Chapitre 2

Étude d'une variable statistique discrète

2.1 Présentation des données

A l'issue de la collecte des données (lors d'une enquête par exemple), les informations recueillies ne sont pas immédiatement exploitables. Il est alors nécessaire de les organiser, les ordonner et les présenter de façon lisible et facilement compréhensible. Pour cela la statistique descriptive offre des techniques pour la représentation des données sous forme de tableaux ou de graphes.

2.1.1 Exemples

2.1.1.1 Exemple 1

Série statistique du nombre d'enfants à charge de 20 employés d'une entreprise : 1 ; 0 ; 1 ; 2 ; 2 ; 5 ; 4 ; 4 ; 3 ; 1 ; 0 ; 1 ; 0 ; 0 ; 0 ; 6 ; 10 ; 7 ; 1 ; 7

On peut regrouper ces donnée sous formes de tableau :

Nombres d'enfants x_i	0	1	2	3	4	5	6	7	10
Effectifs n_i	5	5	2	1	2	1	1	2	1

2.1.1.2 Exemple 2

Langue maternelle des élèves d'une classe de 15 élèves : Mooré ; Mooré ; Dioula ; Mooré ; Français ; Dafing ; Gourmatché, Foulfouldé ; Foulfouldé ; Mooré ; Dioula ; Dioula ; Mooré ; Mooré ; Mooré.

On peut regrouper ces donnée dans un tableau :

Langue maternelle	Mooré	Dioula	Français	Dafing	Gourmatché	Foulfouldé
Effectis	7	3	1	1	1	2

2.1.2 Généralisation

La façon la plus simple de présenter de façon synthétique une série statistique est un tableau présentant en face de chaque modalité le nombre d'individus de l'échantillon qui portent cette modalité.

Considérons les données d'une étude statistique portant sur une population de taille N. Nous supposons que nous avons k modalités de la variable statistique étudiée. Les donnés peuvent se présenter de la façon suivante :

Modalités x_i	x_1	x_2	$\cdots$	x_n
Effectifs n_i	n_1	n_2	$\cdots$	n_n

Exemple 2.1.1 *Une enquête réalisée dans un village porte sur le nombre d'enfants par femme. On note X le nombre d'enfants, les résultats sont données par ce tableau :*

Caractère x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*20*	*28*	*60*	*45*	*35*	*8*	*4*

Tableau 1

2.2 Effectif partiel - effectif cumulé

2.2.1 Effectif partiel

Pour chaque valeur x_i, on pose par définition

$$n_i = Card\{\omega \in \Omega : X(\omega) = xi\}.$$

n_i : le nombre d'individus qui ont le même x_i, ça s'appelle effectif partiel de x_i.

Exemple 2.2.1 *: Dans l'exemple précédent (tableau 1) l'effectif partiel de valeur $x_i = 3$ est 35.*

2.2.2 Effectif cumulé

Pour chaque valeur x_i, on pose par définition

$$N_i = n_1 + n_2 + ... + n_i$$

L'effectif cumulé N_i d'une valeur est la somme de l'effectif de cette valeur et de tous les effectifs des valeurs qui précèdent.

Exemple 2.2.2 *: Dans l'exemple précédent (tableau 1), nous complétons le tableau avec les effectifs cumulés de chaque valeur x_i*

Caractère x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*20*	*28*	*60*	*45*	*35*	*8*	*4*
Effectifs cumulés N_i	*20*	*48*	*108*	*153*	*188*	*196*	*200*

Interprétation : N_i est le nombre d'individus dont la valeur du caractère est inférieur ou égale à x_i. De ce fait, l'effectif total est donné par

$$N = \sum_{i=1}^{n} n_i$$

Dans l'exemple précédent (tableau 1) l'effectif total est $N = 200$.

2.3 Fréquence partielle - Fréquence cumulée

2.3.1 Fréquence partielle

Pour chaque valeur x_i, on pose par définition

$$f_i = \frac{n_i}{N}$$

f_i s'appelle la fréquence partielle de x_i. La fréquence d'une valeur est le rapport de l'effectif de cette valeur par l'effectif total.

