Table des matières

Chapitre 1

Séries Statistique

1.1 Généralités

1.1.1 Population et échantillon

L'orsque on veut étudier les données relatives aux caractéristiques d'un ensemble d'individu ou d'objet (exemple : nombre d'enfants de sexe féminin dans une famille de N enfants, taille des étudiants de l'université, poids de nouveau-nés, etc) il est difficile d'observer toutes les données lorsque leur nombre est élevé.

Au lieu d'examiner l'ensemble qu'on appelle "*population*" (notée $\mathcal{P}$), on examine un nombre restreint qu'on appelle "*échantillon*". Pour être représentatif, l'échantillon doit être pris au hasard, une population peut-être finie ou infinie.

Elle est finie, si elle comporte un nombre déterminé d'individus ou d'objets, par exemple : population d'une ville à un instant donné. Elle est infinie si elle comporte un nombre infini d'individus ou d'objets, par exemple ; l'ensemble des résultats (face ou pile) lors de parties successives de pile ou face avec une même pièce de monnaie.

1.1.2 Description d'un échantillon

Unité statistique ou individu c'est l'élément de base constitutif de la population à laquelle il appartient. Il est indivisible et peut être un pays, un végétal, un humain ou une entreprise. Chaque individu peut être étudié relativement à un ou plusieurs caractères. L'ensemble des données numériques relatives à ces caractères constitue une série statistique ou distribution statistique. Un caractère peut présenter plusieurs modalités. Exemple : le caractère couleur des yeux à les modalités suivantes ; yeux noirs, bruns , verts, gris, etc.

On distingue plusieurs types de caractères, par exemple :

- couleur d'un certain type de fleur
- Poids ou taille nouveau nés.

Caractère quantitatif Un caractère est `quantitatif`, si ses diverses modalités sont mesurables c'est-à-dire, s'il est possible de faire correspondre un nombre à chaque modalité (Exemples :

poids, nombre d'enfants d'une famille ...) ce nombre qui varie d'une modalité à l'autre est appelé variable statistique.

Caractère quantitatif discontinu (discret) Un caractère quantitatif est `discontinu` s'il ne peut prendre que des valeurs isolées, appartenant à un certain intervalle, on dit aussi qu'il est `discret`. (Par exemple : nombre d'enfant dans une famille.)

Caractère quantitatif continu Un caractère quantitatif est `continu` : s'il peut prendre toute valeur appartenant à intervalle de variation ou encore lorsque les valeurs possibles de ce caractère sont des nombres réels. (Par exemple : taille de nouveau nés d'une population donnée.)

Caractère qualitatif Un caractère est quantitatif ; si ses diverses modalités ne sont pas mesurables. Exemple : (sexe, profession, couleurs des yeux ...)

1.1.3 Series statistique dans le cas d'un caractère quantitatif discontinu (discret)

Considérons un échantillon de taille N, c'est à dire composé de N éléments que nous supposerons numérotés de 1 á N, appelons X la valeur du caractère sur lequel porte l'étude, et soient $x_1, x_2, \ldots, x_N$ les valeurs de ce caractère pour les éléments $1, 2, \ldots, N$ de la série.

L'étendue de la série est l'écart qui sépare la plus grande et la plus petite valeur du caractère.

L'effectif total de la série est le nombre N d'éléments constituant l'échantillon étudié.

Lorsque la valeur x_i du caractère se rencontre un nombre n_i de fois dans la série statistique, on dit que n_i, est la répétition de x_i ou **l'effectif partiel** relatif à x_i ou encore la **fréquence absoue** de x_i .

La quantité $f_i = \frac{n_i}{N}$ rapport de la fréquence absolue de x_i à l'effectif total est la **fréquence relative** de x_i.

Il est clair que la somme des fréquences relatives est égale à l'unité ($\sum f_i = 1$).

Exemple 1.1.1. Répartition de 150 grenouilles suivant le nombre de vers trématodes (parasites) qu'elles hébergent.

Nombre des trimatodes par grenouille(x_i)	0	1	2	3	4	5	6
Nombre de grenouilles correspondantes (n_i)	11	22	45	40	19	11	2
Fréquence relative f_i	0.07	0.14	0.30	0.26	0.12	0.07	0.01

1.1.4 Series statistique dans le cas d'un caractère quantitatif continu

Dans le cas d'un caractère continu, le nombre de valeurs distinctes est en principe infini pour éviter une répartition de fréquences trop dispersée, pour permettre une étude commode, on constitue des classes en divisant l'étendue de la série en un certain nombre d'intervalles partiels, chacune des classes groupant ensemble, les mesures relatives à un même intervalle.

Définir une classe revient donc à fixer des limites, ou ce qui est équivalent à fixer le centre de classe et l'étendues de la classe qui est l'écart entre les deux limites.

Les classes sont contiguës et ne se chevauchent pas.

Chaque classe contiendra toutes les valeurs égales ou supérieures à sa limite inférieure mais strictement inférieures à sa limite supérieure. En général, les classe sont détendue égale, mais ceci n'a rien d'impératif. On peut avoir une répartition en classe d'étendue inégale, lorsqu'on s'impose un effectif minimum par classe.

Exemple 1.1.2. Poids de nouveau- nés les poids d'échelonnent entre 2,240 kg et 4,490 kg. Voir le tableau suivant :

classes	centre de classe	effectif n_i	f_i	pourcentage %
[2.2,2.5[	2.350	5	0.031	3.1
[2.5, 2.8[	2.650	11	0.068	6.8
[2.8, 3.1[	2.950	24	0.148	14.8
[3.1, 3.4[	3.230	40	0.248	24.8
[3.4, 3.7[	3.55	42	0.259	25.9
[3.7, 4.0[	3.850	20	0.124	12.4
[4.0, 4.3[	4.150	13	0.080	8.0
[4.3, 4.6[	4.45	6	0.037	3.7
Total		160	1	100

1.1.5 Séries statistique dans le cas d'un caractère qualitatif

Pour représenter les résultats d'une enquète relative à un caractère qualitatif. On groupe les résultats en un nombre de classes égal au nombre de modalitées du caractère étudié.

à chaque classe est associé son effectif n_i (fréquence absolue), ainsi que sa fréquence relative : $f_i = \frac{n_i}{N}$

Exemple 1.1.3. L'analyse du sang de 100 individus à donné les résultats suivants :

Groupe sangun	n_i	f_i	%
O	40	0.40	40
A	43	0.43	43
B	12	0.12	12
AB	5	0.05	5

1.1.6 Effectifs cumulé et fréquence relative cumulée :

Effectif cumulatif : On appelle `effectif cumulé croissant` à la i^{ieme} valeur x_i du caractère la somme notée $N_i \nearrow$ telle que $N_i \nearrow = n_1 + n_2 + ... + n_i$ des effectifs obtenus pour les i premiers valeurs du caractère, de même on appelle `effectif cumulé décroissant` à la i^{ieme} valeur x_i du caractère, la valeur suivante : $N_i \searrow = N - N_i \nearrow$.

Fréquence relative cumulative : On appelle `fréquence relative cumulée croissante` à la i^{ieme} valeur x_i du caractère la somme notée $F_i \nearrow$ telle que $Fi \nearrow = f_1 + f_2 + ... + f_i$ des

fréquences relatives obtenues pour les i premiers valeurs du caractère, de même on appelle appelle `fréquence relative cumulée décroissante` à la i^{ieme} valeur x_i du caractère, la valeur suivante : $F_i \searrow = 1 - F_i \nearrow$.

Exemple 1.1.4. Dans le cas de (**l'exemple 1.1.1**) (N=150).

Nbre de trimatôdes (x_i)	n_i	$N_i \nearrow$	$N_i \searrow$	f_i	$F_i \nearrow$	$F_i \searrow$
0	11	11	139	0.07	0.07	0.93
1	22	33	117	0.14	0.21	0.79
2	45	78	72	0.30	0.51	0.49
3	40	118	32	0.26	0.77	0.23
4	19	137	13	0.12	0.89	0.11
5	11	148	2	0.07	0.96	0.04
6	2	150	0	0.01	1.00	0

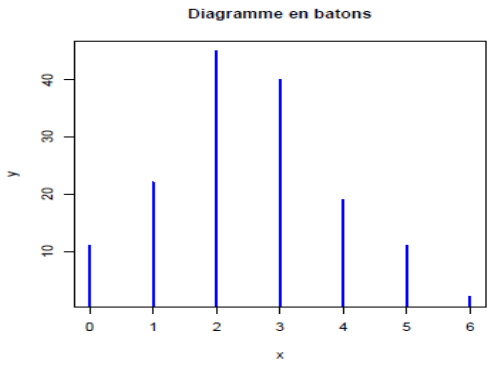
1.2 Représentations Graphiques

1.2.1 Caractère discret

Diagramme en bâtons

C'est un ensemble de bâtons ayant pour abscisses les valeurs $x_1, x_2, \ldots, x_n$ du caractère et en chacun de points d'abscise x_i une ordonnée proportionnelle à l'effectif n_i de x_i.

Exemple 1.2.1. Dans le cas de (**l'exemple 1.1.1**), on obtient le diagramme de bâtons suivant :



Polygone des effectifs

On l'obtient en joignant par des segments de droite les extrémités des bâtons. C'est un graphe linéaire passant par les points ayant pour abscisse x la valeur du caractère étudié et pour ordonner l'effectif correspondant.

Exemple 1.2.2. Dans le cas de (**l'exemple 1.1.1**), on obtient le polygone suivant :

Nombre des trimatodes par grenouille(x_i)	0	1	2	3	4	5	6
Nombre de grenouilles correspondantes (n_i)	11	22	45	40	19	11	2

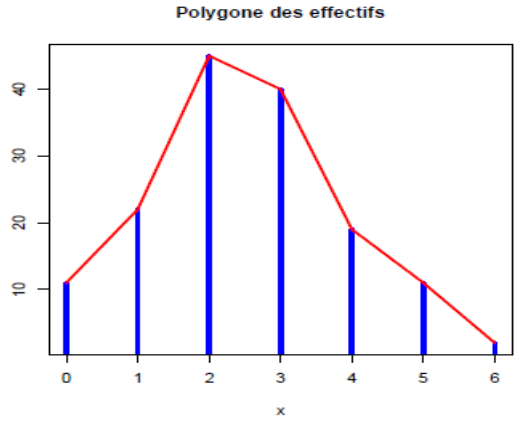
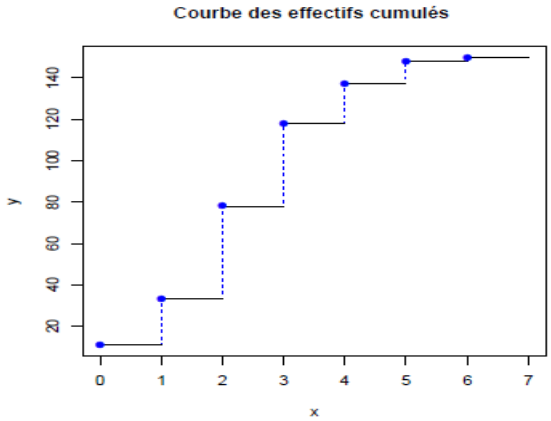


Diagramme cumulatif

Dans le diagramme cumulatif les bâtons ont des longueurs proportionnelles aux effectifs cumulés.

Exemple 1.2.3. Dans le cas de (**l'exemple 1.1.1**), on obtient le Diagramme cumulatif suivant :

Nbre de trimatôdes (x_i)	0	1	2	3	4	5	6
$n_i 0$	11	22	45	40	19	1	2
$N_i \nearrow$	11	33	78	118	137	148	150



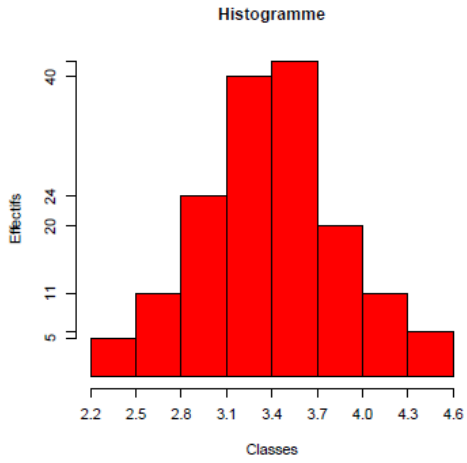
1.2.2 Caractères continus

Histogramme

C'est un ensemble de rectangles ayant pour largeur l'amplitude de la classe et pour hauteur l'effectif de la classe.

