Statistiques descriptives

Chapitre I :

I. Vocabulaire

Les statistiques consistent en diverses méthodes de classement des données
telles que les Tableaux, les histogrammes et les graphiques permettant
d'organiser un grand nbre de données. Elles sont dotées d'un vocabulaire
particulier

1 - Epreuves statistiques :

des statistiques descriptives visent à étudier les caractéristiques
d'un ensemble d'observation comme les mesures obtenues lors d'une expérience

L'expérience est étape préliminaire à toute statistique.

L'objet "entier" avec les observations de manière générale, la méthode
statistique est basée sur le concept suivant :

[illegible] une statistique est une épreuve qu'on
[illegible]

Exp 1 : "la durée de vie d'une lampe" donc imaginons le cas suivant :
un fabricant [illegible] veut avoir de nos jours une lampe
[illegible]

se propose d'étudier

Pour ce faire il va faire fabriquer la plupart des lampes sur
[illegible]

Puis organiser les résultats obtenus

2- Population :

En statistique, on travaille sur des populations. Ce terme vient du fait que la démographie étudie des populations humaines à occuper une place centrale au départ de la statistique notamment au travers des recensements des populations.

Nous définissons la notion de population :

Définition 2 :

On appelle population l'ensemble sur lequel porte notre étude statistique. Cet ensemble est noté Ω

Expl 2 : On considère l'ensemble des étudiants de la Section 2. On s'intéresse au nombre de filles et garçons de type étudiant dans la classe Ω = l'ensemble des étudiants

3- Individu (unité statistique)

une population est composée d'individus.

les individus qui composent une population statistique sont appelés unité statistique

Définition 3 : On appelle individu tout élément de la population Ω

Elle est notée ω (ω dans Ω)

Remarque :

L'ensemble Ω peut être un ensemble de personne, de choses ...

- L'unité statistique est un objet pour lequel nous sommes intéressés à recueillir l'information

4 - Caractères : (variable statistique)

La statistique descriptive traite non pas de l'intégralité de la population donnée. nous nous intéressons aux caractéristiques des unités qui peuvent prendre des plusieurs valeurs.

Définition 4 : On appelle caractère ou variable statistique désignée V.S toute application $X : \Omega \to C$, l'ensemble C est l'ensemble des valeurs des caractères X

C'est ce qui est recueilli pour chacun des individus

ex :
taille, couleur des yeux

Remarque : Soit Ω un ensemble, On appelle et on note Cardinal de Ω Card(Ω) : le nombre de

Card(Ω) nbre d'éléments de Ω = N

Remarque 6: L'ensemble des valeurs est représenté par des chiffres, elle même, elle est partagée en 2 sortes de caractères → discret, → continu

Expte 7: Le salaire d'employé d'une usine:

Modalités: 2000, 3000

Type: Discret

la fragilité des ressorts:

Modalités: [10, 25] N/mm

Type: continu

III - Etude d'une variable statistique discrète

Le caractère statistique peut prendre un nbre fini raisonnable de valeurs (nbre d'enfants, nbre de (cm, nbre)) dans ce cas, le caractère statistique est appelé caractère discret. dans tout le sch de chapitre, nous adoptons la notation suivante $X: \Omega \to \{x_1, x_2, \dots, x_m\}$

avec $\text{card}(\Omega) = N$ est le nbre d'individus dans notre échantillon

nous allons savoir les - ... ci dessous pour illustrer les notions de ce chapitre

Exp 8: une enquête réalisée dans un village porte sur le nbre d'enfants d'une famille. On note X: nbre d'enfants les résultats sont donnés par le Tableau

x_i	0	1	2	3	4	5	6
n_i (effectif)	8	32	66	41	32	9	2

Nous avons :

Ω : ensemble des f

w : une famille

x : nbre d'enfants par f

$X : w \longrightarrow x(w)$

On lit : à la famille w, on associe $x(w)$ = le nbre d'enfant de cette famille

II - 2 Effectif partiel - effectif cumulé

On étudie si un caractère statistique numérique représenté par une liste

Représentation (x_i) désigne les valeurs de caractère avec $i \in \{1 ; n\}$

II.1 - Effectif partiel (toujours absolu)

