Chapitre I Statistiques Descriptives à une dimension

I. Introduction et Définitions

1. **Introduction.** La statistique descriptive a pour but la connaissance des phénomènes collectifs. Elle vise à recueillir et traiter des informations qui sont en général en très grand nombre. La description se fait à travers la présentation des données (la plus synthétique possible), leur représentation graphique et le calcul de résumés numériques. Ensuite à analyser et interpréter les données pour en tirer des conclusions.

Mathématiquement, une statistique est une application X d'un ensemble noté Ω vers un autre ensemble C

$$X : \Omega \rightarrow C$$
$$\omega \rightarrow X(\omega)$$

Exemple 1:
On étudie la situation familiale des enseignants du département d'informatique. Alors Ω est l'ensemble des enseignants du département d'informatique,
et C = {célibataire, marié, divorcé, veuf},
X(ω) = la situation familiale de l'enseignant ω.

Exemple 2: On s'intéresse à l'âge de chacun des 50 salariés d'une entreprise. Nous avons les données brutes suivantes :

36	30	30	56	58	47	30	45
47	18	47	33	26	51	41	33
39	36	41	51	21	33	30	18
56	24	26	41	26	37	26	33
51	56	33	24	51	37	24	37
41	41	45	33	45	33	30	37
45	39						

Alors : Ω est l'ensemble des 50 salariés de l'entreprise,
C = IR
X(ω) = l'âge du salarié ω.

2. Vocabulaire statistique

- **Population** : ensemble d'éléments assez nombreux et qui possèdent une propriété commune à étudier.
- **Individu** (ou entité statistique) c'est l'élément de la population, sur lequel on fait l'étude. Un individu peut être une personne, un animal ou un objet.
- **Echantillon** : une partie représentative de la population
  Il est généralement impossible de réunir l'information relative à tous les individus de la population.
  Parmi les raisons qui justifient un échantillonnage plutôt que de travailler sur la globalité de la population :

- les données à collecter sont illimitées.
- les ressources (humaines, financières,...) disponibles sont limitées.
- l'expérimentation peut être destructive.

- **Caractère** : c'est l'aspect particulier et commun que l'on se propose d'étudier chez les individus.
  En statistiques descriptives à une dimension, on se limite à étudier un seul caractère pour les individus.

  Un caractère peut être **qualitatif** ou **quantitatif**.

- **Caractère qualitatif** :il est non mesurable, il décrit un état. En général, il répond à la question : Comment ...
  Exemples : la situation familiale, la couleur des yeux, la citoyenneté, le sexe, la langue maternelle …
- **Caractère quantitatif** : mesurable, lorsque les données sont numériques. En général, il répond à la question : Combien . . .
  Exemples : nombre d'enfants, nombre de langues parlées, la taille, le poids, le salaire, …

  Un caractère, qu'il soit qualitatif ou quantitatif, prend différentes valeurs appelées **modalités.**

  Caractère qualitatif est dit **ordinal** si les modalités peuvent être ordonnées sinon on dit qu'il est **nominal**.

- **Variable statistique** (vs): c'est la mesure d'un caractère quantitatif.
  Une variable statistique peut être **discrète** ou **continue**.

- Une variable statistique discrète est une variable qui ne prend que des valeurs isolées.

Exemples de v. s. discrètes : nombre d'enfants, nombre de langues parlées

- Une variable statistique continue est une variable qui peut prendre n'importe quelle valeur dans un intervalle de IR.

Exemples de v. s. continues : Les mesures de longueur (taille, largeur, longueur, épaisseur, diamètre…), le temps, le poids (ou masse) et les mesures qui en dépendent (surface, volume, vitesse, densité….), le salaire en général est étudié comme variable statistique continue

Si la v s est continue, on regroupe les données dans des **classes** qui sont des intervalles deux à deux disjoints et dont la réunion englobe l'ensembles des observations. Chaque classe est considérée comme étant une seule modalité.

