ENSIAS

Statistiques descriptives

Lasse9 statistiques descriptives | GHAZOUAN Oumaima

Programme

- Séries simples
- séries doubles
- séries chronologiques

Séries simples

Organigramme

Retourner à la page Programme

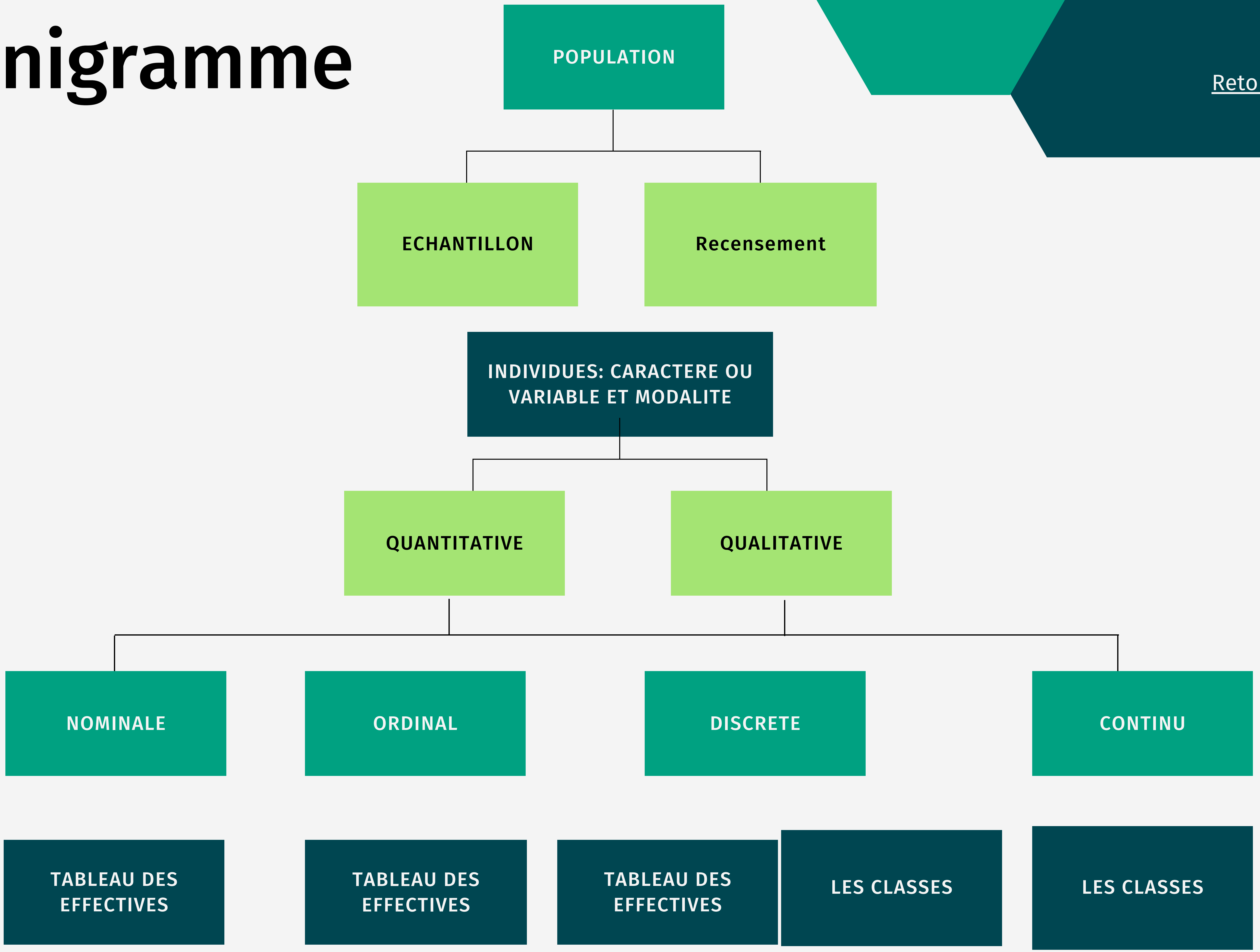


Tableau récapitulatif Pour une variable quantitative continue:

Classes	n_i	a_i	c_i	f_i	f_{icc}	f_{icd}	n_{icc}	n_{icd}	d_i	n_{ic}
[0; 5[										
[5;15[										
[15;30[										

Xi : caractère
Ni : effectif
Nic : effectif cumulé
Nicc: effectif cumulé croissant
Nicd : effectif cumulé décroissant
Fr : fréquence ni/N
N=somme des ni
Pi : pourcentage (n/N)*100

Ci : centre de classe (a+b)/2
Ai : amplitude b-a ou xi-xi-1
di : densité ni /ai
Nic=(Ni /ai)*ppcm(ai)