Déf 7 : Pour chaque valeur x_i on pose par définition :

$n_i = \text{Card}\{w \in \Omega ; X(w) = x_i\}$

Effectif cumulé croissant : pour chaque valeur x_i on appelle effectif partiel de x_i

et la somme N_i

II - 2 - Effectif cumulé

Déf 8 : Pour chaque valeur x_i on pose on note $N_i = n_1 + \dots + n_i = \sum_{k=1}^{i} n_k$

L'effectif cumulé des valeurs est la somme des effectifs d'toutes les valeurs qui précèdent

Effectif décroissant cumulé

Exemple 9:

Dans l'exemple précédent 50 est le nombre de familles ayant un nombre d'enfant inférieur à 1. Nous le regardons dans le tableaux suivant:

x_i	0	1	2	3	4	5	6
N_i	18	50	116	157	189	198	200

Interpretation: 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

$$N = \text{Card}(\Omega) = \sum_{i=1}^{n} x_i$$

Dans notre exemple précédent Nous avons $N = 200$

III - 1-3- Fréquence partielle - Fréquence cumulée:

Définition 9:

. 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 totale.

Remarque:

On peut remplacer f_i par $f_i \times 100$ qui représentent alors un pourcentage.

Interpretation:

f_i: c'est le pourcentage de w telque $X(w) = x_i$

Exemple 10:

Quel est le pourcentage de famille dont le nombre d'enfants égal à 2?

Solution : $f_i = \frac{66}{200} = 0{,}33$

Proposition :

Soit f_i défini comme précédemment Alors $\boxed{\sum_{i=1}^{n} f_i = 1}$

Dém :

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

III - 1 - 4 - Fréquence cumulée :

Définition : Pour chaque valeur x_i, on pose par définition

$$F_i = f_1 + f_2 + \dots + f_i$$

la quantité F_i s'appelle la fréquence Cumulée

Interprétation : F_i est la pourcentage de ω tel que la valeur $X(\omega)$ est inférieur ou égale à x_i

Exemple 11 :

