

Statistiques

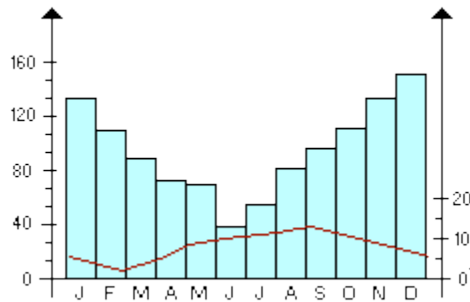
Statistiques descriptives univariées

<http://www.i2m.univ-amu.fr/perso/jean-philippe.preaux/>

1. INTRODUCTION

Qu'est-ce que la STATISTIQUE DESCRIPTIVE ?

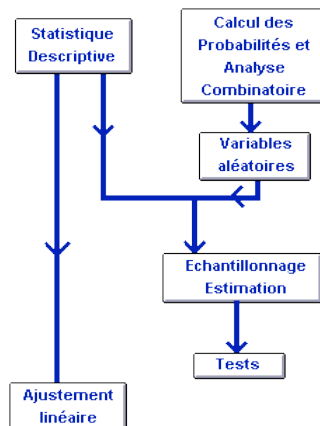
On rencontre, dans des domaines très divers, ce que l'on appelle des statistiques. Ainsi, afin de décrire le climat en un endroit du globe, il est habituel de construire des schémas et de calculer des moyennes annuelles ou des totaux.



Ci-dessus, ce sont les mesures de précipitations (en mm) et de température à Brest, sur une année. Le total annuel des précipitations est de 1126 mm, la température moyenne annuelle de 10.8°C. Ce schéma, ainsi que les moyennes, ont été établis à partir de nombreuses mesures, des données statistiques provenant de relevés de précipitations et de températures sur un grand nombre d'années.

Des données statistiques, sont un ensemble de mesures ou d'observations concernant l'état ou l'évolution d'un phénomène.

Ce que l'on appelle la Statistique Descriptive est l'ensemble des méthodes et techniques mathématiques permettant de présenter, décrire, résumer de telles données.



L'interprétation des résumés obtenus, leur extrapolation éventuelle à un ensemble plus vaste (utilisation de sondages par exemple), et leur utilisation pour prendre des décisions constituent un autre domaine de la Statistique. C'est la Statistique Inférentielle, qui s'appuie sur l'utilisation des probabilités.

L'étude des liens existant entre deux séries d'observations est l'"Ajustement linéaire".

2. VOCABULAIRE DE LA STATISTIQUE

- Un ensemble d'objets ou de personnes d'une étude statistique est appelé population.
- Un élément de cette population est appelé individu.
- L'étude statistique porte sur un caractère.
- Si le caractère est quantitatif, les mesures sont alors les valeurs d'une variable statistique (ex : un âge, une taille...).
- Si le caractère est qualitatif, on est obligé de le quantifier (ex : sexe...).
- La variable est dite discrète si elle ne prend que des valeurs isolées (ex : entières).
- Elle est continue si elle peut prendre toutes la valeurs d'un intervalle (ex : R).
- L'effectif d'une population est le nombre d'individus total de cette population.
- La fréquence d'un caractère est le nombre d'individus possédant ce caractère divisé par l'effectif total de la population.

On présente souvent les données sous forme d'un tableau 'individus $\times$ variables'.

Nom	Sexe	Age	Sit. Maritale	Nbre d'enfant	Département	Salaire/an (€)
Albert	M	35	Marié	3	75	30 000
Sidonie	F	24	Célibataire	0	80	35 000
Kamel	M	28	Célibataire	0	35	28 000
François	M	42	Marié	2	13	18 000
Hélène	F	33	Mariée	1	13	50 000
Magali	F	30	Célibataire	0	35	11 000

- Les variables quantitatives ou numériques, s'expriment par des nombres réels sur lesquels les opérations arithmétiques courantes (somme, moyenne ...) ont un sens.

Exemple : Dans le tableau de données ci-dessus, les variables âge, salaire, nombre d'enfants sont quantitatives.

- Les variables qualitatives prennent des valeurs (ou modalités) sous forme d'expressions littérales ou de nombres sur lesquels les opérations arithmétiques courantes n'ont aucun sens.

Exemple : les variables sexe, situation familiale, département de naissance sont des variables qualitatives.