Exemple 1.2.4. Dans le cas de (**l'exemple 1.1.2**), on obtient l'histogramme suivant :

x_i	[2.2,2.5[	[2.5, 2.8[	[2.8, 3.1[	[3.1, 3.4[	[3.4, 3.7[	[3.7, 4.0[	[4.0, 4.3[	[4.3, 4.6[	Total
c_i	2.35	2.65	2.95	3.23	3.55	3.85	4.15	4.45	
n_i	5	11	24	40	42	20	13	6	160



Remarque 1.2.1. L'histogramme est la représentation graphique d'une variable continue. à chaque classe de la variable, correspond la surface d'un rectangle qui a pour base l'amplitude de classe. (L'amplitude est la différence entre la borne supérieure et la borne inférieure de la classe). Comme c'est la surface des rectangles qui représente les phénomènes étudiés, on remarque que :

- Si les amplitudes sont égales, alors les hauteurs des rectangles sont proportionnelles aux effectifs ou aux fréquences.
- La surface du rectangle représentant la i-ème classe sera ainsi égale à $f_i = n_i/N$.
- Si les amplitudes sont inégales, il faudra corriger la hauteur des rectangles de manière à ce que leur surface corresponde bien à n_i (les effectifs) ou f_i (les fréquences).
- Notons que l'histogramme a une signification statistique et que, dans ce cas, en réalité la somme de toutes les surfaces des rectangles de l'histogramme doit être égale à 1. Autrement dit, toutes les surfaces $S_i = h_i\,(x_i - x_{i-1}) = h_i.a_i = f_i,\ i = 1, \ldots, n$, où h_i est la hauteur correspondant au rectangle de la i-ème classe, $\sum_{i=1}^{n} S_i = \sum_{i=1}^{n} f_i = 1$.

Exemple 1.2.5. Exemple de l'étude de taille en cm d'une classe

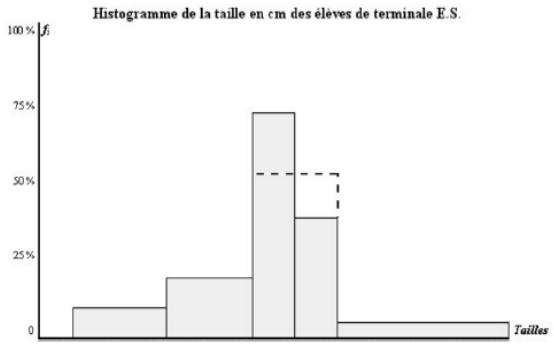
Classes x_i	Effectifs n_i	Fréquences $f_i\%$	Amplitudes a_i
[150-160[	3	8.57%	10
[160-170[	8	22.86%	10
[170-175[	13	37.14%	5
[175-180[	7	20%	5
[180-200[	4	11.43%	20
Total	35	100%	

Ici, les amplitudes ne sont pas égales. Comme c'est la surface des rectangles qui représente l'ampleur du phénomène étudié, il faudra corriger leur hauteur : par exemple, l'amplitude de la classe $[170-175[$ est de 5. Si l'amplitude de référence est 10, il faudra multiplier la hauteur par deux puisque la base de cette clase est égale à la moitié de l'amplitude de référence.

Pour simplifier le calcul, on peut ajouter deux colonnes au tableau de départ (Amplitude de référence = 10) :

Classes x_i	n_i	$f_i\%$	a_i	coefficient correcteur $c = 10/a_i$	Hauteur corrigée $h_i = f_i.c$
[150-160[	3	8.57%	10	1	8.57
[160-170[	8	22.86%	10	1	22.86
[170-175[	13	37.14%	5	2	74.28
[175-180[	7	20%	5	2	40
[180-200[	4	11.43%	20	0.5	5.57
Total	35	100%			

Représentation



Si la troisième et la quatrième classe avaient été regroupées, on aurait eu 20 personnes dans un intervalle, soit environ 57% de l'effectif de la terminale (ce qui est représenté par le trait en pointillés gras sur l'histogramme). On a donc gagné en précision : l'aire est la même, mais en divisant la classe $[170-180[$, on a gagné en précision.

Nombre de classes

En combien de classes partageons-nous les valeurs ? la réponse n'est pas unique. Soit N l'effectif total. Nous pouvons considérer dans ce cours trois réponses à titre d'exemple.

1. Une réponse $k \simeq \sqrt{N}$, $[\sqrt{N}]$ (la partie entière) ou encore $[\sqrt{N}]+1$.
2. Une réponse la formule de **STURGES** $k = 1 + 3.3 \log_{10} N$.
3. Une réponse la formule de **YULE** $k = 2,5\sqrt[4]{N}$.

L'intervalle de classe ou longueur de l'intervalle $a_i := \dfrac{\text{étendu}}{\text{le nombre de classes}} = \dfrac{X_{max} - X_{min}}{k}$.

Remarque 1.2.2.

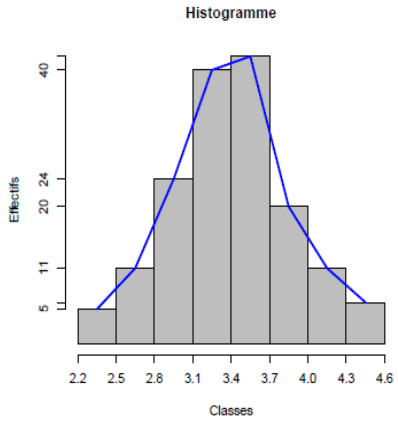
1. Pour le nombre de classes que comptera notre distribution de fréquences :
   (i) Il peut être choisi aribtrairement.
   (ii) Il peut être imposé.
   (iii) Il peut être fixé par une méthode mathématique (règle de **STURGES** ou **YULE**).
2. Si les classes sont d'amplitudes inégales, il faut d'abord réctifier l'effectif avant de tracer l'histogramme, pour celà on doit multiplier chaque classe par la rapport :

$$\frac{\text{amplitude la plus petite}}{\text{amplitude de la classe}}.$$

Polygone des effectifs

Le polygone des effectifs est la ligne brisée joignant les milieux des bases supérieures du différent rectangle adjacents. De même pour les fréquences relatives si en remplace les n_i par les f_i, on obtient le polygone des fréquences relatives.

Exemple 1.2.6. Voir l'exemple précédent.

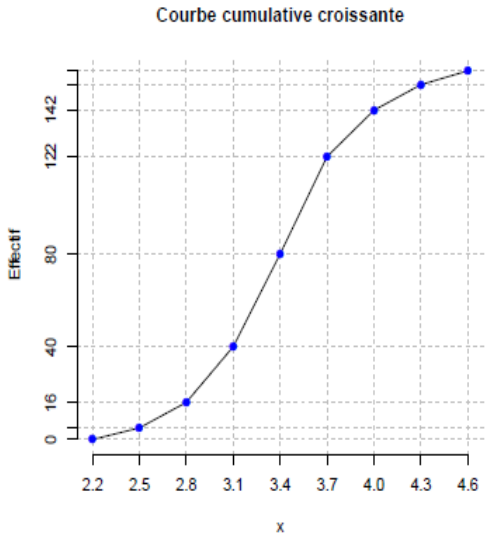


1.2.3 Polygone des effectifs cumulés

S'obtient en posant en ordonnée au droit de chaque limite de classe figurant sur l'axe des abscisses, la somme des effectifs de toutes les classes inférieures, la ligne brisée joignant les points, ainsi s'appelle le polygone des effectifs cumulés croissant.

De même pour les fréquences relatives cumulées si on remplace les $N_i \nearrow$ par les $F_i \searrow$, on obtient le polygone des fréquences relatives cumulées.

Exemple 1.2.7. Dans le cas de (**l'exemple 1.1.2**), on obtient la courbe cumulative suivante :



1.3 Paramètres de Position et Paramètres de Dispersion

1.3.1 Généralités

Les paramètres de position et de dispersion sont un ensemble de valeurs caractéristiques qui permettent une représentation condensée de l'information contenus dans la série statistique.

On distingue deux catégories de valeurs typiques :

- Les paramètres de position : La moyenne, le mode, la médiane, les quartiles donnent l'ordre de grandeur de l'ensemble des mesures,
- Les paramètres de dispersion : l'écart moyen, l'écart-type, semi-interquartile précisent le degré de dispersion des différentes valeurs d'une série autour d'une valeur centrale.

1.3.2 Cas d'une variable qualitative

Diagramme par secteur (diagramme circulaire)

Les diagrammes circulaires, ou semi-circulaires, consistent à partager un disque ou un demi-disque, en tranches, ou secteurs, correspondant aux modalités observées et dont la surface est proportionnelle à l'effectif, ou à la fréquence, de la modalité (voir Figure 1.1).

Les modalités d'un caractère qualitatif n'étant pas ordonnées, on les représente par des graphiques qui utilisent des surfaces : représentation en cercle ou demi-cercle, carrés, tuyaux, etc, ou des volumes : sphères, cônes, cylindres, etc. Comme on ne peut pas leur appliquer les techniques de calcul utilisées avec les nombres, c'est-à-dire que l'on ne peut pas en donner un résumé par quelques chiffres significatifs. L'étude graphique constitue donc une partie importante de l'analyse de ce type de caractères, puisque la description d'une population selon une variable qualitative est totalement résumée dans un tableau de pourcentages ou dans un diagramme circulaire, appelé aussi diagramme en "camembert". Le degré d'un secteur est déterminé à l'aide

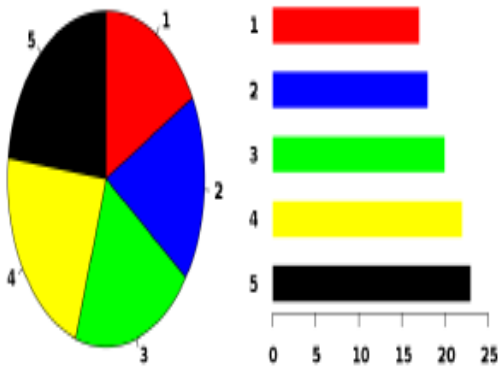


FIG. 1.1 – Diagramme par secteur

de la règle de trois de la manière suivante :

$$\begin{cases} \text{N} \longrightarrow 360^{\circ} \\ \text{n}_i \longrightarrow d_i \text{ (degré de la modalité } i\text{)}. \end{cases}$$

Ainsi $d_i = \dfrac{n_i \times 360}{N}$.

1.4 Pramètres de Position

1.4.1 Moyenne arithmétique

- **Série statistique d'un caractère discret** Soit l'ensemble des mesures d'une même variable X :

$x_1, x_2, \ldots, x_N$, la moyenne arithmétique notée $\bar{x}$ $\bar{x}$ est définie par :

$$m = \bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Exemple 1.4.1. La moyenne arithmétique des valeurs 8, 5, 3, 6,2.

$$\bar{x} = \frac{8+5+3+6+2}{5} = \frac{24}{5} = 4,8.$$

Lorsque les valeurs $x_1, x_2, \ldots, x_N$ se répètent respectivement, on obtient la moyenne arithmétique en comptant chaque valeur x_i autant de fois qu'elle se présente : ceci revient à pondérer la valeur x_i par n_i qui lui correspond. On aura

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$$

tel que $N = \sum_{i=1}^{k} n_i$, $f_i = \frac{n_i}{N}$.

Exemple 1.4.2. Si les valeurs 8,5,3,6,2 se produisent respectivement 1,4,2,2,1 la moyenne arithmétique est

$$\bar{x} = \frac{(8\times 1) + (5\times 4) + (3\times 2) + (6\times 2) + (2\times 1)}{10} = 4,8.$$

- **Série statistique d'un caractère continu** La moyenne arithmétique pour une variable continue est donnée par la formule suivante :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$$

où $c_i = \frac{x_i + x_{i+1}}{2}$ est le centre de la classe $[x_i, x_{i+1}[$.