Graphiques

- Caractère qualitatif

Diagramme en tuyaux d’orgues

Diagramme circulaire / demi-circulaire

- Caractère quantitatif discret

Diagramme en bâtons

Courbe cumulative des fréquences

- Caractère quantitatif continu

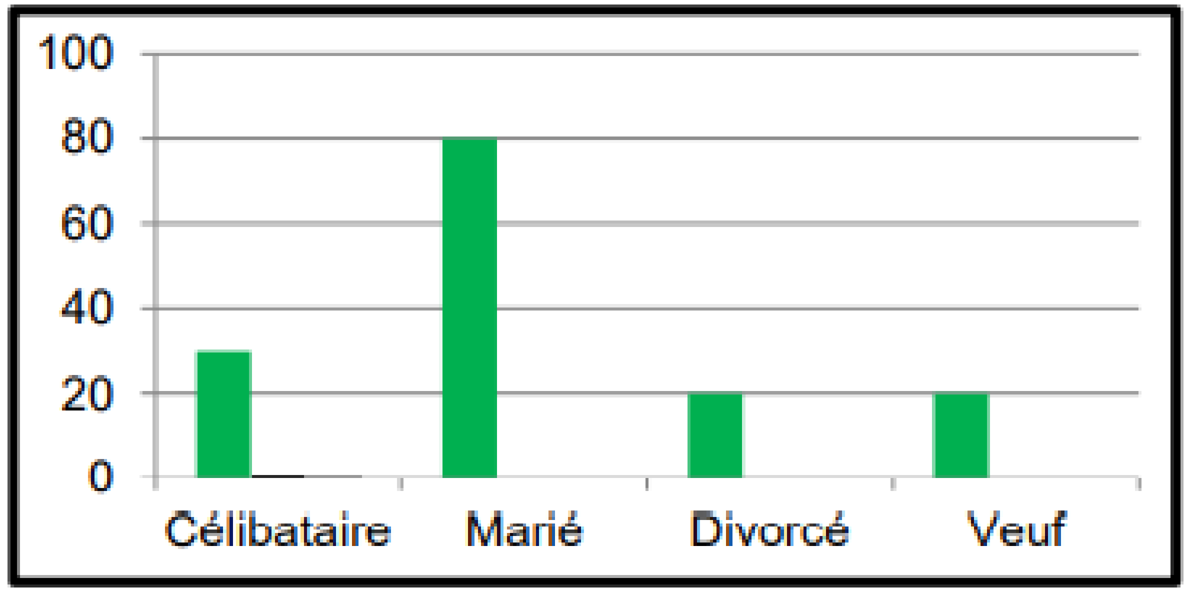
Histogramme

Polygone de fréquences

Courbe cumulative de fréquences

Pour une variable qualitative

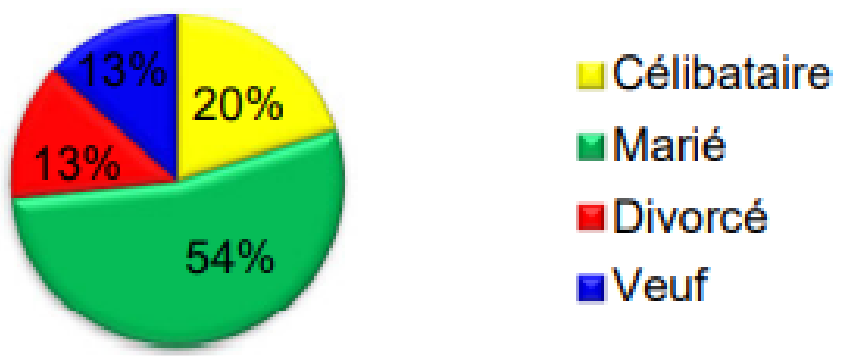
Diagramme des tuyaux d'orgue



La longueur des tuyaux $=n_i$

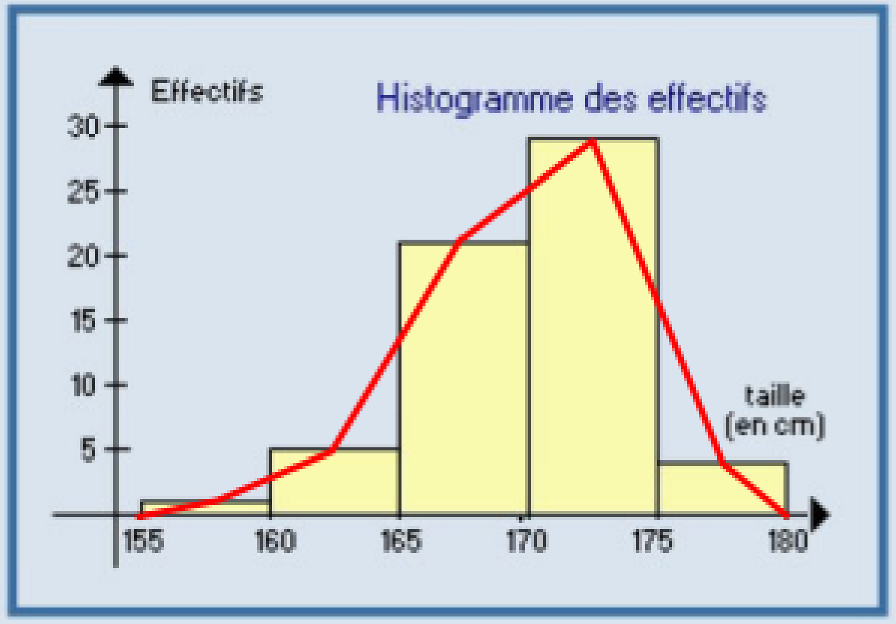
Le graphique à secteurs

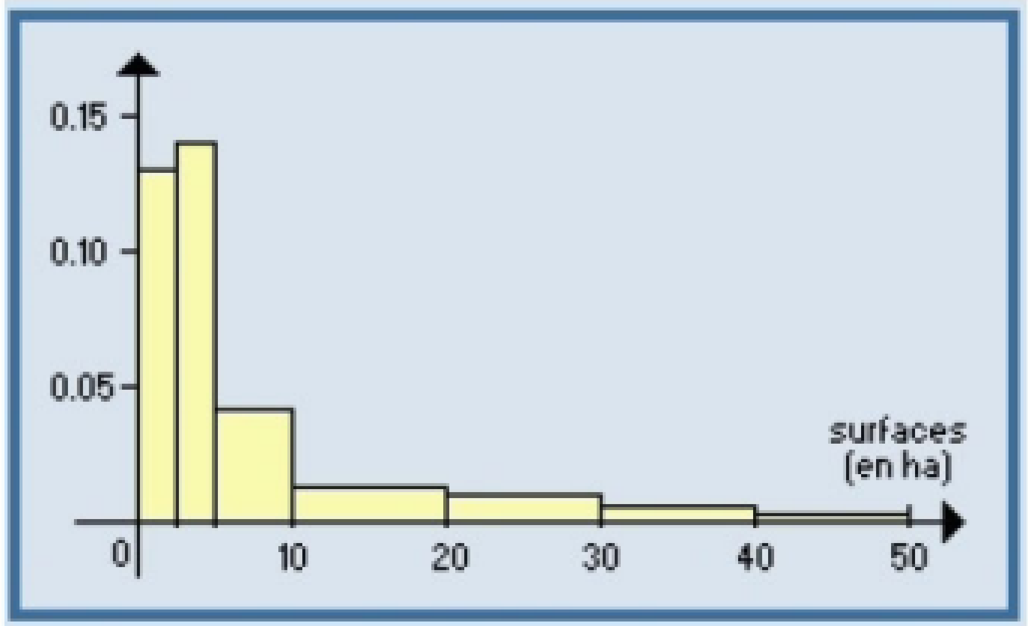
