La statistique descriptive

Définitions

L'ensemble étudié porte le nom de **population**. Cet ensemble est composé d'éléments statistiques appelés **individus**. La population étant définie, elle est observée selon certains critères. Le critère retenu est appelé **caractère**.

Les résultats qu'on détient sont généralement inutilisables sous leur forme brute. On procède donc à des regroupements, à des classements et à l'établissement de tableaux statistiques.

Les individus sont rassemblés par **classes**, ça donne donc des ensembles de données **groupées**. Ce qu'on gagne en simplicité par ce regroupement, on le perd en information. Il est alors commode de formuler l'hypothèse d'une répartition uniforme au sein de chaque classe.

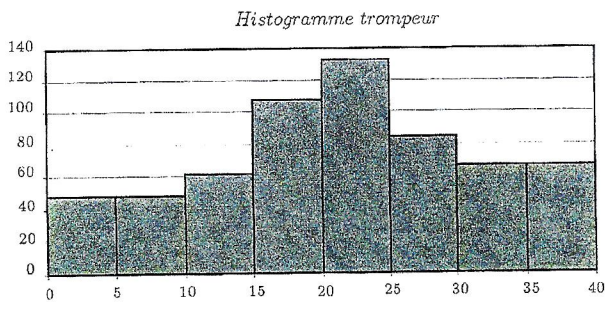
Quand on ne sait pas la fin de la dernière classe (ex. : plus de 30), on convient d'attribuer la même étendue que celle de la première classe.

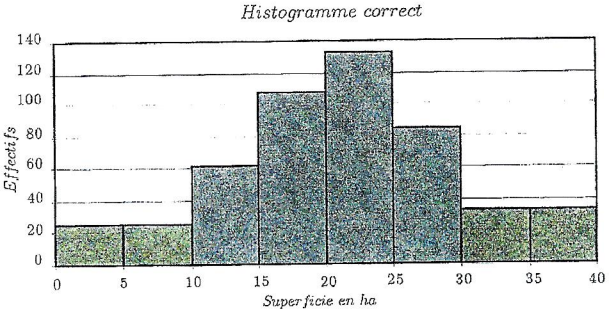
Représentations graphiques

Même qu'un tableau statistique permette déjà une première analyse, Il est souvent très pratique de représenter l'ensemble des données sous la forme de différents graphiques.

Histogramme

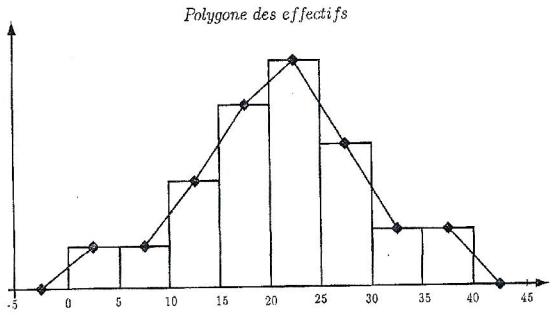
Les classes n'ont pas toujours la même amplitude. Il faut donc tenir compte de ces différences. La surface qui est représentée dans un histogramme doit être proportionnelle à l'effectif de la classe. On doit donc rectifier certains effectifs.





Polygone des effectifs

Il s'agit d'une courbe polygonale telle que la surface comprise entre cette courbe et l'axe des abscisses soit égale à la surface de l'histogramme. Elle est obtenue en joignant les milieux des sommets des rectangles de l'histogramme. Pour la première est la dernière classe, on crée deux classes fictives d'effectifs nuls.



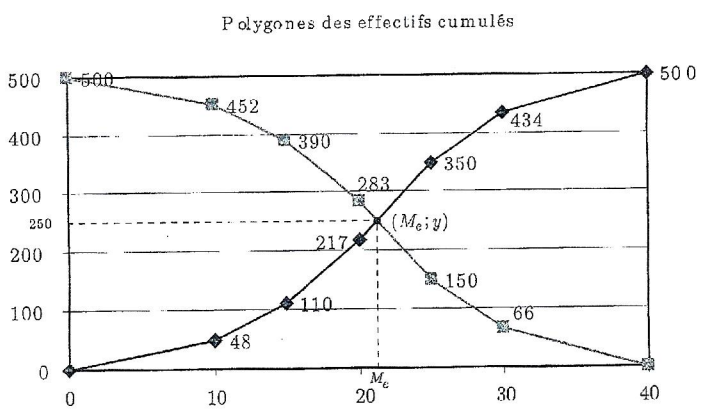
Effectifs cumulés

Aux données de départ, on associe le tableau des effectifs cumulés croissants et décroissants.