Exemple 1.4.3. Soit la distribution

Classes	Effectif n_i	centre de classe c_i	$n_i \times c_i$
[8-10[	1	9	9
[10-12[	2	11	22
[12-14[	4	13	52
[14-16[	6	15	80
[16-18[	5	17	85
[18-20[	2	19	38

$$\bar{x} = \frac{\sum n_i c_i}{N} = \frac{296}{20} = 14,8.$$

Exemple 1.4.4. Dans le cas de (**l'exemple 1.1.1**) :

Nbre de trimatôdes (x_i)	n_i	$n_i x_i$
0	11	0
1	22	22
2	45	90
3	40	120
4	19	76
5	1	55
6	2	12
Total	150	375

$$\bar{x} = \frac{\sum n_i x_i}{N} = \frac{375}{150} = 2,5 \ \ trematodes.$$

Exemple 1.4.5. Dans le cas de (**l'exemple 1.1.2**) :

classes	centre de classe	effectif n_i	$n_i c_i$
[2.2,2.5[	2.350	5	11,75
[2.5, 2.8[	2.650	11	29,15,
[2.8, 3.1[	2.950	24	70,80
[3.1, 3.4[	3.230	40	130,00
[3.4, 3.7[	3.55	42	149,10
[3.7, 4.0[	3.850	20	77,00
[4.0, 4.3[	4.150	13	53,95
[4.3, 4.6[	4.45	6	26,70
		161	=548,45

$$\bar{x} = \frac{\sum n_i c_i}{N} = \frac{458}{161} = 3,4.$$

D'autres moyennes

1. **La moyenne géométrique :** C'est la moyenne applicable à des mesures de grandeurs dont la croissance est géométrique ou exponentielle. La moyenne géométrique conserve le produit des x_i : si on modifie les valeurs de deux observations tout en conservant leur produit, la moyenne géométrique sera inchangée.

   La moyenne géométrique G de la série de valeurs $x_1, \ldots, x_i, \ldots, x_k$ supposées toutes positives (strictement), est définie ainsi lorsque chaque valeur x_i n'intervient qu'une seule fois :

   $$G = \sqrt[N]{\prod_{i=1}^{k} x_i} = \sqrt[N]{x_1 \times x_2 \times \ldots \times x_k} \Longrightarrow \ln G = \frac{1}{N} \sum_{i=1}^{k} \ln x_i$$

   Lorsque la distribution de la variable statistique est donnée par les k couples (x_i, n_i) (moyenne géométrique pndérée), les x_i étant tous positifs ; la moyenne géométrique a pour expression :

   $$G = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}} = \prod_{i=1}^{k} x_i^{f_i} = \sqrt[N]{x_1^{n_1} \times x_2^{n_2} \times \ldots \times x_k^{n_k}} \Longrightarrow \ln G = \frac{1}{N} \sum_{i=1}^{k} n_i \ln x_i = \sum_{i=1}^{k} f_i \ln x_i$$

Exemple 1.4.6. (i) Supposons que pendant une décennie, les salaires aient été multipliés par 2 et que pendant la décennie suivante, ils aient été multipliés par 4 ; le coefficient multiplicateur moyen par décennie est égal à :

$$\sqrt{2.4} = \sqrt{8} \simeq 2,83$$

La moyenne arithmétique ($\bar{x} = 3$) n'est pas égale au coefficient demandé.

Prenons, par exemple, un salaire de $300DA$ au début de la première décennie, il sera de $300.2.4 = 2400DA$ au bout des vingt ans, ce qui équivaut à $300.(2,83)^2$, soit un coefficient multiplicateur moyen de $2,83$ par décennie.

(ii) Au cours des 4 dernières années, le taux de croissance annuels de la production intérieur brut (PIB) ont été les suivants :

$1^{\text{ère}}$ Année $\longrightarrow +7,2\%$

$2^{\text{ème}}$ Année $\longrightarrow +6,3\%$

$3^{\text{ème}}$ Année $\longrightarrow +7,0\%$

$4^{\text{ème}}$ Année $\longrightarrow +4,8\%$

Quel est le taux moyen de croissance de la PIB au cours de ces 4 années ?

$$G = \sqrt[4]{107,2 \times 106,3 \times 107 \times 104,8} \Longrightarrow$$

$$\ln G = \frac{1}{4}\left(\ln 107,2 + \ln 106,3 + \ln 107 + \ln 104,8\right) = 2,027$$

donc $G = 10^{2,027} = 106,3\%$.

Au cours de cette période, le taux moyen de croissance est de $6,3\%$.

(iii) 3 équipes se sont succédées à la direction d'une entreprise pendant la 1-ère période qui a durée 4 ans les bénifices réalisés ont augmenté de 50% par an, pendant la seconde période de 3 ans de 17% par an et pendant la dernière période de 2 ans, les bénifices ont enregistré une baisse de 30% par an.

Quel est le taux de croissance annuel moyen des bénifices réalisé au cours de ces 9 années ?

$G = \sqrt[9]{150^4 \times 117^3 \times 70^2} \Longrightarrow \ln G = \frac{1}{9}\left(4\ln 150 + 3\ln 117 + 2\ln 70\right) = 2,06$

donc $G = 10^{2,06} = 114,8\%$.

L'augmentation moyenne annuelle est donc de $14,8\%$.

2. **La moyenne harmonique** La moyenne harmonique est l'inverse de la moyenne arithmétique des inverses des valeurs. L'inverse de la moyenne harmonique conserve ainsi la somme des inverses des x_i : si on modifie les valeurs de deux observations tout en conservant la somme de leurs inverses, la moyenne harmonique sera inchangée.

$$H = \frac{N}{\sum_{i=1}^{k} \frac{1}{x_i}} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots \frac{1}{x_k}}$$

La moyenne harmonique pondérée est donnée par

$$H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots \frac{n_k}{x_k}}$$

La moyenne harmonique peut être utilisée lorqu'il est possible d'attribuer un sens réel aux inverses des données en particulier pour les taux de change, les taux d'équipement, le pouvoir d'achat, les vitesses. Elle est notamment utilisée dans les calculs d'indices .

Exemple 1.4.7.

(i) On achète des dollars une première fois pour 100 DA au cours de 1,23 DA le dollar, une seconde fois pour 100 DA au cours de 0,97 DA le dollar. Le cours moyen du dollar pour l'ensemble de ces deux opérations est égal à :

$$\frac{200}{\frac{100}{1.23} + \frac{100}{0,97}} \simeq 1,085DA$$

La moyenne arithmétique (= 1,1) ne représente pas le cours moyen du dollar.

(ii) Une entreprise de transport possède 3 camions qui effectues des rotations entre Alger-Oran. Au cours d'une celles-ci, le trajet Alger-Oran a été couvert aux vitesses moyennes suivantes : 40km/h, 60km/h et 80km/h.

– Quelle est la vitesse moyenne ?

$$H = \frac{3}{\frac{1}{40} + \frac{1}{60} + \frac{1}{80}} = 56,6km/h.$$

(iii) Une entreprise de transport possède 10 camions qui font des rotations entre 2 villes. Au cours d'une celles-ci, le trajet a été couvert par ces véhicules aux vitesses moyennes présentées comme suit :

Vitesse moyenne km/h	40	60	70
nbre de camions	4	4	2

– Quelle est la vitesse moyenne globale ?

$$H = \frac{10}{\frac{4}{40} + \frac{4}{60} + \frac{2}{70}} = 51,5km/h.$$

Comparaison des 3 moyennes étudiées On montre que si les x_i sont tous positifs :

$$\min_{1\leq i\leq n} x_i \leq H \leq G \leq \bar{x}\bar{x} \leq \max_{1\leq i\leq n} x_i$$

L'égalité de deux de ces moyennes entre elles entraîne leur égalité dans leur ensemble, et dans ce cas, toutes les valeurs x_i sont égales.

Exemple 1.4.8. Un étudiant à obtenu les notes suivantes : Math=14 (coefficient 3), Stat =8 (coefficient 2) et Physique=6 (coefficient 1).

– Calculer les moyennes pondérées.

1. $\bar{x} = \dfrac{7 \times 3 + 4 \times 2 + 3 \times 1}{6} = \dfrac{32}{6} = 5,33.$
2. $G = \sqrt[6]{7^3 \times 4^2 \times 3^1} = (343 \times 16 \times 3)^{1/6} = 16464^{1/6} = 5,01.$
   $\ln G = \dfrac{1}{6}(3\ln 7 + 2\ln 4 + \ln 3) = 0,7 \Longrightarrow\ G = 10^{0,7} = 5,01.$
3. $H = \dfrac{6}{\dfrac{3}{7} + \dfrac{2}{4} + \dfrac{1}{3}} = \dfrac{6}{1,26} = 4,76.$

Dans ce cas la moyenne arithmétique est la plus indiquée.

1.4.2 La médiane

• **Série statistique d'un caractère discret** Si les valeurs du caractère d'une série statistique sont ordonnées par ordre de grandeurs croissantes ou décroissantes la médiane (notée Me) est la valeur qui se situe au centre de la série ainsi ordonnée.

Si la série possède un nombre impaire de valeurs soit $2n+1$ la médiane sera la $(n+1)^{ieme}$ valeur.

Par exemple : Dans la série de 15 observations

$$1, 2, 4, 4, 4, 5, 6, \mathbf{(7)}, 8, 8, 9, 9, 10, 11, 12.$$

La médiane : $Me = 7$.

Si la série possède un nombre paire de valeurs soit $2n$ la médiane sera la demi-somme de la n^{ieme} et la $(n+1)^{ieme}$ valeur.

Par exepmle : Dans la série de 16 observations

$$1, 2, 4, 4, 4, 5, 6, \mathbf{(7,8)}, 8, 9, 9, 10, 11, 12, 14.$$

La médiane : $Me = \frac{7+8}{2} = 7,5$.

Remarque 1.4.1. La médiane n'a pas toujours de signification au point de vu statistique dans le cas des séries d'un caractère discret, c'est en particulier lorsque plusieurs valeurs du caractère coïncident avec la valeur de la médiane.

Par exemple : Dans la série suivante (20 observations) :

$$2, 2, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 8, 8, 9$$

la médiane est comprise entre la 10^{ieme} et la 11^{ieme} valeur, c'est donc 5 cependant la répétition de la valeur 5 ne permet d'autre conclusion que celle-ci : 9 valeurs sont inférieures à la médiane et 8 valeurs lui sont supérieures : la médiane ne réalise pas d'équipartition effective de la série.

• **Série statistique d'un caractère continu** Pour déterminer la médiane Me, Il faut d'abord déterminer la classe médiane $[x_i, x_{i+1}[$, c'est la classe qui correspond la moitié des effectifs(ou fréquences relative) cumulés. La médiane est obtenue par interpolation définie par :

$$Me = x_i + a_i \left(\frac{\frac{N}{2} - N_i \nearrow}{n_i} \right)$$

où $\begin{cases} N & \text{est l'effectif total} \\ x_i & \text{est la borne inférieure de la classe médiane.} \\ a_i & \text{est l'amplitude de la classe médiane.} \\ N_i \nearrow & \text{est l'effectif cumulé de la classe qui précède la classe médiane.} \\ n_i & \text{est l'effectif de la classe médiane.} \end{cases}$

ou encore

$$Me = x_i + a_i \left(\frac{\frac{1}{2} - F_i \nearrow}{f_i} \right)$$

où $\begin{cases} N & \text{est l'effectif total} \\ x_i & \text{est la borne inferieure de la classe médiane.} \\ a_i & \text{est l'amplitude de la classe médiane.} \\ F_i \nearrow & \text{est la fréquence relative cumulée de la classe qui précède la classe médiane.} \\ f_i & \text{est la fréquence relative de la classe médiane.} \end{cases}$

Exemple 1.4.9. On considère la serie statistique suivante

Classes	Effectif n_i	centre de classe c_i	effectifs cumulés $N_i \nearrow$
[38,40[	11	39	11
[40-42[	28	41	39
[42-44[	16	43	55
[44-46[	25	45	80
[46-48[	15	47	95
[48-50[	5	49	100

La classe médiane est $[x_i, x_{i+1}[= [42, 44[$ *se situent les observations comprisent entre* 39 *et* 55 *donc en particulier l'observation de rang* 50 *qui correspond à la médiane* Me *d'où*

$$Me = 42 + (44 - 42)\frac{50 - 39}{16} = 43,375.$$

`Détermination graphique de la médiane :` *La médiane partage la série en deux groupes de même effectif :*

La médiane est l'abscisse du point ayant pour ordonnée $1/2$ *sur le polygone des fréquences relatives cumulées.*

La médiane est l'abscisse du point ayant pour ordonnée $N/2$ *sur le polygone des effectifs cumulés.*

Par exemple : *Voir la courbe suivante*

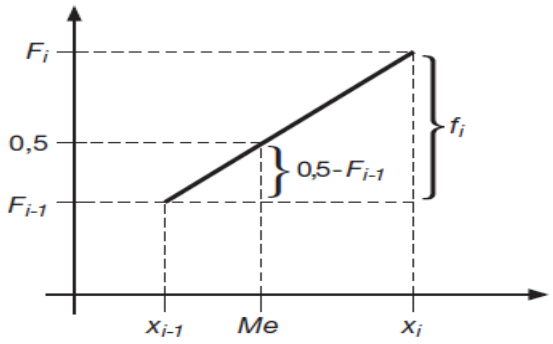


FIG. 1.2 – Médiane graphiquement pour variable continue $F_i \nearrow$ ou $N_i \nearrow$

1.4.3 Les Quartiles

La médiane partage la série en deux parties de même effectif. Les quartiles partage la série en Quatre parties de même effectif.

