UNIVERSITE DE BLIDA 1

FACULTÉ DES SCIENCES

DEPARTEMENT DU TRONC COMMUN MATHEMATIQUES ET INFORMATIQUE(MI)

Cours de Statistique descriptive à une seule variable

Réalisé par : Mme Messaoudi Nadia

Chargée de cours : **1 er année MI Section A**

Courriel : messlina2012@gmail.com

Année universitaire 2019- 2020

Chapitre 3

Représentation des données : Paramètres caractéristiques d'une variable

On cherche souvent à illustrer une distribution de manière beaucoup plus sommaire par quelques nombres caractéristiques. Parmi ceux-ci, les mesures de tendance centrale (aussi appelées paramètres de position) jouent un rôle important. Les plus connues sont : **le mode, la médiane et la moyenne**.
Les mesures de tendance centrale ne suffisent pas à donner une idée de la manière dont les valeurs sont distribuées au voisinage de ces valeurs centrales (en général, la moyenne). Aussi est-il nécessaire d'introduire une mesure de la dispersion. La plus utilisée est l'écart-type, l'intervalle interquartile est aussi trés fréquent. Nous citons également l'étendue, l'écart absolu moyen et la variance.
Dans ce cours, nous présentons trois paramètres de dispersion : **l'étendue**, **l'écart-type** et **la variance**.

3.1 Paramètres de position : mode, médiane, moyenne

Les paramètres de position ou de caractéristique de tendance centrale permettent de résumer et de synthétiser les séries d'observations.

3.1.1 Le mode

Le mode, noté (Mo) est la valeur de la variable statistique la plus fréquente, c'est à dire celle qui correspond au plus grand effectif (ou fréquence).
Dans le cas d'une variable continue en classes, on parle de classe modale : classe ayant l'effectif (ou fréquence) par unité d'amplitude le (la) plus élevé (e).

Remarques :

•*Le mode peut ne pas être unique.*

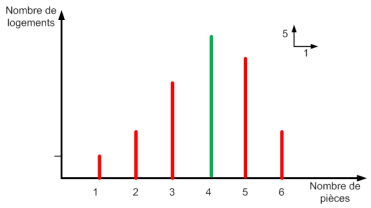
• *Le mode peut être calculé pour tous les types de variables, quantitatives et qualitatives.*

Mode : Cas discret

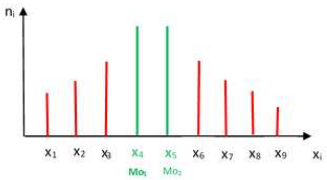
Il correspond à la variable statistique ayant le plus grand effectif (ou fréquence).
Graphiquement, le mode est la valeur de la variable statistique x correspondant au maximum du diagramme en bâtons (ou la valeur x_i correspondant au bâton le plus élevé).

Exemple :

D'aprés le tableau statistique de l'exemple 3 du nombre de pièces par logement. La valeur dominante du nombre de piéces par logement est égal à 4, car il correspond au plus grand effectif soit 30.



(a) Distribution unimodale



(b) Distribution bimodale

FIGURE 3.1 – Mode : cas discret

Définitions

• La distribution statistique est dite unimodale lorsqu'une seule valeur de la variable statistque x correspond au plus grand effectif.

• La distribution statistique est dite bimodale, trimodale ou plurimodale lorsque correspondent respectivement deux valeurs modales, trois valeurs modales ou plusieurs valeurs modales de la variable x à deux, trois ou plusieurs grands effectifs.

Mode : Cas continu

Pour une variable quantitative continue, nous parlons de *classe modale* qui correspond à l'effectif (ou fréquence) le (ou la) plus élevé (e) sur l'histogramme.
La valeur numérique du mode peut être déterminée par le calcul ou graphiquement.
La méthode graphique consiste à construire l'histogramme sur un papier millimétré dans un premier temps, puis de tracer les diagonales AB et CD de la classe modale $[e_i, e_{i+1}[$ et en ensuite de projeter perpendiculairement leur intersection au point G sur le segment de la classe $[e_i, e_{i+1}[$ à partir du quel on obtient la valeur du mode.