Cumulés croissants : somme des effectifs en commençant par le début
Cumulés décroissants : somme des effectifs en commençant par la fin

Ils peuvent être représentés par 2 courbes. On ne se préoccupe pas des différences d'amplitude dans les classes.

L'intersection de ces deux polygones est un point (M_e;y).
M_e : est la médiane
y : est toujours en plein milieu de l'axe vertical



Valeurs centrales

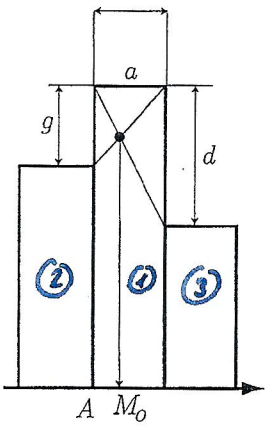
Le mode

Le mode est la valeur du caractère qui correspond à l'effectif le plus grand ou à la fréquence la plus importante. Il ne tient pas compte des valeurs extrêmes.

$$M_o = A + \frac{g}{g+d} \cdot a.$$

1) classe modale
2) classe voisine à gauche
3) classe voisine à droite

A) début de la classe
a) amplitude de la classe
g) différence en l'effectif de la classe modale et l'effectif de la classe voisine à gauche
d) différence en l'effectif de la classe modale et l'effectif de la classe voisine à droite



La médiane

Une moitié des données est plus petite et l'autre moitié est plus grande. Elle ne tient pas compte des valeurs extrêmes.

La médiane M_e d'un ensemble de nombres $X_1, X_2, \ldots, X_N$ rangés par ordre de grandeur croissant est la valeur du milieu ou une valeur comprise entre les valeurs centrales.

Si l'ensemble des nombres (N) est pair → individu réel
Si l'ensemble des nombres (N) est impair → individu virtuel

Graphiquement, la médiane est la 1ère coordonnée du point d'intersection des polygones des effectifs cumulés croissants et décroissants.

La moyenne arithmétique

La moyenne arithmétique $\bar{x}$ est la valeur centrale la plus connue. Elle tient compte des valeurs extrêmes.

$$\bar{x} = \frac{\sum n_i x_i}{\sum n_i}$$

Pour des séries de données groupées, on convient d'affecter à tous les individus d'une classe]a;b] la valeur centrale (a+b)/2

La moyenne géométrique

La moyenne géométrique G de N nombres positifs X_1, X_2, ... , X_N est définie par la racine N-ième de leur produit.

$$G = \sqrt[N]{x_1 \cdot x_2 \cdots x_N} = (x_1 \cdot x_2 \cdots x_N)^{\frac{1}{N}}.$$

La moyenne harmonique

La moyenne harmonique H d'un ensemble de N nombres non nuls X_1, X_2, ... , X_N est définie comme étant l'inverse de la moyenne arithmétique de leurs inverses.

$$H = \frac{1}{\frac{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_N}}{N}} = \frac{N}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_N}}.$$

Relations entre moyennes

$$H \leq G \leq \bar{x}.$$

Mesures de dispersion

Si les valeurs centrales sont généralement nécessaires pour caractériser une série, elles ne sont toutefois pas suffisantes. Deux populations différentes peuvent avoir les mêmes valeurs centrales et différer notablement quant à la dispersion des individus autour de ces valeurs centrales.

L'étendue de la série

L'étendue E de la série est la différence entre les valeurs extrêmes de la série. Elle est donc très sensible aux valeurs extrêmes. Simple à calculer, cette mesure de dispersion n'est pas très fiable puisqu'elle ne tient compte que de deux observations marginales et néglige toutes les autres.

L'écart maximal relatif

$$\frac{E}{\bar{x}} = \frac{\text{Max } x_i - \text{Min } x_i}{\bar{x}}$$

Cette mesure de dispersion, qui n'a pas d'unité, est nulle si toutes les observations sont identiques et augmente quand l'écart extrême s'amplifie. Elle est très sensible aux valeurs extrêmes.