Remarque 2.3.1 *: On peut remplacer f_i par $f_i \times 100$ qui représente alors un pourcentage.*

Exemple 2.3.1 *: Dans l'exemple précédent (tableau 1), nous complétons le tableau avec les fréquences de chaque valeur x_i*

Caractère x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*20*	*28*	*60*	*45*	*35*	*8*	*4*
Fréquences f_i	*0,1*	*0,14*	*0,3*	*0,225*	*0,175*	*0,04*	*0,02*

Propriété 2.3.1 *Soit f_i défini comme précédemment. Alors, on a :* $\sum_{i=1}^{n} f_i = 1$.

Preuve

$\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} \frac{n_i}{N}$ car $f_i = \frac{n_i}{N}$

$\sum_{i=1}^{n} f_i = \frac{1}{N} \sum_{i=1}^{n} n_i$. Or $N = \sum_{i=1}^{n} n_i$ donc

$\sum_{i=1}^{n} f_i = \frac{1}{N} \times N = 1$

2.3.2 Fréquence cumulée

Pour chaque valeur x_i, on pose par définition

$$F_i = f_1 + f_2 + ... + f_i.$$

La quantité F_i s'appelle la fréquence cumulée de x_i.

Interprétation : $F_i =$ est le pourcentage des individus tel la valeur du caractère est inférieure ou égale à x_i.

Exemple 2.3.2 *: Dans l'exemple précédent (tableau 1), nous complétons le tableau avec les fréquences cumulées de chaque valeur x_i*

Caractère x_i	*0*	*1*	*2*	*3*	*4*	*5*	*6*
Effectifs n_i	*20*	*28*	*60*	*45*	*35*	*8*	*4*
Fréquences f_i	*0,1*	*0,14*	*0,3*	*0,225*	*0,175*	*0,04*	*0,02*
Fréquences cumulées F_i	*0,1*	*0,24*	*0,54*	*0,765*	*0,94*	*0,98*	*1*

Nous avons vu que les tableaux sont un moyen souvent indispensable, en tous cas très utile, de classification et de présentation des unités d'une population statistique. Dans le paragraphe suivant, nous allons voir comment on traduit ses tableaux en graphique permettant aussi de résumer d'une manière visuelle les données.

2.4 Représentation graphique des séries statistiques

On distingue plusieurs méthodes de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative). Les représentations recommandées et les plus fréquentes sont les tableaux et les diagrammes (graphe).

2.4.1 Distribution à caractère qualitatif

A partir de l'observation d'une variable qualitative, deux diagrammes permettent de représenter cette variable : le diagramme en barres et le diagramme à secteurs angulaires.

Nous portons en abscisses les modalités, de façon arbitraire. Nous portons en ordonnées des rectangles dont la longueur est proportionnelle aux effectifs, ou aux fréquences, de chaque modalité.

Exemple 2.4.1 *Une étude statistique porte sur la situation matrimoniale des agents d'une entreprise. Nous avons les résultats suivants : 9 agents sont célibataires, 2 agents sont divorcé(e)s, 7 agents sont marié(e)s et 2 agents sont veuf(ve)s.*

Représentons le diagramme en barres des effectifs.

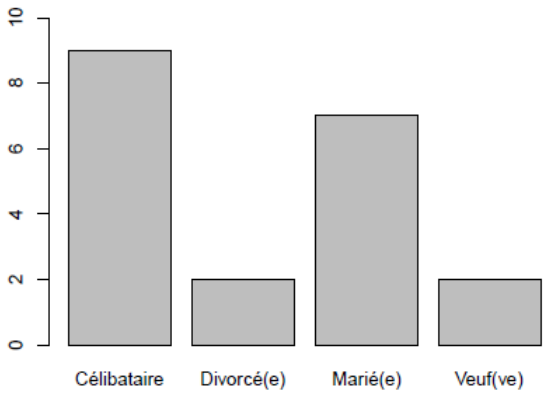


FIGURE 2.1 – Diagramme en barres

Diagramme par secteur (diagramme circulaire)

2.4.1 Distribution à caractère qualitatif

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité (voir Figure suivante).

Le degré d'un secteur est déterminé à l'aide de la règle de trois de la manière suivante :

N $\rightarrow$ 360 dégrés
$n_i \rightarrow d_i$ (degré de la modalité i).
Donc

$$d_i = \frac{360 \times n_i}{N}$$

Regroupons les résultats dans un tableau et trouvons les équivalents des effectifs en angles.