Chaque individu appartient à une modalité ou catégorie d'un ensemble fini. On pourrait noter les modalités de la variable sexe respectivement 0 et 1 au lieu de M et F. Dans ce cas la variable resterait qualitative.

La nature (quantitative ou qualitative) d'une variable est une propriété intrinsèque, qui ne dépend pas de la façon dont on note ses modalités ou valeurs.

Il existe deux sortes de variables quantitatives : les variables discrètes et continues.

- Une variable quantitative est discrète si elle ne peut prendre que des valeurs isolées, généralement entières. Elle est continue si ses valeurs peuvent être n'importe lesquelles d'un intervalle réel.

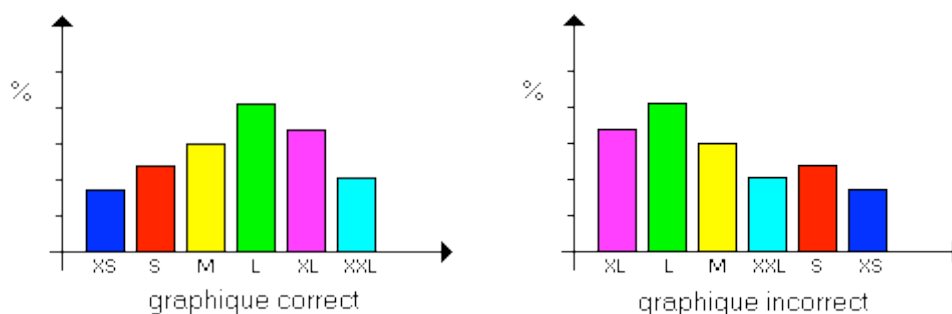
La nature, discrète ou continue, de la variable vient du phénomène lui-même et non de la façon de le mesurer. Par exemple, les tailles mesurées en cm ne seraient que des valeurs entières : 160 cm, 161 cm, 162 cm, etc... Mais la taille est un phénomène continu ; "160 cm" signifie en fait : dans l'intervalle $[159.5 \ 160.5[$. Cette "discrétisation" du phénomène continu n'est liée qu'à la précision de l'appareil de mesure.

On peut de même différencier plusieurs sortes de variables qualitatives.

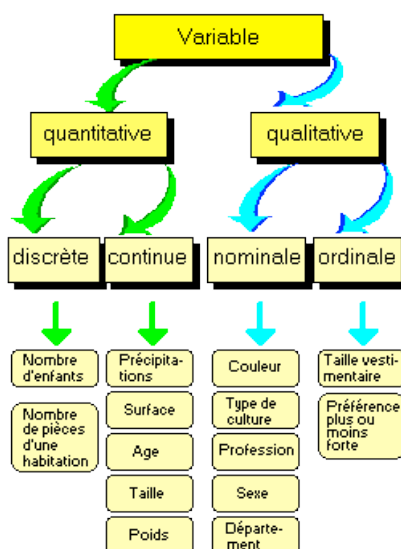
- Une variable est ordinaire si l'ensemble des catégories est munie d'un ordre total. Sinon elle est nominale.

Exemple : taille de vêtement : $XS < S < M < L < XL < XXL$

Il faudra, notamment pour les représentations graphiques, présenter toujours les modalités d'une variable ordinaire dans l'ordre.



En résumé :



3. REPRÉSENTATIONS DES DONNÉES STATISTIQUES

3.1. Type de données statistiques.

Les données statistiques peuvent être de 3 types :

3.1.1. *Une série statistique.* Une fois fixé un ordre sur les individus, ou numérotation, on représente un caractère comme une suite de valeurs :

$$(v_1, v_2, \dots, v_n)$$

Pour une population de n individus, la valeur v_i est la valeur du caractère mesurée pour l'individu numéroté i .

C'est sous cette forme brute que sont le plus souvent représentées les données après leur acquisition.

3.1.2. *Distribution observée.* On regroupe toutes les individus ayant même valeur du caractère (par exemple toutes les familles ayant 0 enfant, 1 enfant, 2 enfants). On obtient ainsi un effectif n_i correspondant à un groupe x_i de même valeur v_i .

Exemple :

Nombre d'enfants : x_i	Nombre de familles : n_i
0	6
1	4
2	5
3	2
4	1
Total :	$n = 18$

$$\{(x_i, n_i) \mid i = 1, \dots, k\}$$

est une distribution observée de la variable x ; n_i est le nombre d'individus pour lesquels le caractère prend la valeur x_i . C'est l'effectif de la valeur x_i .