Une telle mesure, qui décrit une dispersion relative, permet de comparer les dispersions de deux populations dont les caractères sont différents.

Les intervalles interquantiles

On appelle **quantiles d'ordre n** les valeurs du caractère qui partagent l'effectif total de la série en n groupes d'effectifs égaux.

Les quartiles

C'est-à-dire les quantiles d'ordre 4, notés Q_1, Q_2 et Q_3. Un quart de l'effectif total possède un caractère inférieur à Q_1. Le deuxième quartile $Q_2 = M_e$ n'est autre que la médiane. Enfin, les trois quarts de la population se trouvent en dessous de la valeur définie par le troisième quartile Q_3.

Les déciles

D_1, D_2, ... , D_9 partagent l'effectif total en 10 groupes égaux. Le décile D_5 est égal à la médiane.

Les centiles

C_1, C_2, ... , C_{99} partagent la population en 100 groupes d'effectifs égaux.

Pour trouver ces quantiles, il faut utiliser l'algorithme.

L'intervalle interquartile

L'intervalle interquartile $I_Q = Q_3 - Q_1$ est défini par la différence des quartiles extrêmes. Il exclut les 50% des valeurs marginales inférieures et supérieures.

L'intervalle interdécile

L'intervalle interdécile $I_D = D_9 - D_1$ définit un intervalle comprenant les 80% de la population.

Les intervalles interquantiles se prêtent bien à une mesure de dispersion relative.

L'intervalle interquartile relatif

$$\boxed{\frac{I_Q}{M_e} = \frac{Q_3 - Q_1}{M_e}}$$

L'intervalle interdécile relatif

$$\boxed{\frac{I_D}{M_e} = \frac{D_9 - D_1}{M_e}}$$

L'écart absolu moyen

L'écart absolu moyen e_a est la moyenne arithmétique des valeurs absolues des écarts de tous les termes de la série par rapport à leur moyenne $\bar{x}$.

$$e_a = \frac{\sum n_i \cdot |c_i - \bar{x}|}{\sum n_i} = \frac{3339}{500} = 6{,}678 \text{ ha.}$$

On peut introduire une mesure de dispersion relative : l'écart moyen relatif.

$$\frac{e_a}{\bar{x}}$$

L'écart-type

C'est certainement la mesure de dispersion la plus utilisée. Etant donnée une population, on peut calculer la moyenne des carrés des écarts entre toutes les données et leur moyenne arithmétique. Cette mesure est appelée variance V. L'écart type est dont la racine carrée de la variance.

$$\sigma = \sqrt{V} = \sqrt{\overline{(x_i - \bar{x})^2}}$$

Le calcul de la variance (et donc de l'écart-type) n'est pas toujours commode. Ce dernier peut toutefois être simplifié de la manière suivante. La variance V est obtenue en calculant la différence entre la moyenne $\overline{x^2}$ des carrés des données x_i et le carré de leur moyenne $\bar{x}^2$.

$$V = \overline{x^2} - \bar{x}^2.$$

Classes x_i	Centres c_i	Effectifs n_i	Carrés des c_i c_i^2	Produits $n_i \cdot c_i^2$
[0; 10[	5	48	25	1200
[10; 15[	12,5	62	156,25	9687,5
[15; 20[	17,5	107	306,25	32768,75
[20; 25[	22,5	133	506,25	67331,25
[25; 30[	27,5	84	756,25	63525
[30; 40[	35	66	1225	80850
Total		**500**		**255362,5**

On en déduit que $\overline{x^2} = \frac{255362{,}5}{500} = 510{,}725$. Comme $\bar{x} = 21$, $\bar{x}^2 = 441$ et il suit que

$$V = \overline{x^2} - \bar{x}^2 = 510{,}725 - 441 = 69{,}725.$$

Notons que l'écart-type sera d'autant plus grand plus grand que la série est dispersée autour de la moyenne. Il est nul que dans le cas où toutes les données sont identiques.