Situation matrimoniale	Célibataire	Divorcé(e)	marié(e)	veuf(ve)
Effectif n_i	9	2	7	2
Angles en dégré d_i	162	36	126	36

$$d_1 = \frac{360 \times 9}{20} = 162$$

$$d_2 = \frac{360 \times 2}{20} = 36$$

$$d_3 = \frac{360 \times 7}{20} = 126$$

$$d_4 = \frac{360 \times 2}{20} = 36$$

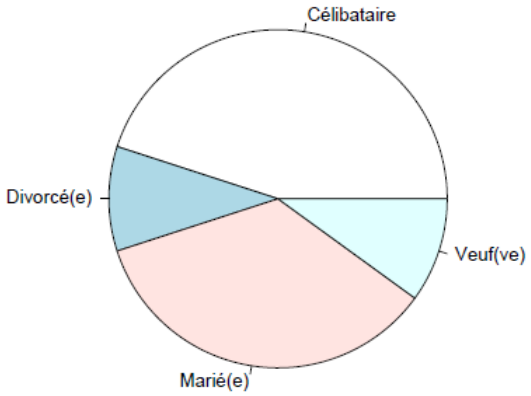


FIGURE 2.2 – Diagramme circulaire

2.4.2 Distribution à caractère quantitatif discret

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représenter cette variable : le diagramme en bâtons et le diagramme cumulatif (voir ci-dessous).

Diagramme à bâtons

On veut représenter cette répartition sous la forme d'un diagramme en bâtons. À chaque marque correspond un bâton. Les hauteurs des bâtons sont proportionnelles aux effectifs représentés

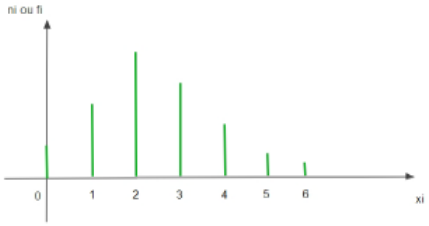


FIGURE 2.3 – Diagramme en bâtons

2.4.3 Représentation sous forme de courbe et fonction de répartition

Nous avons déjà abordé les distributions cumulées d'une variable statistique. Nous allons dans cette partie exploiter ses valeurs cumulées pour introduire la notion de la fonction de répartition. Cette notion ne concerne que les variables quantitatives.

Soit la fonction

$$F_x : \mathbb{R} \to [0, 1]$$

définie par $F_x(x) :=$ pourcentage des individus dont la valeur du caractère est $\leq x$. Cette fonction s'appelle la fonction de répartition du caractère X.

Remarque 2.4.1 *Pour tout $i \in \{1, ..., n\}$, on a*

$$F_x(x_i) = F_i.$$

La courbe de F_x passe par les points (x_1, F_1), (x_2, F_2), ... et (x_n, F_n).

En se basant sur notre exemple (tableau 1), la courbe de F_x est représentée sur

$$\mathbb{R} =]-\infty; 0[\cup [0; 1[\cup [1; 2[\cup ... \cup [5; 6[\cup [6; +\infty[.$$

Dans ce cas, nous avons

- $F_x(x) = 0 \quad$ si $x < 0$;
- $F_x(x) = 0,1 \quad$ si $0 \leq x < 1$;

- $F_x(x) = 0,24 \quad \text{si } 1 \leq x < 2\,;$
- $F_x(x) = 0,54 \quad \text{si } 2 \leq x < 3\,;$
- $F_x(x) = 0,765 \quad \text{si } 3 \leq x < 4\,;$
- $F_x(x) = 0,94 \quad \text{si } 4 \leq x < 5\,;$
- $F_x(x) = 0.98 \quad \text{si } 5 \leq x < 6\,;$
- $F_x(x) = 1 \quad \text{si } x \geq 6$

Cette courbe s'appelle "la courbe cumulative des fréquences". La courbe cumulative est une courbe en escalier représentant les fréquences cumulées relatives.