L'effectif de la population est alors égal à $n = \sum_{i=1}^k n_i$.

La fréquence de la valeur x_i est alors $\frac{n_i}{n}$.

3.1.3. *Distribution groupée.* C'est un peu la même chose qu'une distribution observée à la différence près que le groupe devient une classe, c'est un intervalle. On parle donc de distribution groupée.

Exemple :

Classes de taille (en cm)	Effectifs
$[159; 165[$	6
$[165; 168[$	8
$[168; 171[$	20
$[171; 174[$	20
$[174; 177[$	5
$[177; 179[$	1

On regroupe par taille les personnes observées. L'ensemble des couples formés, d'une part, des classes construites et, d'autre part, des effectifs de classe associés (qui représentent les nombres d'observations appartenant à ces classes) est qualifié de distribution groupée.

$$\left\{ ([e_i; e_{i+1}[, n_i) \mid i = 1, \dots, k \right\}$$

est une distribution groupée de la variable x ; n_i est le nombre d'individus pour lesquels le caractère prend une valeur dans l'intervalle $[e_i; e_{i+1}[$ que l'on appelle une classe. L'entier n_i est l'effectif de la classe $[e_i; e_{i+1}[$.

L'effectif de la population est alors égal à $n = \sum_{i=1}^k n_i$.

La fréquence de la classe $[e_i; e_{i+1}[$ est alors $\frac{n_i}{n}$.

Le choix des classes est parfois délicat. A priori tout choix tel qu'on ait des classes contigües recouvrant l'ensemble des valeurs est correct : chaque individu appartient à une classe et une seule ; mais il faut dans la mesure du possible des classes d'amplitudes égales (intervalles de même longueur), et en nombre suffisant pour ne pas schématiser trop grossièrement les données.

3.2. Effectifs cumulés ; fréquences cumulées.

Étant donnée une distribution observée ou groupée :

$$\left\{ (x_i, n_i) \mid i = 1, \dots, k \right\} \quad \text{ou} \quad \left\{ ([e_i; e_{i+1}[, n_i) \mid i = 1, \dots, k \right\}$$

Les effectifs cumulés croissants sont pour tout $j = 1, \dots, k$:

$$N_j = n_1 + n_2 + \dots + n_j = \sum_{i=1}^j n_i$$

N_j est le nombre d'individus dont la valeur du caractère est inférieure ou égale à x_j (respectivement strictement inférieure à e_{j+1}).

On définit aussi les fréquences cumulées croissantes :

Avec les mêmes notations, en notant pour tout $i = 1, 2, \dots, k$:

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

la fréquence de la valeur x_i (respectivement de la classe $[e_i; e_{i+1}[$, les fréquences cumulées croissantes sont pour tout $j = 1, \dots, k$:

$$F_j = f_1 + f_2 + \dots + f_j = \sum_{i=1}^j f_i$$

F_j est la proportion d'individus dont la valeur du caractère est inférieure ou égale à x_j (respectivement strictement inférieure à e_{j+1}).

De manière analogue on définit aussi les effectifs et fréquences, cumulé(e)s décroissant(e)s :

Avec les mêmes notations :

Les effectifs cumulés décroissants sont pour tout $j = 1, \dots, k$:

$$N'_j = n_j + n_{j+1} + \dots + n_k = \sum_{i=j}^k n_i$$

N'_j est le nombre d'individus dont la valeur du caractère est supérieure ou égale à x_j (respectivement supérieur ou égale à e_j).

Avec les mêmes notations, les fréquences cumulées décroissantes sont pour tout $j = 1, \dots, k$:

$$F'_j = f_j + f_{j+1} + \dots + f_k = \sum_{i=j}^k f_i$$

F'_j est la proportion d'individus dont la valeur du caractère est supérieure ou égale à x_j (respectivement supérieure ou égale à e_j).

Exemple :

Nombre d'enfants : x_i	Nombre de familles : n_i	N_i	N'_i	F_i	F'_i
0	6	6	18	6/18	1
1	4	10	12	10/18	12/18
2	5	15	8	15/18	8/18
3	2	17	3	17/18	3/18
4	1	18	1	1	1/18
Total :	$n = 18$				

3.3. Les graphiques.

Pour visualiser une distribution statistique, il est généralement plus parlant d'utiliser un graphique, à la place ou en complément du tableau.