La fréquence relative cumulée jusqu'au premier quartile Q_1 est au plus $1/4$ alors que la fréquence relative cumulée des valeurs qui sont supérieures à Q_1 est au plus $3/4$.

Le second quartile se confond avec la médiane.

Exemple 1.4.10.

$$\begin{array}{cccccccc} 4 & 5 & 6 & 11 & 13 & 14 & 16 \\ & | & & | & & | & \\ & Q_1 & & Me & & Q_3 & \\ & & & || & & & \\ & & & Q_2 & & & \end{array}$$

Pour la série

Exemple 1.4.11.

$$\begin{array}{cccccccccc} 4 & 5 & 6 & 7 & 11 & 13 & 14 & 15 & 16 & 17 \\ & & | & & | & & | & & & \\ & & Q_1 & & Me & & Q_3 & & & \end{array}$$

Le nombre inférieurs à Q_1 est donné par $1/4 \times 9 = 2.25$ c'est à dire deux termes et Q_3 est le symétrique de Q_1 par rapport à la médiane d'où : $Q_1 = 6$, $Q_2 = 11$, et $Q_3 = 14$.

1.4.4 Détermination graphique des quartiles

Le premier quartile ou le quartile inférieur à Q_1 est l'abscisse du point d'ordonnée $1/4$ sur le polygone des fréquences relatives cumulées.

Le quartile Q_3 est l'abscisse du point d'ordonnée $3/4$ sur le polygone des fréquences relatives cumulées.

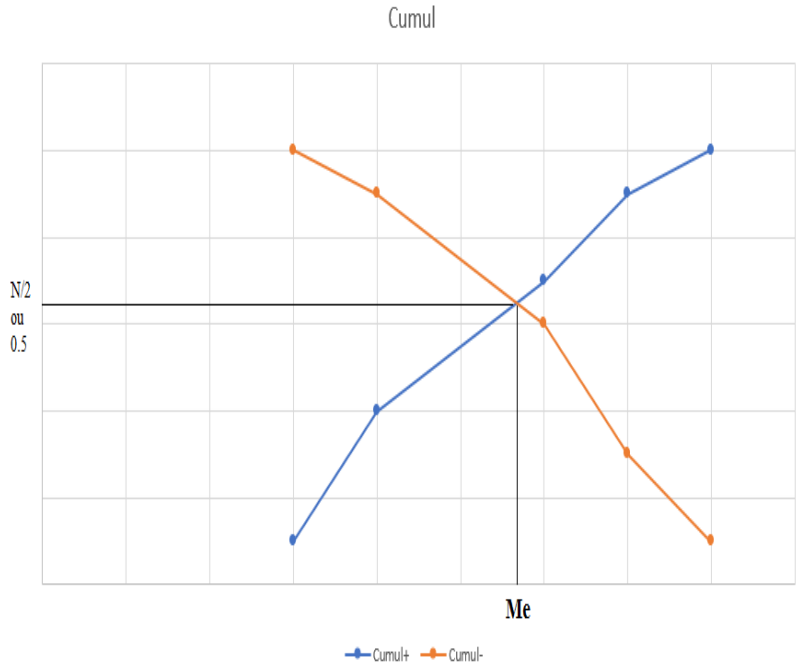


FIG. 1.3 – Détermination graphique de la médiane pour une variable continue $F_i \nearrow\searrow$ ou $N_i \nearrow\searrow$

1.4.5 Le Mode

Le **Mode** (noté Mo) d'une série statistique est la valeur de la variable statistique la plus fréquente.

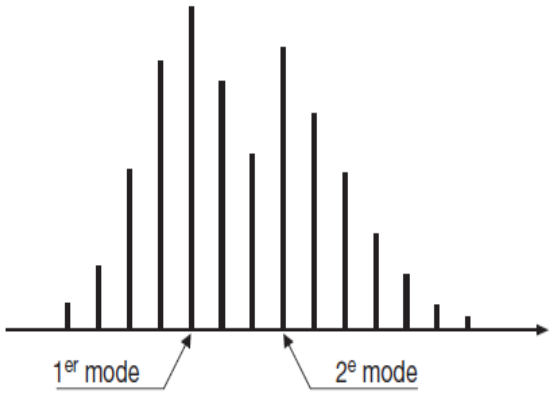


FIG. 1.4 – Représentation graphique du mode (cas discret).

Exemple Dans le cas de l'exemple 1.1.1. le mode est $Mo = 2$.

Si le polygone ne présente qu'un seul point, la série est dite unimodale. S'il présente plusieurs modes on dira alors qu'elle est multimodale.

Dans le cas d'une série statistique à classes : Il faut déterminer d'abord la clase modale

$[x_i, x_{i+1}[$, puis le mode est donné par la formule d'interpolation suivante

$$\text{Mo} = x_i + a_i \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

où

$\begin{cases} x_i & \text{est la borne inférieure de la classe modale.} \\ a_i & \text{est l'amplitude de la classe modale.} \\ \alpha_1 & \text{est la différence entre l'effectif de la classe modale est l'effectif de la classe qui la précède,} \\ \alpha_2 & \text{est la différence entre l'effectif de la classe modale est l'effectif de la classe qui la suivre.} \end{cases}$

Exemple 1.4.12. Dans le cas de l'exemple 1.1.2(Nouveau-nés).

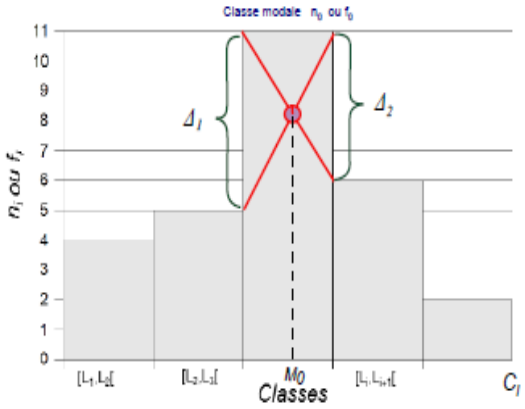


FIG. 1.5 – Détermination graphique du mode (cas continu).

L'expression du mode donnée ci-dessus est déterminée à partir de l'intersection des deux segments représentés dans la Figure 1.5. Cette notion n'est pas unique.

- Déciles divisent la population en 10 parties égales $(D_1, D_2, \ldots, D_{10})$, les déterminations approchées sont du même type que celles indiquées pour les quartiles. On peut remarquer que $D_5 = Q_2 = Me$.
- Percentiles divisent la population en 100 parties égales $(P_1, P_2, \ldots, P_{100})$, les déterminations approchées sont du même type que celles indiquées pour les quartiles. On peut remarquer que $P_{50} = D_5 = Q_2 = Me$.

Remarque :

• L'intervalle interdéciles $D = D_9 - D_1$ contient 80% des observations en laissant 10% des observations à droite et à gauche. Autrement dit on élimine 10% des valeurs se trouvant aux extrémités des distributions.

1.5 Paramètres de Dispersion

1.5.1 Variance

Introduction et exemple

La *variance et l'écart-type* sont des valeurs qui indiquent la dispersion des données par rapport à la moyenne.

1.5.2 Variance

Définition 1.5.1. Soit X une série statistiques telles que $x_1, x_2, \ldots, x_p$ sont les p valeurs de cette série, et $n_1, n_2, \ldots, n_p$ les effectifs associé à ces valeurs. Soit $N = n_1 + n_2 + \ldots + n_p$ l'effectif total. La variance de cette série statistique est la moyenne des carrés des écarts à la moyenne. Ce nombre noté $V(X)$, vaut donc :

- **Série statistique d'un caractère discret**

$$V(X) = \frac{n_1(x_1 - \bar{x})^2 + \ldots + n_p(x_p - \bar{x})^2}{n_1 + \ldots + n_p} = \frac{1}{N}\sum_{i=1}^{p} n_i(x_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(x_i - \bar{x})^2$$

La variance peut être aussi calculée en utilisant la formule suivante :

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

- **Série statistique d'un caractère continu**

$$V(X) = \frac{n_1(c_1 - \bar{x})^2 + \ldots + n_p(c_p - \bar{x})^2}{n_1 + \ldots + n_p} = \frac{1}{N}\sum_{i=1}^{p} n_i(c_i - \bar{x})^2 = \sum_{i=1}^{p} f_i(c_i - \bar{x})^2$$

La variance peut être aussi calculée en utilisant la formule suivante :

$$V(X) = \left(\frac{1}{N}\sum_{i=1}^{p} n_i c_i^2\right) - \bar{x}^2 = \left(\sum_{i=1}^{p} f_i c_i^2\right) - \bar{x}^2$$

Pour calculer la variance, il faut avoir calculer préalablement la moyenne.

Ecart-type

Définition 1.5.2. L'écart-type d'une série statistique, noté $\sigma(x)$ est la racine carrée de la variance :

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

Valeurs centrales des variables centrées

1. Moment centré d'ordre k noté μ_k est donné par

$$\text{variable discrète} \quad \mu_k(x) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \bar{x})^k = \sum_{i=1}^{n} f_i(x_i - \bar{x})^k$$

$$\text{variable continue} \quad \mu_k(x) = \frac{1}{N}\sum_{i=1}^{n} n_i(c_i - \bar{x})^k = \sum_{i=1}^{n} f_i(c_i - \bar{x})^k$$

2. Moment d'ordre k noté m_k est donné par

$$\text{variable discrète} \quad m_k(x) = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^k = \sum_{i=1}^{n} f_i x_i^k$$

$$\text{variable discrète} \quad m_k(x) = \frac{1}{N}\sum_{i=1}^{n} n_i c_i^k = \sum_{i=1}^{n} f_i c_i^k$$

On remrque que :

$$\mu_2(x) = Var(x) = \frac{1}{N}\sum_{i=1}^{n} n_i(x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2 = m_2 - m_1^2$$

$$\mu_3(x) = m_3 - 3m_1m_2 + 2m_1^3$$

$$\mu_4(x) = m_4 - 4m_1m_3 + 6m_1^2m_2 - 3m_1^4$$

Mesure de dispersion relative

On l'appelle aussi coefficient de dispersion relative ou le coefficient de variation et on le note par CV, il permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations et de comparer la dispersion dans différentes série. Il donne une bonne idée du degré d'homogénéité d'une série. Il faut qu'il soit le plus faible possible ($< 15\%$ en pratique).

$$CV = \frac{\sigma}{\bar{x}} = \frac{\sqrt{Var(x)}}{\bar{x}}$$

Coefficient interquartile

De par la définition des quartiles, l'intervalle interquartile $[Q1, Q3]$ contient 50% des observations. Sa longueur, notée $e_Q = Q_3 - Q_1$ (étendue InterQuartile), est un indicateur de dispersion, à partir de là on défini le coefficient interQuartile noté C_Q donné par

$$C_Q = \frac{Q_3 - Q_1}{Me} \times 100$$

Etendue

L'étendue d'une série statistique est la différence entre la plus grande et la plus petite valeurdu caractère.

Ecart-absolu moyen

Soit $\bar{x}$ la moyenne des valeurs $x_1, x_2, \ldots, x_n$, l'écart absolue moyen est la moyenne des écarts absolus.

$$\bar{e} = \frac{\sum_i |x_i - \bar{x}|}{N}$$

si n_i désigne la fréquence absolue de x_i l'écart absolu moyen est :

$$\bar{e} = \frac{\sum_i n_i |x_i - \bar{x}|}{N}$$

Ecart-absolu interquartile

L'ecart interquartile est la différence entre le troisième et le premier quartile.

$$e_Q = Q_3 - Q_1$$

1.6 Paramètres de Forme

Les caractéristiques de forme permettent de préciser l'allure de la courbe des effectifs ou des fréquences sans besoin de les tracer par rapport à une courbe idéale dite **courbe normale**. Nous définissons 2 sortes de caractéristiques.