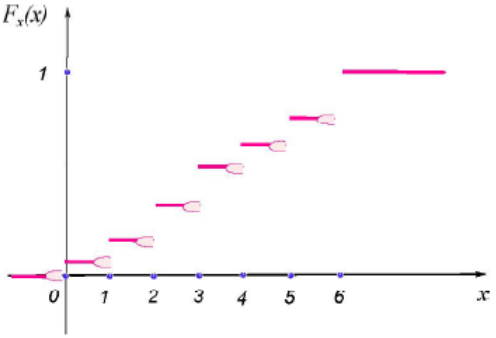


FIGURE 2.4 – Représentation d'une variable quantitative discrète par la courbe cumulative.

Propriété 2.4.1 *La fonction de répartition satisfait, pour* $i \in \{1, ..., n\}$,

- *l'égalité,* $F_x(x_i) = F_i$,
- *l'expression,*

$$F_x(x) = \begin{cases} 0, & si \quad x < x_1, \\ F_1, & si \quad x_1 \leq x < x_2, \\ F_i, & si \quad x_i \leq x < x_{i+1}, \\ 1, & si \quad x \geq x_n \end{cases} \tag{2.1}$$

2.5 Paramètres de positions

Les caractéristiques de tendance centrale sont des valeurs numériques, calculées à partir d'une série (ou d'une distribution) statistique et qui permettent de déterminer la valeur typique ou l'ordre de grandeur de la distribution. Les principales caractéristiques de tendance centrale sont : le mode, la médiane et la moyenne.

2.5.1 Le mode

Le mode d'une V.S est la valeur qui a le plus grand effectif partiel (ou la plus grande fréquence partielle) et il est dénoté par M_0.

Exemple 2.5.1 *:*

- *Dans l'exemple précédent (tableau 1) le mode est $M_0 = 2$ car la valeur 2 a l'effectif le plus élevé qui 60.*
- *Dans l'exemple précédent où le caractère était la langue maternelle, le mode est le "mooré" qui avait l'effectif le plus élevé.*

Remarque 2.5.1 *Le mode d'une série n'est pas nécessairement unique. Il peut ne pas exister.*

Exemple 2.5.2

- *la série $\{1; 7; 2; 4; 5; 3\}$ n'a pas de mode.*
- *la série $\{2; 1; 2; 2; 3; 1; 5; 4; 4; 5; 4\}$ a deux modes à savoir 2 et 4.*

2.5.2 La médiane

On appelle médiane la valeur M_e de la V.S X qui vérifie la relation suivante :

$$F_x(Me^-) < 0.5 \leq F_x(M_{e^+}) = F_x(M_e).$$

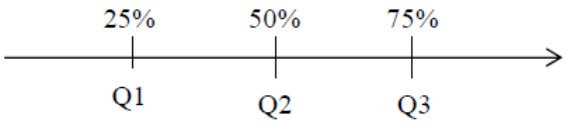
La médiane partage la série statistique en deux groupes de même effectif. C'est la valeur qui sépare une série d'observations ordonnées en ordre croissant ou décroissant, en deux parties comportant le même nombre d'observations.

2.5.3 Généralisation de la notion de médiane-Les quantiles.

2.5.3.1 Les quartiles

Ce sont les valeurs du caractère qui partagent la série en quatre sous-ensembles de tailles égales. Ils sont au nombre de 3 : Q_1, Q_2 et Q_3.

- Q_1 : 25% de valeurs inférieures et 75% de valeurs supérieures.
- Q_2 : 50% de valeurs inférieures et 50% de valeurs supérieures, Q_2 est la médiane.
- Q_3 : 75% de valeurs inférieures et 25% de valeurs supérieures.



2.5.3.2 Les quintiles

Ils divisent la série en cinq sous-ensembles de tailles égales, soit 20%. Ils sont au nombre de quatre.

2.5.3.3 Les déciles

Les déciles $D_i, i = 1, \cdots, 9$, divisent la série en dix sous-ensembles de tailles égales, soit 10%.

2.5.3.4 Les centiles

Les centiles $C_i, i = 1, \cdots, 99$, sont les valeurs du caractère qui partagent la série en 100 sous-ensembles de tailles égales, soit 1%.

2.5.4 Détermination des quantiles.

Les quantiles sont déterminés de la même manière que la médiane par méthode graphique à partir de la courbe des fréquences cumulées ou par extrapolation linéaire (voir cas de la médiane).