3.3.1. Variable Qualitative.

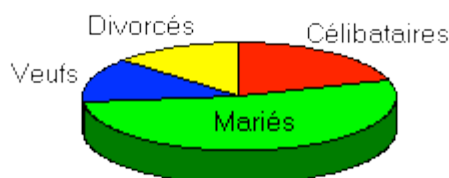
Dans le cas d'une variable qualitative, les modalités ne peuvent pas être représentées sur un axe, selon une échelle donnée, car elles ne sont pas numériques.

On utilise surtout dans ce cas des diagrammes circulaires (ou "camemberts") : chaque modalité est représentée par un secteur circulaire dont l'angle (et donc la surface) est proportionnel à son effectif. Le rayon du cercle est arbitraire.

Une représentation équivalente consiste à construire un diagramme en barres (ou "histogramme") : chaque rectangle a une base constante et une hauteur proportionnelle à l'effectif n_i ou à la fréquence f_i .

Exemple.

Modalité	Effectifs
Marié	80
Célibataire	30
Veuf	20
Divorcé	20
Total :	150



3.3.2. Variables quantitatives.

Dans le cas d'une variable quantitative, les valeurs observées, numériques, seront toujours placées selon un axe, en suivant une échelle précise.

- Le cas des séries chronologiques est particulier : l'ordre des individus étant primordial, on n'effectue pas de tri à plat, et on représente directement les données

brutes en ordonnée, l'échelle du temps étant placée en abscisse. Le temps étant continu, on relie par des segments de droite les points obtenus.

Si un phénomène saisonnier apparaît (même type de variations d'année en année par exemple), il est possible de superposer plusieurs graphiques, ou de les remplacer par des moyennes.

- Pour une variable discrète, après un tri à plat conduisant à la distribution observée, on représente celle-ci par un diagramme en bâtons les x_i sont placés suivant une échelle sur l'axe des abscisses, et les effectifs n_i sont matérialisés par un "bâton" de longueur n_i (axe des ordonnées).
- Pour une variable continue, après choix des classes et un tri à plat conduisant à la distribution groupée, on représente celle-ci par un diagramme en barres les e_i sont placés suivant une échelle sur l'axe des abscisses, et les effectifs n_i sont matérialisés par une "barres" de hauteur n_i (axe des ordonnées) et de largeur l'intervalle $[e_i, e_{i+1}[$.

4. LES PARAMÈTRES

Il existe plusieurs types de paramètre pour caractériser une série statistique :

- (1) paramètres de tendance centrale : moyenne, médiane, mode
- (2) paramètres de dispersion : étendue, variance, écart-type

4.1. Paramètres de tendance centrale.

4.1.1. La moyenne.

La moyenne arithmétique d'une série brute numérique $X = (x_1, x_2, \dots, x_n)$ est le quotient de la somme des observations par leur nombre :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

La moyenne arithmétique d'une série X s'obtient à partir d'une distribution observée $\{(x_i, n_i) \mid i = 1, \dots, k\}$ par la formule :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \sum_{i=1}^k f_i x_i.$$

La moyenne s'exprime toujours dans la même unité que les observations x_i . Elle peut être décimale, même si les x_i sont entiers par nature.

4.1.2. La médiane.

Pour une série statistique :

Si la série brute des valeurs observées est triée par ordre croissant (série ordonnée) :

$$x(1) \leq x(2) \leq \dots \leq x(n)$$

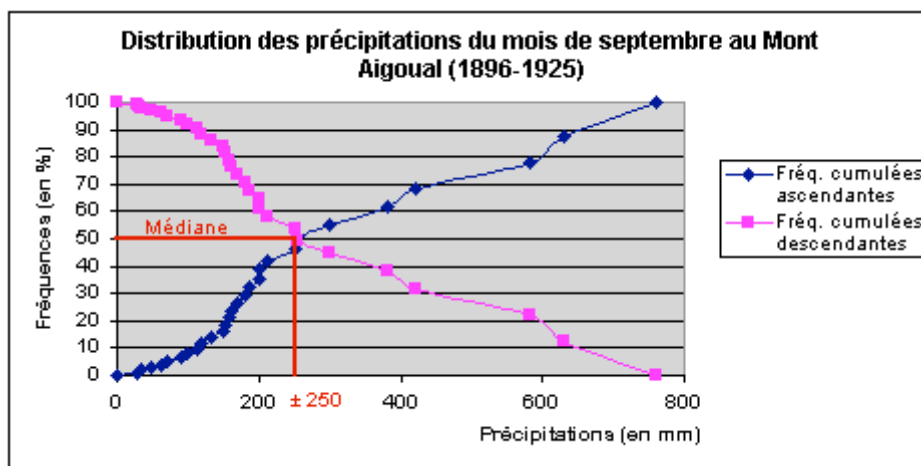
la médiane M est la valeur du milieu, telle qu'il y ait autant d'observations "au-dessous" que "au-dessus".