- Caractéristiques d'asymétrie.
- Caractéristiques d'aplatissement.

– **Asymétrie** Une distribution est symétrique si les 3 caractéristiques de tendance sont confondues (Figure 1.6).

(i) $Mo = Me = \bar{x}$.

(ii) Q_1 et Q_3 équidistant de la médiane (situés à égale distance).

(iii) D_1 et D_9 équidistant de la médiane (situés à égale distance).

La position respective de ces élélements va permettre le calcul de coefficient d'asymétrie, qui sont nul en cas de symétrie et s'éloignent plus de la valeur 0 que la distribution est plus asymétrique.

Les principaux coefficient d'asymétrie ne sont valable que si les séries statistiques sont unimodales (l'existence d'un mode unique).

1. Coefficient de Pearson noté β_1 et $\beta_2 = F_1^2$ est nul pour une distribution symétrique, négatif pour une distribution unimodale étalée vers la gauche, positif pour une distribution unimodale étalée vers la droite (Figure 1.6), donné par :

$$\beta_1 = \frac{\bar{x} - Mo}{\sigma}$$

(i) Si $\beta_1 = 0$ on a une distribution symétrique.

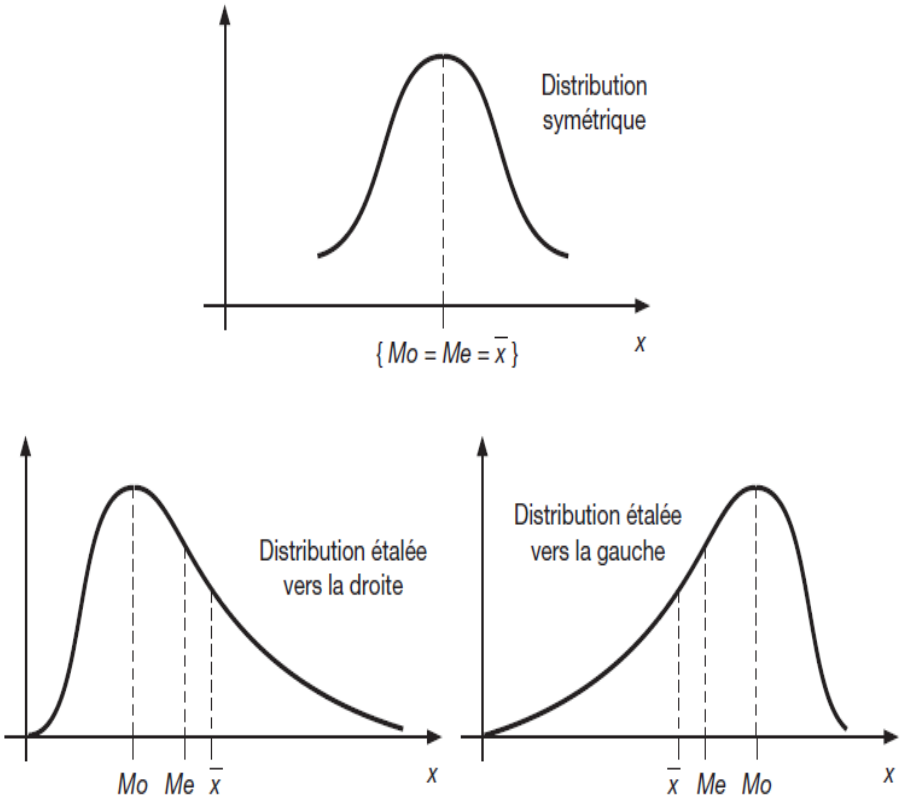


FIG. 1.6 – Positions respective du mode, de la médiane et de la moyenne

(ii) Si $\beta_1 > 0$ on a un étalement à droite, la distribution est plus accentué à gauche, autrement dit Mo est à gauche de Me (courbe oblique à gauche).

(iii) Si $\beta_1 < 0$ on a un étalement à gauche, la distribution est plus accentué à droite, autrement dit Mo est à droite de Me (courbe oblique à droite).

$$F_1^2 = \beta_2 = \frac{\mu_3^2}{(Var(x))^3} = \frac{\mu_3^2}{(\mu_2^3}$$

(i) Si $\beta_2 = 0$ on a une distribution symétrique.

(ii) Si $\beta_1 \neq 0$ on a une courbe asymétrique

$$\begin{cases} \mu_3 > 0, & \text{étalement à droite (Mo est à gauche de Me)}\,; \\ \mu_3 < 0, & \text{étalement à gauche (Mo est à droite de Me).} \end{cases}$$

2. Coefficient de Fisher noté F_1 ou γ_1 il est nul pour une distribution symétrique, négatif pour une distribution unimodale étalée vers la gauche, positif pour une distribution unimodale étalée vers la droite (Figure 1.7), donné par :

$$\gamma_1 = F_1 = \frac{\mu_3}{\sigma^3}$$

(i) Si $\gamma_1 = F_1 = 0$ on a une distribution symétrique.

(ii) Si $\gamma_1 = F_1 > 0$ on a un étalement à droite, la distribution est plus accentué à gauche, autrement dit Mo est à gauche de Me (courbe oblique à gauche).

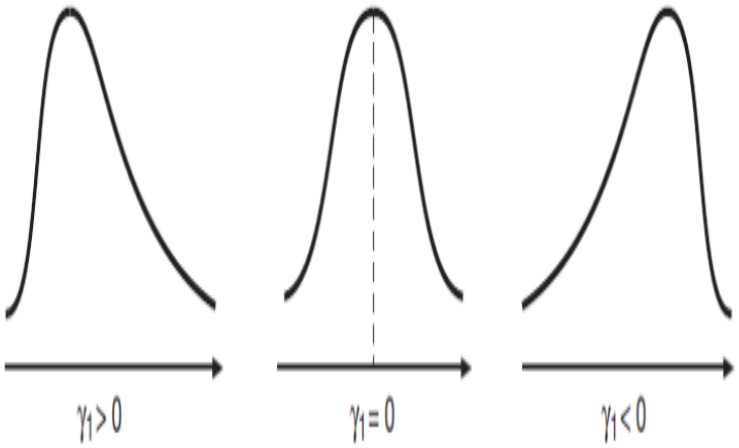


FIG. 1.7 – Signe de coefficient d'asymétrie

(iii) Si $\gamma_1 = F_1 < 0$ on a un étalement à gauche, la distribution est plus accentué à droite, autrement dit Mo est à droite de Me (courbe oblique à droite).

3. Coefficient d'asymétrie de Yule : Le coefficient d'asymétrie de Yule est basé sur les positions des 3 quartiles (1er quartile, médiane et troisième quartile), et est normalisé par la distance interquartile :

$$C_y = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_2}.$$

(i) Si $C_y = 0$ courbe symétrique.

(ii) Si $C_y > 0$ étalement à droite, Mo est à gauche de Me.

(iii) Si $C_y < 0$ étalement à gauche, Mo est à droite de Me.

– **Aplatissement**

Les coefficients d'aplatissement de Kurtosis mesurent l'aplatissement d'une distribution ou l'importance des **"Queues"** d'une distribution, ces coefficients nous renseignent sur l'aplatissement relatif d'une distribution comparée à la distribution de la loi normale ; leurs formules sont indiquées ci-dessous :

L'aplatissement est mesuré par le coefficient d'aplatissement de Pearson

$$\beta_3 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{(Var(x))^2}$$

en général on a toujours $\mu_4 > Var(x)$.

Ou le coefficient d'aplatissement de Fisher, pour une distribution normale, ce coefficient est nul ; une valeur positive indique une distribution plus pointue que la loi normale ; une valeur négative indique à l'inverse une distribution plus aplatie.

$$F_2 = \beta_3 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{(Var(x))^2} - 3$$

C'est un coefficient sous dimension, invariant par changement de variable et nul pour les distributions symétriques. Ce coefficient est nul pour une distribution normale, positif ou négatif selon que la distribution est plus ou moins aplatie que la distribution normale de même moyenne et de même écart-type.

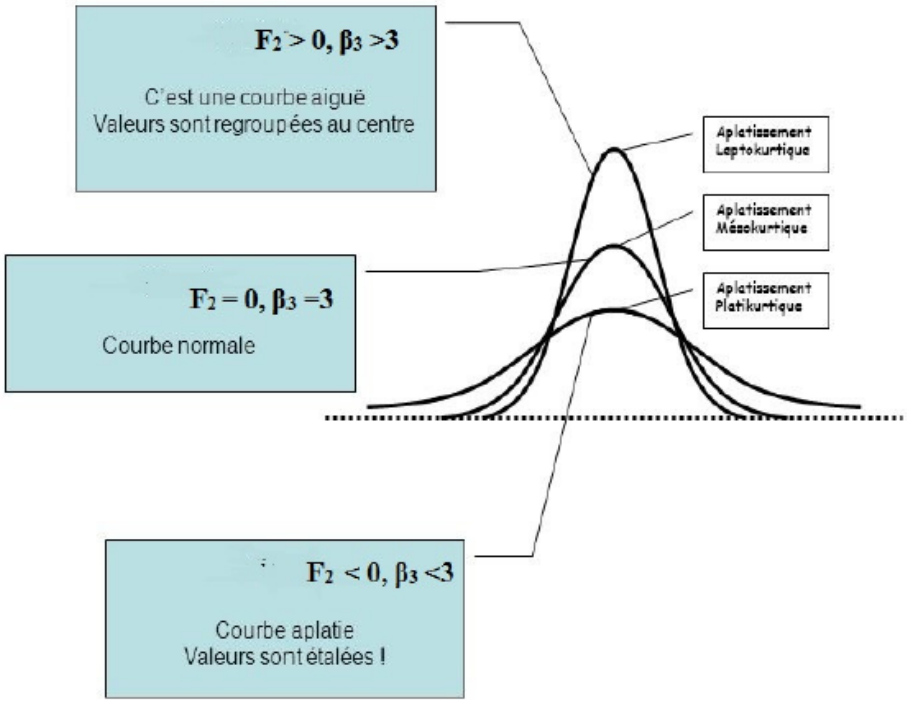


FIG. 1.8 – Courbe normale, Leptokurtique et Platikurtique

Donc

(i) Si $\beta_3 = 3$ ou $F_2 = 0$ on a une courbe normale ou Mésokurtique.

(ii) Si $\beta_3 > 3$ ou $F_2 > 0$ on a une courbe Leptokurtique ou moins aplatie. Elle est plus pointue et possède des queues plus longues. Autrement dit la concentration des valeurs de la série autour de la moyenne est forte.

(iii) Si $\beta_3 < 3$ ou $F_2 < 0$ on a une courbe Platikurtique ou plus aplatie. Elle est plus arrondie et possède des queues plus courtes. Autrement dit la concentration des valeurs de la série autour de la moyenne est faible.

Remarque 1.6.1. Les coefficients d'asymétrie et d'aplatissement sont invariants par changement d'origine et d'échelle, mais ils sont sensibles aux fluctuations d'échantillonnage puisqu'ils font intervenir des moments d'ordre élevé.

