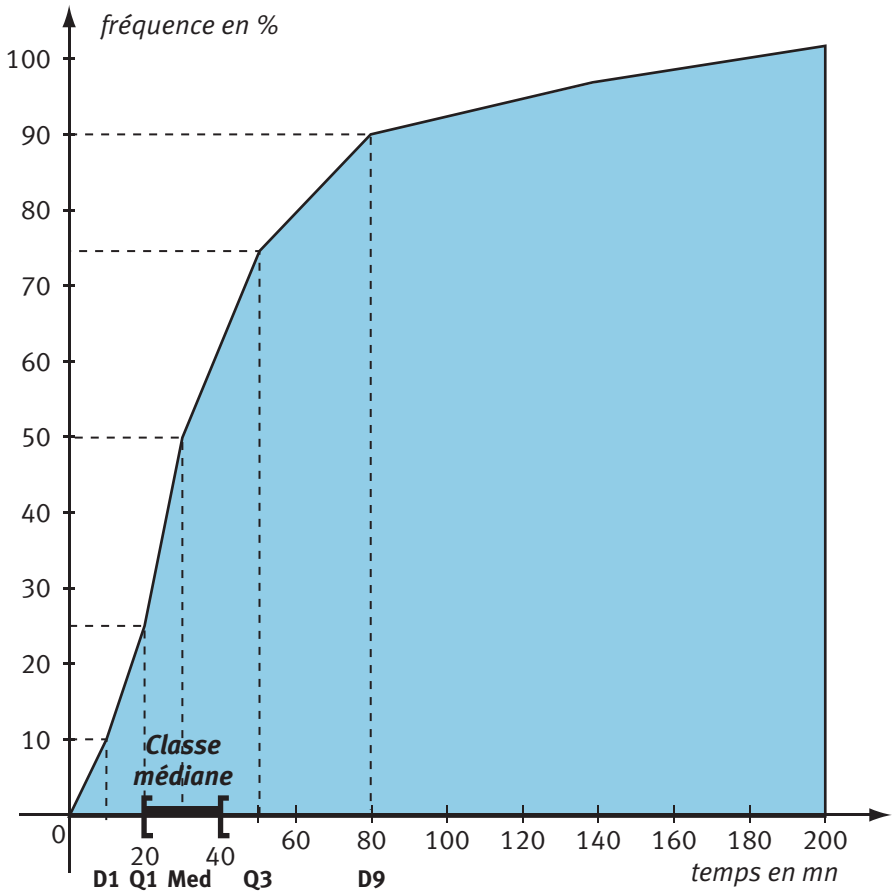
Dans la courbe des fréquences cumulées, les fréquences cumulées sont lues sur l'axe des ordonnées.

On considère donc l'ordonnée 50%.

Puis on lit l'abscisse du point correspondant de la courbe, c'est cette abscisse qui fournit une valeur approchée de la médiane.

On peut procéder de manière analogue pour les quartiles Q_1 et Q_3 en considérant les abscisses des points de la courbe d'ordonnée 25% et 75%, et pour les déciles D_1 et D_9 en considérant les abscisses des points de la courbe d'ordonnée 10% et 90%.

▶ **Exemple** Dans le cas de l'entreprise A, on obtient ainsi



On lit donc que la médiane vaut à peu près 33 min, le premier quartile 18 min, le troisième 52 min, le premier décile 7min et le neuvième 82 min.

5 Avec une calculatrice ou un tableur

Les calculs faits dans le cours sont développés pour vous permettre de **comprendre** les notions.

Mais dans la pratique, y compris dans les exercices et les devoirs (sauf avis contraire), vous effectuerez ces calculs à l'aide de votre calculatrice ou d'un ordinateur.

On s'intéresse ici à la détermination de la médiane et des quartiles d'une série statistique.

Les écrans suivants correspondent à la série statistique de l'activité 1 :

Nombre de voyages en autobus	1	2	3	4	5	6	7	8	9	10
Fréquence cumulée croissante en %	6%	12%	22%	36%	48%	66%	76%	84%	94%	100%

a) Avec une calculatrice Casio 25+

Les procédures sont identiques ou très voisines pour les autres modèles de Casio

▶ Saisie On saisit les données.

Dans le menu général, on sélectionne l'icône STAT (ou LIST). Sur l'écran apparaît alors l'éditeur de listes.

On saisit les valeurs x_i du caractère dans une liste, List 1 par exemple, et les effectifs correspondants dans une autre liste, List 2 par exemple.

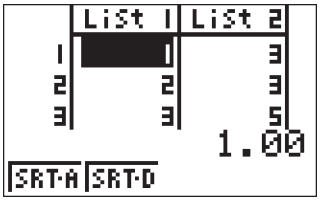
▶ Calcul En bas de l'éditeur de listes se trouve un menu déroulant horizontal.

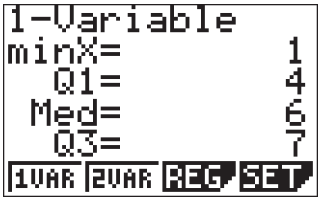
On active le sous-menu CALC puis SET

Sur la ligne **1Var Xlist** on indique List 1, et sur la ligne **1Var Freq** on indique List 2, pour indiquer les valeurs puis les effectifs.

On tape alors EXIT. Sélectionner enfin le menu 1 VAR.

Des paramètres de la série statistique apparaissent à l'écran ; parmi eux, en utilisant la touche ↓, on trouve la médiane **Med**, et les quartiles $\mathbf{Q_1}$ et $\mathbf{Q_3}$.





▶ Graphique On peut aussi faire apparaître un diagramme en boîte.

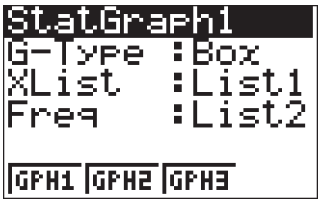
Dans l'éditeur de listes on active le sous-menu GRPH, puis le menu SET et PH1.

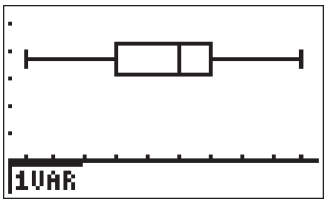
On indique alors sur la ligne **G-Type** le type de graphique qui est souhaité, en validant l'option MedBox du menu horizontal du bas de l'écran, puis on complète la ligne **XList** avec List 1, pour indiquer la liste des valeurs, et la ligne **Frequency** avec List 2, pour indiquer la liste des effectifs.

On valide l'écran.

On affiche alors le graphique en validant GRPH, puis GPH1.

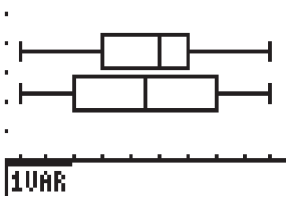
Pour visualiser l'axe horizontal et ses graduations il faut éventuellement adapter la fenêtre.