C'est-à-dire que si n est impair, soit $n = 2p + 1$, $M = x(p + 1)$.

Si n est pair, soit $n = 2p$, toute valeur de l'intervalle médian $[x(p); x(p + 1)]$ répond à la question.

Ce paramètre ne peut être défini que si les valeurs de la variable sont mesurées sur une échelle ordinale ou quantitative (mais pas nominale).

Pour une distribution observée : la médiane peut être trouvée graphiquement à l'aide des fréquences cumulées croissantes et décroissantes :



1

4.1.3. Les quartiles et déciles.

Les p -quantiles sont $(p-1)$ valeurs qui découpent une série statistique rangée dans l'ordre croissant en p intervalles contenant chacun le même nombre de valeurs : Effectif total/ p , à la décimale près lorsque l'Effectif n'est pas un multiple de p .

- Lorsque $p = 2$ on retrouve la médiane.
- Lorsque $p = 4$ on parle de quartiles. Dans une série statistique d'effectif au moins 3, il y a 3 quartiles :
 - Le premier quartile sépare les 25% des valeurs les plus faibles des 75% des valeurs les plus hautes.
 - Le deuxième quartile est la médiane : il sépare les 50% des valeurs les plus faibles aux 50% des valeurs les plus hautes.
 - Le troisième quartile sépare les 75% des valeurs les plus faibles des 25% des valeurs les plus hautes.

Exemple : série ordonnée 1, 3, 6, 7, 9, 10, 11, 16, 20.

Effectif = 9.

Calcul du premier quantile : On divise l'effectif par 4 : $9/4 = 2,25$.

Lorsque le résultat n'est pas entier on prend l'arrondi entier par excès.

Ici 3 : le premier quartile est le 3ème élément : 6.

Calcul du 2ème quantile : il s'agit de la médiane.

Ici le 2ème quartile est la médiane 9.

Calcul du 3ème quartile : On procède comme pour le calcul du 1er quartile en multipliant l'effectif par $3/4$ puis en arrondissant par excès à une valeur entière : $9 \times 3/4 = 27/4 \approx 7$.

Ici le 3ème quartile est le 7ème élément 11.

$$1 \leq 3 \leq (\text{1er quartile}) \quad 6 \leq 7 \leq (\text{médiane}) \quad 9 \leq 10 \leq (\text{3ème quartile}) \quad 11 \leq 16 \leq 20$$

1. (Sources : Météo France, in SAINT-JULIEN 1999, p.17).

4.1.4. Mode.

Pour une distribution observée :

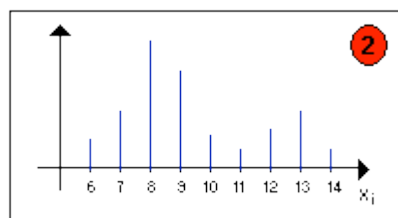
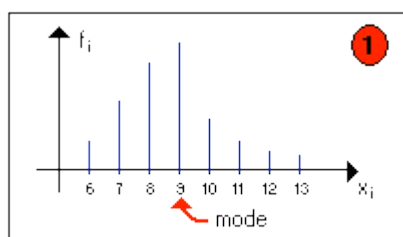
Une distribution est unimodale si elle présente un maximum marqué, et pas d'autres maxima relatifs (sur le diagramme en bâtons ou l'histogramme).

Dans ce cas, l'abscisse du maximum, correspondant à la valeur la plus fréquente de la série, est appelée le mode.

Pour une distribution groupée :

Il n'est pas toujours bien défini dans le cas d'une variable continue, le découpage en classes peut être trompeur ; la classe modale est la classe pour laquelle l'histogramme passe par un maximum.

Exemple : la distribution observée de gauche est unimodale ; le mode est 9. Celle de droite n'est pas unimodale.