Effectifs



$$\alpha_i = \frac{n_i}{n} \times 360 = f_i \times 360$$

Pour une variable quantitative





Fonction de répartition

La fonction de répartition est donnée par:

F_i = fréquence de nombre de famille qui ont moins de x_i enfants n_{icc}/n

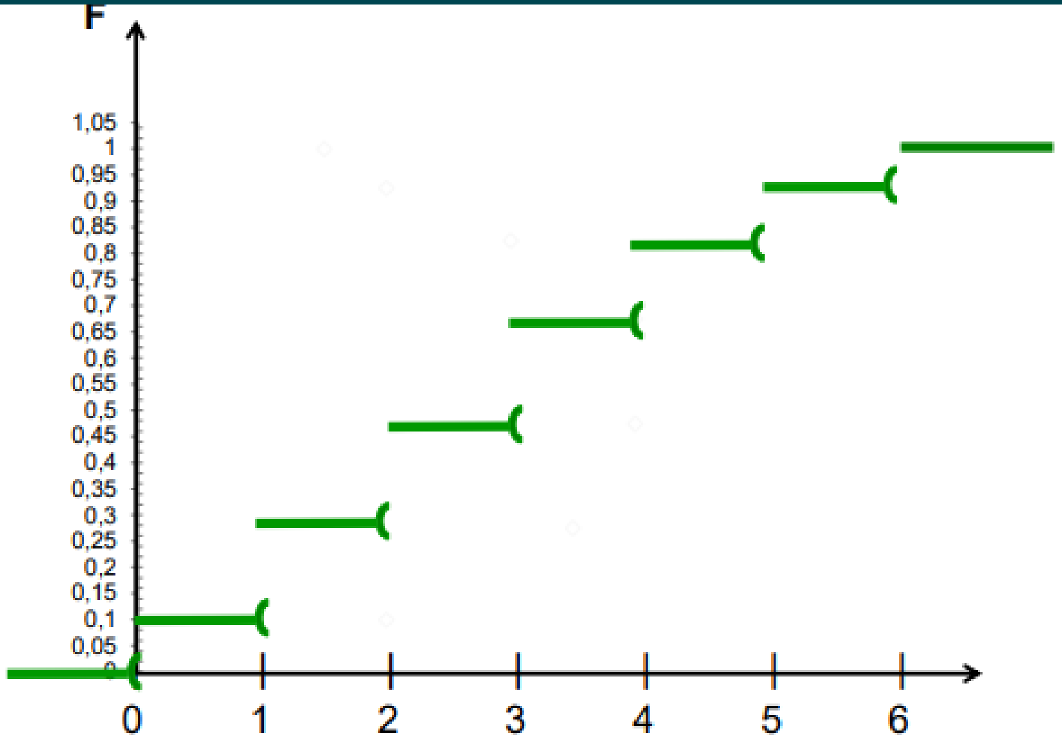
$$F(x)=\begin{cases} 0 & si \quad x < x_1 \\ F_i & si \quad x_i \le x < x_{i+1} \\ 1 & si \quad x \ge x_k \end{cases}$$

Exemple: le nombre d'enfants par famille,

xi	ni	Ni	Fi
0	4	4	0,1
1	7	11	0,275
2	8	19	0,475
3	8	27	0,675
4	6	33	0,825
5	4	37	0,925
6	3	40	1