Remarque

On peut afficher deux diagrammes en boîte simultanément.
Par exemple ici, on a rentré en List 3 les mêmes valeurs x_i qu'en List 1, puis on a mis partout l'effectif $n_i = 1$ en List 4.
Sur l'écran dont l'image est donnée ci-dessus, on active GPH2, on choisit successivement MedBox List 3, et List 4.
Après EXIT on choisit SEL qui permet de choisir les deux graphiques en sélectionnant ON pour GPH1 et pour GPH2.
Et enfin DRAW permet d'obtenir l'écran ci-dessus.



b) Avec une TI 82Stats.fr

Les procédures sont identiques ou très voisines pour les autres modèles TI.

▶ Saisie

Il faut d'abord saisir les données

Appuyer sur la touche STATS, puis choisir le menu EDIT, suivi de entrer.

On tape chaque valeur du caractère x_i dans une liste, par exemple L1, et chaque effectif ou fréquence n_i dans une autre liste, par exemple L2, et on termine par entrer.

▶ Calculs

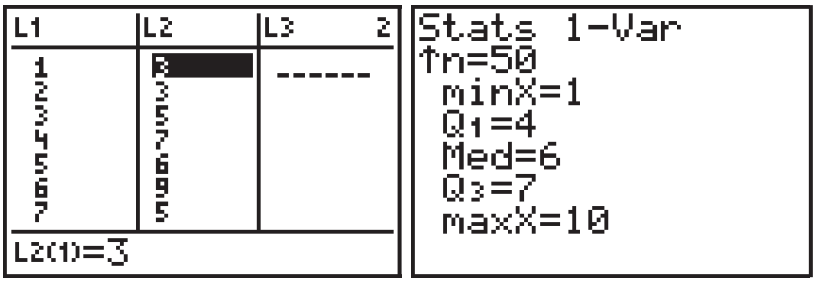
Appuyer de nouveau sur la touche STATS, puis choisir le menu CALC, suivi de entrer.

Sur l'écran apparaît alors l'indication **Stats 1-Var.**

Taper alors L1 , L2 pour indiquer, dans l'ordre, la liste des valeurs et celle des effectifs (attention : pour obtenir L1, il faut taper sur les touches 2nde puis 1, et après la virgule on fait de même pour L2).

Appuyer sur entrer.

Des paramètres de la série statistique apparaissent à l'écran, parmi eux, en utilisant la touche ↓, on trouve la médiane **Med** et les quartiles $\mathbf{Q_1}$ et $\mathbf{Q_3}$.



▶ Graphiques

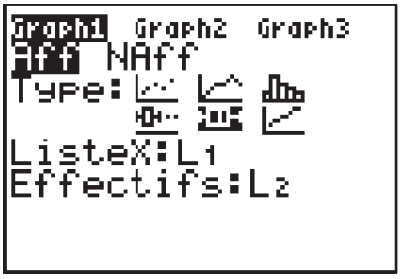
On peut représenter une série statistique par un diagramme en boîte après avoir saisi les données.

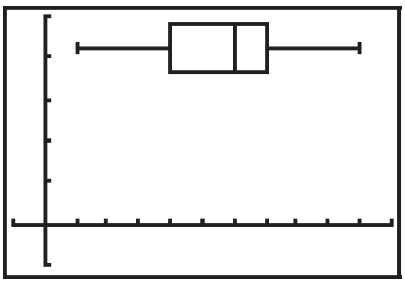
Appuyer sur la touche graph stats (touche 2nde de la touche $f(x)$), puis sur entrer (ce qui sélectionne le dessin n°1 : **Graph1**).

On place le curseur sur **ON** ou (**Aff**) que l'on valide par entrer, puis sur le type de graphique (ou) que l'on valide par entrer (remarque il y a ici deux types de diagramme en boîte, on choisira plutôt le même que sur l'écran ci-dessus, au milieu de la deuxième ligne).

On renseigne alors la ligne **ListeX** avec L1 (touche 2nde puis 1), pour indiquer la liste des valeurs, et la ligne **Effectifs** avec L2, pour indiquer la liste des effectifs.

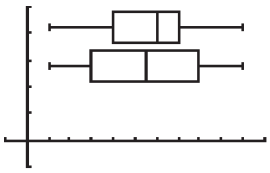
On affiche alors le graphique en appuyant sur la touche graphe. Pour visualiser l'axe horizontal et ses graduations il faut éventuellement adapter la fenêtre.





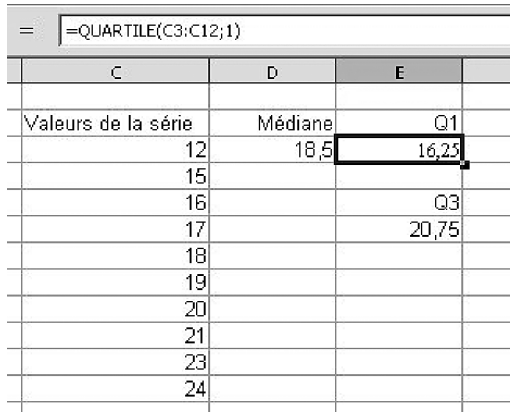
Remarque

Il est possible d'afficher simultanément deux diagrammes en boîte en utilisant aussi Graph2 : on procède de la même manière que pour Graph1 en choisissant On (ou Aff) et en précisant les listes concernées.
Par exemple ici, on a rentré List 3 les mêmes valeurs x_i qu'en List 1 , puis on a mis partout l'effectif $n_i = 1$ en List 4 .



c) Avec un tableur

Pour déterminer la médiane et les quartiles, on utilise les fonctions statistiques présentes dans la plupart des tableurs lorsque **la série est donnée par une seule colonne**, c'est-à-dire **que tous les effectifs sont égaux à 1.**



Si tous les effectifs ne sont pas égaux à 1, il n'est pas possible d'utiliser les fonctionnalités d'un tableur pour déterminer la médiane et les quartiles.
Voici l'exemple d'une série statistique où tous les effectifs sont égaux à 1.
On sélectionne la plage de cellule concernée.
Pour les quartiles, on doit préciser 1 ou 3 en respectant la syntaxe du logiciel.
Pour le premier quartile de cette série statistique de 10 termes, on devrait trouver le troisième terme, c'est-à-dire 16.
Ici $Q_1 = 16{,}25$. Il s'agit d'OpenOffice et on observe que ce quartile n'est pas une valeur de la série statistique, ce logiciel n'utilise pas la même définition que le cours. On rappelle que c'est peu gênant dans la pratique réelle des statistiques où les effectifs sont importants.

C Exercices d'apprentissage

Pour ces exercices, il est vivement conseillé d'utiliser une calculatrice ou un tableur ou le logiciel sinequanon.

Exercice 1

Une pharmacie de garde a enregistré le nombre d'appels reçus pendant 1000 nuits entre 20h et 6h du matin. Les résultats sont les suivants :

Nombre d'appels x_i	0	1	2	3	4	5	6	7	8	9	10	11
Nombre de nuits n_i	14	70	155	185	205	150	115	65	30	5	1	5

Déterminer la médiane et les quartiles de cette série, puis faire un diagramme en boîte.

Exercice 2

Deux sauteurs à la perche ont relevé leurs performances lors de leurs 25 derniers sauts.