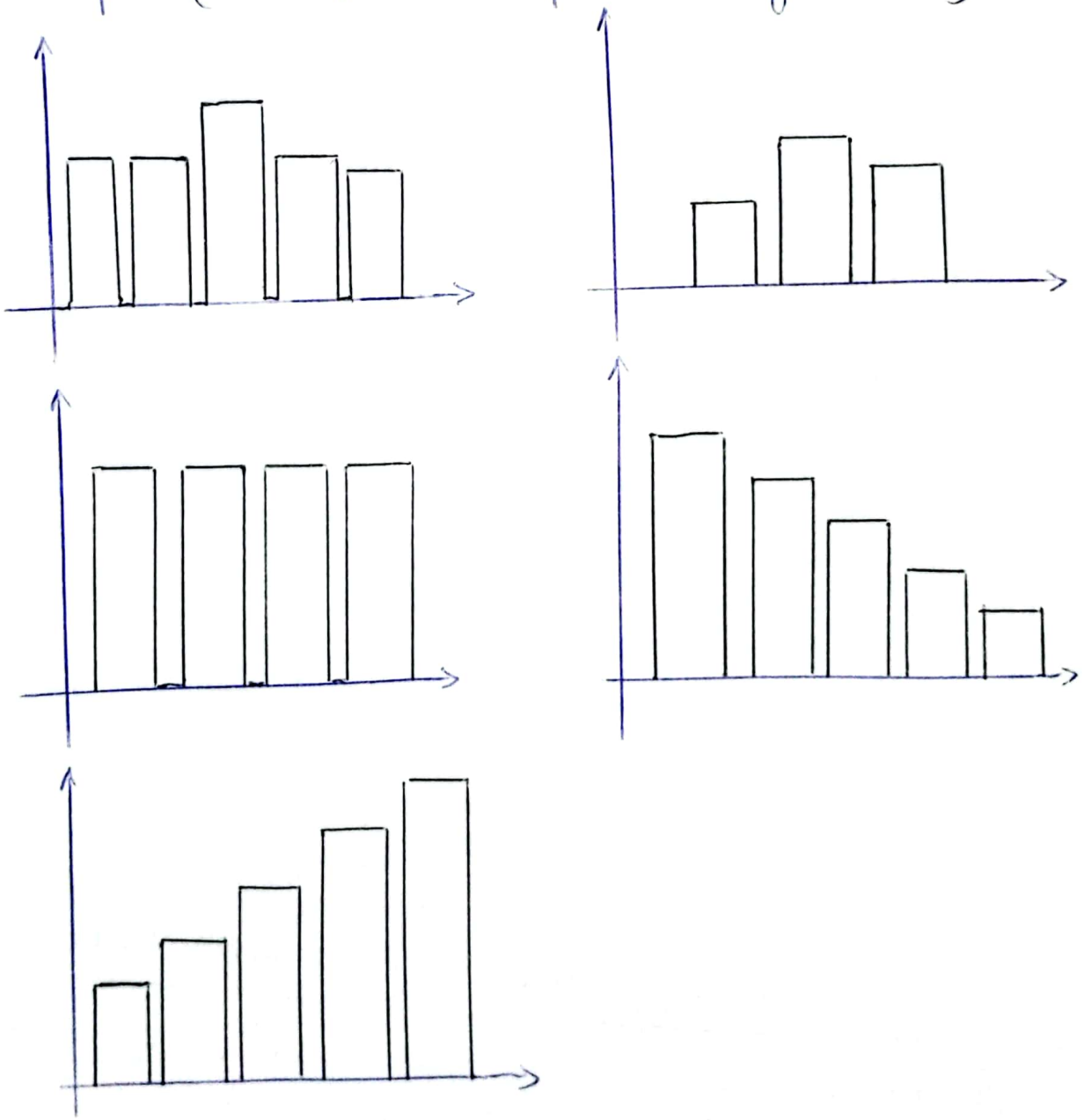
* Dans l'exemple précédent 0,785 représente 78,5% de familles dont le nombre d'enfant est inférieur ou égale à 3.
* Dans un Deuxième exemple, Nous nous intéressons aux nombres d'erreurs d'assemblage sur un ensemble d'appareils

Nombre d'erreurs	Nombre d'appareils	Fréquence cumulées
0	101	0,26
1	140	0,61
2	92	0,84
3	42	0,94
4	18	0,99
5	3	1

e_Representation graphique des séries statistiques :

distingue le méthode de représentation d'une variable statistique en fonction de la nature de cette variable (qualitative ou quantitative)
Les représentations récommandé et le plus frequent sont les tableaux et les diagrammes (Graphes).
Le graphique est un support visuel qui permet :

- **La synthèse:** visualise d'un seul le principal caractéristique (mais on perd une quantité d'information)



- La découvert

Met en évidence tendance

- Le contrôle: on apperçoit les anomalie sur un graphique dans un tableau

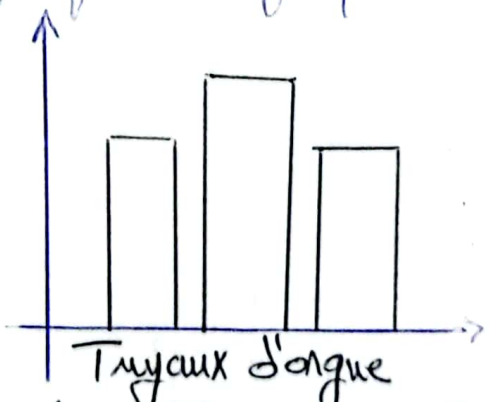
- La recherche des régularités:

Régularités dans le mouvement, répetition du phenomène

- III.2.1 Distribution à caractère qualitatif:

A partir de l'observation d'une variable qualitative, des diagrammes permettent de representent cette variable: les diagrames en bandes. (dit ~~(tuyaux d'orgue)~~ tuyaux d'orgue) et le diagrame à secteurs angulaire (dit Camembert)