Si la distribution présente 2 ou plus maxima relatifs, on dit qu'elle est bimodale ou plurimodale. Cela signifie que la population est hétérogène du point de vue de la variable observée. La population est composée de plusieurs sous-populations ayant des caractéristiques de tendance centrale différentes.

4.2. Paramètres de dispersion.

Deux distributions peuvent, tout en ayant des caractéristiques de tendance centrale voisines, être très différentes.

Ainsi la moyenne annuelle des températures dans une zone océanique peut être égale à la moyenne annuelle des températures dans une zone continentale, pourtant les distributions sont très différentes.

Dans le premier cas les variations de température autour de la moyenne sont assez faibles, dans le second cas elles sont beaucoup plus importantes.

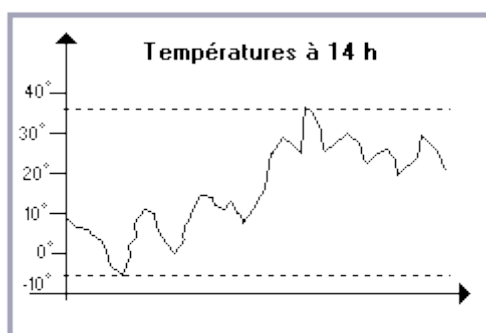
Il est donc nécessaire de mesurer la dispersion des valeurs autour des tendances centrales.

4.2.1. L'étendue.

L'étendue (ou amplitude) d'une série statistique est la différence entre la valeur maximum et la valeur minimum de la série.

$$R = \max(X) - \min(X).$$

Exemple :



$$X_{\max} = 36^{\circ} \text{C}$$

$$X_{\min} = -4^{\circ} \text{C}$$

Ici l'étendue est $R = 40^{\circ}\text{C}$.

Facile à déterminer, l'étendue ne dépend que des 2 observations extrêmes qui sont parfois le fait de situations exceptionnelles. Il est donc difficile de considérer l'étendue comme une mesure stable de la dispersion.

Ce paramètre se calcule aisément dans le cas d'une distribution observée, mais ne peut être connue avec exactitude si on ne dispose que d'une distribution groupée.

4.2.2. La variance.

La variance d'une série statistique $X = (x_1, x_2, \dots, x_n)$ est la moyenne des carrés des écarts à la moyenne.

$$V(X) = \frac{1}{n} \sum_{k=1}^n (x_i - \bar{X})^2$$

Remarques.

La variance étant une somme de carrés :

- Elle est toujours positive.
- Elle est nulle si et seulement si la série ne prend qu'une valeur constante.

La variance mesure la dispersion des valeurs de X autour de sa valeur moyenne.

Elle s'exprime dans le carré de l'unité de mesure de X .

Pour une distribution groupée :

$$V(X) = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{X})^2 = \sum_{i=1}^k f_i (x_i - \bar{X})^2$$

Il est plus commode de calculer la variance à l'aide de la formule de Koenig-Huygens :

Formule de Koenig-Huygens.

$$V(X) = \left(\frac{1}{n} \sum_{i=1}^k x_i^2 \right) - \bar{X}^2$$

En considérant la série statistique $X^2 = (x_1^2, x_2^2, \dots, x_n^2)$:

$$V(X) = \bar{X^2} - \bar{X}^2$$

Démonstration.

$$\begin{aligned}
 V(X) &= \frac{1}{n} \sum_{k=1}^n (x_i - \bar{X})^2 \\
 &= \frac{1}{n} \sum_{k=1}^n (x_i^2 - 2x_i\bar{X} + \bar{X}^2) \\
 &= \frac{1}{n} \sum_{i=1}^k x_i^2 - 2\bar{X} \frac{1}{n} \sum_{i=1}^k x_i + \frac{1}{n} \sum_{k=1}^n \bar{X}^2 \\
 &= \left(\frac{1}{n} \sum_{i=1}^k x_i^2 \right) - 2\bar{X}^2 + \bar{X}^2 \\
 &= \left(\frac{1}{n} \sum_{i=1}^k x_i^2 \right) - \bar{X}^2 \\
 &= \overline{X^2} - \bar{X}^2
 \end{aligned}$$