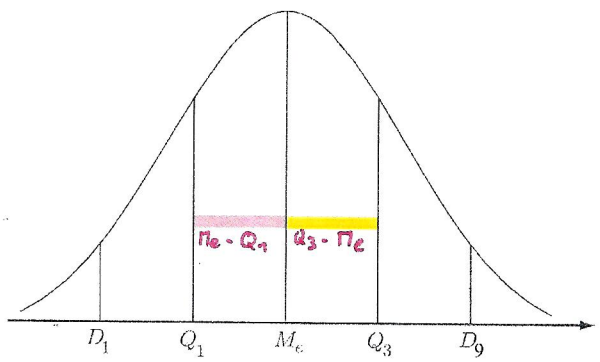
Le coefficient de variation

Il est utilisé pour faciliter les comparaisons entre les séries. Ce coefficient C est donnée par le rapport entre l'écart-type et la moyenne arithmétique.

$$C = \frac{\sigma}{\bar{x}}$$

Mesures de symétrie

La répartition d'une population autour de sa médiane peut présenter une symétrie plus ou moins prononcée. La position des quantiles par rapport à la médiane permet de porter un jugement sur la symétrie d'une distribution. Une population parfaitement symétrique s'étend de façon identique de part et d'autre de la médiane.



Le coefficient d'asymétrie

$$k_a = \frac{(Q_3 - M_e) - (M_e - Q_1)}{(Q_3 - M_e) + (M_e - Q_1)} = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2M_e}{I_Q}.$$

Ce coefficient est nulle pour une distribution symétrique ; sa valeur absolue augmente avec la dissymétrie de la population.

Toujours : $-1 \leq K_a \leq 1$

Quand $(Q_3 - M_e) = (M_e - Q_1)$ → $K_a = 0$

Quand $(Q_3 - M_e) > (M_e - Q_1)$ → $0 \leq K_a \leq 1$ → asymétrie à droite

Quand $(Q_3 - M_e) < (M_e - Q_1)$ → $-1 \leq K_a \leq 0$ → asymétrie à gauche

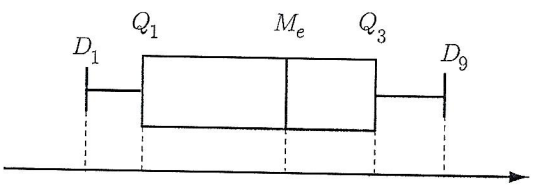
Le coefficient de symétrie

$$k_s = \frac{(Q_3 - M_e)}{(M_e - Q_1)}$$

Ce coefficient est égal à 1 pour une distribution symétrique. Il s'écarte de l'unité lorsque la dissymétrie augmente.

La boîte à moustache

C'est une représentation codifiée des quantiles D_1, Q_1, M_e, Q_3 et D_9 qui donne une information graphique concernant la symétrie de la distribution.



Mesures de concentration

La notion de concentration est une notion connexe de celle de dispersion.

2 conditions sont nécessaires pour parler de concentration
- L'addition des différentes modalités du caractère doit avoir un sens.
- Le partage de la masse globale du caractère doit être possible.

La médiale

La médiale M_L est la valeur du caractère qui partage en deux la masse globale du caractère.

La comparaison des valeurs de la médiane et de la médiale constitue une mesure de concentration. Lorsque l'écart entre la médiale et la médiane est important par rapport à l'étendue de la série, la concentration est forte. Si la distribution est au contraire fortement égalitaire, la concentration sera faible et l'écart entre médiale et médiane sera petit.

On calcule la médiale, comme la médiane, mais en considérant la masse cumulée du caractère.

Exemple

Classes x_i	Centres c_i	Effectifs n_i	Surfaces globales par classes	Surfaces cumulées
[0; 10[	5	48	240	240
[10; 15[	12,5	62	775	1015
[15; 20[	17,5	107	1872,5	2887,5
[20; 25[	22,5	133	2992,5	5880
[25; 30[	27,5	84	2310	8190
[30; 40[	35	66	2310	10500

On cherche la surface d'exploitation au-dessous de laquelle on trouvera des exploitations partageant la moitié de la surface totale (10500 ha), c'est-à-dire 5250 ha. Ça se trouve donc dans la classe [20;25].

$$M_L = 20 + 5 \cdot \frac{2363}{2992,5} \cong 23,95 \text{ ha.}$$

Remarquons que ces exploitations sont au nombre de 217 + (3.95/5) x 133 = 322 et représente donc 64.4%.

Pour la médiane, nous avions trouvé M_e = 21,26 ha. Ainsi, l'écart $M_L - M_e$ = 2,69 entre médiale et médiane représente seulement 6.73% de l'étendue. La concentration est donc relativement faible.