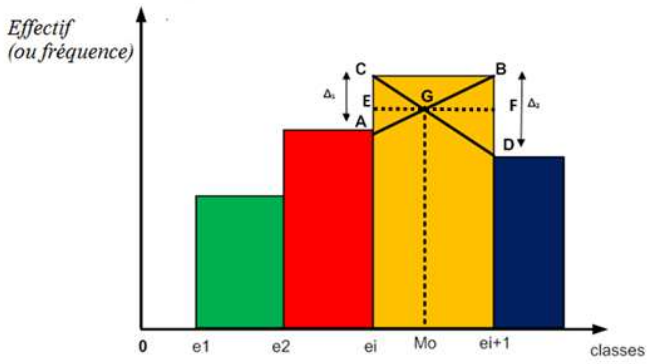


FIGURE 3.2 – Histogramme : Mode

Détermination analytique de la valeur du mode

On considère $[e_i, e_{i+1}[$ la classe modale. D'aprés l'histogramme donné par la figure 3.2, en appliquant le théorème de Thalès, les triangles DGB et AGC sont semblables, dont on a en valeurs algébriques les rapports suivants :

$$\frac{GE}{AC} = \frac{FG}{DB}$$

Ces rapports algébriques sont équivalents à :
GE équivalents à $[Mo - e_i]$; FG équivalents à $[e_{i+1} - Mo]$
AC équivalents à $[\Delta_1]$; DB équivalents à $[\Delta_2]$
En remplacant GE, AC, FG et DB par leurs équivalentes, l'égalié devient alors :

$$\frac{Mo - e_i}{\Delta_1} = \frac{e_{i+1} - Mo}{\Delta_2} \Longrightarrow (Mo - e_i)\Delta_2 = (e_{i+1} - Mo)\Delta_1$$

En développant les deux membres de l'égalié, on obtient :

$$Mo(\Delta_1 + \Delta_2) = e_{i+1}\Delta_1 + e_i\Delta_2$$

Comme $e_{i+1} = e_i + a_i$, on peut écrire :

$$Mo(\Delta_1 + \Delta_2) = (e_i + a_i)\Delta_1 + e_i\Delta_2 \Longrightarrow Mo(\Delta_1 + \Delta_2) = e_i(\Delta_1 + \Delta_2) + a_i\Delta_1$$

D'où :

$$Mo = e_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times a_i$$

avec :

e_i : extrémité inférieure de la classe modale.

e_{i+1} : extrémité supérieure de la classe modale.

a_i : amplitude de la classe modale.

Δ_1 : la différence entre les effectifs (ou fréquences) de la classe modale et celle de la classe précédente.

Δ_2 : la différence entre les effectifs (ou fréquences) de la classe modale et celle de la classe suivante.

Propriétés du mode

• Le mode est simple et facile à calculer, pour caractériser la distribution statistique, on préfère souvent la moyenne arithmétique.

• Lorsque les amplitudes sont inégales, pour la détermination du mode, on doit corriger les effectifs (ou fréquences) de la distribution statistique. Les effectifs à corriger sont à diviser par autant d'unités d'amplitudes dont l'unité d'amplitude "a" est le PGCD de toutes les amplitudes des classes.

Exemple : Cas amplitudes toutes égales

Considérons l'exemple d'ancienneté du personnel cadre d'une entreprise :

Durée en années	[0,4[	[4,8[	[8,12[	[12,16[	[16,20[
n_i	8	22	8	7	5

La classe modale est $[4, 8[$ puisque elle correspond au plus grand effectif, soit 22 employés. Sa valeur est égale :

$$Mo = 4 + \frac{(22-8)}{(22-8)+(22-8)} \times 4 = 6$$

L'ancienneté du personnel la plus fréquente est 6 ans.

La détermination du mode est illustrée par le graphique suivant :

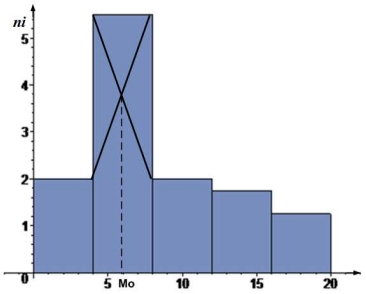


FIGURE 3.3 – Histogramme : Mode

Exemple : Cas d'amplitudes inégales

Soit la distribution d'âge (année) des personnes hospitalier dans un service d'hôpital :

Durée en années	[10,20[	[20,30[	[30,50[	[50,80[	[80,90[
Effectif	12	14	32	47	15
a_i	10	10	20	30	10
$n_i^c = \frac{n_i}{a_i} \times a$	12	14	16	15.66	15

On remarque, que les classes n'ont pas la même amplitude, pour cela on calcul l'unité d'amplitude. Elle vaut : $au = a = PGCD(10, 20, 30) = 10$.

Pour cela, on doit procéder à la correction des effectifs (ou fréquences). La classe modale est donc $[30, 50[$ correspondant à l'effectif corrigé le plus élevé $n_i^c = 16$ et non pas $[50, 80[$. Le mode est égal à :

$Mo = e_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times a_i = 30 + \frac{(16-14)}{(16-14)+(16-15.66)} \times (50-30) = 47.09$ L'âge le plus fréquent des patients dans le service d' hôpital est environ 47 ans.