1er sauteur

Hauteur	4,70	4,80	4,85	4,90	4,95	5,00	5,05	5,10	5,20
Nombre de sauts	1	1	1	3	12	4	1	1	1

2e sauteur

Hauteur	4,60	4,70	4,75	4,80	4,85	4,90	4,95	5,00	5,05	5,10	5,15	5,20
Nombre de sauts	3	2	2	3	2	2	1	3	2	1	1	3

Déterminer la médiane et les quartiles de chacune de ces deux séries.

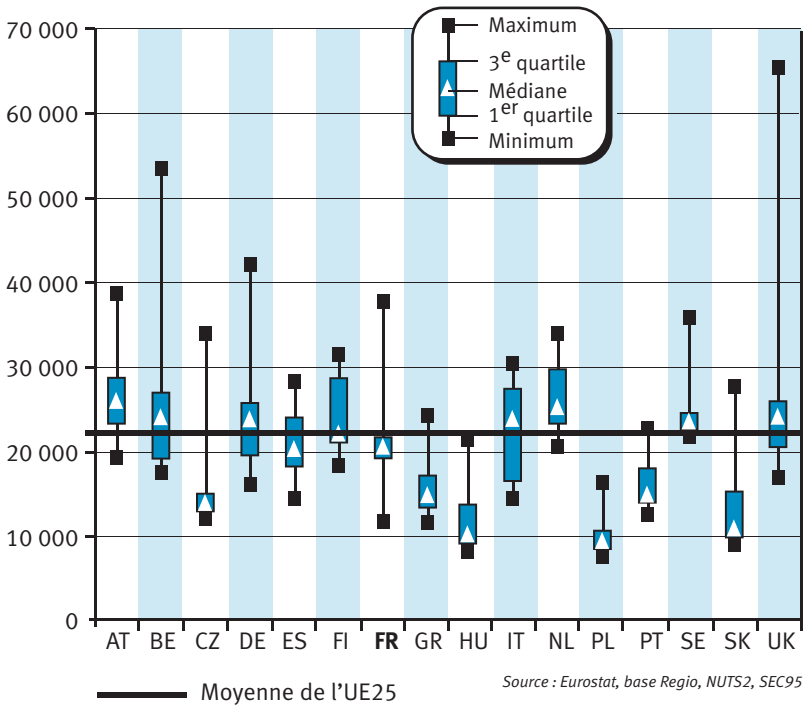
Construire les deux diagrammes en boîte et comparer l'ensemble des performances des deux sportifs.

Exercice 3

Dans le numéro 97-98 de la revue Economie Lorraine on trouve le graphique ci-dessous, construit à partir de données Eurostat de le Communauté européenne pour l'année 2004.

Ce graphique concerne le PIB (Produit Intérieur Brut) par habitant en SPA (standards de pouvoir d'achat, c'est-à-dire une monnaie commune qui élimine les différences de prix entre les pays, permettant des comparaisons significatives).

Pour chaque pays on a représenté un diagramme en boîte construit à partir des régions (par exemple le diagramme de la France est construit à partir des PIB moyens des 26 régions).



Source : Eurostat, base Regio, NUTS2, SEC95

AT : Autriche	ES : Espagne	HU : Hongrie	PT : Portugal
BE : Belgique	FL : Finlande	IT : Italie	SE : Suède
CZ : Tchéquie	**FR : France**	NL : Pays-bas	SK : Slovaquie
De : Allemagne	GR : Grèce	PL : Pologne	UK : Royaume-Uni

1. Dans quel pays se trouve la région ayant le PIB par habitant le plus élevé ? le moins élevé ?

2. Dans quel pays l'écart interquartile est-il le plus grand ? le plus petit ?

3. Donner deux propriétés particulières au diagramme de la France.

4. Quelles est la propriété commune des diagrammes de la Belgique, de l'Allemagne, de l'Italie et de la Suède ?

Exercice 4

D'après l'INSEE, les revenus annuels (en milliers d'euros) des salariés en 2007 se répartissent suivant le tableau ci-dessous qui donne les valeurs des déciles des deux séries :

Déciles	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9
Femmes	1,8	5	8,7	12	14,5	16,6	19,1	22,6	28,9
Hommes	2,8	8,2	13,2	15,6	17,7	20	23,1	27,8	37,2

1. Les déciles permettent de déterminer des classes. Pour les femmes, donner le tableau indiquant ces classes et les fréquences correspondantes.
2. Représenter dans un même repère les courbes des fréquences cumulées croissantes (on prendra, pour les deux séries, 1 pour valeur minimale et 45 pour valeur maximale).
   Quelle courbe est « à gauche de l'autre », « au dessus de l'autre » ? Quelle signification cela a-t-il ?
3. Déterminer graphiquement des valeurs approchées des quartiles des deux séries, et construire les diagramme en boîte des deux séries.

Exercice 5

Les données extrêmes d'une série qui se différencient trop des autres (beaucoup trop grandes ou beaucoup trop petites) sont appelées « valeurs aberrantes ».
Le statisticien américain John W.Tukey (1915-2000) a proposé un critère pour isoler les valeurs aberrantes : on appellera valeur aberrante toute valeur qui se situera à plus de 1,5 fois l'écart interquartile $Q_3 - Q_1$ avant Q_1 ou après Q_3.
Le taux de chômage pour le deuxième trimestre 2009 pour les 22 régions françaises en % est fournie par l'INSEE par le tableau suivant :

Alsace	Aquitaine	Auvergne	Bourgogne	Bretagne	Centre	Champagne Ardennes	Corse	Franche Comté	Ile de France	Languedoc roussillon
8,3	8,9	8,4	8,4	7,8	8,3	10,0	8,4	9,6	7,8	12,5

Limousin	Lorraine	Midi Pyrénées	Nord Pas de calais	Basse Normandie	Haute Normandie	Pays de Loire	Picardie	Poitou Charentes	Provence Alpes Côte d'Azur	Rhône Alpes
7,8	10,0	9,0	12,7	9,0	10,2	8,2	10,9	9,0	10,5	8,6

Pour la France métropolitaine, ce taux est de 9,1%.

1. Montrer que la valeur 12,7 (qui correspond à la région Nord-Pas-de-Calais) peut-être qualifiée d'aberrante avec la définition donnée dans l'information ci-dessus.
2. Pouvez vous expliquer économiquement le résultat de la région Nord-Pas-de-Calais ?
3. Construire sans tenir compte de cette valeur le diagramme en boîte de la série ci-dessus.

3 Moyenne, écart-type

A Activités

▶ **Activité 3**

Pendant la semaine du 13 au 17 septembre 2010, on a relevé les températures minimales et les températures maximales à Brest (d'après les données de Météo-France).

Date	lundi	mardi	mercredi	jeudi	vendredi	samedi	dimanche
Température minimale en °C	8,8	12,2	13,5	12,7	8,5	7,7	5,2
Température maximale en °C	19,5	19,9	18,6	17,8	18	17,3	18,1

Les températures maximales semblent plus « régulières » que les températures minimales.

Le but de cette activité est d'introduire une nouvelle caractéristique d'une série statistique pour mesurer sa dispersion autour de la moyenne. On pourra alors comparer la « régularité » de deux séries.

1. Dans les quatre premières questions, on considère seulement les températures minimales.

   Calculer la température minimale moyenne $\overline{x}$.