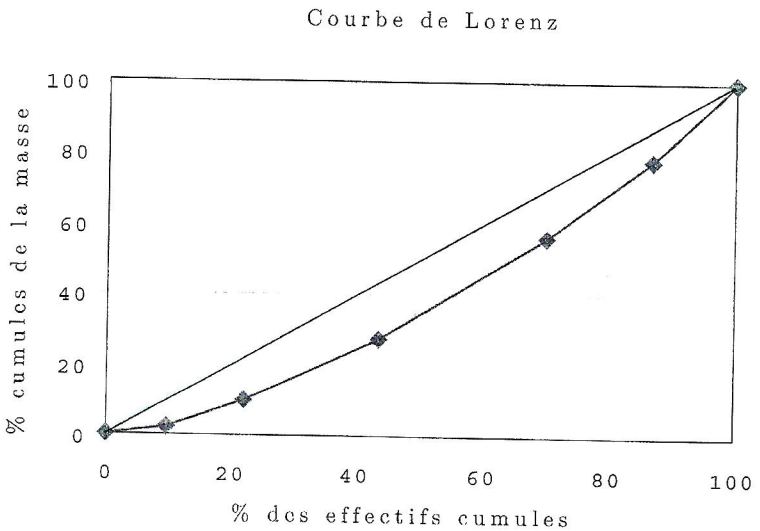
La courbe de Lorenz

On mesure la concentration d'une population en la comparant avec une distribution où la masse du caractère serait également répartie entre les individus. Dans une répartition parfaitement égalitaire, un certain pourcentage des individus est titulaire d'un pourcentage égal de la masse totale du caractère.

La courbe de Lorenz est une courbe polygonale formée de points (x;y) portant l'information suivante : x% des individus se partagent y% de la masse totale du caractère. Ainsi, cette ligne polygonale lie les points (0;0) et (100;100) et se compose de plusieurs segments, sauf si la distribution est parfaitement égalitaire.

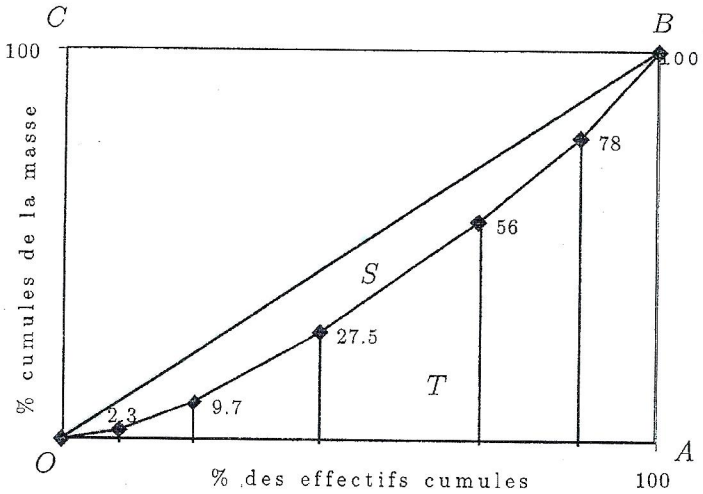
Pour représenter cette courbe, on a besoin du pourcentage cumulé des effectifs et du pourcentage cumulé de la masse globale du caractère.

Centres de classes	Effectifs cumulés	% cumulés des effectifs	Masses cumulées	% cumulés de la masse
5	48	9,6	240	2,3
12,5	110	22	1015	9,7
17,5	217	43,4	2887,5	27,5
22,5	350	70	5880	56
27,5	434	86,8	8190	78
35	500	100	10500	100



Le coefficient de Gini

Plus la courbe de Lorenz est proche de la diagonale du carré, plus la concentration est faible ; plus elle s'en éloigne, plus la concentration est importante.