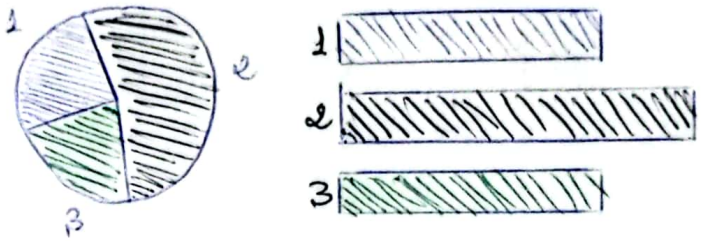
* Tuyaux d'orgue: Nous portons en abscisse les modalités, de façons arbitraire, nous portons en ordonnées ~~(les)~~ des rectangles dont la longueurs de ses effectifs, ou de frequance de chaque modalité.



* Diagramme par secteur (Diagramme Circulaires):

Les diagramme circulaires ou semi-circulaires consistent à partager un disque ou un demi-disque en tranche ou secteur

correspondant au modalité observé et dans la surface est proportionnelle à l'effectif ou à la fréquence de la modalité.



* Diagramme par secteur :

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

$$N \longrightarrow 360°$$
$$n_i \longrightarrow d_i \text{ (degré de la modalité)}$$

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

2-2 - Distribution d'un Caractère quantitatif discret :

A partir de l'observation d'une variable quantitative discrète, deux diagrammes permettent de représentant cette variable le diagramme en batons et le diagramme Cummulatif.

Pour l'illustration, nous prenons l'exemple précédent de départ (nombre d'enfants par famille) Nous rappelons le tableau statistique associé.

x_i	0	1	2	3	4	5	6
n_i	13	32	66	41	32	9	2

Diagramme à bâtons :

on veut représenter cette répartition sous la forme d'un diagramme

en bâtons. A chaque marque correspond un bâtons. Les hauteurs sont proportionnelles aux effectifs représenté.

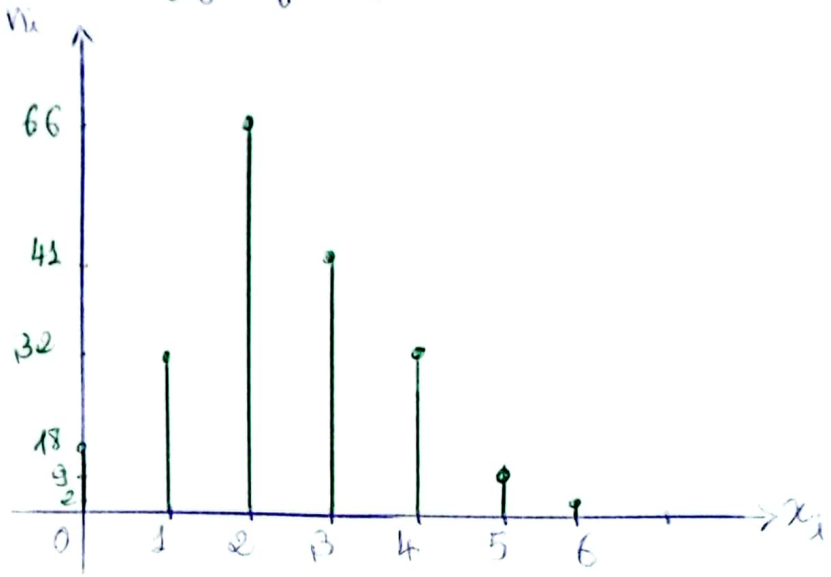


Diagramme à bâtons

2-3-ème. Représentation sous forme de Courbe et fonction de répartition:

Nous avons déjà aborder les distribution cumulée d'une variable statist
Nous allons dans cette partie exploiter ce valeurs cumulée pour intro
la notions de la fonction de répartition. Cette fonction ne concerne
que les variables quantitative.