2. Dans le tableau suivant on indique les différences avec la moyenne (on dit aussi « l'écart à la moyenne »).

Température minimale en °C : x_i	8,8	12,2	13,5	12,7	8,5	7,7	5,2
Écart : $x_i - \overline{x}$							

   Qu'observe-t-on quand on calcule la moyenne de ces différences ?

3. Ce qui précède amène à ne considérer que des quantités positives.

   Pour cela, on peut utiliser les valeurs absolues ou les carrés. Les carrés, moins naturels, ont cependant étaient choisis car les propriétés mathématiques sont ensuite beaucoup plus intéressantes.

Température minimale en °C : x_i	8,8	12,2	13,5	12,7	8,5	7,7	5,2
Écart : $x_i - \overline{x}$							
Carré de l'écart à la moyenne : $\left(x_i - \overline{x}\right)^2$							

Compléter ce tableau, puis calculer la moyenne des carrés des écarts à la moyenne $\overline{x}$.

Le nombre obtenu s'appelle **la variance** de la série statistique, on le note *V*.

4 Pour compenser l'utilisation des carrés et se ramener à une quantité représentant une grandeur de même nature que les termes de la série statistique, on calcule maintenant la racine carrée de la variance V.

Ce nouveau nombre s'appelle **l'écart-type** de la série statistique, on le note *s*.

Calculer l'écart-type *s* de la série statistique des températures minimales.

5 Calculer la variance *V'* et l'écart-type *s'* de la série statistique des températures maximales.

Comparer les deux écarts-types *s* et *s'*.

▶ Activité 4 On reprend l'exemple du nombre des voyages en autobus.

Nombre de voyages en autobus : x_i	1	2	3	4	5	6	7	8	9	10
Effectif : n_i	3	3	5	7	6	9	5	4	5	3
Carré de l'écart à la moyenne $\left(x_i - \overline{x}\right)^2$										

Déterminer la moyenne $\overline{x}$, puis compléter la dernière ligne du tableau.

Calculer ensuite l'écart-type, attention : ici, les effectifs ne sont pas tous égaux à 1 comme dans l'activité précédente.

▶ Activité 5 Avec une série à caractère continu : on reprend l'exemple du temps consacré au sport dans l'entreprise A.

Montant des achats (en €)	[0 ; 20[	[20 ; 40[	[40 ; 60[	[60 ; 100[	[100 ; 140[	[140 ; 200]
Effectif n_i	35	41	30	12	5	2
Carré de l'écart à la moyenne						

En utilisant les centres des classes, déterminer la moyenne puis compléter le tableau.

Déterminer ensuite l'écart-type de cette série statistique.

B Cours

1 La moyenne et ses propriétés

a) Rappel de la définition

Supposons donnée une série statistique à caractère quantitatif discret.

On note N l'effectif total, x_i les valeurs du caractère et n_i les effectifs correspondants.

Si on considère une série statistique à caractère quantitatif continu, on appliquera alors tout ce qui est défini pour une série discrète en utilisant le centre de chaque classe et l'effectif correspondant.

■ Définition

La moyenne $\bar{x}$ de la série est le nombre défini par :

$$\bar{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N}.$$

On peut aussi écrire $\bar{x} = \sum_{i=1}^{i=p} \frac{n_ix_i}{N} = \sum_{i=1}^{i=p} f_ix_i$.

Remarque

La somme $n_1x_1 + n_2x_2 + \ldots + n_px_p$ est égale à la somme de toutes les valeurs de la série (puisque x_1 est compté n_1 fois, etc.).
Et, en multipliant par N, on obtient une égalité qui est très importante dans le paragraphe suivant.

A savoir

$N\bar{x} = n_1x_1 + n_2x_2 + \ldots + n_px_p$
Cette égalité signifie que la **moyenne multipliée par l'effectif** est égale à la **somme des valeurs de la série**.

b) Calcul de la moyenne d'une série à partir des moyennes de deux sous-groupes

La remarque précédente permet de démontrer le théorème suivant :

■ Théorème

Si une population d'effectif total N est partagée en deux sous-groupes, l'un d'effectif P pour lequel la moyenne est $\overline{x'}$, et l'autre d'effectif Q pour lequel la moyenne est $\overline{x''}$

la moyenne $\overline{x}$ de la population entière est donnée par l'égalité:

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{P}\overline{\boldsymbol{x}'} + \boldsymbol{Q}\overline{\boldsymbol{x}''}}{\boldsymbol{P}+\boldsymbol{Q}}.$$

Démonstration

La moyenne $\overline{x}$ de la série est égale au quotient

$$\frac{\textit{somme de toutes les valeurs de la série}}{\textit{effectif total}}.$$

Pour le premier sous-groupe la somme des valeurs vaut $P\overline{x'}$, pour le second elle vaut $Q\overline{x''}$, donc pour la série entière la somme de toutes les valeurs est égale à $P\overline{x'} + Q\overline{x''}$.

Et l'effectif total est égal bien sûr à $P+Q$, on obtient ainsi le résultat annoncé.

Remarque

On peut exprimer cette égalité en utilisant les fréquences : $\overline{x} = f'\overline{x'} + f''\overline{x''}$.

En effet, N étant l'effectif total, on a $P+Q=N$, la fréquence du premier groupe est $f' = \frac{P}{P+Q}$ et la fréquence du second groupe est $f'' = \frac{Q}{P+Q}$.

On a donc : $\overline{x} = \frac{P\overline{x'} + Q\overline{x''}}{P+Q} = \frac{P}{P+Q}\overline{x'} + \frac{Q}{P+Q}\overline{x''} = f'\overline{x'} + f''\overline{x''}$.

▶ Exemple Une entreprise est installée sur deux sites.

Sur le premier site, la moyenne des salaires est égale à 1600 € et 35 personnes y travaillent.

Sur le second site, la moyenne des salaires est égale à 1900 € et 21 personnes y travaillent.

Le théorème précédent permet de calculer la moyenne des salaires sur l'ensemble des deux sites.

Les données sont donc : $P=35$, $\overline{x'}=1600$, $Q=21$, $\overline{x''}=1900$.

La moyenne $\overline{x}$ de la série est donnée par :

$$\overline{x} = \frac{35\times 1600 + 21\times 1900}{35+21} = 1712{,}5.$$

La moyenne des salaires dans cette entreprise est donc égale à 1712,5 €.

c) Effet de structure

▶ Exemple

On appelle A′ l'entreprise de l'exemple précédent.

Supposons qu'une seconde entreprise B′ soit aussi sur deux sites.

Dans le premier, la moyenne des salaires est 1650 € et, dans le deuxième, la moyenne est 1950 €.

On est tenté de penser que la moyenne $\overline{y}$ des salaires dans l'entreprise B′ est supérieure à la moyenne des salaires dans l'entreprise A′.

Pour le vérifier, il est nécessaire de compléter les données concernant l'entreprise B′ : le salaire moyen est 1650 € pour un effectif de 50 personnes, et le salaire moyen est 1950 € pour un effectif de 10 personnes.

On a alors :

$$\overline{y} = \frac{50 \times 1650 + 10 \times 1950}{50 + 10} = 1700.$$

Le salaire moyen est donc 1700 € dans l'entreprise B′, il est inférieur à celui de l'entreprise A′ !