■

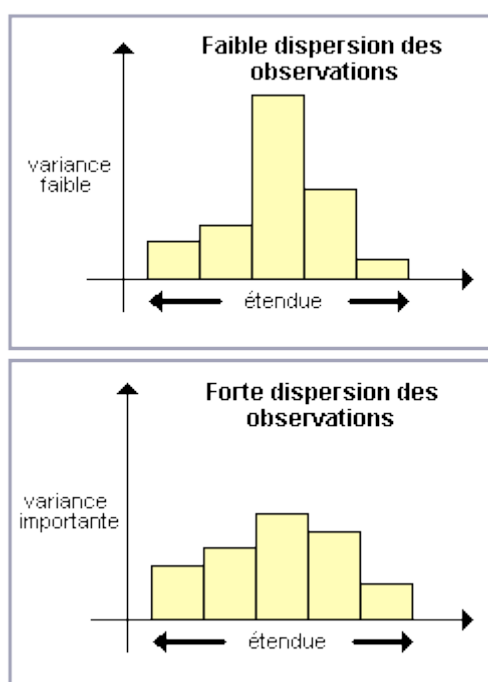
4.2.3. L'écart-type.

C'est la racine carrée de la variance ; en particulier pour une série statistique :

L'écart-type d'une série statistique $X = (x_1, x_2, \dots, x_n)$ est :

$$\sigma_X = \sqrt{\frac{1}{n} \sum_{k=1}^n (x_i - \bar{X})^2}$$

Elle s'exprime, contrairement à la variance, dans la même unité que les x_i .



Une variance (et donc un écart-type) est d'autant plus faible que les données sont groupées autour de $\bar{X}$, car en moyenne les écarts sont plus faibles. variance et écart-type sont nuls si et seulement si les données sont identiques. On calcule généralement l'écart-type en complément de la moyenne.

La connaissance de la moyenne et de l'écart-type permet d'appliquer :

Inégalité de Bienaymé-Tchébichev.

Pour toute population de moyenne $\bar{X}$ et d'écart-type σ_X , la proportion d'individus appartenant à des intervalles de la forme : $]\bar{X} - p\sigma_X; \bar{X} + p\sigma_X[$ est supérieure ou égale à $1 - \frac{1}{p^2}$, pour tout $p > 0$.

Cette inégalité précise en quoi l'écart-type est une mesure de la dispersion autour de la valeur moyenne.

Démonstration. Si $V(X) = 0$, tous les éléments de X sont égaux à $\bar{X}$ est la conclusion est vraie. Supposons donc dans la suite que $V(X) \neq 0$.

Remarquons que :

$$x_i \in]\bar{X} - p\sigma_X; \bar{X} + p\sigma_X[\iff |x_i - \bar{X}| < p\sigma_X \iff (x_i - \bar{X})^2 < p^2 V(X)$$

À partir d'une distribution groupée, $\{(x_i, n_i) \mid i = 1, \dots, k\}$ ordonnée de sorte que :

$$\begin{cases} (x_i - \bar{X})^2 < p^2 V(X) & \text{pour tout } i = 1, 2, \dots, t \\ (x_i - \bar{X})^2 \geq p^2 V(X) & \text{pour } i = t+1, \dots, k \end{cases}$$

$$\begin{aligned} V(X) &= \sum_{i=1}^k f_i \times (x_i - \bar{X})^2 \\ &= \underbrace{\sum_{i=1}^t f_i \times \underbrace{(x_i - \bar{X})^2}_{< p^2 V(X)}}_{\geq 0} + \sum_{i=t+1}^k f_i \times \underbrace{(x_i - \bar{X})^2}_{\geq p^2 V(X)} \\ \implies V(X) &\geq \sum_{i=t+1}^k f_i \times \underbrace{(x_i - \bar{X})^2}_{\geq p^2 V(X)} \\ \implies V(X) &\geq p^2 V(X) \times \sum_{i=t+1}^k f_i && \text{car } f_i \geq 0 \\ \xRightarrow[V(X) \neq 0]{} 1 &\geq p^2 \times \sum_{i=t+1}^k f_i \xRightarrow[p > 0]{} \sum_{i=t+1}^k f_i \leq \frac{1}{p^2} \end{aligned}$$

Or $\sum_{i=t+1}^k f_i$ est la proportion d'éléments qui ne sont pas dans l'intervalle $]\bar{X} - p\sigma_X; \bar{X} + p\sigma_X[$, et elle est au plus de $1/p^2$. La proportion d'éléments dans $]\bar{X} - p\sigma_X; \bar{X} + p\sigma_X[$, qui est $\sum_{i=1}^t f_i = 1 - \sum_{i=t+1}^k f_i$, est donc au moins de $1 - 1/p^2$. ■