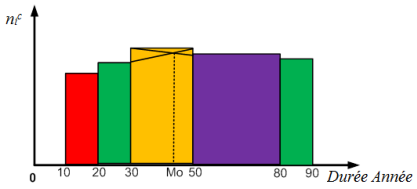


FIGURE 3.4 – Histogramme : Mode

3.1.2 La médiane

La médiane, notée *Me* est définie comme étant la valeur de la variable qui divise la totalité des effectifs (ou fréquences) en deux parties égales.
Dans le cas d'une variable continue, on parle de la classe médiane.

Médiane : Cas discret

Lorsque la variable est discrète, on rencontre deux possibiliés :
a) Série statistique à valeurs non répétitives : Pour calculer la médiane, on commence par ranger en ordre croissant ou décroissant les N valeurs prises par la variable statistique discrète.

- Si le nombres d'observations est impair $N = 2p + 1$ avec $p = \dfrac{N-1}{2}$:

$$x_1,\ x_2,...,x_p,\ x_{p+1},\ x_{p+2},...,x_{2p+1}$$

La médiane est :

$$Me = x_{p+1}$$

Exemple : Les notes d'examen d'un groupe d'étudiants sont les suivantes : 1,1,1,3,4,6,8,8,10. $N = 9, p = 4$. La valeur médiane de la série ordonnée est $Me = x_5 = 4$

- Si le nombres d'observations est pair $N = 2p, p = \dfrac{N}{2}$:

$$x_1,\ x_2,\ ....,x_p,\ x_{p+1},....,x_{2p}$$

$$Me = \frac{x_p + x_{p+1}}{2}$$

Exemple : Soit les notes d'un examen des étudiants : 1,2,4,6,9,10. $N = 6, p = 3$.
La valeur médiane est égale à : $Me = \dfrac{x_3 + x_4}{2} = \dfrac{4+6}{2} = 5$

a) Série statistique à valeurs répétitives :
Lorsqu'on dispose d'une série statistique discrète (x_i, n_i), représentée par un tableau statistique, on peut déterminer la médiane en utilisant les effectifs cumulés (ou fréquence cumulées) de la manière suivante :
Utiliser la colonne des n_i cumulés (ou F_i cumulées) .

Situer où se trouve $\frac{N}{2}$ (ou $F(Me) = 0.5$). Deux cas peuvent se présentés :

- *Cas 1 : (voir figure 3.5 (a))* : S'il existe un x_j telle que :

$$n_{j-1}cum < \frac{N}{2} < n_jcum$$

(ou

$$F_{j-1} < 0.5 < F_j)$$

alors $Me = x_j$.

-*Cas 2 : voir figure 3.5 (b)* : S'il existe une valeur x_j telle que :

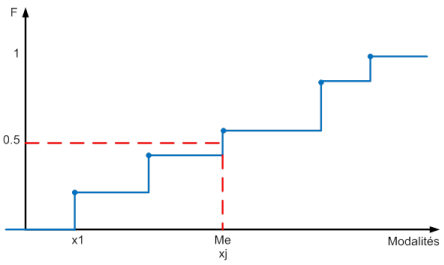
$$n_jcum = \frac{N}{2}$$

(ou 0.5), on parle dans ce cas d'intervalle médiane $Me \in [x_j, x_{j+1}[$, sa valeur vaut $Me = \frac{x_j+x_{j+1}}{2}$.

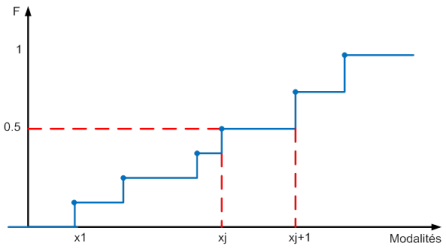
Détermination de la médiane graphiquement

A l'aide de la construction du tableau on a pu déterminer la médiane, celle-ci peut être déterminée par la construction graphique. En effet, la représentation graphique de la distribution statistique (n_i, x_i) consiste à reporter sur un diagramme à deux dimensions sur l'axe des abscisses les valeurs (x_i) de la variable statistique et sur l'axe des ordonnées les fréquences cumulées F(x) (ou effectifs cumulés N(x)). Ainsi, chaque point reporté sur le graphe a pour coordonnées,$p_i = (x_i, F(x_i))$ ou $(x_i, N(x_i))$, en joignant les points p_i, on obtient la courbe cumulative en escalier dont la médiane Me correspond à l'ordonnée $F(Me) = 0.5$ (ou $N(Me) = \frac{N}{2}$) (voir figure 3.5).