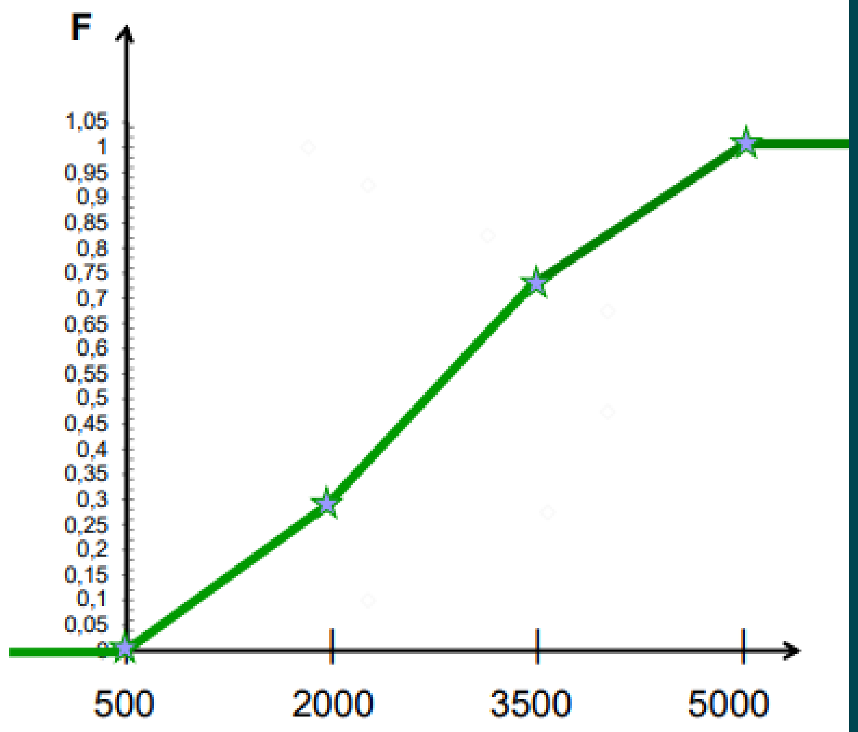
F(0)= fréquence de nombre de famille moins de 0 enfants = 0/40
F(1) = fréquence de nombre de famille moins de 1 enfants = 4/40
F(2) = fréquence de nombre de famille moins de 2 enfants = 11/40
F(3) = fréquence de nombre de famille moins de 3 enfants = 19/40
F(4) = fréquence de nombre de famille moins de 3 enfants = 27/40
F(5) = fréquence de nombre de famille moins de 3 enfants = 33/40
F(6) = fréquence de nombre de famille moins de 6 enfants = 37/40
F(7) = fréquence de nombre de famille moins de 7 enfants = 40/40

xi	ni	fi	Fi
0	4	0,1	0,1
1	7	0,175	0,275
2	8	0,2	0,475
3	8	0,2	0,675
4	6	0,15	0,825
5	4	0,1	0,925
6	3	0,075	1



Représentation graphique de la fonction de répartition d'une variable discrète

xi	ni	fi	Fi
[500;2000[	30	0,3	0,3
[2000;3500[	45	0,45	0,75
[3500;5000[	25	0,25	1



Représentation graphique de la fonction de répartition d'une variable classée ou qualtitative continue

Mode

Le mode de cette série statistique est la modalité de la variable correspondant à l'effectif le plus élevé

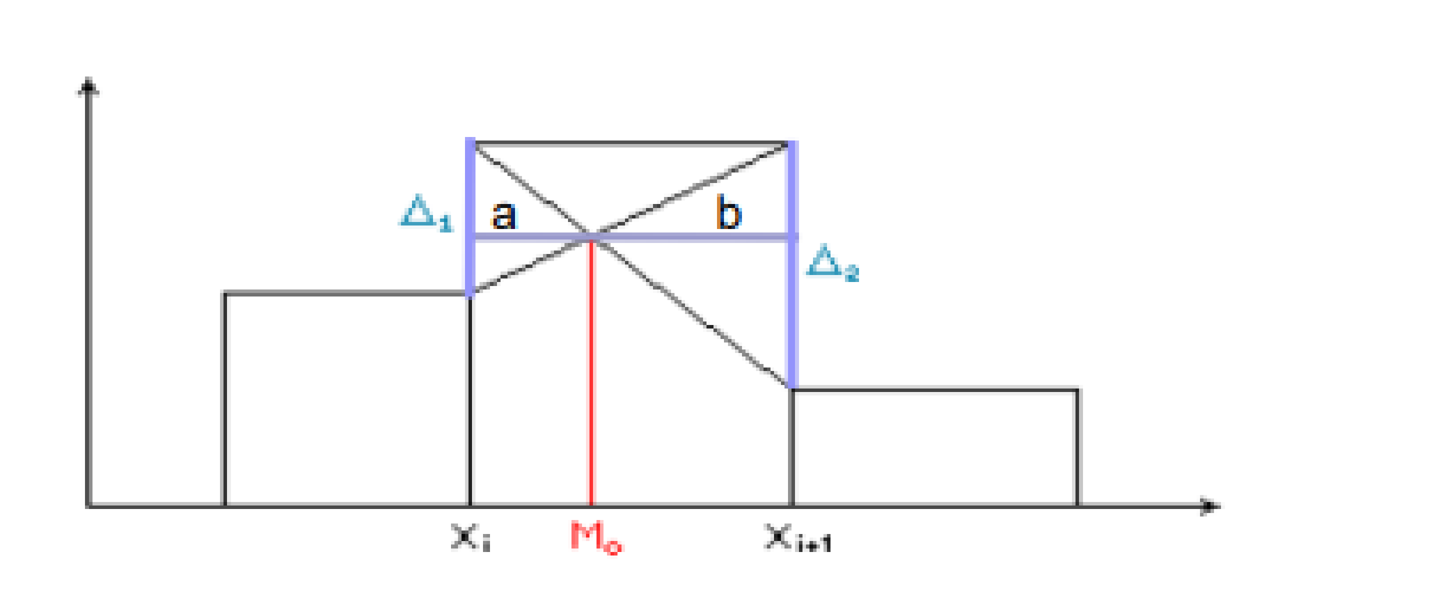
Les paramètres de position

Pour les variables quantitatives classées

on parle d'abord de la classe modale:

- Si les classes sont d'égales amplitudes, la classe modale sera la classe où l'effectif est le plus élevé.
- Si les classes sont d'inégales amplitudes, la classe modale sera la classe où:

La densité ou La densité de fréquence est la plus élevée



$$Mo = \frac{x_i\Delta_2 + x_{i+1}\Delta_1}{\Delta_1 + \Delta_2} = x_i + \frac{\Delta_1(x_{i+1} - x_i)}{\Delta_1 + \Delta_2}$$

Médiane

pour les variables quantitatives

Les paramètres de position

Cas d’une variable non classée

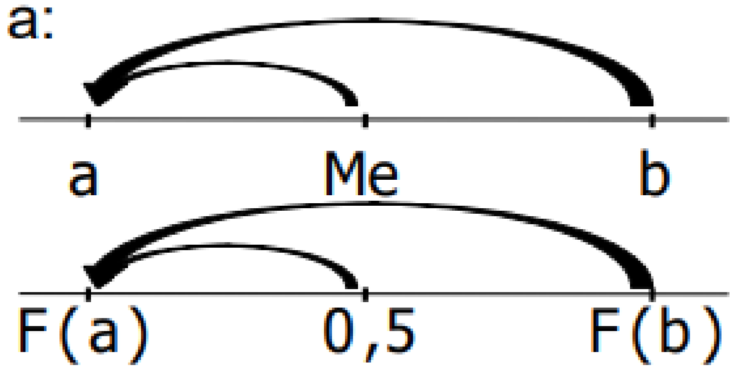
$x_1<x_2<......;<x_n$

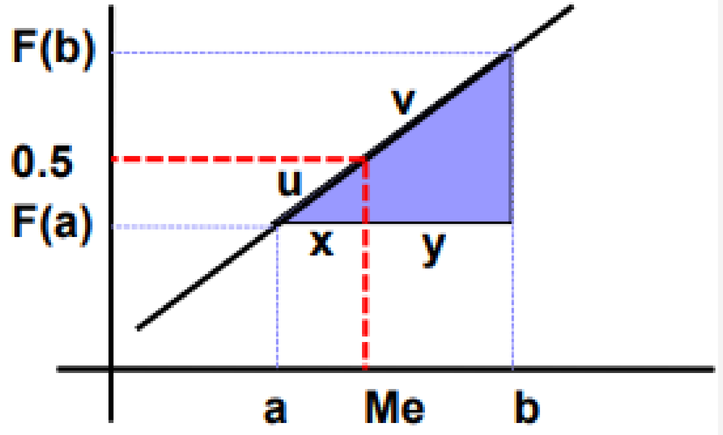
- Si n est impair et égal 2p+1 la médiane sera: x_{p+1}.
- Si n est pair et égal 2p, la médiane est: $x_p+x_{p+1}/2$

Cas d’une variable classée

Soient a et b les bornes inférieurs et supérieurs de la classe contenant la médiane, F(a) et F(b) les valeurs des fréquences cumulées croissantes en a et b, alors:

On a:





$$\frac{x}{y}=\frac{u}{v}$$

$$\frac{Me-a}{b-a}=\frac{0,5-F(a)}{F(b)-F(a)} \Longrightarrow Me=a+(b-a)\times\frac{0,5-F(a)}{F(b)-F(a)}$$

Moyenne

pour les variables classées on untilise ni

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

pour les variables non classées on utilise ci

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} \frac{n_i}{n} c_i = \sum_{i=1}^{k} f_i c_i$$

Les paramètres de position

Arithmétique	Géométrique	Harmonique	Quadratique
$\overline{X} = \sum_{i=1}^{k} f_i x_i$	$\overline{X}_g = \prod_{i=1}^{k} x_i^{f_i}$	$\overline{X}_h = \frac{1}{\sum_{i=1}^{k} f_i \frac{1}{x_i}}$	$\overline{X}_q = \sqrt{\sum_{i=1}^{k} f_i x_i^2}$

Comparaison des moyenne

$$\overline{X}_h \leq \overline{X}_g \leq \overline{X} \leq \overline{X}_q$$

La variance V

la dispersion des valeurs autour de la moyenne

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

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

L’écart-type σ

mesure la dispersion des valeurs d'un échantillon par rapport a la moyenne mais son interprétation dépend de l’échelle de la variable

Les paramètres de dispertion

Etendue Vmax-Vmin

Quartile [Q1 ;Q3]

des caractéristiques de position partageant la série statistique

Intervalles interquartiles Q3 – Q1

Décile [D1 ;D9]

Ecart interdécile D9-D1

Exercice

Les paramètres de dispertion

* Cas d'une Caractère Continue :

Ex : On a 100 DH

Donc : $Q_1 = 25\%$ $\quad Q_3 = 75\%$

$Q_2 = 50\%$ $\quad Q_4 = 100\%$

xi	ni	nicc
10-30	4	4
30-50	16	20
50-70	20	40
70-80	17	57
N	57	

* Rang $= \frac{N}{2}$

* Q_1 : Rang $= \frac{N \times 25}{100} = \frac{57 \times 25}{100} = 14,25$

La classe de $Q_1 = [30-50[$

$Q_1 = 30 + (50-30) \times \frac{14,25-4}{20-4}$

$\boxed{Q_1 = 32,03}$

* Q_2 : Rang $= \frac{N \times 50}{100} = \frac{57 \times 50}{100} = 28,5$

La classe de $Q_2 = [50-70[$

$Q_2 = 50 + (70-50) \times \frac{28,5-20}{40-20}$

$\boxed{Q_2 = 29,75}$

* Decile : $D_1, D_2, D_3 \ldots\ldots D_{10}$

$\% = \boxed{10\%}$

$D_1 = \frac{N \times 10}{100}$

$D_2 = \frac{N \times 20}{100}$

Boîte à moustaches

Le diagramme en boîte à moustaches ou box-plot permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles.