Soit la statistique $X : \Omega \rightarrow C$

$\omega_i \rightarrow X(\omega_i) = x_i$

On appelle fréquence partielle (ou effectif partiel) de la modalité $x_i \in X(\Omega) \subset C$, le cardinal de l'ensemble $X^{-1}(\{x_i\})$ notée n_i. C'est le nombre d'individus qui ont la même modalité x_i .
Exemple : Prenons l'exemple de situation familiale des enseignants du département d'informatique. $x_i = X(\omega_i)$ = la situation familiale de l'enseignant ω_i. Si on a 100 enseignants au département d'informatique, on obtient une série statistique de 100 valeurs.

Marié, marié, célibataire, marié, marié, marié, marié, célibataire, célibataire, célibataire, célibataire, marié, marié, veuf, marié, marié, célibataire, …..

Se contenter d'énumérer les 100 valeurs, l'information ne sera pas pratique.

Une façon commode de représenter les résultats consiste à créer une distribution statistique des fréquences. On reprend l'ensemble des 4 modalités observées (les 4 situations familiales) et pour chacune, on donne le nombre n_i d'individus (enseignants) qui ont cette situation.

x_i (modalités)	n_i
Marié	n_1
Célibataire	n_2
Divorcé	n_3
Veuf	n_4

$\sum n_i = 100$

Pour un caractère qualitatif, les modalités sont classées selon l'ordre décroissant des effectifs.

On a : $\sum n_i$ = N = effectif total

On peut établir la distribution des fréquences relatives partielles $f_i = \frac{n_i}{N}$ dans laquelle chaque fréquence relative est exprimée en proportion (comprise entre 0 et 1) ou en pourcentage (compris entre 0 et 100) de l'effectif.

$$\sum_{i=1}^{k} f_i = 1$$

Si le caractère est quantitatif, on définit l'effectif cumulé n_{ic} de la modalité x_i par

$$n_{ic} = \sum_{j=1}^{i} n_j$$

De même, on définit la fréquence relative cumulée F_i de la modalité x_i par

$$F_i = \sum_{j=1}^{i} f_j = \frac{n_{ic}}{N}$$

II Représentation d'une série statistique

II.1 Représentation dans un tableau :

- Le titre est ainsi libellé : répartition (ou distribution) de tels individus selon tel caractère. En bas du tableau on indique la source d'où proviennent les informations, on peut ajouter la date et le lieu.
- Le corps du tableau:

- pour une série statistique qualitative, il comporte 3 colonnes : on met les modalités x_i dans la 1ère , dans la seconde les fréquences n_i et dans la 3ième les fréquences relatives en pourcentages ($100 \times f_i$)
- pour une série statistique quantitative discrète (x_i , n_i)$_{i=1,\ldots,k}$, le corps du tableau est analogue à celui d'une série qualitative et on ajoute une 4ième colonne pour les fréquences cumulées et une 5ième pour les fréquences relatives cumulées.
- Pour une série statistique quantitative continue, il faut définir au préalable le nombre de classes et leur positionnement. Certaines règles sont utiles :
  Les classes ($[a_1 , a_2[$, $[a_2 , a_3[$, . . . , $[a_k , a_{k+1}[$) sont des ensembles mutuellement disjoints et leur réunion englobe l'ensemble des données.
  Le nombre de classes k ne doit être ni trop petit ni trop grand et doit dépendre du nombre de données N : $5 \leq k \leq 15$
  Le nombre moyen de données par classe = $N/k \geq 5$
  S'il est possible, pour des raisons pratiques, on prend des classes de même amplitude (longueur) e
  Dans ce cas $e = \frac{w+1}{k} = \frac{(x_{max} - x_{min})+1}{k}$ et donc $k = \frac{w+1}{e}$

On mentionne dans la première colonne les classes, les autres colonnes sont les mêmes que pour une série discrète. On peut ajouter une colonne pour les centres des classes.