Si la valeur de la médiane correspond à un intervalle médian $[x_j, x_{j+1}[$ sur le tableau statistique, sur la courbe cumulative F(x) elle correspond à un palier horizontal $Me \in [x_j, x_{j+1}[$, ce qui montre la figure 3.5 (b). La détermination de la médiane graphiquement est donnée par la courbe cumulative en escalier :



(a) Cas 1 : Médiane



(b) Cas 2 : Médiane

FIGURE 3.5 – Courbe cumulative en escalier : Médiane

Exemple : Reprenons l'exemple 3 de la distribution du nombre de pièces par logement.

Nombre de pièces	1	2	3	4	5	6	Total
Effectif n_i	5	10	20	30	25	10	100
Fréquence f_i	0.05	0.1	0.2	0.3	0.25	0.1	1
$F_i cum \nearrow$	0.05	0.15	0.35	0.65	0.9	1	

On remarque que $0.35 < F(Me) = 0.5 < 0.65$. Donc la médiane est égale à 4 pièces par logement.

Interprétation de la médiane : 50% *des logements ont le nombre de pièces inférieur à 4 et l'autre* 50% *ont le nombre de pièces supérieur à 4.*

Détermination de la médiane graphiquement :

On tracer la courbe cumulative des $Ficum \nearrow$. A partir de la valeur de $F(Me) = 0.5$, on fait une projection sur la courbe cumulative, ensuite à partir de ce point d'intersection on fait une projection perpendiculairement sur l'axe des abscisses, on obtient ainsi la médiane.

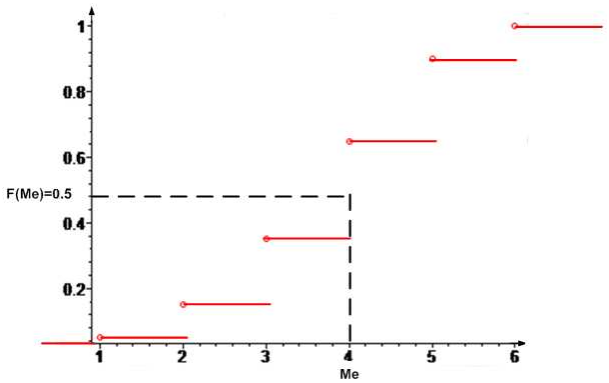


FIGURE 3.6 – Courbe en escalier : Médiane

Exemple : Soit la distribution du nombre d'enfants de 50 employés d'une entreprise :

Nombre d'enfants	0	1	2	4	5	6	Total
Effectif n_i	2	10	13	10	9	6	50
$n_i cum \nearrow$	2	12	25	35	44	50	

La valeur de la médiane de la variable statistique correspond à $\frac{N}{2} = \frac{50}{2} = 25$(ou 0.5) est comprise entre 2 et 4. On parle ici, d'intervalle médiane $Me \in [2, 4[$. Une valeur approchée de la médiane est égal à :

$$Me = \frac{2+4}{2} = 3 enfants$$

.

Médiane graphiquement :

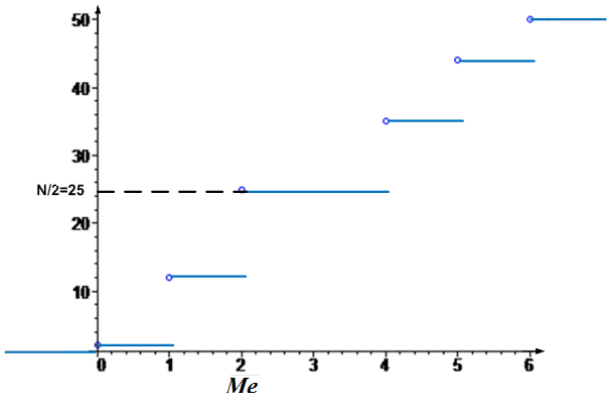


FIGURE 3.7 – Courbe en escalier : Médiane $Me \in [2, 4[$

D'aprés la courbe des effectifs cumulés croissants, la valeur de $N/2 = 25$ coincide avec le pallier $[2, 4[$ qui est considéré comme intervalle médian.

Médiane : cas d'une variable quantitative continue

Lorsque la variable statistique x est continue, on peut déterminer la médiane,à partir de la construction du tableau des fréquences cumulées $F(x_i)$ (ou effectifs cumulés $N(x_i)$), on va repérer parmi les valeurs x_i celle qui correspond à 0.5 (ou N/2). Ainsi, on peut déterminer la médiane par le calcul ou graphiquement.