La partie centrale de la distribution est représentée par une boîte de largeur arbitraire et de

longueur la distance interquartile, la médiane est tracée à l'intérieur.

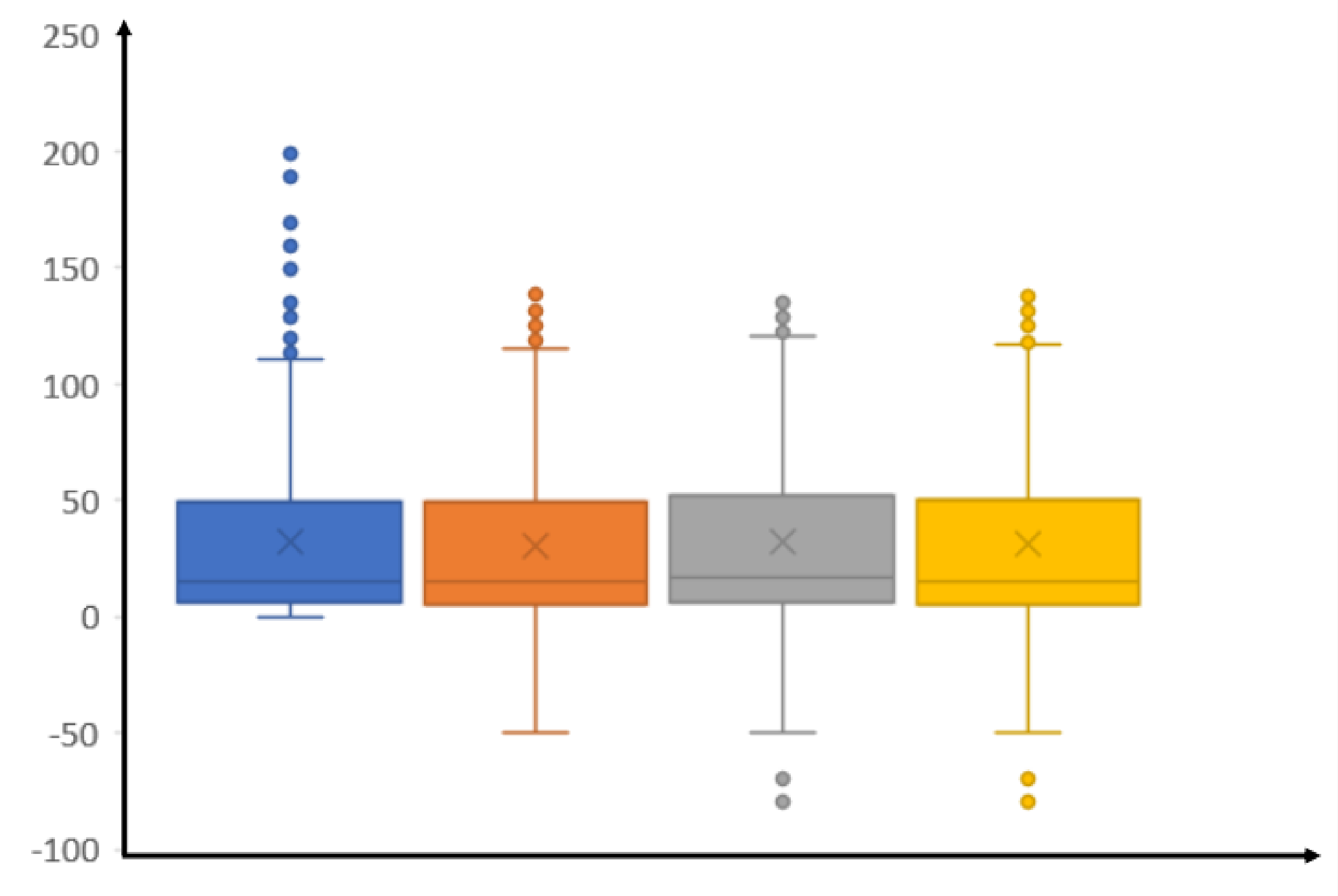
La boîte rectangle est complétée par des moustaches correspondant aux valeurs suivantes:

– Valeur supérieure : Min(la plus grande la valeur; Q3 + 1,5(Q3 – Q1))

– Valeur inférieure : Max(la plus petite valeur; Q1 - 1,5(Q3– Q1))

Les valeurs extérieures « aux moustaches » sont représentées

par des étoiles et peuvent être considérées comme aberrantes.