On appelle surface de concentration, l'aire S du domaine compris entre la diagonale du carrée OABC et la courbe de Lorenz. On définit alors comme mesure de concentration, ledit coefficient de Gini G qui exprime le rapport entre la surface de concentration et l'aire du triangle OAB.

$$G = \frac{\text{surface de concentration}}{\text{aire du triangle } OAB} = \frac{S}{5000}. \rightarrow \frac{100 \cdot 100}{2}$$

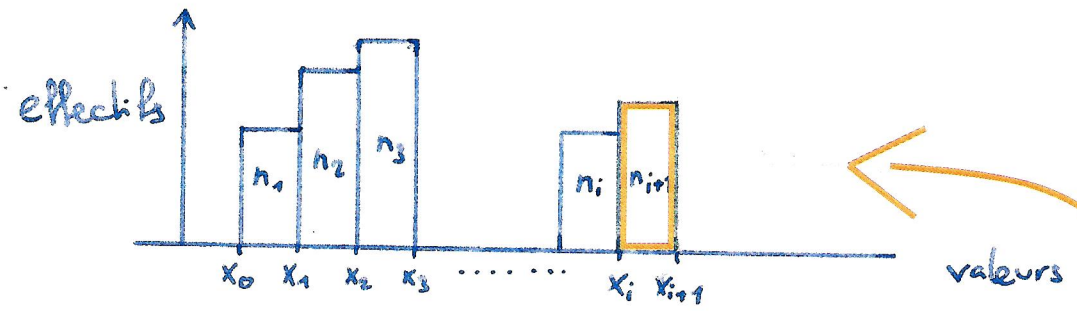
Ce coefficient varie en 0 et 1. Il est nul lorsque la courbe de Lorenz est confondue avec la diagonale ; il est égal à 1 quand la courbe est formée des côtés OA et AB du carré OABC.

L'aire T comprise entre la courbe de Lorenz et l'axe horizontal est formée de triangle et de trapèzes.

Aire du triangle : (base x hauteur) / 2
Aire du trapèze : ((Grande base + petite base) x hauteur) / 2

Le coefficient de Gini est l'un des indicateurs les plus opérationnels et utilisés. Il est insensible à l'unité de mesure car il utilise des pourcentages. On peut donc faire des comparaisons.

Algorithme



Comment estimer la $p^{ème}$ valeur ?

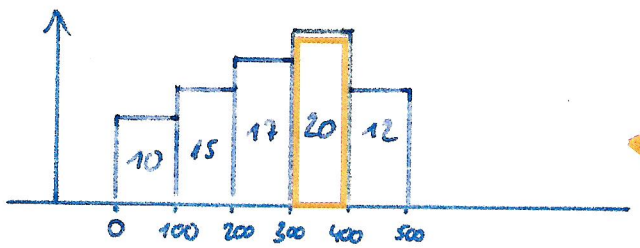
1) trouver i tel que

$$n_1 + \dots + n_i \leq p < n_1 + \dots + n_i + n_{i+1}$$

2) en supposant que les n_{i+1} valeurs sont distribuée uniformément entre x_i et x_{i+1}

$$p^{ème}\ \text{valeur} = x_i + \frac{x_{i+1} - x_i}{n_{i+1}} \cdot \left(p - (n_1 + \dots + n_i)\right)$$

Exemple



Où est la $60^{ème}$ valeur ?

1) $\underbrace{10 + 15 + 17}_{42} \leq 60 < \underbrace{10 + 15 + 17 + 20}_{62}$

2) $60^{ème}$ valeur : $300 + \frac{400 - 300}{20} \cdot (60 - 42)$

Moyennes

Données simples : $x_1 ; x_2 ; \dots ; x_n$

Moyenne arithmétique : $\dfrac{x_1 + \dots + x_n}{n} = \bar{x}$

Moyenne géométrique : $\sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = G$

Exemple :

Taux d'inflation :

		inflation moyenne
2007	→ 2%	x
2006	→ 1,5%	x
2005	→ 1,4%	x

Inflation moyenne sur les 3 ans ?

$$S \xrightarrow{3\text{ans}} (1+0{,}02)(1+0{,}015)(1+0{,}014)\,S = (1{,}02)(1{,}015)(1{,}014)\,S$$

$$S \longrightarrow (1+x)(1+x)(1+x)\,S = (1+x)^3 S$$

$$(1+x)^3 = (1{,}02)(1{,}015)(1{,}014)$$

$$1+x = \sqrt[3]{1{,}02 \cdot 1{,}015 \cdot 1{,}014} = 1{,}0163\dots$$

Taux moyen $(1+x)$ obtenu comme la moyenne géométrique des (1 + taux années)

Moyenne harmonique : $\dfrac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n}} = H$

Exemples ⇒ p. 6 et 7 du cours

Type de données

1) $x_1, x_2, x_3, \dots x_i, \dots x_N$ (données individuelles)

2)

Fréquence	n_1	n_2	n_3	n_i	...
Valeurs	x_1	x_2	x_3	x_i	...