Dans la détermination de la médiane par la méthode de calcul deux cas peuvent se présenter :

• Cas 1 : Si la moitié de l'effectif total $\frac{N}{2}$ (ou fréquence 1/2) correspond à une valeur de l'extrémité de la classe, *la médiane* est une valeur exacte.

Exemple Soit la répartition des salaires horaires des employés d'une entreprise en $(10DA)$:

Salaire horaire	[10,20[	[20,30[	[30,40[	[40,60[	[60,70[	[70,80[	Total
Effectif n_i	10	15	25	30	15	5	100
$n_i cum \nearrow$	10	25	50	80	95	100	

La valeur $\frac{N}{2} = \frac{100}{2} = 50$ (ou 0.5) correspond à l'extrémité supérieure de la classe $[30, 40[$. Donc Me =400 Da/heure.

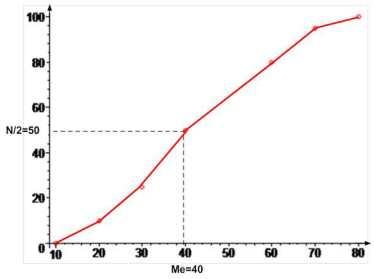


FIGURE 3.8 – Courbe cumulative : Médiane

Interprétation de la médiane : *on peut dire que la moitié des employés ont un salaire horaire inférieur à 400 DA/h et l'autre moitié ont un salaire horaire qui dépasse 400Da/h.*

Remarque : *Contrairement à la détermination du mode, pour le calcul de la médiane on ne corrige pas les effectifs (ou fréquences).*

• Cas 2 : Si la moitié de l'effectif total $\frac{N}{2}$ (ou fréquence totale 0.5) est comprise entre deux extrémités de classe e_i et e_{i+1}, on a un intervalle médian ou une classe médiane $Me \in [e_i, e_{i+1}[$.
La portion de courbe de la classe médiane est donnée par le graphique suivant :

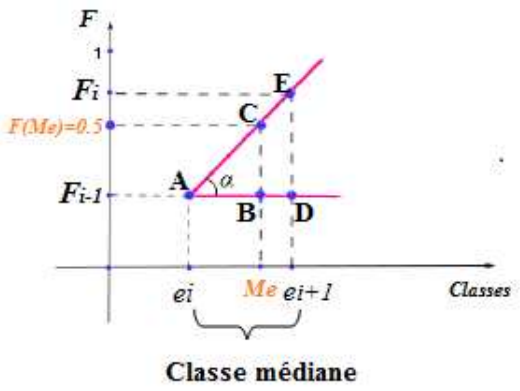


FIGURE 3.9 – Détermination de la médiane par Thalès

On considère F_i et F_{i-1} comme étant les valeurs des fréquences cumulées correspondant à e_i et e_{i+1} les extrémités de la classe médiane. D'après le théorème de Thalès les triangles ABC et ADE sont semblables, donc on a les valeurs algébriques :

$$\frac{AB}{AD} = \frac{BC}{DE} \Longleftrightarrow \frac{AB}{BC} = \frac{AD}{DE}$$

D'aprés la figure 3.9, les rapports algébriques sont équivalents à :

$\frac{Me-e_i}{e_{i+1}-e_i} = \dfrac{0.5 - f_{i-1}cum\nearrow}{f_i}$

D'où :

$$Me = e_i + \frac{0.5 - f_{i-1}cum\nearrow}{f_i} \times a_i$$

avec :

f_i est la fréquence de la classe médiane,

a_i est l'amplitude de la classe médiane $[e_i, e_{i+1}[$,

$f_{i-1}cum\nearrow$ fréquence cumulée croissante de la classe qui précèdent la classe médiane.

Si on utilise les effectifs cumulés, la formule deviendrait alors la suivante :

$$Me = e_i + \frac{(N/2) - n_{i-1}cum\nearrow}{n_i} \times a_i$$

Exemple :

Durée en années	[10,20[	[20,30[	[30,50[	[50,80[	[80,90[
Effectif n_i	12	14	32	47	15
$n_icum\nearrow$	12	26	58	105	120

On remarque que : $58 < \dfrac{N}{2} = \frac{120}{2} = 60 < 105$. Donc la classe médiane est la première classe qui dépasse par excès $\dfrac{N}{2}$, alors $Me \in [50, 80[$. Sa valeur :

$$Me = 50 + \frac{60 - 58}{47} \times (80 - 50) = 51,276 ans$$

Détermination graphique de la médiane :cas continu

Contrairement à la variable discrète, la représentation graphique de la courbe cumulative dans le cas continu est identique dans les deux cas en raison de la continuité de la distribution.