Exemple: pour la série st sur l'âge des salariés de l'exemple 2,

36	30	30	56	58	47	30	45	45	39	45	33	30	37
47	18	47	33	26	51	41	33	41	41	45	33	24	37
39	36	41	51	21	33	30	18	51	56	33	24	51	37
56	24	26	41	26	37	26	33						

On peut l'étudier comme une série statistique discrète (car il n'y a que 15 modalités du caractère âge),

x_i	n_i
18	2
21	1
24	3
26	4
30	5
33	7
36	2
37	4
39	2
41	5
45	4
47	3
51	4
56	3
58	1

N= $\sum n_i$ = 50

On peut aussi regrouper les valeurs dans des classes et l'étudier comme une série statistique continue

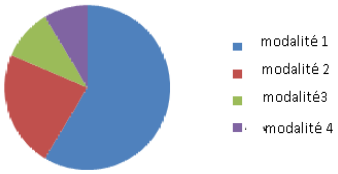
X	n_i
[18 , 25[	6
[25 , 32[	9
[32 , 39[	13
[39 , 46[	11
[46 , 53[	7
[53 , 60[	4

N= $\sum n_i$ = 50

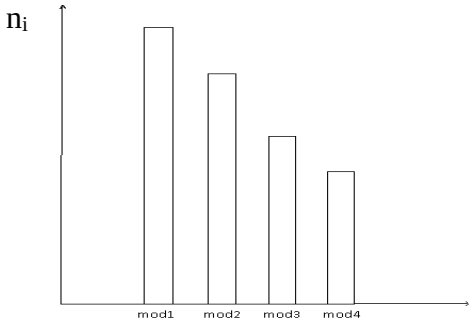
II.2 Représentation graphique

- **Représentation d'une série qualitative**

La représentation par secteurs: chaque modalité est représentée par un secteur (une portion) du disque. La surface (et donc l'angle au centre α_i) du secteur est proportionnelle à la fréquence de la modalité. $\alpha_i = 360° \times f_i$

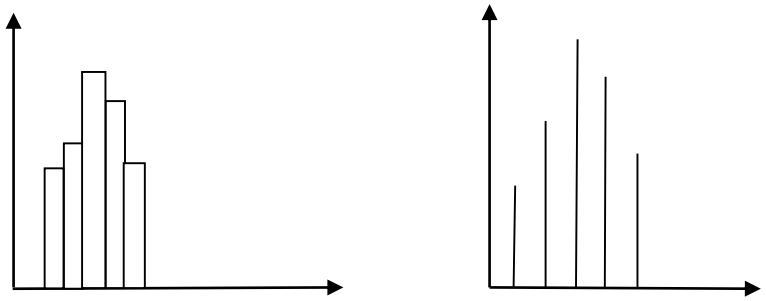


La représentation par tuyaux d'orgues: les modalités sont représentées sur un repère cartésien par des rectangles de base constante et des hauteurs proportionnelles aux fréquences.



Représentation d'une série quantitative : Il existe deux types de représentations :
Le **diagramme différentiel:** il correspond à la représentation par rapport aux fréquences partielles (ou fréquences relatives partielles).
Le **diagramme intégral:** il correspond à la représentation par rapport aux fréquences cumulées (ou fréquences relatives cumulées).

- Le diagramme différentiel d'une série discrète est un diagramme en bâtons. Sur un repère cartésien, de chaque point de coordonnées $(x_i , 0)$ est tracé un bâton de longueur proportionnelle à n_i ou f_i
- Le diagramme différentiel d'une série continue est appelé histogramme : c'est la figure obtenue en traçant de chaque base $[a_i , a_{i+1}[$ un rectangle de surface (**et non pas la hauteur**) proportionnelle à n_i ou f_i

