Analyse Des Données

Ouazza Ahmed

École Supérieure de Management, d'Informatique et de télécommunications
SUP MTI

2023-2024

Plan

- Rappel
- Analyse en Composante Principales ACP
- Analyse Factorielle des Correspondances AFC
- Méthodes de classification

Rappel:

Statistique descriptive univariée

Introduction

La statistique peut être définie comme un ensemble de principes et de méthodes scientifiques pour recueillir, classer, synthétiser et communiquer des données numériques en vue de leur utilisation pour en tirer des conclusions et prendre des décisions.

La statistique est utilisée en plusieurs domaines, par exemple : **Comptabilité**, **finance** (bilans ou comptes de résultats, gestion du capital, opérations avec les banques), **Biologie** (évolution d'une maladie), **Production** (gestion des stocks ou du matériel, contrôle de la qualité), **Achats**, **ventes** (statistiques des ventes, études de marché),...

Introduction

Une étude statistique concerne soit :

• Une seule variable : on parle de statistique à une dimension, ou statistique **univariée**.

• Deux variables : on parle de statistique à deux dimensions ou **bivariée**.

• Plus de deux variables : on parle de statistique multidimensionnelle ou statistique **multivariée**.

Notions de base

1. Populations, unité statistique et échantillon

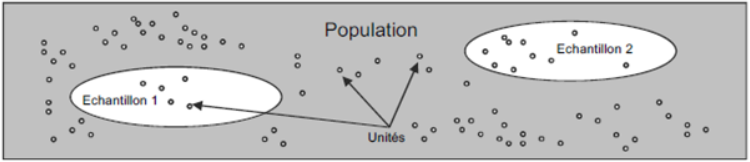
• On appelle population un ensemble d'individus: personnes, objets ou éléments sur lesquels on veut effectuer une étude statistique.

• Les individus qui composent une population statistique sont appelés unités statistiques.

• Un sous ensemble de la population est appelé échantillon.

Notions de base

Les relations qui existent entre la population, les échantillons et les unités statistiques sont résumées dans le schéma ci-dessous:



Exemple:

On considère comme population l'ensemble des étudiants d'une université. L'unité statistique dans ce cas est un(e) étudiant(e). Les étudiants du premier semestre représentent un échantillon.

2. Caractères

• Caractère ou variable : C'est la propriété commune de la population étudiée, qui est observée ou mesurée sur les individus de cette population statistique.

Exemple:

Etude de la taille des étudiants, étude du nombre d'enfants dans une famille, étude de la couleur des voitures,...

• Modalité : On appelle une modalité la valeur que peut prendre un caractère.

Exemple:

– Population étudiée : Les étudiants du premier semestre de la filière Economie et Gestion,

– Variable statistique (caractère) : mention obtenue au Baccalauréat

– Modalités : Passable, Assez-bien, Bien, Très Bien.

2.1 Caractères qualitatifs et quantitatifs

Il existe deux grandes catégories de caractères : les caractères **qualitatifs** et les caractères **quantitatifs**.

Variable quantitative:

Une variable statistique est dite quantitative lorsque ses modalités sont mesurables. Selon la forme des valeurs de la variable, on distingue deux types de caractères quantitatifs: discrets et continus:

Notions de base

• Une variable quantitative discrète est une variable qui peut prendre uniquement certaines valeurs d'un intervalle de nombres réels. Généralement, les valeurs admissibles ne sont que les nombres entiers.

Exemple :

Le nombre d'enfants par famille.

Le nombre d'accidents par mois.

Notions de base

• Une variable est quantitative continue si elle peut prendre toutes les valeurs d'un intervalle. (Nombre de valeurs possibles est infini).

Exemple :

-Salaire d'un fonctionnaire.

-Âge d'un étudiant.

-Taille ou poids d'un bébé.

Variable qualitative:

Une variable statistique est dite qualitative lorsque ses modalités ne sont pas mesurables. On distingue deux types de variables qualitatives : Nominale et ordinale.

• Une variable est qualitative nominale si ses modalités ne peuvent pas être ordonnées.

Exemple :

-Sexe (les modalités sont : masculin et féminin).

-Couleur des cheveux (les modalités sont : blanc, brun, noir...)

-Groupe sanguin (les modalités sont : A, B, AB et O).

-Etat matrimonial (les modalités sont : célibataire, marié, veuf et divorcé).

Notions de base

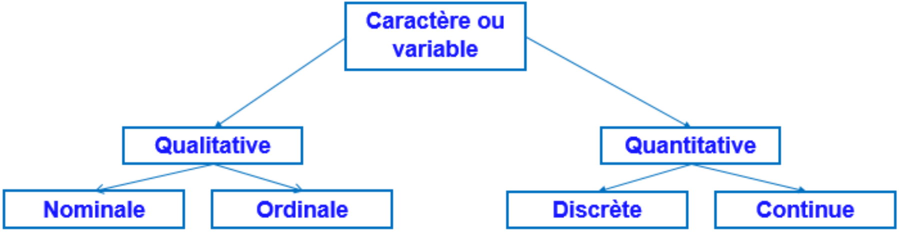
• Une variable est qualitative ordinale si ses modalités ne sont pas des valeurs numériques et elles peuvent être ordonnées.