Exemple 1.6.1. La répartition de 200 assurés en fonctions du nombre X d'accident d'auto-

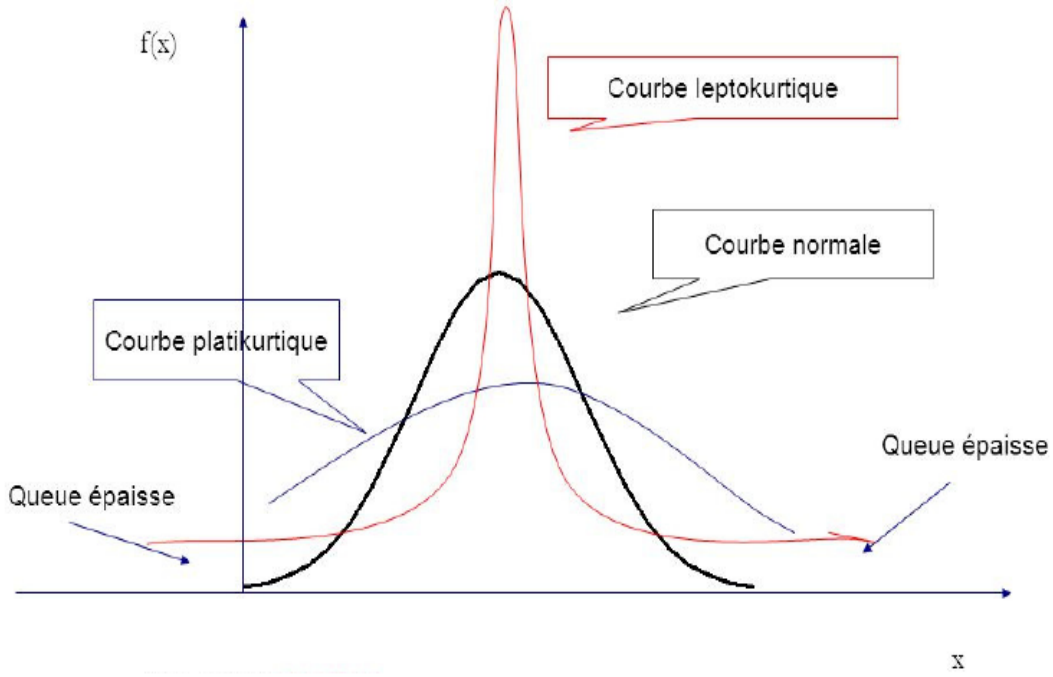


FIG. 1.9 – Courbe normale, Leptokurtique et Platikurtique

mobiles signalés à leurs compagnie d'assurance en 5 ans est donnée par le tableau suivant :

x_i (nbre d'accidents)	0	1	2	3	4	5	6	Total
n_i (nbre d'assurés)	747	718	376	118	31	6	4	200
$n_i x_i$	0	718	752	354	124	30	24	2002
$x_i - \bar{x}$	-1	0	1	2	3	4	5	
$n_i x_i^2$	0	718	1504	1062	496	150	144	4074
$n_i(x_i - \bar{x})^4$	747	0	376	1888	2511	1536	2500	9558
$N_i \nearrow$	747	1465	1841	1959	1990	1996	2000	
$N_i \searrow$	2000	1253	535	159	41	10	4	

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{7} n_i x_i = \frac{2002}{2000} = 1,001 \simeq 1 \text{ accident/assuré.}$$

$$Var(x)\frac{1}{N}\sum_{i=1}^{7} n_i x_i^2 - \bar{x}^2 = \frac{40742}{2000} - (1,001)^2 = 2,037 - 1,002 = 1,035$$

$$\sigma = \sqrt{Var(x)} = \sqrt{1,035} = 1,02 \simeq 1 \text{ accident.}$$

$$\mu_4 = \frac{1}{N}\sum_{i=1}^{7} n_i(x_i - \bar{x})^4 = \frac{9558}{2000} = 4,779.$$

$\frac{1}{4}N = 500$, $\frac{1}{2}N = 1000$, $\frac{3}{4}N = 1500$, Q_1 se situe à la 500 observation $\Longrightarrow Q_1 = 0$, $Q_2 = Me = 1$ et $Q_3 = 2$.

Calcule des coefficients d'asymétrie et d'aplatissement

• Coefficient de Yule

$$C_y = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} = \frac{2 - 2(1) + 0}{2 - 0} = \frac{0}{2} = 0$$

On remarque que le mode $Mo = 0$.

• Coefficient d'asymétrie

$$\beta_1 = \frac{\bar{x} - Mo}{\sigma} = \frac{1,001}{1,02} = 0,98 \simeq 1 > 0$$

Donc on a une forte asymétrie (étalement à droite).

• Coefficient d'aplatissement

$$\beta_3 = \frac{\mu_4}{Var(x)^2} = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{Var(x)^2} = \frac{4,779}{(1,02)^2} = \frac{4,779}{1,082} = 4,42 > 3.$$

Donc la distribution est moins aplatie (courbe Leptokurtique).

Chapitre 2

Statistique Double

2.1 Généralités

Il est très courant que l'étude statistique ne porte pas que sur un seul caractère, mais plusieurs caractères simultanément pour une même population. On étudie par exemple, un ensemble de salariés non plus seulement selon leur salaire, mais encore selon leur ancienneté. Les tableaux de données seront alors des tableaux à deux dimensions. L'étude statistique peut porter en même temps sur deux caractères qui peuvent être de même nature ou de nature différente (qualitatif-quantitatif discret, qualitatif-quantitatif continu ...). Dans ce chapitre on présentera ces tableaux, ainsi que le traitement statistique associé.

2.1.1 Tableaux à double entrée

Les tableaux à double entrée présentant pour chaque couple de modalités des deux caractères étudiés, l'effectif des individus, présentant ces deux modalités simultanément.

Construction d'un tableau de contingence

Considérons une population de N unités statistiques décrites simultanément selon deux variables statistiques X et Y prenant les modalités (ou valeurs) suivantes :

$$X : x_1, x_2, \ldots, x_p$$

$$Y : y_1, y_2, \ldots, y_q.$$

Le nombre d'individus présentant les modalités x_i et y_j simultanément est noté : n_{ij}. Le tableau est de la forme :

On désigne par N l'effectif total de la population de référence.

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

L'effectif $n_{i\bullet}$ (total de la ligne i) est le nombre total d'individus présentant la modalité x_i du caractère X indépendamment des modalités de Y.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_q	Colonne marginale $n_{i\bullet}$
x_1	n_{11}	n_{12}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2j}	$\cdots$	n_{2q}	$n_{2\bullet}$
$\vdots$							
x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{iq}	$n_{i\bullet}$
$\vdots$							
x_p	n_{p1}	n_{p2}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	$n_{p\bullet}$
Ligne marginale $n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet q}$	$N = n_{\bullet\bullet}$

TAB. 2.1 – Tableau Contingence de X et Y

En effectuant la somme des termes de chaque ligne, on définit les termes $n_{i\bullet}$ de la colonne marginale

$$\sum_{j=1}^{q} n_{ij} = n_{i1} + n_{i2} + \ldots + n_{ij} + \ldots + n_{iq} = n_{i\bullet}.$$

$n_{i\bullet}$ est l'effectif de la population qui présente les modalités x_i du caractère X, $\forall$ les modalités du caractère Y.

L'effectif $n_{\bullet j}$ (total de la colonne j) est le nombre total d'individus présentant la modalité y_i du caractère Y indépendamment des modalités de X.

En effectuant la somme des termes de chaque colonne, on définit les termes $n_{\bullet j}$ de la ligne marginale

$$\sum_{i=1}^{p} n_{ij} = n_{1j} + n_{2j} + \ldots + n_{ij} + \ldots + n_{pj} = n_{\bullet j}.$$

n_i est l'effectif de la population qui présente les modalités y_j du caractère Y, indépendamment du caractère X.

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

On appelle fréquence du couple de modalités (x_i, y_j) (ou encore fréquence totale), la proportion f_{ij} d'individus présentant simultanément les deux modalités :

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

Remarque : La somme de toutes les fréquences de couples = 1.

Exemple 2.1.1. Soit la répartition des salariés d'une entreprise selon le nombre d'enfant(X) et le salaire mensuel (Y) en 10^4 DA.

$n_{\bullet\bullet} = 60$ Le nombre total des salariés de l'entreprise.

$n_{22} = 4$ Salariés ont chacun 2 enfants et gagnent un salaire compris entre 6×10^4 et 10^5 DA.

$n_{13} = 2$ Salariés ont chacun 1 enfant et gagnent un salaire compris entre 10^5 et 1×10^4 DA.

$n_{2\bullet} = 18$ Salariés ont 2 enfants quelque soit leurs salaires.

Y(salaire) / X(nbre d'enfants)	$2-6$	$6-10$	$10-16$	Colonne marginale $n_{i\bullet}$
1	15	8	2	25
2	13	4	1	18
3	11	3	3	17
Ligne marginale $n_{\bullet j}$	39	15	6	60

TAB. 2.2 – Table contingence X : le nombre d'enfant, Y : le salaire mensuel

$n_{\bullet 2} = 15$ salariés ont un salaire compris entre 6×10^4 et 10^5 DA quelque soit leurs nombre d'enfants.

$f_{22} = \dfrac{n_{22}}{n_{\bullet\bullet}} = 4/60,\ f_{2\bullet} = \dfrac{n_{2\bullet}}{n_{\bullet\bullet}} = 18/60,\ f_{\bullet 2} = \dfrac{n_{\bullet 2}}{n_{\bullet\bullet}} = 15/60.$

2.1.2 Distributions Marginales (Conditionnelle)

Considérons la colonne de droite du tableau de contingence. Les effectifs $n_{i\bullet}$ représentent les individus présentant la modalité x_i indépendamment des modalités du second caractère étudié Y.

La notion de série conditionnelle est essentielle pour comprendre l'analyse de la régression. Un tableau de contingence se compose en autant de séries conditionnelles suivant chaque ligne et chaque colonnes.

On dit qu'ils définissent la **distribution marginale (conditionnelle)** de X.

(Cette série statistique est une série statistique à un seul caractère).

Série marginale (conditionnelle) de X : $(x_i, n_{i\bullet});\ i \in [1, p]$.

X	$n_{i\bullet}$	$f_{i\bullet}$
x_1	$n_{1\bullet}$	$f_{1\bullet}$
x_2	$n_{2\bullet}$	$f_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i\bullet}$	$f_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$
x_p	$n_{p\bullet}$	$f_{p\bullet}$
$\sum$	$n_{\bullet\bullet}$	$f_{\bullet\bullet} = 1$

On définit alors la fréquence marginale (conditionnelle) de la modalité x_i par :

$$f_{i\bullet} = \frac{n_{i\bullet}}{N}; \quad \sum_{i=1}^{p} f_{i\bullet} = 1.$$

Nbre d'enfants (X)	$n_{i\bullet}$	$f_{i\bullet}$
1	25	0.417
2	18	0.3
3	17	0.283
$\sum$	60	1

De la même façon on définit **la distribution marginale (conditionnelle)** de Y en considérant la dernière ligne du tableau.

Série marginale (conditionnelle) de Y : $(y_j, n_{\bullet j});\ j \in [1, q]$.

Y	$n_{\bullet j}$	$f_{\bullet j}$
y_1	$n_{\bullet 1}$	$f_{\bullet 1}$
y_2	$n_{\bullet 2}$	$f_{\bullet 2}$
$\vdots$	$\vdots$	$\vdots$
y_i	$n_{\bullet j}$	$f_{\bullet j}$
$\vdots$	$\vdots$	$\vdots$
y_q	$n_{\bullet q}$	$f_{\bullet q}$
$\sum$	$n_{\bullet\bullet}$	$f_{\bullet\bullet} = 1$

On définit alors la fréquence marginale de la modalité y_j par :

$$f_{\bullet j} = \frac{n_{\bullet j}}{N}; \quad \sum_{j=1}^{q} f_{\bullet j} = 1.$$

Salaire (Y) en ($10^4 X$ DA)	$n_{\bullet j}$	$f_{\bullet j}$
2-6	39	0.65
6-10	15	0.25
10-16	6	0.1
$\sum$	60	1

2.2 Caractéristiques des séries à deux variables

Dans le cas où les variables X et Y sont des variables quantitatives, on peut associer à chacune des séries marginales (conditionnelles) définies par le tableau de contingence des caractéristiques de tendance centrale et de dispersion.

On considère un tableau de contingence comme celui défini en (reftable0)

X prend les valeurs $x_1, x_2, \ldots, x_p$;

Y prend les valeurs $y_1, y_2, \ldots, y_q$.

(Les x_i ou y_j sont les centres de classes dans le cas où X et Y sont des variables quantitatives continues).

2.3 Caractéristiques marginales (conditionnelle)

2.3.1 Moyennes marginales (conditionnelle)

Moyenne marginale (conditionnelle) de X

La moyenne marginale (conditionnelle) de X notée $\bar{\bar{X}}$ correspond à la valeur moyenne du caractère X possédé par les individus de la population indépendamment du caractère Y.

la moyenne marginale (conditionnelle) : $\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{p} n_{i\bullet} x_i = \sum_{i=1}^{p} f_{i\bullet} x_i$.

x_i	x_i^2	$n_{i\bullet}$	$n_{i\bullet}x_i$	$n_{i\bullet}x_i^2$
1	1	25	25	25
2	4	18	36	72
3	9	17	51	153
$\sum$		60	112	250

TAB. 2.3 – Distribution marginale (conditionnelle) de X

Moyenne marginale (conditionnelle) de Y

La moyenne marginale (conditionnelle) de Y notée $\bar{\bar{Y}}$ correspond à la valeur moyenne du caractère Y possédée par les individus de la population indépendamment du caractère Y.

la moyenne marginale (conditionnelle) : $\bar{\bar{Y}} = \frac{1}{N}\sum_{j=1}^{q} n_{\bullet j}y_i = \sum_{j=1}^{q} f_{\bullet j}y_j$

2.3.2 Variances marginales (conditionnelle)

Variance marginale (conditionnelle) de X

La variance marginale (conditionnelle) de X notée $V(X)$ ou $Var(X)$ est une mesure de la dispersion des individus de la population selon X et indépendamment de Y.