(données groupées)

3)

Classes	1	2	i	...
Fréquence	n_1	n_2	n_i	...
Centre	c_1	c_2	c_i	...

(données par classe)

N = nbre total
i = nbre quelconque

Mesures de centralités

Mode

1) Valeur(s) la (les) plus grande(s)
2) Le ou les n_i les plus grands fournissent le ou les mode(s)
3) Déterminer la classe modale (n_i le plus grand) puis formule de la page 5

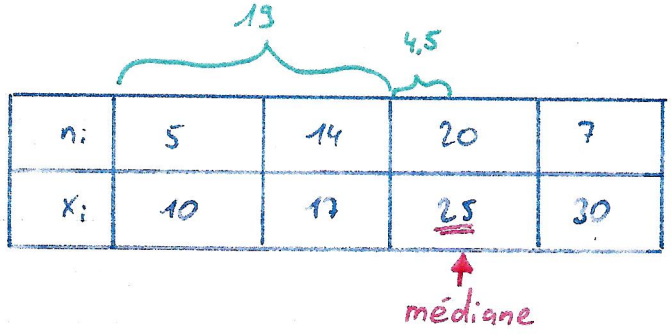
Médiane

1) N impair : $x_1, \dots, x_{\frac{N-1}{2}}, \boxed{x_{\frac{N-1}{2}+1}}, x_{\frac{N-1}{2}+2}, \dots x_N$

N pair : $x_1, \dots, x_{\frac{N}{2}}, x_{\frac{N}{2}+1}, \dots, x_n$

$$\boxed{\frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2}}$$

2)



n_i	5	14	20	7
x_i	10	17	25	30

→ 46 observations

médiane

$\frac{x_{23} + x_{24}}{2} = x_{23,5}$

3) Algorithme

Moyenne arithmétique

1) $\frac{x_1 + x_2 + \dots + x_N}{N}$

2) $\frac{n_1 \cdot x_1 + n_2 \cdot x_2 + \dots + n_i \cdot x_i}{n_1 + n_2 + \dots + n_i} = \frac{\sum n_i \cdot x_i}{\sum n_i}$

3) $\frac{\sum n_i \cdot c_i}{\sum n_i}$

Moyenne géométrique

1) $\sqrt[N]{x_1 \cdot x_2 \cdot \dots \cdot x_N}$

2) $\sqrt[\sum n_i]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots x_i^{n_i} \cdot \dots}$

3) $\sqrt[\sum n_i]{c_1^{n_1} \cdot c_2^{n_2} \dots \cdot c_i^{n_i} \dots}$

Moyenne harmonique

1) $$\frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N}}$$

2) $$\frac{n_1 + n_2 + \ldots + n_i}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_i}{x_i}} = \frac{\sum n_i}{\frac{n_i}{x_i}}$$

3) $$\frac{\sum n_i}{\sum \frac{n_i}{c_i}}$$

Mesures de dispersion

Etendue

1) $x_N - x_1$
2) idem
3) Borne supérieure dernière classe - Borne inf. première classe

Ecart absolu moyen

1) $$\frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \ldots + |x_N - \bar{x}|}{N} = \frac{\sum |x_i - \bar{x}|}{N}$$

2) $$\frac{n_1 \cdot |x_1 - \bar{x}| + n_2 \cdot |x_2 - \bar{x}| + \ldots + n_N \cdot |x_N - \bar{x}|}{\sum n_i} = \frac{\sum n_i \cdot |x_i - \bar{x}|}{\sum n_i}$$

2) $$\frac{\sum n_i \cdot |c_i - \bar{x}|}{\sum n_i}$$ ← attention valeur absolue

Variance / Ecart-type

Variance = V

Ecart-type = $\sigma = \sqrt{V}$

1) $$V = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N}$$

2) $$V = \frac{n_1 (x_1 - \bar{x})^2 + \ldots + n_i (x_i - \bar{x})^2 + \ldots}{\sum n_i} = \frac{\sum n_i (x_i - \bar{x})^2}{\sum n_i}$$

3) $$V = \frac{\sum n_i (c_i - \bar{x})^2}{\sum n_i}$$