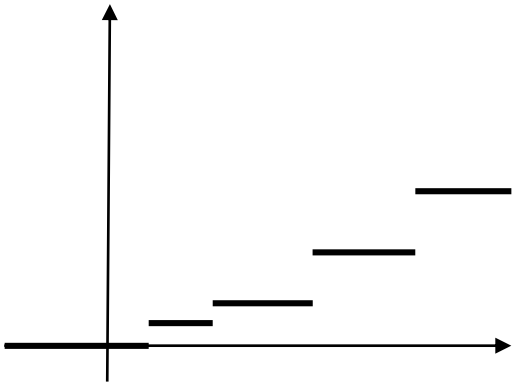


Histogramme Diagramme en bâtons

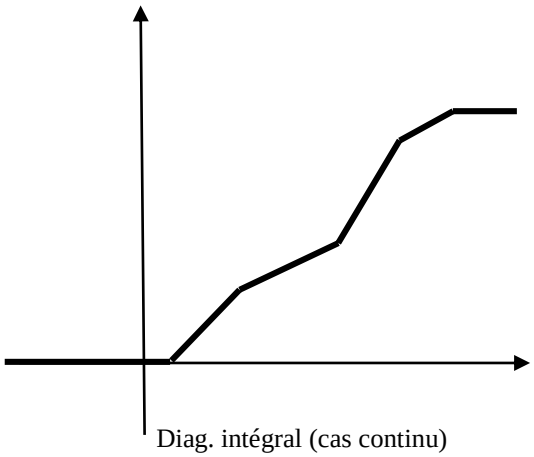
- Le diagramme intégral (ou courbe cumulative) pour une série discrète, est la représentation graphique (en escalier) de la fonction de répartition définie par :

$$F(x) = \sum_{j=1}^{i} f_j = f_1 + f_2 + \cdots + f_i \quad si \quad x_i \leq x < x_{i+1}$$

- Le diagramme intégral pour une série continue: sur un repère cartésien, on représente chaque classe $[a_i , a_{i+1}[$ par un point de coordonnées (a_{i+1} , n_{ic}). On joint les points successifs par des segments de droites pour obtenir le polygone des fréquences cumulées. On polit ensuite ce polygone pour obtenir la courbe cumulative (le diagramme intégral)



Diag. intégral (cas discret)



Diag. intégral (cas continu)

Exercice

Un fabricant de céréales fait une enquête pour savoir si ses boites de céréales contiennent effectivement 500g comme indiqué sur le contenant. Il vérifie donc un échantillon de 1000 boites sorties de l'usine en une journée. On a les données suivantes :

X	[490,496[	[496,498[	[498,500[	[500,502[	[502,504[	[504,510[
n_i	33	168	415	293	75	16

Déterminer la population étudiée, le caractère et sa nature.
Tracer les diagrammes différentiel et intégral.

1. la population : les boites de céréales produites par ce fabricant en une journée
   le caractère : le poids des boites de céréales
   la nature du caractère : quantitative continue

2. Le diagramme différentiel

   Les hauteurs $h_i = \frac{kn_i}{e_i} = \frac{2n_i}{e_i}$ (k= pgcd (e_i)=2)

X	[490,496[	[496,498[	[498,500[	[500,502[	[502,504[	[504,510[
n_i	33	168	415	293	75	16
n_{ic}	33	201	616	909	984	1000
e_i	6	2	2	2	2	6
h_i	11	168	415	293	75	16/3
x_i	493	497	499	501	503	507

III. Paramètres de tendance centrale pour une série statistique à caractère quantitatif

1. Le mode (M_o) : c'est la valeur de la vs qui a la plus grande fréquence partielle.
Si la vs est continue, on définit d'abord la classe modale. C'est la classe qui a la plus grande fréquence moyenne par unité d'intervalle.

$$\mathbf{M_o} = a_i + \frac{\Delta_1}{\Delta_1+\Delta_2} e_i$$

où

a_i : borne inférieure de la classe modale
e_i : amplitude de la classe modale
Δ_1 : fréquence moyenne de la classe modale – fréquence moyenne de la classe précédente
Δ_2 : fréquence moyenne de la classe modale - fréquence moyenne de la classe suivante

2. La médiane (M_e) : c'est la valeur de la vs qui partage en 2 parties égales les observations constituants la série préalablement rangées par ordre croissant ou décroissant

Pour une série statistique discrète $x_1, x_2, \ldots, x_N$ où N est l'effectif total

Si N est impair : $M_e = x_{\frac{N+1}{2}}$

Si N est pair : $M_e = \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1})$

Pour une série statistique continue, on détermine la classe médiane. La $i^{ème}$ classe $[a_i,a_{i+1}[$ est la classe médiane si $F_{i-1} \leq 1/2 < F_i$ (ou $n_{(i-1)c} \leq N/2 < n_{ic}$)

$$M_e = a_i + \frac{\frac{N}{2} - n_{(i-1)c}}{n_i} e_i \qquad \text{ou bien} \qquad M_e = a_i + \frac{\frac{1}{2} - F_{i-1}}{f_i} e_i$$

Exemple

Calculer le mode et la médiane pour l'exercice 1 des boites de céréales.