Ce paradoxe s'explique par la comparaison des effectifs : dans l'entreprise B′, les effectifs des groupes sont 50 et 10 (le premier groupe est donc cinq fois plus nombreux que le second), alors que dans l'entreprise A′ les effectifs des groupes sont 35 et 21 (l'effectif du premier groupe est inférieur au double du second).

Les effectifs ne sont pas répartis de la même façon dans les deux entreprises.

On l'observe encore mieux avec les fréquences.

Dans l'entreprise A′, le premier groupe correspond à 62,5% de l'effectif total, le second groupe à 37,5%.

Dans l'entreprise B′, le premier groupe correspond à environ 83,3% de l'effectif total, le second groupe à environ 16,7%. Dans cette entreprise B′, le salaire moyen est tellement « tiré » vers 1650 €, le salaire du premier groupe, que le salaire moyen dans l'entreprise B est inférieur à celui de l'entreprise A.

■ Définition

Dans l'expression $\overline{x} = \frac{P\overline{x'} + Q\overline{x''}}{P+Q} = f'\overline{x'} + f''\overline{x''}$, il est possible que $\overline{x}$ diminue alors que $\overline{x'}$ et $\overline{x''}$ augmentent car la valeur du quotient dépend aussi des changements des valeurs des effectifs P et Q (et donc des fréquences) : ce résultat paradoxal s'appelle un **effet de structure**.

❷ Ecart-type

On donne ici un indicateur numérique mesurant la dispersion d'une série statistique autour de sa moyenne. On généralise ce qui a été fait dans les activités.

■ Définition

La variance de la série statistique est définie par :

$$V = \frac{n_1(x_1 - \bar{x})^2 + n_2(x_2 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{n_1 + n_2 + \ldots + n_p} = \sum_{i=1}^{i=p} \frac{(n_i x_i - \bar{x})^2}{N} =$$

$$= \sum_{i=1}^{i=p} f_i(x_i - \bar{x})^2.$$

L'écart type s de la série est défini par :

$$s = \sqrt{V}.$$

▶ **Commentaire** La variance est égale à : la moyenne des carrés des écarts à la moyenne de la série.

L'écart-type est donc égal à :

la racine carrée…de la moyenne…des carrés…des écarts à la moyenne de la série.

Propriétés

La **variance** et **l'écart type** sont nécessairement des **nombres positifs.**

Remarque

On a utilisé des carrés, puis pour « compenser » on a pris la racine carrée du résultat. On obtient **l'écart-type** qui est un donc un paramètre représentant bien une même grandeur (euros, centimètres…) que les valeurs du caractère.
S'il donne une bonne indication sur la dispersion de la série, il n'est malheureusement pas interprétable ou représentable aussi facilement que les quartiles et l'écart interquartile.
Dans la suite du cours de statistiques-probabilité vous constaterez que l'écart-type est un indicateur très utilisé car il possède de très nombreuses propriétés mathématiques au delà des statistiques descriptives.
(Les quartiles et l'écart interquartile sont eux plus faciles à comprendre mais on ne les utilisera qu'en statistique descriptive.)

Résumé d'une série statistique

On peut alors ainsi **résume**r une série statistique par le couple **(moyenne ; écart-type).**

▶ **Exemple** Dans l'exemple de l'activité 1, la série des températures minimales à Brest a pour moyenne 9,8°C et pour écart-type environ 2,8°C.

Remarque

Par sa définition, l'écart-type n'est pas simple à calculer.
Dans la pratique, vous utiliserez une calculatrice ou un tableur ou le logiciel sinequanon (des explications sont données plus loin).
On dispose, d'une formule plus simple que celle de la définition, mais dans laquelle on ne voit plus la signification de la variance.
Elle est donnée ci-dessous : on remarque que la moyenne $\bar{x}$ n'apparaît plus qu'une seule fois ce qui diminue les approximations.

■ Théorème

$$V = \frac{1}{N}\left(\sum_{i=1}^{i=p} n_i x_i^2\right) - \bar{x}^2 \text{ et } s = \sqrt{V}.$$

Nous admettrons cette propriété.

Remarque

Cette égalité permet de dire que :
la variance est égale à … la moyenne … des carrés … moins … le carré … de la moyenne.

❸ Détermination de la moyenne et de l'écart-type d'une série avec une calculatrice ou un tableur

a) Calculer la moyenne et l'écart-type d'une série statistique à l'aide d'une calculatrice Casio GRAPH 25 ou d'une TI-82 Stats.fr.

La liste des paramètres de la série statistique est obtenue comme on l'a vu dans le chapitre sur la médiane et l'écart interquartile.

La moyenne $\bar{x}$ est facile à lire.

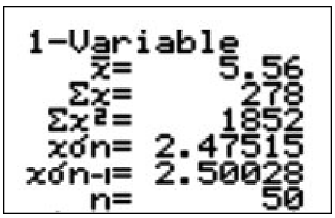
Il faut faire plus attention pour bien lire l'écart-type.

En effet, les mêmes tableaux sont utilisés ailleurs en statistique et un autre paramètre (que nous n'utiliserons pas) apparaît et il risque d'être confondu avec l'écart-type qui nous intéresse ici.

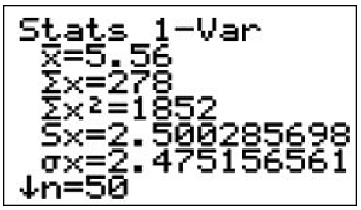
Il y a deux valeurs très proches qui sont nommées $x\sigma n$ et $x\sigma n-1$ ou encore σx et Sx (ou sx sur d'autres modèles de calculatrice).

L'écart-type est la plus petite de ces deux valeurs, $\boldsymbol{x\sigma n}$ pour la calculatrice Casio utilisée ici, $\boldsymbol{\sigma x}$ pour la calculatrice TI.

Casio :



TI :



b) Calculer la moyenne et l'écart-type d'une série statistique à l'aide d'un tableur.

Premier cas

Lorsque toutes les valeurs de la série sont énumérées dans une colonne, c'est-à-dire lorsque **tous les effectifs sont égaux à 1**, on utilise les fonctions statistiques présentes dans la plupart des tableurs.

Comme pour les calculatrices, il faut faire attention : l'écart-type dont nous avons besoins est celui d'une population (et non pas d'un échantillon).
Ici, avec OpenOffice, on choisira ECARTTYPEP.

Deuxième cas

Les effectifs ne sont pas tous égaux à 1, les valeurs sont présentées avec leur effectif (ou fréquence) dans deux colonnes, il faut faire les calculs intermédiaires avec le tableur.

Moyenne

On calcule dans la colonne C les produits des valeurs (colonne A) par leur effectif (colonne B) en écrivant dans la cellule C2 : =A2*B2, et en « étirant » la formule vers le bas jusqu'à la dernière valeur.

Dans deux cellules libres (par exemple B13 et C13) on calcule les sommes des colonnes B et C (effectif total et somme de toutes les valeurs) en écrivant : =SOMME(B2:B11) et =SOMME(C2:C11).