Coefficient de variation

comparer la distribution autour de la moyenne de deux variables statistiques de natures différentes: Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la moyenne est grande.

$$C_V = \frac{\sigma_X}{\overline{X}}$$

Les paramètres de dispertion

Le moment d'ordre r et moment centré d'ordre r

$$m_r = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^r = \sum_{i=1}^{k} f_i x_i^r$$

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^r = \sum_{i=1}^{k} f_i (x_i - \overline{x})^r$$

Le moment d'ordre r et moment centré d'ordre r

Les paramètres de dispertion

$$m_r = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^r = \sum_{i=1}^{k} f_i x_i^r$$

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i\left(x_i - \bar{x}\right)^r = \sum_{i=1}^{k} f_i\left(x_i - \bar{x}\right)^r$$

Coefficient d’asymétrie de Fisher et Coefficient d’aplatissement de Fisher

Paramètres de forme

’une variable suit une loi normale ou de Gauss si :
Le coefficient d’asymétrie doit être inférieur à |1|
Le coefficient d’aplatissement ou encore de concentration doit être inferieur à |1,5|

Il est défini par: $$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Si $\gamma_1 = 0$, la distribution est **symétrique** autour de la moyenne.

Si $\gamma_1 < 0$, la distribution est plus étalée **vers la gauche**.

Si $\gamma_1 > 0$, la distribution est plus étalée **vers la droite**.

Il est défini par: $$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Si $\gamma_2 = 0$, L’aplatissement est le même que celui de la **loi Normale** (de Gauss).

Si $\gamma_2 < 0$, la concentration des valeurs autour de la moyenne est faible: la distribution est **plus aplatie** que la loi Normale.

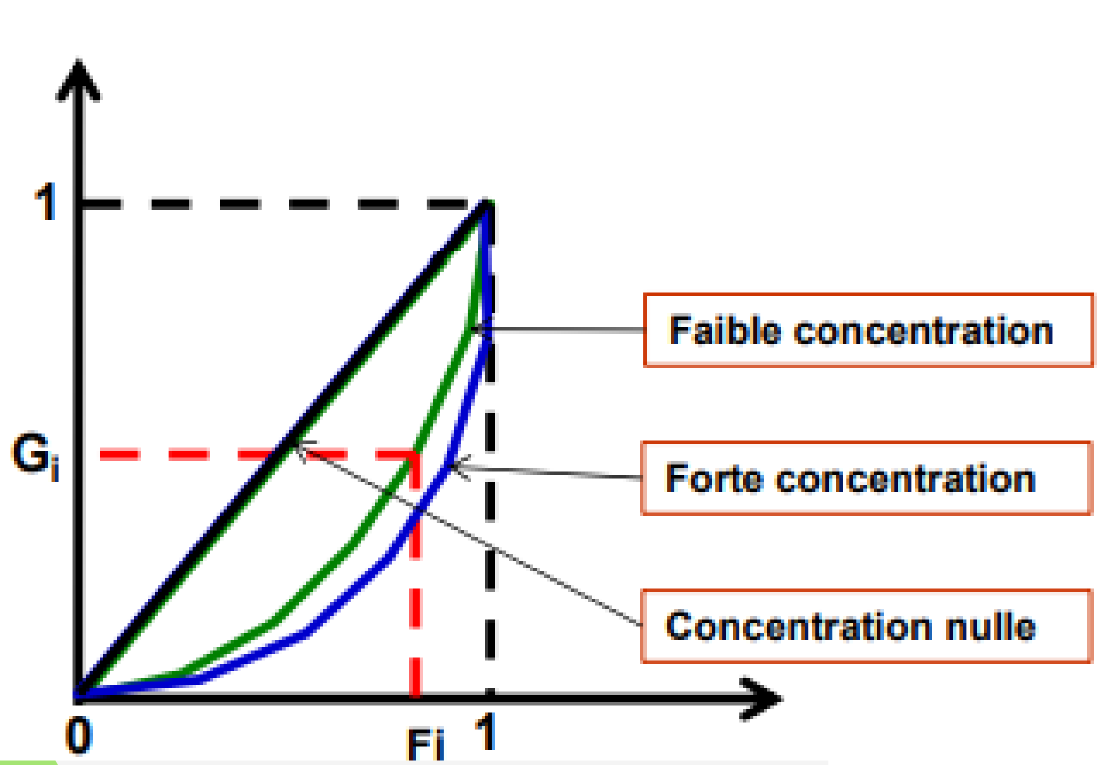
Si $\gamma_2 > 0$, la concentration des valeurs autour de la moyenne est forte: la distribution est **moins aplatie** que la loi Normale.

Exercice d'application page 777

Indice de concentration de Gini

La $i^{ème}$ classe $[x_{i-1}, x_i[$ a, pour centre, c_i et, pour effectif, n_i.

- $s_i = n_i c_i$ la masse de caractère X dans la classe $[x_{i-1}, x_i[$.
- $S = \sum_{i=1}^{k} s_i$ la masse globale de X
- $g_i = \frac{s_i}{S}$ la fréquence de la masse de X possédée par les individus dans la classe $[x_{i-1}, x_i[$.
- $G_i = \sum_{j=1}^{i} g_j$ La masse cumulée relative à la classe $[x_{i-1}, x_i[$.