Exemple :

-Mention obtenue au Baccalauréat (les modalités sont: Passable, Assez-bien, Bien, Très Bien).

Satisfaction d'un service client (les modalités sont: Très insatisfait, insatisfait, neutre, satisfait et très satisfait).

Notions de base



Effectifs et fréquences

Définition

L'effectif total est le nombre d'individus appartenant à la population statistique étudiée. L'effectif total sera noté N.

Exemple :

Considérons un groupe comprenant trente étudiants et observons l'âge des étudiants dans cette population.

L'effectif total de la population statistique étudiée est $N = 30$.

Définition

L'effectif d'une modalité x_i d'un caractère X est le nombre d'individus présentant cette modalité. L'effectif correspondant à la $i^{\text{ème}}$ modalité du caractère X est noté n_i .

Effectifs et fréquences

Exemple :

Considérons de nouveau le groupe de $N = 30$ étudiants et construisons un tableau pour regrouper les différentes informations que l'on a sur leur âge.

La première information que l'on va noter dans ce tableau est l'effectif de chaque âge observé :

Age	Effectif n_i
18	2
19	4
20	10
21	11
22	3
Total	**30**

Effectifs et fréquences

Propriété et notation

De façon générale, pour une variable qui a k modalités, l'effectif total N est égal à la somme des effectifs de chaque modalité du caractère, ce que l'on peut écrire :

$$n_1 + n_2 + \cdots + n_k = N$$

Définition

On considère les modalités d'un caractère X variant de 1 à k, l'effectif cumulé, noté N_i, d'une modalité i est le nombre d'individus de la population présentant une modalité d'indice inférieur ou égal à i.

Effectifs et fréquences

Exemple:

Age	Effectif n_i	Effectif cumulé N_i
18	2	2
19	4	6
20	10	16
21	11	27
22	3	30
Total	**30**	–

Définition

La fréquence d'une modalité est la proportion d'individus de la population totale qui présentent cette modalité : elle est obtenue en divisant l'effectif de cette modalité du caractère par l'effectif total et notée f_i soit :

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

Effectifs et fréquences

Exemple : Considérons l'exemple précédent. On a regroupé les fréquences correspondant à l'âge des étudiants dans le tableau suivant :

Age	Effectif n_i	Fréquence f_i
18	2	$\frac{2}{30} = 0.067$
19	4	$\frac{4}{30} = 0.133$
20	10	$\frac{10}{30} = 0.333$
21	11	$\frac{11}{30} = 0.367$
22	3	$\frac{3}{30} = 0.1$
Total	**30**	1

Effectifs et fréquences

Les notions présentées ci-dessus peuvent être regroupées dans un tableau récapitulatif (tableau statistique) comme suit :

Modalités ou valeurs de X	Effectifs n_i	Effectifs cumulés N_i	Fréquences f_i	Fréquences cumulées F_i
x_1	n_1	N_1	f_1	F_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	N_i	f_i	F_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	N_k	f_k	F_k

Groupement des données en classes

Lorsque la variable étudiée est continue ou quand la variable statistique est discrète mais prenant trop de valeurs, il est pratique de regrouper l'ensemble des valeurs en intervalles statistiques (ou classes). Le tableau récapitulatif associé à une variable quantitative continue est donné comme suit :

Classe	Effectifs n_i	Amplitude a_i $a_i = x_i - x_{i-1}$	Centre de classe c_i $c_i = \frac{x_i + x_{i-1}}{2}$	Effectifs corrigés $n_i^\star$ $n_i^\star = \frac{n_i}{a_i}$
$[x_0, x_1[$	n_1	a_1	c_1	$n_1^\star$
$[x_1, x_2[$	n_2	a_2	c_2	$n_2^\star$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[x_{k-1}, x_k[$	n_k	a_k	c_k	$n_k^\star$

Groupement des données en classes

Exemple:

Les données présentées dans le tableau statistique suivant correspondant à la tranche d'âge de $N = 140$ personnes.

Classe	Eff n_i	Amplitude a_i	Centre c_i	Fréq f_i	Fréq Cum F_i	Eff cum N_i	Eff corrigé $n_i^\star$
$[20, 25[$	9	5	22.5	0.06	0.06	9	1.8
$[25, 30[$	17	5	27.5	0.12	0.19	26	3.4
$[30, 35[$	36	5	32.5	0.26	0.44	62	7.2
$[35, 40[$	27	5	37.5	0.19	0.64	89	5.4
$[40, 50[$	45	10	45	0.32	0.96	134	4.5
$[50, 60[$	6	10	55	0.04	1	140	0.6

Représentations graphiques

Il est parfois indispensable de recourir à la présentation graphique des données pour visualiser la distribution statistique d'une variable.

Il existe plusieurs types de graphiques, selon le type de données.