La moyenne s'obtient alors en divisant la somme des valeurs par l'effectif total, en écrivant dans une cellule libre (par exemple C15) : =C13/B13.

Ecart-type

On calcule les produits $n_i(x_i - \overline{x})^2$ dans la colonne D en écrivant dans la cellule D2 : =(A2-C13)^2, et en « étirant » la formule vers le bas jusqu'à la dernière valeur. Le symbole $ sert à « figer » la valeur « 15 » car la cellule C15 est celle qui contient la moyenne.

Dans une cellule libre (par exemple D13) on calcule la somme de la colonne D en écrivant : =SOMME(D2:D11).

Dans une cellule libre (par exemple D15) la variance s'obtient alors en écrivant =D13/B13. L'écart type s'obtient alors en écrivant dans une cellule libre (D17) : =RACINE(D15).

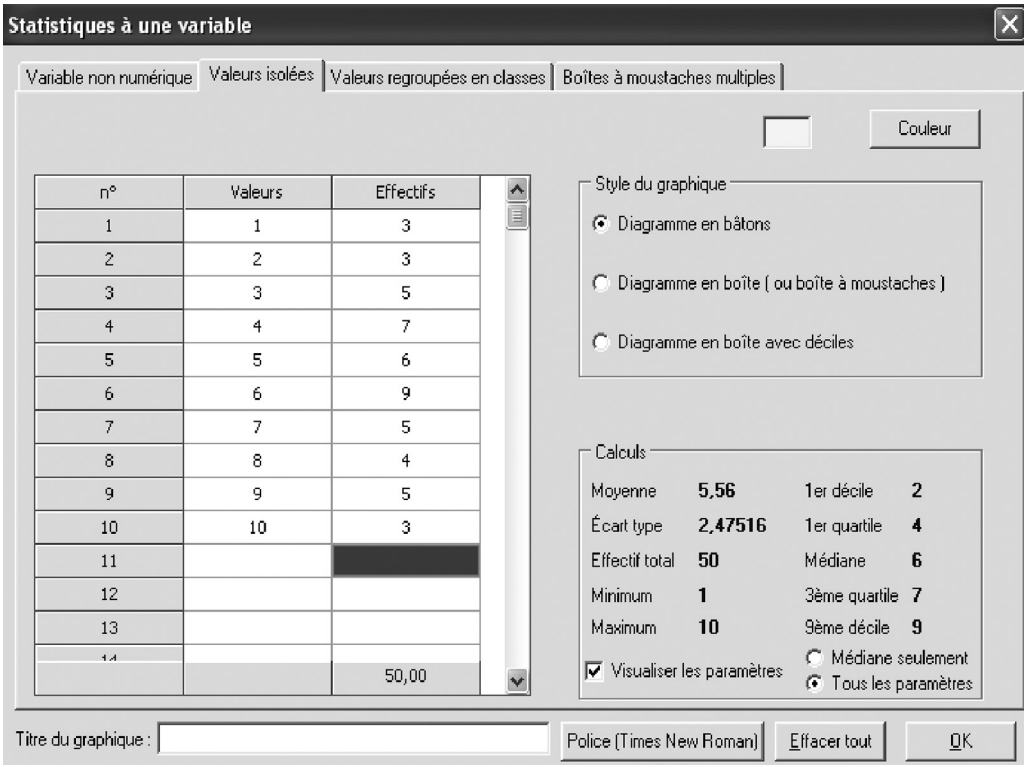
Deuxième méthode : pour limiter le nombre d'approximations dues à la moyenne, on peut utiliser l'égalité $V = \frac{1}{N}\sum_{i=1}^{i=p} n_i x_i^2 - \overline{x}^2$ (cellule G15).

G15 : =G13/B13-PUISSANCE(C15;2)

	A	B	C	D	E	F	G	H	I
1	x_i	n_i	n_ix_i	$n_i(x_i\text{-moy})^2$		deuxième	$n_ix_i^2$		
2	1	3	3	62,3808		méthode pou	3		
3	2	3	6	38,0208		l'écart-type :	12		
4	3	5	15	32,768			45		
5	4	7	28	17,0352			112		
6	5	6	30	1,8816			150		
7	6	9	54	1,7424			324		
8	7	5	35	10,368			245		
9	8	4	32	23,8144			256		
10	9	5	45	59,168			405		
11	10	3	30	59,1408			300		
12		N =	somme(n_ix_i)	somme($n_i(x_i\text{-moy})^2$)			somme($n_ix_i^2$)		
13		50	278	306,32			1852		
14			moyenne =	variance = [somme($n_i(x_i\text{-moy})^2$)]/N =			variance = [somme($n_ix_i^2$)]/N-moy^2 =		
15			5,56	6,1264			6,1264		
16				écart-type = racine(variance) =			écart-type = racine(variance) =		
17				2,475156561			2,475156561		
18									

c) Avec le logiciel Sinequanon

Les paramètres se lisent directement après avoir introduit les données.



C Exercices d'apprentissage

Exercice 6

Un élève a 12 de moyenne aux quatre premiers devoirs de l'année.

1. Si le cinquième devoir est noté 15, quelle sera sa nouvelle moyenne ?

2. Quelle est la note minimale du cinquième devoir pour que la moyenne aux cinq devoirs soit au minimum égale à 13 ?

Exercice 7

Dans une chaîne de magasins de vêtements, 60 % de ses magasins sont destinés aux hommes et 40 % sont destinés aux femmes.

Le chiffre d'affaire moyen des magasins pour hommes est de 1,1 million d'euros, celui des magasins pour femmes de 1,4 million d'euros.

1. Calculer le chiffre d'affaire moyen par magasin dans cette chaîne.

2. Le chiffre d'affaire de chaque magasin augmente de 5 %.
   Quel est le nouveau chiffre d'affaire moyen par magasin de cette chaîne ?

3. Le chiffre d'affaire de chaque magasin pour homme augmente de 5 % et celui de chaque magasin pour femme de 7 %.
   a) Sans faire de calcul, dire si le chiffre d'affaire moyen augmente de 6 %, plus de 6 % ou moins de 6 %.
   b) Calculer le nouveau chiffre d'affaire moyen par magasin de cette chaîne. Quel est le pourcentage d'augmentation de ce chiffre d'affaire moyen ?

Exercice 8

1. Une salle de spectacle a vendu pour une soirée 150 places à 12 € et 100 places à 10 €, quel est le prix moyen d'une place ?

2. Donner un exemple montrant un effet de structure. Pour cela on suppose que, pour une autre soirée, les deux prix augmentent de 1 € : les places seront donc vendues 13€ et 11 €. Chercher deux nombres entiers a et b non nuls tels que, si a places à 13 € ont été vendues ainsi que b places à 11 €, alors le prix moyen d'une place pour le second spectacle est inférieur au prix moyen d'une place pour le premier spectacle.

Exercice 9

On reprend la situation de l'exercice 2 du chapitre 3.

Deux sauteurs à la perche ont relevé leurs performance au cours des derniers mois.