La variance marginale (conditionnelle) :

$V(X) = \frac{1}{N}\sum_{i=1}^{p} n_{i\bullet}\left(x_i - \bar{\bar{X}}\right)^2 = \left(\frac{1}{N}\sum_{i=1}^{p} n_{i\bullet}x_i^2\right) - \bar{\bar{X}}^2 = \sum_{i=1}^{p} f_{i\bullet}\left(x_i - \bar{\bar{X}}\right)^2 = \left(\sum_{i=1}^{p} f_{i\bullet}x_i^2\right) - \bar{\bar{X}}^2$.

L'écart type marginale (conditionnelle) : $\sigma_X = \sqrt{V(X)}$.

Variance marginale (conditionnelle) de Y

La variance marginale (conditionnelle) de Y notée $V(Y)$ est une mesure de la dispersion des individus de la population selon Y et indépendamment de X.

La variance marginale (conditionnelle) :

$V(Y) = \frac{1}{N}\sum_{i=1}^{q} n_{\bullet j}\left(y_j - \bar{\bar{Y}}\right)^2 = \left(\frac{1}{N}\sum_{j=1}^{q} n_{\bullet j}y_j^2\right) - \bar{\bar{Y}}^2 = \sum_{i=1}^{q} f_{\bullet j}\left(y_j - \bar{\bar{Y}}\right)^2 = \left(\sum_{j=1}^{q} f_{\bullet j}y_j^2\right) - \bar{\bar{Y}}^2$

L'écart type marginale (conditionnelle) : $\sigma_Y = \sqrt{V(Y)}$.

Exemple 2.3.1. Soit le tableau de contingence (2.2) de l'exemple 2.1.1,

- La distribution marginale (conditionnelle) de X est donnée par le tableau suivant

  $\bar{\bar{X}} = \frac{1}{N}\sum_{i=1}^{p} n_{i\bullet}x_i = \frac{112}{60} = 1.86.$

  la variance marginale (conditionnelle) $V(X) = \frac{1}{N}\sum_{i=1}^{p}(n_{i\bullet}x_i^2) - \bar{\bar{X}} = 250/60 - (1.86)^2 = 0.71.$

y_j	y_j^2	$n_{.j}$	$n_{.j}y_j$	$n_{.j}y_j^2$
4	16	39	156	624
8	64	15	120	960
13	169	6	78	1014
$\sum$		60	354	2598

TAB. 2.4 – Distribution marginale (conditionnelle) de Y

L'écart type marginale (conditionnelle) : $\sigma_X = \sqrt{V(X)} = \sqrt{0.71} = 0.84$.

- La distribution marginale (conditionnelle) de Y est donnée par :

  la moyenne marginale (conditionnelle) : $\bar{\bar{Y}} = \frac{1}{N}\sum_{j=1}^{q} n_{.j}y_j = \frac{354}{60} = 5.9$.

  la variance marginale (conditionnelle) : $V(Y) = \frac{1}{N}\sum_{i=1}^{q}(n_{.j}y_j^2) - \bar{\bar{Y}} = 2598/60 - (5.9)^2 = 8.49$.

  L'écart type marginale (conditionnelle) : $\sigma_Y = \sqrt{V(Y)} = \sqrt{8.49} = 2.91$.

Chapitre 3

Ajustement Linéaire et Corrélation

Dans les chapitres précédents, nous avons présenté les méthodes qui permettent de résumer et représenter les informations relatives à une variable. Un même individu peut être étudié à l'aide de plusieurs caractères (ou variables). Par exemple, les salaries en regardant leur ancienneté et leur niveau d'étude, la croissance d'un enfant en regardant son poids et sa taille. Dans la suite, nous introduisons l'étude globale des relations entre deux variables (en nous limitant au cas de deux variables) X et Y.

3.1 Données et nuages de points

On note $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ la série des observations relevant deux caractères quantitatifs x et y pour les N individus d'une population par exemples : la taille et le poids d'un groupe d'étudiants.

Des unités étant convenablement choisies sur chacun des axes, on peut représenter l'individu i de la population précédente par le point : (x_i, y_i) du plan XY. Figurant ainsi les N individus, on obtient le nuage de points associé à la série statistique. Les nuages de points associés à des séries statistiques à deux caractères peuvent présenter sous différentes formes :

Les points du nuage 1 sont presque alignés, tandis que le nuage 2 laisse simplement apparaître une direction d'allongement privilégiée. Dans ces deux cas, on dit que le nuage présente un caractère linéaire. Le nuage 3 ne manifeste pas de structure particulière ; le nuage 4, enfin, semble se placer approximativement selon une courbe régulière. L'ajustement linéaire est la recherche de la droite résumant le mieux la structure du nuage. Une telle recherche n'a donc d'intérêt que pour des nuages de l'un des deux premiers types.

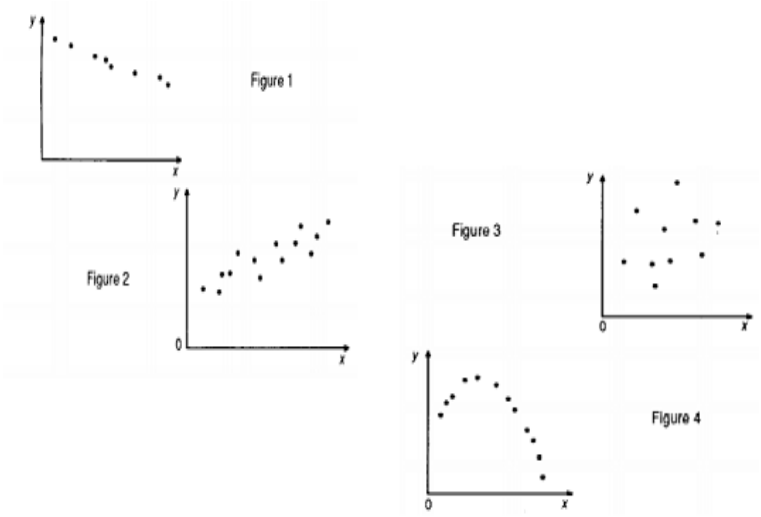


FIG. 3.1 – Représentation sous forme de nuage de points

3.2 Caractéristiques d'un couple de deux variables quantitatives

3.2.1 Moyenne d'une somme de deux variables statistiques

On montre sans difficulté le résultat suivant

1. $x \bar{+} y = \bar{x} + \bar{y}$.
2. $\forall a, b, c \in \mathbb{R}, \quad ax + \bar{b}y + c = a\bar{x} + b\bar{y} + c$.

3.2.2 Présentation

Les séries statistiques à deux variables peuvent être présentées de deux façons.

Présentation 1

Le cas où $n_{ij} = 1, \quad \forall i, j$.

On rassemblera les données comme dans le tableau suivant

X	x_1	x_2	...	x_k
Y	y_1	y_2	...	y_k

Présentation 2

Soit la variable statistique Z donnée par le couple (X, Y). Soient $x_1, \ldots, x_p$ et $y_1, \ldots, y_q$ les valeurs prises respectivement par X et Y. Dans ce cas, nous définissons les valeurs de Z comme suite, pour i allant de 1 à p et pour j allant de 1 à q, $z_{ij} := (x_i, y_j)$. La variable statistique Z prend $p \times q$ valeurs. Lors de cette étude, nous avons le tableau à double entrée (ou tableau de contingence) suivant (discrète ou continue)

X \ Y	y_1 ou c'_1	$\cdots$	y_j ou c'_j	$\cdots$	y_q ou c'_q	Marginale % à X ($n_{i\bullet}$)
x_1 ou c_1	n_{11} ou f_{11}	$\cdots$	n_{1j} ou f_{1j}	$\cdots$	n_{1q} ou f_{1q}	$n_{1\bullet}$ ou $f_{1\bullet}$
$\vdots$						
x_i ou c_i	n_{i1} ou f_{i1}	$\cdots$	n_{ij} ou f_{ij}	$\cdots$	n_{iq} ou f_{iq}	$n_{i\bullet}$ ou $f_{i\bullet}$
$\vdots$						
x_p ou c_p	n_{p1} ou f_{p1}	$\cdots$	n_{pj} ou f_{pj}	$\cdots$	n_{pq} ou f_{pq}	$n_{p\bullet}$ ou $f_{p\bullet}$
Marginale % à Y ($n_{\bullet j}$)	$n_{\bullet 1}$ ou $f_{\bullet 1}$	$\cdots$	$n_{\bullet j}$ ou $f_{\bullet j}$	$\cdots$	$n_{\bullet q}$ ou $f_{\bullet q}$	$N = n_{\bullet\bullet}$

TAB. 3.1 – Tableau Contingence de X et Y

Cette représentation on l'a notera "présentation 2". A chaque couple (x_i, y_j), on a n_{ij} est l'effectif qui représente le nombre d'individus qui prennent en même temps la valeur x_i et y_j.

3.2.3 Covariance entre deux variables statistiques

On associe aux deux caractères quantitatifs X et Y une caractéristique globale appelée la covariance et définie par :

$$Cov(X,Y) = \frac{1}{N}\sum_i\sum_j n_{ij}\left(x_i - \bar{\bar{X}}\right)\left(y_j - \bar{\bar{Y}}\right) = \sum_i\sum_j f_{ij}\left(x_i - \bar{\bar{X}}\right)\left(y_j - \bar{\bar{Y}}\right).$$

On vérifie qu'on peut calculer la covariance par la formule :

$$Cov(X,Y) = \left(\frac{1}{N}\sum_i\sum_j n_{ij}x_iy_j\right) - \bar{\bar{X}}\times\bar{\bar{Y}} = \left(\sum_i\sum_j f_{ij}x_iy_j\right) - \bar{\bar{X}}\times\bar{\bar{Y}}.$$

Propriété 3.2.1. *1.* $Cov(X,Y) = Cov(Y,X)$.

2. $Cov(X,X) = Var(X)$.

3. $\forall a,b,c,d \in \mathbb{R}$, $Cov(aX+b,cY+d) = abCov(X,Y)$.

4. $|Cov(X,Y)| \leq \sqrt{Var(X)\times Var(Y)}$.

5. $Var(constante) = 0$.

3.3 Ajustement linéaire

Dans le cas où on peut mettre en évidence l'existence d'une relation linéaire significative entre deux caractères quantitatifs continus X et Y (la silhouette du nuage de points est étirée dans une direction), on peut chercher à formaliser la relation moyenne qui unit ces deux variables à l'aide d'une équation de droite qui résume cette relation. Nous appelons cette démarche l'ajustement linéaire.

3.3.1 Ajustement graphique

Lorsque le nuage présente un caractère linéaire, on peut tenter de tracer. (à main levée) la droite qui résume le mieux la structure du nuage. La subjectivité du procédé est évidente.

3.3.2 Corrélation

Coefficient de corrélation linéaire r ou ρ

Le coefficient de corrélation permettent de donner une mesure synthétique de l'intensité de la relation entre deux caractères et de son sens lorsque cette relation est monotone. Le coefficient de corrélation de Pearson permet d'analyser les relations linéaires (voir cidessous).

En fait, on calcule plus fréquemment le coefficient de corrélation linéaire, noté r ou ρ.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

on voit sur celte expression que les variables X et Y jouent des rôles symétriques.

Le coefficient r mesure non seulement la qualité de l'une ou l'autre régression, mais, plus généralement, le caractère linéaire du nuage de points (le degré de liaison linéaire entre X et Y) voir Figure 3.2, ou encore l'intensité de la liaison linéaire entre les deux variables (en particulier lors qu'aucune des deux ne paraît devoir "expliquer" l'autre, on a les deux caractéristiques suivantes (voir Figures 3.3-3.4).