La courbe est construite en portant sur l'axe des abscisses les classes de valeurs et sur l'axe des ordonnées les fréquences cumulées F(x)(ou effectifs cumulés N(x)), la valeur de la médiane correspond à l'ordonnée F(x)=0.5 (ou N(x)=N/2).

La courbe des effectifs cumulés croissants $n_i cum \nearrow$ de l'exemple précédent est donnée par la figure suivante :

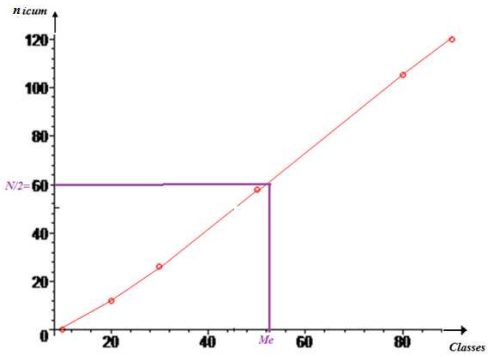


FIGURE 3.10 – Courbe cumulative : Médiane cas continu

A partir de l'ordonnée $(N/2 = 60)$, on fait une projection sur la courbe, puis à partir de ce point on projette perpendiculairement sur l'axe des classes, on obtient ainsi la médiane,$Me = 51 ans$.

Propriétés de la médiane

• L'interprétation de la médiane est simple. Elle indique que 50 % des observations sont supérieures à la valeur médiane et 50 % lui sont inférieures.

• La médiane n'est pas influencée par des classes d'amplitudes inégales.

• La médiane n'est pas affecter par les valeurs extrêmes de la distribution statistique.

3.1.3 Moyenne arithmétique

Définition : La moyenne arithmétique notée $\bar{x}$ (se lit x barre) d'une variable statistique x se définit comme étant le rapport de la somme des valeurs prises par cette variable sur leur nombre d'apparition N. On distingue deux types de moyennes : la moyenne arithmétique simple et la moyenne arithmétique pondérée.

• Moyenne arithmétique simple :

Lorsque les valeurs prises par la variable x n'apparaissent qu'une seule fois chacune, elle vaut :

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \sum_{i=1}^{N} \frac{x_i}{N}$$

Exemple :

Soient les notes de 4 étudiants dans un contrôle : 2,2,3,5. la note moyenne des étudiants est :

$$\bar{x} = \frac{2+2+3+5}{4} = 3$$

• Moyenne arithmétique pondérée :

Lorsque les valeurs prises par la variable statistique x : $(x_1, x_2, ..., x_k)$ ont des fréquences d'apparition respectives $(n_1, n_2, ..., n_k)$, la moyenne arithmétique correspondante est pondérée et égale à :

$$\bar{x} = \sum_{i=1}^{k} \frac{n_i x_i}{N} = \sum_{i=1}^{k} f_i x_i$$

Moyenne : cas de variable statistique discrète

Lorsque la variable statistique est discrète à valeurs répétitives on rajoute une nouvelle ligne $(n_i \times x_i)$ dans le tableau statistique pour faciliter les calculs de la moyenne arithmétique.

Exemple : Soit la distribution de logement selon le nombre de pièces :

Nombre de pièces x_i	1	2	3	4	5	6	Total
Effectif n_i	5	10	20	30	25	10	$\Sigma n_i = 100$
$n_i x_i$	5	20	60	120	125	60	$\Sigma n_i x_i =$ 390

Donc $\bar{x} = \frac{390}{100} = 3.9$. Le nombre moyen de pièces par logement est environ 4.

Ou bien :

Nombre de pièces x_i	1	2	3	4	5	6	Total
Effectif n_i	5	10	20	30	25	10	$\Sigma n_i = 100$
Fréquence f_i	0.05	0.1	0.2	0.3	0.25	0.1	1
$f_i x_i$	0.05	0.20	0.6	0.12	0.125	0.6	$\Sigma f_i x_i = 3.9$

Donc $\bar{x} = \sum_{i=1}^{6} f_i x_i = 3.9$ Ainsi, en utilisant les fréquences f_i la moyenne arithmétique est identique, ce qui signifie que les deux formules de calcul sont équivalentes.

Moyenne : cas de variable statistique continue

Les modalités du caractètre sont constituées de classe de valeurs et afin de calculer la moyenne, on est amené à choisir les centres x_i de chaque classe $[e_i, e_{i+1}[$.

$$x_i = \frac{e_i + e_{i+1}}{2}$$

Cette convention signifie que toutes les observations sont uniformément réparties à l'intérieur des classes de valeurs.