Classe modale [498 , 500[(classe ayant le plus grande h_i)

$$Mo = a_i + \frac{\Delta_1}{\Delta_1 + \Delta_2} e_i = 498 + \frac{415 - 168}{(415 - 168) + (415 - 293)} 2 = 499.34\,g$$

La troisième classe [498 , 500[(i=3) est la classe médiane (c'est la première classe ayant un effectif cumulé supérieur à N/2)

$$Me = a_i + \frac{N/2 - n_{(i-1)c}}{n_i} e_i = 498 + \frac{500 - 201}{415} 2 = 499.44\,g$$

3. Quartiles, quintiles, déciles et centiles

La médiane est une valeur telle que 50% des données sont plus petites qu'elle i.e. elle partage la distribution en 2 parties égales. On peut généraliser cette idée et partager la distribution des fréquences en quatre parties égales on obtient les 3 quartiles Q_1 , Q_2 et Q_3. Si on partage la distribution des fréquences en cinq parties égales on obtient les 4 quintiles q_1 , q_2 , q_3 et q_4. Si on partage la distribution des fréquences en dix parties égales on obtient

les 9 déciles d_1 , d_2 , . . .,d_9. Si on partage la distribution des fréquences en cent parties égales on obtient les 99 centiles c_1 , c_2 , . . .,c_{99}.

Le centile d'ordre α, c_α, est défini par :

- Pour une vs discrète

Si $\frac{N\alpha}{100}$ est entier alors $c_\alpha = \frac{1}{2}(x_{\frac{N\alpha}{100}} + x_{\frac{N\alpha}{100}+1})$

Si $\frac{N\alpha}{100}$ n'est pas entier, c_α est la donnée x_i dont le rang i est l'entier qui suit $\frac{N\alpha}{100}$.

- Pour une vs continue, on détermine la classe $[a_i , a_{i+1}[$ contenant c_α . C'est la 1ère classe où la fréquence cumulée dépasse $\frac{N\alpha}{100}$

$$c_\alpha = a_i + \frac{\frac{N\alpha}{100} - n_{(i-1)c}}{n_i} e_i \quad \text{ou} \quad c_\alpha = a_i + \frac{\frac{\alpha}{100} - F_{i-1}}{f_i} e_i$$

Les quartiles sont les 25ème, 50ème et 75ème centiles.

$Q_1 = c_{25}$, $Q_2 = c_{50}$ et $Q_3 = c_{75}$

- Les quintiles sont les 20ème , 40ème , 60ème et 80ème centiles .
  $q_1 = c_{20}$, $q_2 = c_{40}$, $q_3 = c_{60}$ et $q_4 = c_{80}$

- Les déciles sont les 10ème , 20ème , . . . , 90ème centiles.
  $d_1 = c_{10}$, $d_2 = c_{20}$, . . . , $d_9 = c_{90}$

Exemple

Pour l'exercice des boites de céréales, quels sont les rangs centiles des masses 495g et 505g ?

rg(495) ?

rg(495) = α ⇔ c_α = 495

495 ∈[490 , 496[

$$495 = 490 + \frac{N\alpha/100 - 0}{33} \times 6 \quad \text{d'où} \quad \alpha = 2.75 \approx 3 \quad \text{donc} \quad rg(495) = 3$$

505 ∈[504 , 510[

rg(505) ?

rg(505) = α ⇔ c_α = 505

$$505 = 504 + \frac{N\alpha/100 - 984}{16} 6 \quad \text{d'où} \quad rg(505) = \alpha = 98.66 \approx 99$$

Si on juge comme tolérable une erreur de 5g, quelle est la proportion des boites acceptables ?

P(495 ≤ X ≤ 505) = rg(505) – rg(495)] = 99 – 3 = 96 %

4. La moyenne arithmétique ($\overline{X}$)

La moyenne arithmétique est la valeur que devraient avoir toutes les données pour que leur somme totale soit inchangée.

$$\overline{X} = \sum_{i=1}^{k} f_i x_i = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$

Pour une vs continue, les x_i sont les centres des classes.