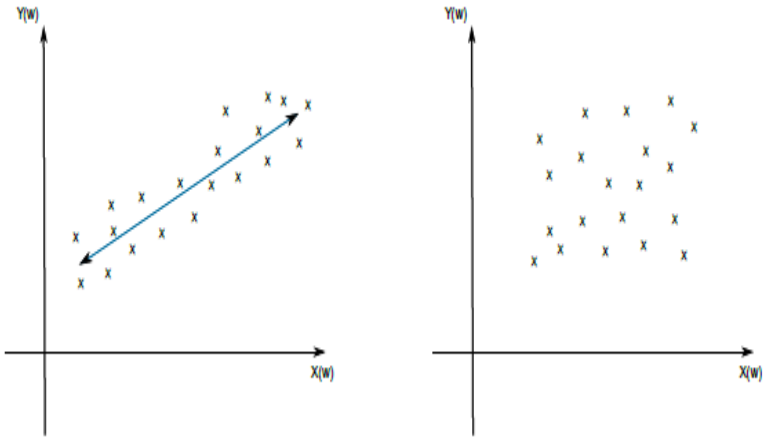


FIG. 3.2 – À gauche, le coefficient de corrélation est proche de 1. À droite, le coefficient de corrélation est proche de 0.

Remarque 3.3.1. • Plus le module de $r = \rho_{XY}$ est proche de 1 plus X et Y sont liées linéairement.

• Plus le module de $r = \rho_{XY}$ est proche de 0 plus il y a l'absence de liaison linéaire entre X et Y .

Son emploi est précisé par les propriétés suivantes :

Proposition 3.3.1.

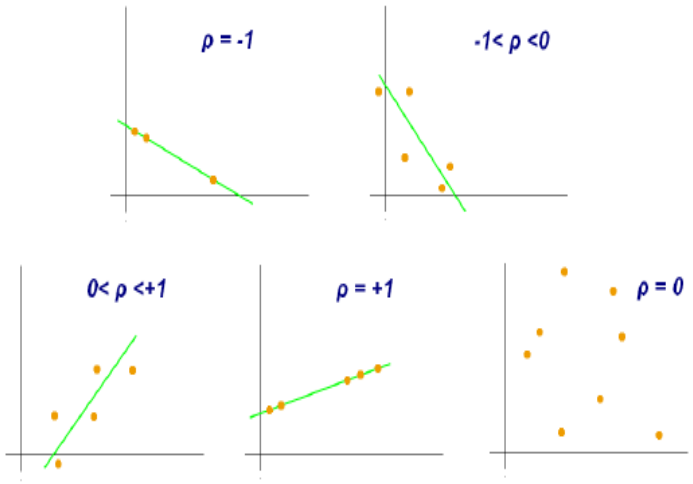


FIG. 3.3 – Exemples de diagrammes de dispersion avec différentes valeurs de coefficient de corrélation.

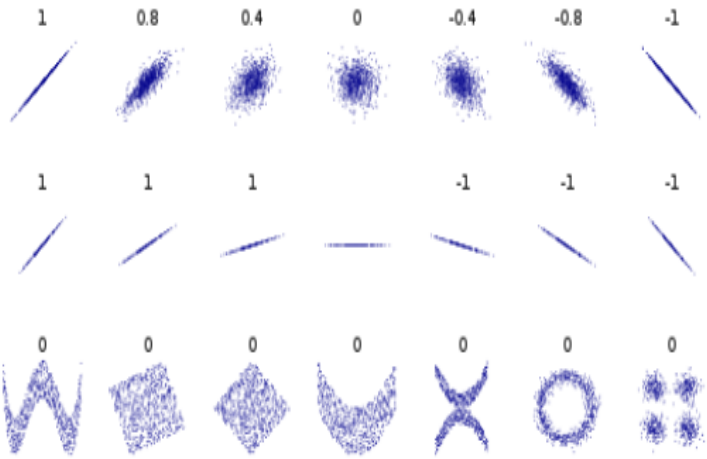


FIG. 3.4 – La corrélation reflète la non-linéarité et la direction d'une relation linéaire mais pas la pente de cette relation ni de nombreux aspects des relations non linéaires (en bas). La figure au centre a une pente de 0, mais dans ce cas, le coefficient de corrélation est indéfini car la variance de Y est nulle.

1. $|r| \leq 1$.
2. *r vaut $+1$ (respectivement $-l$) lorsque les points sont alignés sur une droite ascendante, traduisant une variation dans le même sens des deux caractères (respectivement descendante, pour une variation de sens contraire).*
3. *r est proche de $+1$ (respectivement de -1) lorsque les caractères montrent une liaison linéaire marquée et croissante (respectivement décroissante). Autrement dit, la régression est a priori intéressante, et les deux droites de régression ne seront guère éloignées.*
4. *r est proche de 0 en l'absence de liaison linéaire apparente, la régression linéaire est alors peu justifiée.*

Remarque 3.3.2. • r peut être calculé en premier lieu (c'est-à-dire avant les droites de régression) et par exemple dans le dernier cas ne pas donner suite à une régression.

- Par définition, si $r = 0 \Longrightarrow Cov(X, Y) = 0$.
- X et Y indépendantes (non corrélées)$\Longrightarrow r = 0$. La réciproque est fausse.
- $r = 0 \Longleftrightarrow$ pas de liaison linéaire, mais possibilité d'une liaison d'un autre type.

3.3.3 Droite de régression

Les points (x_i, y_i) forment un nuage dont on cherche une approximation dans un but de simplification. Mais qui dit simplification dit déformation : nous voudrions qu'elle soit minimale, encore faut-il préciser ce que l'on entend par là. Disons tout de suite que le choix du critère sera arbitraire même si l'on tente de le justifier par des considérations plus ou moins $\ll$ *intuitives* $\gg$. On peut vouloir par exemple :

- Préserver au mieux les distances entre points.
- Préserver au mieux les angles des droites joignant les points.

L'idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite.

Il n'existe pas de moyen de satisfaire à toutes ces exigences à la fois. Pour cela, on utilise la méthode des moindres carrées pour choisir la **meilleure droite au sens des moindres carrés**. Cette méthode vise à expliquer un nuage de points par une droite qui lie Y à X, c'est à dire

$$Y = aX + b,$$

telle que la distance entre le nuage de points et droite soit minimale. Cette distance matérialise l'erreur, c'est à dire la différence entre le point réellement observé et le point prédit par la droite. Si la droite passe au milieu des points, cette erreur sera alternativement positive et négative, la somme des erreurs étant par définition nulle. Ainsi, la méthode des moindres carrés consiste à chercher la valeur des paramètres a et b qui minimise la somme des erreurs élevées au carré.

Méthode des moindres carrés

Il s'agit de déterminer la droite D d'équation $\{y = ax + b\}$ telle que

$$\sum_{i=1}^{n} e_i^2 = F(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \quad \text{soit minimum,}$$

avec e_i est l'erreur commise sur chaque observation, c'est à dire

$$e_i = |y_i - y_i^*| = |y_i - (ax_i + b)|.$$

• x_i est la valeur observée de la variable explicative X pour l'individu i.

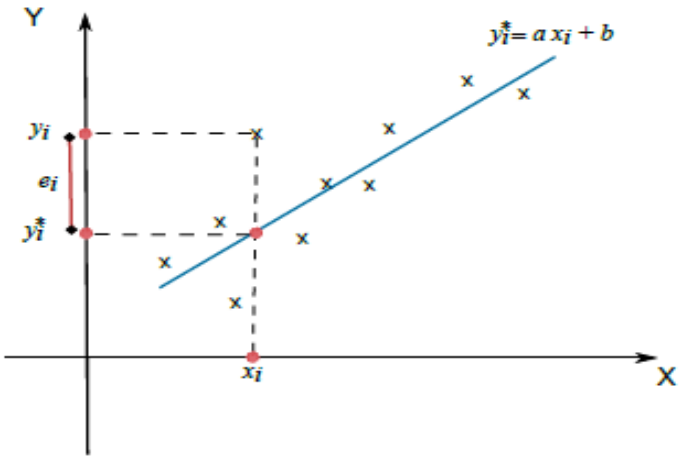


FIG. 3.5 – La droite la plus proche possible de chacun des points.

• y_i est la valeur observée de la variable à expliquer Y pour l'individu i.

• $y^* = ax + b$ est la valeur théorique ou ajustée la variable Y associée à la valeur observée x de la variable X.

On considère le plus souvent que l'un des caractères, ou l'une des variables, dépend de l'autre (par exemple, la consommation dépend du revenu) ; soit Y le premier caractère, ou variable à expliquer, et X le second, ou la variable explicative. On cherche une expression de Y en fonction de X, de la forme $y = ax + b$, qui approche "le mieux" les données. D'un point de vue géométrique, cela revient à chercher la droite d'équation $y = ax + b$, qui traduit le mieux l'aspect linéaire du nuage de points.

Nos inconnues sont a et b. Commençons par chercher le minimum de $F(a, b)$ relativement à b lorsque a est fixé. On peut écrire $F(a, b)$ comme un trinôme du second degré en b :

$$\begin{aligned} F(a,b) &= \sum_{i=1}^{n} \left((y_i - ax_i) - b\right)^2 = \sum_{i=1}^{n} \left((y_i - ax_i)^2 - 2b(y_i - ax_i) + b^2\right) \\ &= \sum_{i=1}^{n} (y_i - ax_i)^2 - 2b \sum_{i=1}^{n} (y_i - ax_i) + nb^2. \end{aligned}$$

Quand a est fixé, le dernier membre constitue une fonction de b qui atteint pour $b = \widehat{b}$ telle que $\dfrac{\partial F}{\partial b}(a, \widehat{b}) = 0$ soit :

$$\begin{aligned} \frac{\partial F}{\partial b}(a, \widehat{b}) &= -2 \left(\sum_{i=1}^{n} (y_i - ax_i) - nb \right) = 0 \\ &\Longrightarrow \quad \widehat{b} = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i) = \bar{y} - a\bar{x}. \end{aligned}$$

• 1^{re} conséquence : la droite des moindres carrés passe par le point de coordonnées qu'on appelle parfois le centre de gravité ou point moyen du nuage.

Notre problème est maintenant de trouver le minimum de $F(a,\widehat{b})$ relativement à a :

$$\begin{aligned} F(a,\widehat{b}) &= \sum_{i=1}^{n}\left((y_i-\bar{y})-a(x_i-\bar{x})\right)^2 \\ &= \sum_{i=1}^{n}(y_i-\bar{y})^2-2a\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})+a^2\sum_{i=1}^{n}(x_i-\bar{x})^2, \end{aligned}$$

ce qui peut encore s'écrire :

$$F(a,\widehat{b}) = n\left(a^2Var(X)-2aCov(X,Y)+Var(Y)\right).$$

Le coefficient de a^2 étant positif ou nul, ce trinôme du second degré en a atteint son minimum relativement à a pour $a=\widehat{a}$ avec :

$$\widehat{a}=\frac{Cov(X,Y)}{Var(X)}.$$

Ainsi le couple $(\widehat{a},\widehat{b})$ avec $\widehat{b}=\bar{y}-\widehat{a}\bar{x}$ réalise le minimum de la fonction F.

- 2^{me} conséquence : la droite des moindres carrés a pour équation $y=\widehat{a}x+b$ soit

$$y-\bar{y}=\frac{Cov(X,Y)}{Var(X)}.(x-\bar{x}).$$

On posera pour tout i variant de 1 à n : $\widehat{y_i}=\widehat{a}x_i+\widehat{b}$, $\widehat{y_i}$ est la valeur estimée de Y par la droite des moindres carrés lorsque $X=x_i$.

Remarque 3.3.3. Le coefficient de corrélation rho_{XY} permet de justifier le fait de l'ajustement linéaire. On adopte les critères numériques suivants (voir Figure 3.6),

- Si $\rho_{XY}<0.7$ alors l'ajustement linéaire est refusé (droite refusée).
- Si $\rho_{XY}\geq 0.7$ alors l'ajustement linéaire est est accepté (droite acceptée).

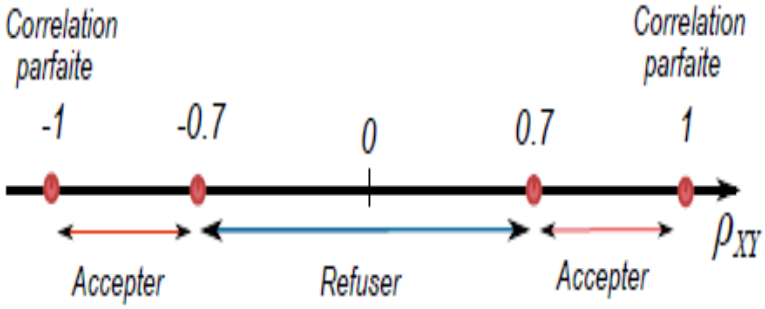


FIG. 3.6 – La zone d'acceptation ou de refus de l'ajustement linéaire.