Exemple

Durée en années	[10,20[	[20,30[	[30,50[	[50,80[	[80,90[
Centre x_i	15	25	40	65	85
Effectif n_i	12	14	32	47	15
Fréquence f_i	0.1	0.117	0.267	0.391	0.125
$f_i x_i$	1.5	2.925	10.68	25.415	10.625

L'âge moyen est égal à : $\bar{x} = \sum_{i=1}^{5} f_i x_i = 51.145 ans.$

Définition : on appelle changement d'origine et d'unité l'opération consistant à multiplier (ou diviser) toutes les observations par la même quantité $b \in \Re$ puis à ajouter la même quantité $a \in \Re$ à toutes les observations x_i, d'où les nouvelles observations :

$$y_i = a + bx_i, i = 1, ..., k$$

Propriétés de la moyenne

• L'écart par rapport à la moyenne simple est nul :

$$\sum_{i=1}^{k}(x_i - \bar{x}) = 0$$

Preuve : $\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} \bar{x} = N\bar{x} - N\bar{x} = 0$

• Si chaque effectif est pondéré par la même constante C, la moyenne reste inchangée :

$$\sum_{i=1}^{k} \frac{Cx_i}{NC} = \bar{x}$$

• Si on effectue un changement d'origine et d'unité sur un variable x, alors sa moyenne est affectée du même changement d'origine et d'unité . Si $y_i = a + bx_i$ alors $\bar{y} = a + b\bar{x}$

• Généralement, on utilise la moyenne comme paramètre de tendance centrale. Lorsque la série contient des valeurs aberrantes, la moyenne sera influencée, dans ce cas on utilise la médiane et quelque fois le mode.

• Lorsque la distribution est bimodale (ou plurimodale) la moyenne arithmétique perd toute signification.

• Si on connait seulement la somme des valeurs de la série statistique et le total des effectifs, la moyenne peut être calculée contrairement au mode et à la médiane ($\sum x_i = N\bar{x}$).

3.2 Paramètres de dispersion

L'analyse des séries statistique à l'aide des paramètres de position est insuffisante. A titre d'exemple, les deux séries suivantes concernant la distribution des salaires horaires (en DA) des ouvriers de deux entreprises :

Entreprise E_1 : 14, 15, 18, 25,30, 35 et 38.

Entreprise E_2 : 18, 20, 22, 25, 29,30 et 31.

On calcule les valeurs centrales pour les deux entreprises E_1 et E_2, on obtient les valeurs $(Me_1 = 25, \bar{x}_1 = 25)$ et $(Me_2 = 25, \bar{x}_2 = 25)$ respectivement. Cela ne signifie pas pour autant qu'elles sont identiques, ou qu'elles se dispersent de la même manière autour d'une (ou ces) valeur(s) centrale(s).

Pour les comparer, on doit recourir aux mesures de dispersion. Elles nous permettre de savoir laquelle des deux séries est la plus dispersée ou la plus concentrée, en mesurant l'écart des observations par rapport à une valeur centrale (généralement la moyenne arithmétique). Ainsi on présente : l'étendue d'une série, l'écart-type et la variance.

3.2.1 L'étendue

Considérons une série de k valeurs rangées dans l'ordre croissant : $x_1, x_2, ..., x_k$.
L'étendue notée e est :

$$e = max(x_1, ..., x_k) - min(x_1, ..., x_k)$$

Exemple
Reprenons l'exemple 3 du nombre de pièces par logement, l'étendue est égal à : $e = 6-1 =$ 5 pièces.

• Dans le cas d'une variable statistique continue, l'étendue représente la différence entre la valeur de la borne supérieure de la dernière classe et celle de la borne inférieure de la première classe, il vaut :

$$e = e_{k+1} - e_1$$

Dans l'exemple de distribution d'âge des personnes hospitalier dans un service d'hôpital, l'étendue vaut : $e = 90 - 10 = 80$ ans.
Remarque : *l'étendue est influencé lorsque la série statistique comporte des valeurs aberrantes*. Comme illustre l'exemple suivant :
Soient les deux séries statistiques :
Série A : 3,4,5,6,7,8,9,10.
Série B :18,19,20,21,22,23,24,25.
Les deux séries ont même étendue (e=7) et pourtant elles sont complètement différentes, c'est l'inconvénient majeur de l'étendue.

3.2.2 Ecart-type et variance

On définit en premier l'écart-type, ensuite la variance .
L'écart -type $\sigma(x)$ est la racine carrée de la moyenne des carrés des écarts par rapport

à la moyenne arithmétique $\bar{x}$.