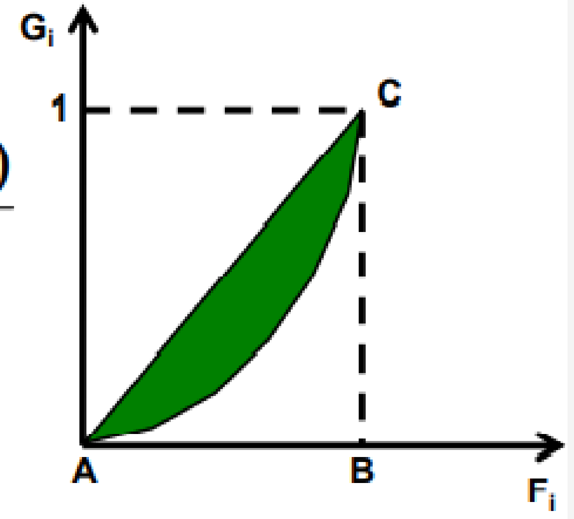
Paramètre de concentration

L'indice de Gini est égal à:

$$I_G = \frac{aire\ de\ concentration\ (en\ vert)}{aire\ du\ triangle\ ABC}$$

$$L'aire\ du\ triangle\ ABC = \frac{1 \times 1}{2} = 0{,}5$$

$$Donc\ I_G = 2 \times (aire\ de\ concentration)$$



'une variable suit une loi normale ou de Gauss si :
Le coefficient d'asymétrie doit être inférieur à |1|
Le coefficient d'aplatissement ou encore de concentration doit être inferieur à |1,5|

Séries doubles

Tableau de contingence

Modalités de Y

X \ Y	y_1	…	y_j	…	y_q	Total
x_1	n_{11}				n_{1q}	$n_{1.}$
…						…
x_i			n_{ij}			$n_{i.}$
…						…
x_p	n_{p1}				n_{pq}	$n_{p.}$
Total	$n_{.1}$	…	$n_{.j}$	…	$n_{.q}$	$n_{..}$

Modalités de X

Distribution marginale de X

Distribution marginale de Y

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

$$n_{\bullet\bullet} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

Eléments d'un tableau de contingence

Les effectifs

Les effectifs partiels: nij
Les effectifs marginaux

Les fréquences

Les fréquences partielles

$$f_{ij} = \frac{n_{ij}}{n_{\bullet\bullet}}$$

Les fréquences conditionnelles

- de X selon Y

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}}$$

- de Y selon X

$$f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$$

Les fréquences marginales

$$f_{i\bullet} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \quad \text{et} \quad f_{\bullet j} = \frac{n_{\bullet j}}{n_{\bullet\bullet}}$$

Relations entre les fréquences marginales et conditionnelles

$$f_{i\bullet} \times f_{j/i} = f_{\bullet j} \times f_{i/j} = f_{ij}$$

Indépendance de deux variables

Deux variables X et Y sont totalement indépendantes si les fréquences fi/j conditionnelles ne dépendent plus de j.

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = \ldots\ldots = \frac{n_{iq}}{n_{\bullet q}} = \frac{\sum_{j=1}^{q} n_{ij}}{\sum_{j=1}^{q} n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} = f_{i\bullet}$$

$$\Rightarrow \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \Rightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}} \Rightarrow \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \times \frac{n_{\bullet j}}{n_{\bullet\bullet}} \text{ donc } \boxed{f_{ij} = f_{i\bullet} \times f_{\bullet j}}$$

Exercice

Calculer les fréquences partielles et les fréquences marginales.
Montrer que les caractères X et Y sont indépendants.

X \ Y	y_1	y_2	Total
x_1	3	5	8
x_2	6	10	16
Total	9	15	24

La moyenne marginale, La variance marginale , L'écart-type, La covariance

$$\bar{x} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i$$ avec $$n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

$$V(x) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} (x_i - \bar{x})^2$$ ou $$V(x) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i^2 - \bar{x}^2$$

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

$$\operatorname{cov}(x, y) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

Exercice

Pour 25 ménages, les âges de l'époux et de l'épouse, relevés sur le registres d'un état civil sont les suivants:

(22,17); (23,18); (24,17); (24,18); (24,20); (24,21); (25,18); (25,19); (25,20); (26,18); (26,19); (26,21); (26,23); (27,19); (27,21); (28,21); (28,22); (30,22); (30,23); (31,24); (31,25); (34,24); (35,24); (35,25); (36,25);

Sachant que chaque couple (x_i,y_j) représente respectivement l'âge de l'époux et l'âge de l'épouse au moment de mariage.

1. Ranger les données en classes de même amplitude 5, qui commencent par 20 pour X et par 15 pour Y.
2. Calculer l'âge moyenne des époux et des épouses.
3. Calculer la variance de l'âge d'épouse, et son écart-type.
4. Calculer la covariance des deux variables

Etude de liaison entre deux variables

Nulle(aucune influence), totale, relative

Notion de corrélation

même sens, sens inverse

Coefficient de corrélation

- Si r est proche de 1, il y a une forte corrélation positive entre X et Y (même sens de variation)
- Si r est proche de -1, il y a une forte corrélation négative entre X et Y (différence du sens de variation).
- Si r=0, X et Y sont non corrélées : il n'y a pas d'association linéaire entre X et Y.
- Si r = ±1, alors chacune de ces deux variables peut définir l'autre d'une façon exacte

$$r = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Ajustement par la méthode des moindres carrés

trouver les coefficients a et b de la droite de régression y=ax+b, qui minimisent la distance quadratique entre et qui revient à minimiser: S(a,b)

$$S(a,b)=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$$

$$\hat{a}=\frac{\operatorname{cov}(X,Y)}{V(X)}=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{\frac{1}{n}\sum_{i=1}^{n}x_iy_i-\overline{xy}}{\frac{1}{n}\sum_{i=1}^{n}x_i^n-\overline{x}^2};\quad \hat{b}=\overline{y}-\hat{a}\overline{x}$$

Exercice

Soit X la note des mathématiques sur 20 points et Y la note de statistique sur 20 points pour 10 étudiants:

X	2	4	6	6	9	10	11	12	13	18
Y	3	6	6	7	9	10	10	11	14	14

1. Donner la droite des moindres carrés de Y en X ,
2. Donner la droite des moindres carrés de X en Y ,