1er sauteur

Hauteur	4,70	4,80	4,85	4,90	4,95	5,00	5,05	5,10	5,20
Nombre de sauts	1	1	1	3	12	4	1	1	1

2e sauteur

Hauteur	4,60	4,70	4,75	4,80	4,85	4,90	4,95	5,00	5,05	5,10	5,15	5,20
Nombre de sauts	3	2	2	3	2	2	1	3	2	1	1	3

Déterminer maintenant la moyenne et l'écart-type de chaque série.

Comparer l'ensemble des performances des deux sportifs en utilisant ces deux indicateurs.

Exercice 10

On reprend les données de l'exercice 1 du chapitre 2.

Une pharmacie de garde a enregistré le nombre d'appels reçus pendant 1000 nuits entre 20h et 6h du matin. Les résultats sont les suivants :

Nombre d'appels x_i	0	1	2	3	4	5	6	7	8	9	10	11
Nombre de nuits n_i	14	70	155	185	205	150	115	65	30	5	1	5

1. Déterminer la moyenne et l'écart-type de cette série statistique.
2. Déterminer le nombre de nuits pour lesquelles le nombre d'appels appartient à l'intervalle $[\overline{x} - s ; \overline{x} + s]$; quelle est la fréquence correspondante ?
3. Même question avec l'intervalle $[\overline{x} - 2s ; \overline{x} + 2s]$.

4 Synthèse

On peut résumer une série statistique en déterminant **une mesure de tendance centrale** et **la caractéristique de dispersion associée.**

Deux possibilités ont été étudiées : la médiane avec l'écart interquartile et la moyenne avec l'écart-type.

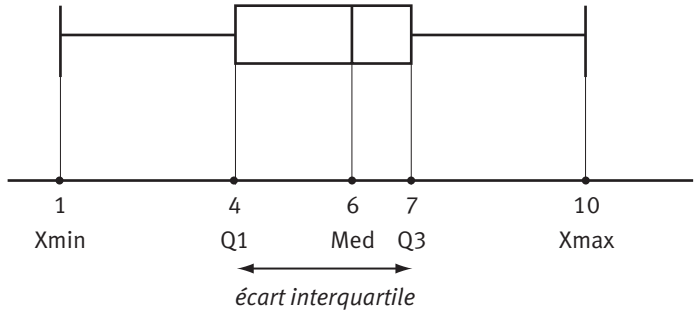
1 La médiane et les quartiles partagent la série statistique en quatre groupes de même effectif environ.

Ces paramètres sont assez simples à expliquer à des non-statisticiens.

Ils ne changent pas si les valeurs extrêmes sont un peu modifiées, on dit qu'ils sont « robustes ».

La médiane, les quartiles, l'écart interquartile permettent ainsi de décrire assez simplement une série statistique.

La représentation graphique par un diagramme en boîte donne immédiatement sur une image 5 (ou 7) paramètres, ce qui favorise les comparaisons.



2 On peut aussi résumer une série statistique par **sa moyenne et son écart-type**.

Pour une série statistique, la moyenne $\overline{x}$ est définie par l'égalité

$\overline{x} = \dfrac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N}$, et l'écart-type s est défini par

$s = \sqrt{V}$ avec $V = \dfrac{n_1(x_1 - \overline{x})^2 + n_2(x_2 - \overline{x})^2 + \ldots + n_p(x_p - \overline{x})^2}{N}$,

ou encore $V = \dfrac{1}{N}\left(\sum_{i=1}^{i=p} n_i x_i^2\right) - \overline{x}^2$

Ces deux paramètres sont moins simples que les précédents, mais ils sont très utiles.

- Si on connaît les effectifs P et Q et les moyennes partielles de deux sous-groupes de la série, on peut en déduire la moyenne de la série entière car $\bar{x} = \dfrac{P\overline{x'} + Q\overline{x''}}{P+Q}$.

- La relation précédente permet d'expliquer les étonnants effets de structure.

- La moyenne et l'écart-type ont des propriétés mathématiques très riches, ce qui les rend indispensables dans l'étude ultérieure des statistiques.

3. Dans la pratique, **il est indispensable de savoir déterminer la médiane, les quartiles, la moyenne et l'écart-type d'une série statistique avec une calculatrice ou avec un tableur ou avec le logiciel sinequanon.**

5 Exercices d'approfondissement

Pour ces exercices, il est vivement conseillé d'utiliser une calculatrice ou un tableur ou le logiciel sinequanon.

Exercice I

1. Voici la liste des notes obtenues par une classe au premier trimestre.

10 – 15 – 18 – 5 – 11 – 6 – 9 – 12 – 12 – 17 – 4 – 7 – 10 – 8 – 9 –14 – 16 – 7 –11 –15 – 11 – 10.

Déterminer la médiane, les quartiles, puis la moyenne et l'écart-type.

2. Même question pour le second trimestre pour lequel les notes sont :

11 – 14 – 15 – 5 – 11 – 9 – 10 – 13 – 12 – 15 – 5 – 8 – 10 – 8 – 9 – 13 – 14 – 8 –13 – 13 –10 –11.

3. En utilisant les paramètres de position et les paramètres de dispersion qui ont été déterminés, comparer les deux séries statistiques

Exercice II

Le tableau ci-dessous donne, pour l'année 2008, le nombre de médecins généralistes et le nombre de médecins spécialistes pour 100 000 habitants (données de l'INSEE).

Région	Nombre de médecins généralistes pour 100 000 hab.	Nombre de médecins spécialistes pour 100 000 hab.		Nombre de médecins généralistes pour 100 000 hab.	Nombre de médecins spécialistes pour 100 000 hab.
Alsace	169	184	Midi-Pyrénées	173	179
Aquitaine	171	178	Nord-Pas-de-Calais	165	138
Auvergne	159	138	Basse-Normandie	143	138
Bourgogne	152	134	Haute-Normandie	141	132
Bretagne	157	179	Pays de la Loire	142	136
Centre	135	131	Picardie	140	116
Champagne-Ardenne	152	131	Poitou-Charentes	159	133
Corse	165	153	Provence-Alpes-Côte d'Azur	188	218
Franche-Comté	158	137	Rhône-Alpes	161	172
Ile-de France	175	230	Guadeloupe	139	114
Languedoc-Roussillon	176	185	Guyane	99	71
Limousin	177	159	Martinique	138	121
Lorraine	154	151	La Réunion	149	123

Comparer ces deux séries en déterminant pour chacune la moyenne et l'écart-type, puis en faisant les deux diagrammes en boîte.

Exercice III

Voici un tableau obtenu à partir des données de l'INSEE.

Faire de même qu'à l'exercice précédent avec la série des données de 1995 et avec celle de 2009.

Pourcentage de femmes élues au Parlement dans quelques pays du monde								
Pays	**1995**	**2009**		**1995**	**2009**		**1995**	**2009**
Afrique du Sud	25	45	Espagne	16	36	Pays-Bas	31	41
Algérie	7	8	États-Unis	11	17	Pologne	13	20
Allemagne	26	33	Finlande	34	42	Portugal	9	28
Argentine	22	42	**France**	**6**	**18**	République tchèque	10	16
Australie	10	27	Grèce	6	17	Royaume-Uni	10	20
Autriche	24	28	Hongrie	11	11	Russie	13	14
Belgique	12	35	Inde	8	11	Rwanda	4	56
Brésil	7	9	Irlande	13	13	Sénégal	12	22
Cameroun	12	14	Italie	15	21	Suède	40	47
Canada	18	22	Japon	3	11	Suisse	18	29
Chine	21	21	Lituanie	7	18	Tunisie	7	28
Corée du Sud	2	14	Luxembourg	20	20	Turquie	2	9
Cuba	23	43	Malte	2	9	Viêt Nam	19	26
Danemark	33	38	Mexique	14	28			
						Monde	**12**	**19**

Exercice IV

Dans un lycée, on a rendu les copies d'un contrôle commun aux élèves des trois classes de Première ES.