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Lorsque les observations $(x_1, x_2, ..., x_k)$ sont pondérées par des effectifs $(n_1, x_2, ..., n_k)$, l'écart -type pondéré s'écrit :

$$\sigma(x) = \sqrt{\frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2}$$

En utilisant les fréquences, l'écart type devient :

$$\sigma(x) = \sqrt{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}$$

Remarque : *Dans le cas d'une série regroupée par classes, on utilise les $x_i, i = 1, ..., k$ qui représente les centres des classes.* **La variance**

La variance d'une variable x est le carré de l'écart type.

$$V(x) = \sigma(x)^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Lorsque la série statistique est répétitive $(x_i, n_i), i = 1...k$, la variance s'écrit :

$$V(x) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2$$

Ou bien :

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

L'écart type et la variance sont des indicateurs de dispersion de même nature puisque le premier est la racine carré de l'autre.

Calcul par la formule développée :

Afin d'éliminer les deux opérations de calculs $(x_i - \bar{x})$ et $(x_i - \bar{x})^2$, une formule simplifiée et rapide est dévelopée comme suit :

On a par définition :

$$V(x) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2$$

En développant V(x), on obtient :

$$V(x) = \frac{1}{N}\sum_{i=1}^{k} n_i(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

En faisant sortir toutes les constantes qui ne dépendent pas du signe de sommation Σ, on obtient :

$$V(x) = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - 2\bar{x}\frac{\sum_{i=1}^{k} n_i x_i}{N} + \bar{x}^2\frac{\sum_{i=1}^{k} n_i}{N}$$

Comme $\sum_{i=1}^{k} n_i = N$, on aura :

$$V(x) = \frac{\sum_{i=1}^{k} n_i x_i^2}{N} - 2\bar{x}\bar{x} + \bar{x}^2$$

D'où :

$$V(x) = (\frac{1}{N}\sum_{i=1}^{k} n_i x_i^2) - \bar{x}^2$$

ou

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

Propriétés de la variance :

• La variance est toujours positive ou nulle.

• La variance d'une constante est nulle.

• La variance est la différence entre la moyenne des carrés de la variable x et le carré de la moyenne.

• Soit $y_i = a + bx_i$, on effectue un changement d'origine et d'unité sur un variable x dont la moyenne est $\bar{y} = a + b\bar{x}$. La variance vaut : $V(y) = b^2V(x)$.Par conséquent, l'écart type : $\sigma(y) = |b|\sigma(x)$

Pour l'application de la formule développée, on considère deux types d'exemples, l'un est discret et l'autre est continu.

Exemple : cas discret

On reprend l'exemple 3 du nombre de pièces par logement :

Nombre de pièces x_i	1	2	3	4	5	6	Total
Effectif n_i	5	10	20	30	25	10	$\Sigma n_i = 100$
$n_i x_i$	5	20	60	120	125	60	$\Sigma n_i x_i = 390$
$n_i x_i^2$	5	40	180	480	625	360	$\Sigma n_i x_i^2 = 1690$

D'où :

$$V(x) = (\frac{1}{N}\sum_{i=1}^{6} n_i x_i^2) - \bar{x}^2 = \frac{1690}{100} - 3.9^2 = 1.69$$

Par conséquent, l'écart type : $\sigma(x) = \sqrt{1.69} = 1.3$

Remarque :

Si on avait utilise les fréquences f_i certaines lignes de calculs seraient devenues :

La ligne $n_i x_i$ serait remplacée par $f_i x_i$

La ligne $n_i x_i^2$ serait remplacée par $f_i x_i^2$. Donc les résultats de la moyenne arithmétique et de la variance seront donnes directement par le total des lignes respectives.

Exemple : cas continu

Considèrons l'exemple précédent :

Durée en années	[10,20[	[20,30[	[30,50[	[50,80[	[80,90[	Total
Centre x_i	15	25	40	65	85	
Effectif n_i	12	14	32	47	15	$N = 120$
Fréquence f_i	0.1	0.117	0.267	0.391	0.125	1
$f_i x_i$	1.5	2.925	10.68	25.415	10.625	$\Sigma f_i x_i = 51.145$
$f_i x_i^2$	22.5	73.125	427.2	1651.975	903.125	$\Sigma f_i x_i^2 = 3077.925$

La moyenne $\bar{x}$ =51.145 ans

La variance égale à :

$$V(x) = (\sum_{i=1}^{5} f_i x_i^2) - \bar{x}^2 = 3077.925 - 51.145^2 = 462.114 ans^2$$

L'écart type : $\sigma(x) = \sqrt{462.114} = 21.497$ ans