Soit la fonction F_x. $\mathbb{R} \longrightarrow [0,1]$, définie par: $F_x(x) =$ poids

$F_x(x) =$ pourcentage des individu dont la valeur de caractere $\leq x$

Cette fonction s'appelle la fonction de repartition du caractère x

Remarque:

Pour tout $i \in \{1, \dots, n\}$ on a $F_x(x_i) = F_i$

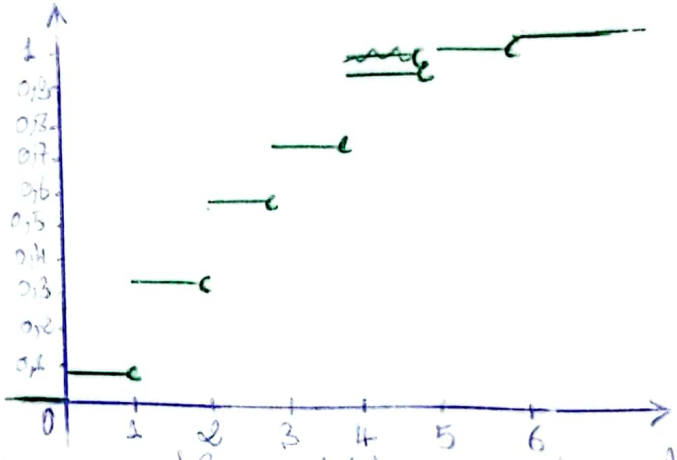
La courbe de F_x passe par les points: $(x_1, F_1), (x_2, F_2) \dots$ et (x_n, F_n)

En se basant sur notre exemple, la courbe F_x représentée ci-dessous sur

$\mathbb{R} =]-\infty, 0[\cup [0,1[\cup \dots \cup [6, +\infty[$

1) Dans ce cas nous avons:

- si $x < 0$, alors $F_n(x) = 0$
- si $x \in [0,1[$, alors $F_n(x) = 0,09$ $\left(= \frac{18}{200}\right)$
- si $x \in [1,2[$; $F_n(x) = \frac{32}{200} = 0,16 + 0,09 = 0,25$
- si $x \in [2,3[$; $F_n(x) = \frac{66}{200} = 0,33 + 0,25 = 0,58$
- si $x \geq 6$, alors $F_n(x) = F_n(x) = 0,205 + 0,58 = 0,785$
- si $x \in [3,4[$; $F_n(x) = 0,205 + 0,58 = 0,785$
- si $x \in [4,5[$; $F_n(x) = 0,16 + 0,785 = 0,945$
- si $x \in [5,6[$; $F_n(x) = 0,045 + 0,945 = 0,99$
- si $x \geq 6$; $F_n(x) = 0,01 + 0,99 = 1$



- Representation d'une variable quantitative discrète par la fct cumulative

Proposition :

La fonction de répartition satisfait pour $i \in \{1, \dots, n\}$

- l'egalité : $F_n(x_i) = F_i$
- l'expression : $F_n(x_i) = \begin{cases} 0 ; & \text{si } x < x_1 \\ F_1 ; & \text{si } x_1 \leq x < x_2 \\ F_i ; & \text{si } x_i \leq x < x_{i+1} \\ 1 ; & \text{si } x \geq x_n \end{cases}$

- 2 - 4 - Paramètre de Position : (Caractéristique de tendance Centrale)

Les indicateurs statistique de tendance centrale (dit aussi de Position)

Considéré fréquamment sont la moyenne, la médiane et le

• **Le mode:** Le mode d'une variable statistique est la valeur qui grand effectif partiel (pour la plus grande fréquence partiel) il est de M_0.

Exemple 11: Dans l'exemple précédent le mode est égale qui correspond au plus grand effectif.

Remarque: on peut avoir plus d'un mode ou rien.

• **La médiane:** on appelle médiane la valeur M_e de la variable statistique qui vérifié la relation suivant $F_x(M_e^-) < 0,5 \leq F_x(M_e^+) = F_x(M_e$

La médiane partage la série statistique en deux groupe de même effectif

Exemple 12:

* $F_x(0) = 0 < 0,5 \not\leq F_x(0^+) = 0,09$

n'est pas satisfait donc la médiane est différente de 0.