Pour chacune des classes on a déterminé les paramètres suivants (m désigne la médiane) :

$1ES_A$: l'effectif est $N = 30$ et

$x_{min} = 2, Q_1 = 8, m = 11, Q_3 = 13, x_{max} = 18, \overline{x} = 11{,}5$ et $s = 3{,}5$.

$1ES_B$: l'effectif est $N' = 28$ et

$x'_{min} = 5, Q'_1 = 9{,}5, m' = 12, Q'_3 = 13, x'_{max} = 15, \overline{x'} = 12{,}3$ et $s' = 2{,}7$.

$1ES_C$: l'effectif est $N'' = 33$ et

$x''_{min} = 4, Q''_1 = 7, m'' = 10, Q''_3 = 15, x''_{max} = 17, \overline{x''} = 12$ et $s'' = 4{,}1$.

On veut faire un bilan général pour l'ensemble des élèves de ces trois classes.

Quel(s) indicateur(s) numérique(s) peut-on déduire des données précédentes ?

Exercice V

Le tableau suivant donne le montant (en tonnes) des ventes d'une ferme d'élevage de saumons sur une période de 15 ans.

année	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Montant x_i en tonnes	51	60	68	75	67	80	85	88	81	93	92	91	100	105	107

1. Construire le graphique représentant l'évolution de ces ventes ;
2. On remplace chacune des valeurs de la série (à partir de la 2ième) par la moyenne de cette valeur avec les deux qui l'entourent.

   Par exemple, $y_2 = \dfrac{51+60+68}{3}$ soit $y_2 = 59,7$.

   Calculer de même $y_3, ..., y_{14}$ et construire l'évolution de ces moyennes.
3. Quelle tendance peut-on mettre ainsi en évidence ?

Exercice VI

Moyennes mobiles

On appelle moyenne mobile centrée d'ordre k, pour k impair, la série obtenue en remplaçant la valeur x_i de rang i de la série par la moyenne arithmétique de x_i et des $k-1$ valeurs qui l'entourent :

Ordre 3 : $y_i = \dfrac{x_{i-1}+x_i+x_{i+1}}{3}$

Ordre 5 : $y_i = \dfrac{x_{i-2}+x_{i-1}+x_i+x_{i+1}+x_{i+2}}{5}$ *, etc.*

La série de moyennes mobiles permet de lisser la série chronologique initiale en gommant les irrégularités comme on a pu le constater sur l'exercice précédent où l'on a calculé des moyennes mobiles centrées d'ordre 3.

Le tableau ci-dessous donne l'indice des prix d'une matière, année par année, de 2000 à 2011.

Année	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Indice	100	80	110	135	95	105	140	160	120	110	80	105

1. A l'aide d'un tableur, calculer les moyennes mobiles d'ordre 3 et 5 de cette série.
2. Insérer un diagramme montrant les trois courbes ; celle de la série initiale, celles des moyenne mobiles d'ordre 3 et celles des moyennes mobiles d'ordre 5. Que pouvez vous constater concernant ces courbes ?

Exercice VII *Courbes de Lorenz*

1. Les fonctions f et g sont définies sur $[0\,;1]$ par $f(x)=0{,}2x^2+0{,}8x$ et $g(x)=0{,}8x^2+0{,}2x$

   a) Etudier les variations de f et g sur $[0\,;1]$ et construire leurs courbes représentatives dans un repère orthonormé d'unité 10 cm.

   b) Construire sur le même graphique la droite d'équation $y=x$ restreinte à $[0\,;1]$.

2. Les courbes représentatives des fonctions f et g sont des courbes de Lorenz de deux pays F et G.

   Elles illustrent la répartition du patrimoine des ménages dans chacun des pays.

   En abscisse, x représente le pourcentage des personne les plus pauvres par rapport à la population totale, et en ordonnée, y représente le pourcentage du patrimoine total qu'ils possèdent.

   Exemple de lecture : $f(0{,}2)=0{,}168$ signifie que 20% des personnes les plus pauvres possèdent 16,8% du patrimoine total.

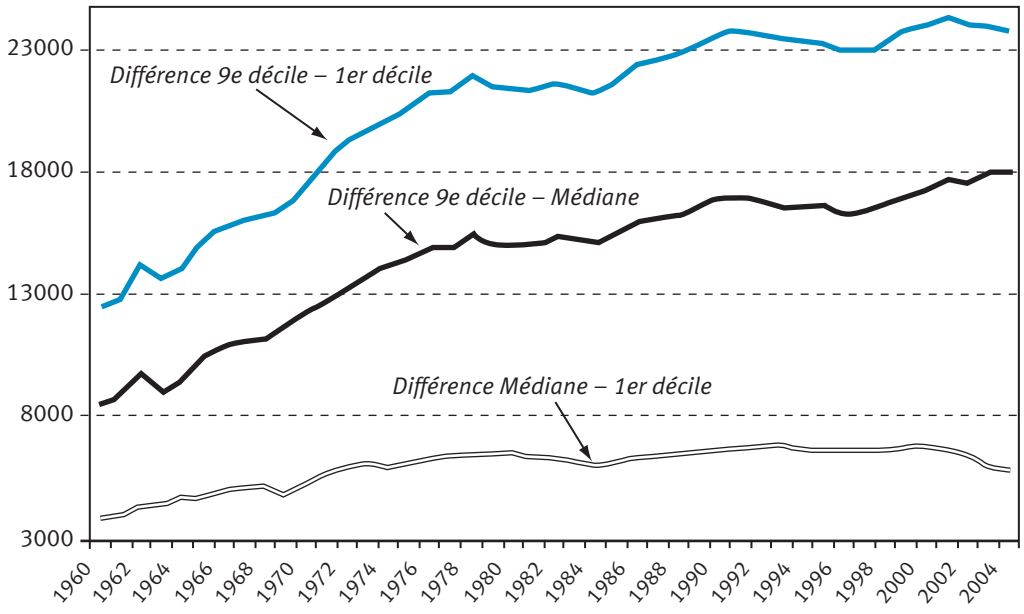
   a) Sachant que, pour chacun de ces pays, le patrimoine total des ménages est d'environ 165000 €, déterminer pour chacun des pays la médiane, les premiers et troisième quartiles, les premiers et neuvième déciles de la série des patrimoine des ménages.

   b) Construire les diagrammes en boîte correspondant à chacun des pays, les moustaches des boîtes s'arrêtant au premier et au neuvième décile. Commentez.

Exercice VIII

Commenter le graphique ci-dessous.

Échelle absolue des salaires en France : différences absolues de salaires annuels réels (en € 2005)



Source : Insee, Dads.