Les quartiles sont les valeurs dont les fréquences cumulées sont respectivement :

$$F(Q_1) = \frac{1}{4} = 25\%, \quad F(Q_2) = \frac{1}{2} = 50\%, \quad F(Q_3) = \frac{3}{4} = 75\%$$

De même

$$F(D_i) = \frac{i}{10}, \quad i = 1, \cdots, 9$$

$$F(C_i) = \frac{i}{100}, \quad i = 1, \cdots, 99$$

2.5.5 La moyenne arithmétique

La moyenne arithmétique d'un ensemble de données est la somme des valeurs obtenues divisée par le nombre d'observations. Elle est notée X pour une variable notée X.

Il existe deux façons courantes de calculer la moyenne arithmétique.

2.5.5.1 Moyenne arithmétique simple

Sa formule est

$$\overline{x}_A = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{2.2}$$

où les x_i sont les valeurs observées et N est le nombre d'observations ou la taille de la population.

Cette formule est utilisée dans le cas où les données sont présentées sous forme de série.

2.5.5.2 La moyenne arithmétique pondérée

Sa formule est

$$\overline{x}_A = \frac{\sum_{i=1}^{n} n_i x_i}{N} = \frac{1}{N}\sum_{i=1}^{n} n_i x_i \tag{2.3}$$

où les x_i sont les modalités (différentes valeurs) de la variable et n_i les effectifs de ces modalités et n le nombre de modalités de la variable.

Remarque 2.5.2 *:*

- *Cette formule est intéressante dans le cas où les données sont présentées sous forme d'un tableau de distribution des effectifs (ou des fréquences).*
- *La formule peut aussi s'écrire de la façon suivante :*
  $\overline{x}_A = \sum_{i=1}^{n} \frac{n_i}{N} x_i = \sum_{i=1}^{n} f_i x_i$ *où les* $f_i = \frac{n_i}{N}$ *sont les fréquences des modalités.*
- *La formule (2.3) diffère de la formule (2.2) par le fait que le calcul se fait dans le cas (2.3) sur les n valeurs distinctes de la variable et non sur les N individus. Les valeurs sont alors pondérées par les effectifs.*

2.5.6 Généralisation de la notion de moyenne

2.5.6.1 Moyenne géométrique

Elle est utilisée dans le cas d'une variable positive (strictement > 0). Sa formule est :

- La moyenne géométrique simple est donnée par :
$$\overline{x}_G = \sqrt[N]{x_1 \times x_2 \times \cdots \times x_N}$$
- La moyenne géométrique pondérée est donnée par :
$$\overline{x}_G = \sqrt[N]{x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_n^{n_n}}$$

Remarque 2.5.3 *La moyenne géométrique est utilisée dans le cas des variables positives présentant une évolution géométrique telle que par exemple la population. Elle permet le calcul du taux de croissance moyen, du coefficient multiplicateur moyen. Par exemple, si une variable X croît au cours de N périodes à des taux* $t_1; t_2 \cdots; t_n$*, alors le taux de croissance moyen annuel est :*

$$\overline{t} = \sqrt[n]{(1+t_1) \times (1+t_2) \times \cdots \times (1+t_n)}$$

2.5.6.2 Moyenne harmonique

- La moyenne harmonique simple est donnée par :
$$\overline{x}_H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

- La moyenne harmonique pondérée est donnée par :

$$\overline{x}_G = \frac{N}{\sum_{i=1}^{n} \frac{n_i}{x_i}}$$

Remarque 2.5.4 *:*

- *La moyenne harmonique ne peut être calculée que lorsque la série a des valeurs non nulles.*
- *Elle est utilisée pour le calcul des durées moyennes, des distances moyennes, et de certains ratios.*

2.5.6.3 Moyenne quadratique

- La moyenne quadratique simple est donnée par :

$$\overline{x}_Q = \sqrt{\frac{1}{N} \sum_{i=1}^{N} x_i^2}$$

- La moyenne quadratique pondérée est donnée par :

$$\overline{x}_G = \sqrt{\frac{1}{N} \sum_{i=1}^{n} n_i x_i^2}$$

Remarque 2.5.5 *On utilise la moyenne quadratique pour le calcul des écarts quadratiques moyens*

$$\overline{x}_G = \sqrt{\frac{1}{N} \sum_{i=1}^{n} n_i (x_i - m)^2}$$

où m est une mesure de tendance centrale. Si m est la moyenne, $\overline{x}_G$ est l'écart-type de la série.