Exemple
Calculer le poids moyen d'une boite de céréale (exercice)

X	[490,496[	[496,498[	[498,500[	[500,502[	[502,504[	[504,510[	
n_i	33	168	415	293	75	16	
x_i	493	497	499	501	503	507	
$n_i x_i$	33×493	168×497	415×499	293×501	75×503	16×507	∑=499480

$$\overline{X} = \frac{1}{N}\sum n_i.\ x_i = \frac{499480}{1000} = 499{,}45\ g$$

5. **La moyenne géométrique (G) :**
La moyenne géométrique est la valeur que devraient avoir toutes les données pour que leur produit soit inchangé.

$$G = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}$$

Exemple : une quantité positive Q_0 évolue de 4 % une première année puis 7 % l'année suivante. Quel est le taux annuel moyen d'évolution t ?

Q_0 quantité initiale

Après la 1è année

$Q_0 \longrightarrow Q_1 = Q_0 + \frac{4}{100} Q_0 = (1 + \frac{4}{100}) \times Q_0 = 1{,}04 \times Q_0$

Après la 2è année

$Q_1 \longrightarrow Q_2 = Q_1 + \frac{7}{100} Q_1 = (1 + \frac{7}{100}) \times Q_1 = 1{,}07 \times Q_1 = 1{,}07 \times 1{,}04 \times Q_0$

On pose $c_1 = 1 + \frac{4}{100} = 1{,}04$ et $c_2 = 1 + \frac{7}{100} = 1{,}07$

Alors après les 2 années, la quantité est $Q_2 = c_1 c_2 Q_0 = 1{,}04 \times 1{,}07 \times Q_0$
c_1 et c_2 sont les coefficients multiplicateurs correspondants aux taux 4% et 7 % des 2 années.

Soit t le taux annuel moyen d'évolution cad si la quantité initiale Q_0 évolue de t % la première année puis aussi de t % l'année suivante alors quantité Q_2 après deux années doit être la même.
Après la première année $Q_1 = Q_0 + \frac{t}{100} Q_0 = (1 + \frac{t}{100}) Q_0$
Après la seconde année $Q_2 = Q_1 + \frac{t}{100} Q_1 = (1 + \frac{t}{100}) Q_1 = (1 + \frac{t}{100})^2 Q_0$
On doit avoir : $Q_2 = (1 + \frac{t}{100})^2 Q_0 = 1{,}04 \times 1{,}07 \times Q_0$

En posant $c = 1+\frac{t}{100}$ le coefficient multiplicateur correspondant à t , on a alors $c^2=(1+\frac{t}{100})^2 = c_1c_2$ d'où $c = \sqrt{c_1c_2} =\sqrt{1{,}04 \times 1{,}07} = 1{,}05489$ c'est la moyenne géométrique des coefficients multiplicateurs c_1 et c_2

et comme $c = 1 + \frac{t}{100}$ alors $t = (1{,}05489 - 1)\times100=5{,}489$

Exercice 2 Les variations annuelles du prix d'un certain produit (durant la période 1980-1990) sont données par :

	variations annuelles t_i (en %)	**Coefficients multiplicateurs $c_i = 1+t_i/100$**
1980	0,7	1,007
1981	2,3	1,023
1982	2,1	1,021
1983	-0,3	0,997
1984	-0,1	0,999
1985	-0,4	0,996
1986	1,1	1,011
1987	0,3	1,003
1988	1,1	1,011
1989	0,1	1,001
1990	1,0	1,010

Le coefficient multiplicateur $c = 1+\frac{t}{100}$ correspondant au taux de variation annuel moyen t est égal à la moyenne géométrique des coefficients c_i

$c=(1{,}007\times1.023\times1{,}021\times0{,}997\times0{,}999\times0{,}996\times1{,}011\times1{,}003\times1{,}011\times1{,}001\times1{,}010)^{1/11}$ =
=1,007144476

Comme $c= 1+\frac{t}{100}$ alors $t = 100\times(c - 1) = 0.71\%$

6. **La moyenne harmonique (H):**

$$H = \frac{N}{\Sigma \frac{n_i}{x_i}}$$

Exemple : si un train fait un trajet aller-retour entre 2 villes à la vitesse constante V_1 pour l'aller et la vitesse constante V_2 pour le retour. La vitesse moyenne du trajet est

$V_{moy} = \frac{2d}{t_{aller}+t_{retour}} = \frac{2d}{\frac{d}{V_1}+\frac{d}{V_2}} = \frac{2}{\frac{1}{V_1}+\frac{1}{V_2}} = H$ c'est la moyenne harmonique

Paramètres de dispersion

- **L'étendue (W)** : $W = x_{max} - x_{min}$

- **La variance V(X)** : c'est la moyenne arithmétique des carrés des écarts à la moyenne.