* $F_x(2^-) = 0,25 < 0,5 \leq F_x(2^+) = 0,58$

donc $M_e = 2$

• **La moyenne:** on appelle moyenne des la quantité $\bar{x} = \frac{1}{N}\sum_{i=1}^{n} n_i x$

$= \sum_{i=1}^{n} F_i x$

avec $N = Card(\Omega)$.

On peut donc exprimer et calculer la moyenne dite "Arithmétique" avec des effectif ou avec des fréquence.

Exemple 13: Si $\bar{x} = 2,46$, alors nous avons en moyenne une famille à 2,46 enfant.

• La valeur de la moyenne est abstraite. Comme dans l'exemple précédent $\bar{x} = 2,46$ est un chiffre qui ne correspond pas à un fait concrais

• La moyenne arithmétique dont on vient la formule est dite **moyenne pondérée**: c'est là signifie de chaque valeur de la variable

multiplié par le coefficient, ici par l'effectif n_i qui correspond, dans ce cas chaque valeur x_i de la variable intervient dans le calcul de la moyenne, autant de fois qu'elle a été observé. On parle de la moyenne arithmétique simple quand on effectue par des comparaisons. Par exemple si 5 étudiants ont pour âge respective 18, 19, 20, 21 et 22 ans, leur âge moyenne est donné par :

$$\frac{18+19+20+21+22}{5} = 20 \text{ ans (moyenne simple)}.$$

Remarque : nous mentionnons qu'il existe d'autre moyenne que la moyenne arithmétique.

- 2 - 5 - Paramètre de dispersion : (variabilité)

- Les indicateurs statistique de dispersion usuel sont : l'étendu la variance et l'écart-type.
- La différence entre la plus grande valeur et la plus petite valeur du caractère donner par la quantité $e = x_{max} - x_{min}$ s'appelle l'étendu (la variable) de la variable statistique X.

Le calcul de l'étendu est très simple, il donne une première idée de la dispersion des observation.

* **La variance :** on appelle variance de cette série statistique X, le nombre

$$Var(X) = \sum_{i=1}^{n} f_i (\bar{x} - x_i)^2$$

on dit que la variance est la moyenne des carré des écart à la moyenne de $\bar{X}$.

Le théorème suivant (Théorème de König-Huygens) donne une identité remarquable reliant la variance et la moyenne, parfois plus pratique dans le calcul de la variance

Théorème:

Soit (x_i, n_i) une série statistique de moyenne $\bar{x}$ et de variance $Var(X)$

Alors $$\boxed{Var(X) = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2}$$

Dém:

$$Var(X) = \sum_{i=1}^{n} f_i (\bar{x} - x_i)^2 = \sum_{i=1}^{n} f_i (\bar{x}^2 - 2\bar{x}x_i + x_i^2)$$

$$= \sum_{i=1}^{n} f_i \bar{x}^2 - 2\bar{x} \sum_{i=1}^{n} x_i f_i + \sum_{i=1}^{n} f_i x_i^2$$

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

$\sum f_i x_i = \bar{x}$

$$= \bar{x}^2 - 2\bar{x}^2 + \sum_{i=1}^{n} f_i x_i^2 = \sum_{i=1}^{n} f_i x_i^2 - \bar{x}^2$$

Remarque:

(1) Dans l'utilisation de la formule du théorème précédent (1) il faut veiller à remplacer $\bar{x}$ par sa valeur approché la plus précis aussi

- **L'écart-type:**

La quantité $\sigma_X = \sqrt{Var(X)}$ s'appelle l'écart type de V.S. X ou de la série statistique

Remarque:

Les paramètres σ_X mesure **la distance moyenne** entre $\bar{x}$ et le valeur de X. Il sert à mesurer la dispersion d'une série statistique autour de sa moyenne.

- Plus il est petit, plus ses caractère est concentrique autour de la moyenne (on dit que la série est **homogène**).
- Plus il est grand, plus ses caractère sont dispersé autour de la moyenne (on dit que la série est **hétérogène**).