2.5.6.4 Comparaison des moyennes

On démontre que :

$$\overline{x}_H \leq \overline{x}_G \leq \overline{x}_A \leq \overline{x}_Q$$

pour une série à valeurs positives non nulles.

2.6 Les caractéristiques de dispersion

2.6.1 L'étendue

La différence entre la plus grande valeur et la plus petite valeur du caractère, donnée par la quantité

$$x_{max} - x_{min}$$

s'appelle l'étendue de la V.S X. Le calcul de l'étendue est très simple. Il donne une première idée de la dispersion des observations. C'est un indicateur très rudimentaire et il existe des indicateurs de dispersion plus élaborés

Remarque 2.6.1 *: La signification de l'étendue est claire et sa détermination facile. Cependant, elle présente des inconvénients sérieux. En effet, ne dépendant que des valeurs extrêmes qui sont souvent exceptionnelles voire aberrantes et non pas de tous les termes, elle est sujette à des fluctuations considérables d'un échantillon à un autre.*

2.6.2 Intervalle inter-quartile

C'est la différence entre le 3^e et le 1^{er} quartile.

$$I_Q = Q_3 - Q_1$$

On définit de la même façon l'intervalle inter-décile ($I_D = D_9 - D_1$) et l'intervalle inter-centile ($I_C = C_{99} - C_1$).

Remarque 2.1 *:*

- *L'utilisation de ces intervalles permet d'éliminer l'influence des valeurs extrêmes qui sont des valeurs rares ou aberrantes.*
- *La perte de l'information du fait de la diminution des observations qu'elle entraîne est compensée par l'homogénéité des données dans l'intervalle inter-quartile.*

2.6.3 Ecart absolu moyen

C'est la moyenne des écarts absolus entre chaque observation et la moyenne.

- L'écart absolu moyen simple est donné par la formule :

$$EM = \frac{1}{N}\sum_{i=1}^{N}|x_i - \overline{x}|$$

- L'écart absolu moyen pondéré est donné par la formule :

$$EM = \frac{1}{N}\sum_{i=1}^{n} n_i|x_i - \overline{x}|$$

Remarque 2.6.2

- *On peut aussi calculer l'écart absolu moyen à partir de la médiane*

$$EM = \frac{1}{N}\sum_{i=1}^{n} n_i|x_i - M_e|$$

- *L'écart absolu moyen mesure la dispersion des valeurs observées d'une variable statistique autour d'une valeur centrale. Une valeur faible de l'écart absolu moyen traduit une faible dispersion des valeurs autour de la valeur centrale. Cependant la comparaison de cette caractéristique pour deux séries est difficile car sa valeur dépend de l'ordre de grandeur (échelle ou unité de mesure) des observations.*

2.6.4 Variance et écart-type

La variance est la moyenne des écarts (élevés au carré) des valeurs observées par rapport à la moyenne arithmétique de la série. On la note $V(X)$ pour une variable notée X.

- La variance simple est calculée à partir de la formule

$$V(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2.$$

- La variance pondérée est calculée à partir de la formule

$$V(X) = \frac{1}{N}\sum_{i=1}^{n} n_i\,(x_i - \overline{x})^2.$$

- L'écart-type est la racine carrée de la variance. On le note σ_X. Sa formule est :

$$\sigma_X = \sqrt{V(X)}$$

Remarque 2.6.3 *:*

- *Il existe une formule encore plus simple pour calculer la variance dite théorème de Koenigs*

$$V(X) = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \overline{x}^2.$$

- *Le paramètre σ_X mesure la distance moyenne entre $\overline{x}$ et les valeurs de X. Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.*
  - ⋆ *Plus il est petit, plus les caractères sont concentrés autour de la moyenne (on dit que la série est homogène).*
  - ⋆ *Plus il est grand, plus les caractères sont dispersés autour de la moyenne (on dit que la série est hétérogène).*