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

- **L'écart-type**

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

- **L'écart absolu:**

$$E_a(X) = \frac{1}{N}\sum_{i=1}^{k} n_i \left|x_i - \overline{X}\right|$$

- **Le coefficient de variation** : Le coefficient de variation est une mesure relative de la dispersion des données autour de la moyenne. Il permet de comparer le degré de variation d'un échantillon à un autre, même si les moyennes sont différentes.

  $cv = \frac{\sigma_X}{\overline{X}}$ c'est un coefficient sans unité.

  Si cv > 0.15 (ou 15%) alors la série est dispersée

- **Le coefficient de dissymétrie** :
  Lorsque la distribution est symétrique, la moyenne et la médiane sont égales.
  Lorsqu'elle est dissymétrique, la moyenne se déplace plus rapidement que la médiane dans le sens de l'étalement.

  La distance entre ces deux mesures de tendance centrale (la moyenne et la médiane), pondérée par l'écart type est appelé coefficient de dissymétrie.

  $CD = \frac{3(\overline{X} - M_e)}{\sigma_X}$

  Le signe de ce coefficient indique le type de dissymétrie (positive ou négative)
  Ce coefficient est nul lorsque la distribution est symétrique

  Si CD > 0 alors la distribution est étalée vers la droite (asymétrie à gauche).

- **L'écart interquartile :** $EIQ = Q_3 - Q_1$ c'est la longueur de l'intervalle interquartile $[Q_1, Q_3]$ qui contient 50% des données situées au centre de la distribution.

  L'écart semi-interquartile est : $ESIQ = \frac{Q_3 - Q_1}{2}$

IV. Changement de variable

Soit Y une nouvelle variable transformée de X définie par :

$Y = \frac{X-b}{a}$ où a et b sont 2 constantes et $a \neq 0$

On a alors

1. $\overline{X} = a\,\overline{Y} + b$ où $\overline{Y}$ est la moyenne de la variable transformée Y définie par $\overline{Y} = \frac{1}{N}\sum_{i=1}^{k} n_i y_i$; $y_i = \frac{x_i - b}{a}$
2. $V(X) = a^2 V(Y)$ où V(Y) est la variance de Y définie par

$$V(Y) = \frac{1}{N}\sum_{i=1}^{k} n_i (y_i - \overline{Y})^2 = \frac{1}{N}\sum_{i=1}^{k} n_i y_i^2 - \overline{Y}^2$$

Si a et b sont bien choisis alors les calculs de $\overline{Y}$ et V(Y) sont plus faciles que les calculs directs de $\overline{X}$ et V(X).

En pratique, on prendra a = pgcd (x_i) et b = le mode si la vs est discrète

Si la vs est continue, on prendra a = pgcd (e_i) et b = le centre de la classe modale

En utilisant le changement de variable $Y = \frac{X - 499}{2}$, calculer la moyenne $\overline{X}$ et la variance V(X) de l'exercice 1 des boites de céréale.

Soit le changement de variable

$$Y = \frac{X - b}{a} = \frac{X - 499}{2}$$

On a alors $y_i = \frac{x_i - b}{a} = \frac{x_i - 499}{2}$

X	[490,496[	[496,498[	[498,500[	[500,502[	[502,504[	[504,510[	
n_i	33	168	415	293	75	16	
x_i	493	497	499	501	503	507	
y_i	-3	-1	0	1	2	4	
$n_i y_i$	-99	-168	0	293	150	64	∑=240

$$\overline{Y} = \frac{1}{N}\sum n_i . y = \frac{240}{1000} = 0.24$$

X=aY+b d'où $\overline{X} = a\overline{Y} + b = 2\overline{Y} + 499$=499,48 g