Cas d'une variable qualitative

Dans le cas d'une variable qualitative, les modalités ne peuvent pas être représentées sur un axe, selon une échelle donnée, car elles ne sont pas numériques. On utilise dans ce cas des diagrammes en barres, des diagrammes circulaires et demi-circulaires.

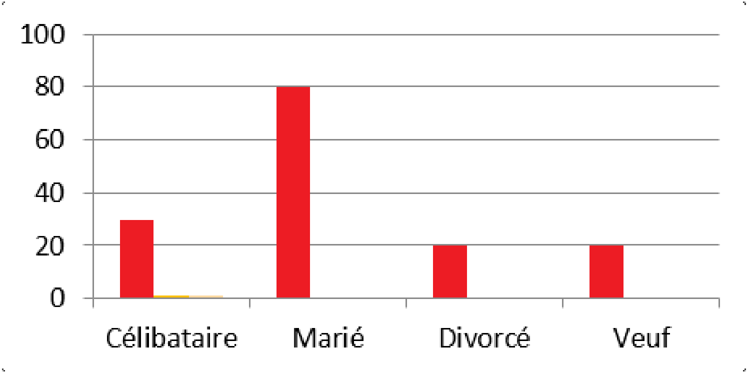
Exemple:

On considère le tableau statistique suivant :

État matrimonial	Effectif n_i	Effectif cumulé N_i
Célibataire	30	30
Marié	80	110
Divorcé	20	130
Veuf	20	150
Total	**150**	–

Représentations graphiques

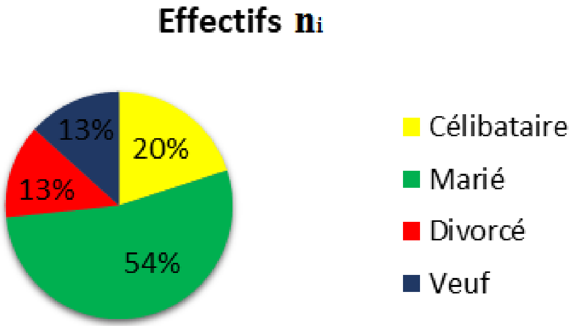
- **Diagramme en barres:**



Avec la longueur des barres = Effectif n_i

Représentations graphiques

- **Diagramme circulaire:**



Avec l'angle pour chaque modalité est donné par :

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

Cas d'une variable quantitative discrète

Souvent un caractère quantitatif discret est représenté par un diagramme en bâtons des effectifs (ou des fréquences) et les polygones des effectifs (ou des fréquences).

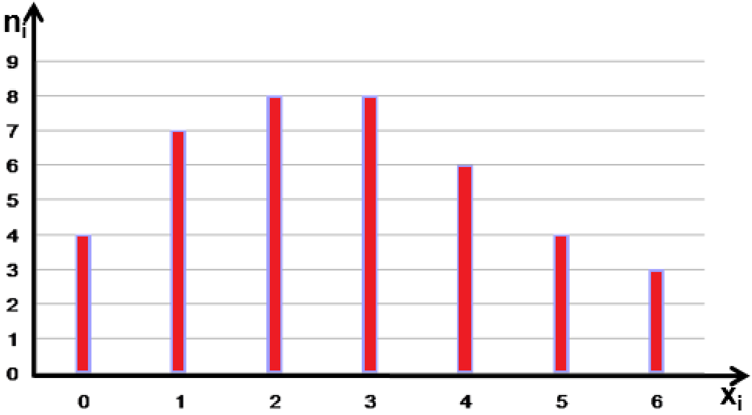
Exemple:

Sur $N = 40$ familles, on a compté le nombre d'enfants pour chaque famille. Les résultats obtenus sont regroupés dans le tableau suivant:

Nombre d'enfants x_i	Effectif n_i	Effectif cumulé N_i
0 (sans enfant)	4	4
1	7	11
2	8	19
3	8	27
4	6	33
5	4	37
6	3	40
Total	**40**	–

Représentations graphiques

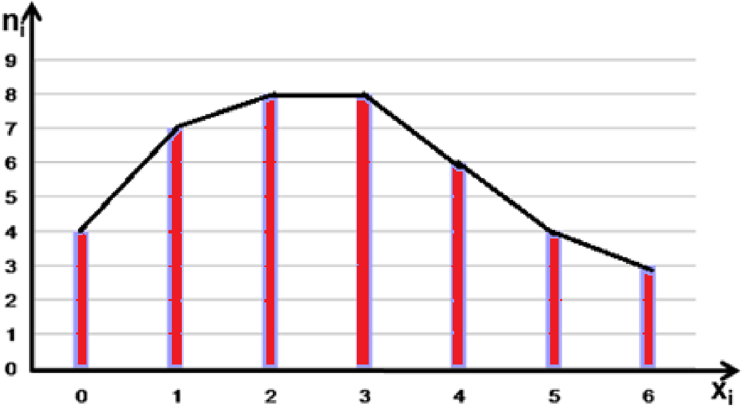
Le **diagramme en bâtons** associé est le suivant :



Avec la longueur des bâtons = Effectif n_i

Représentations graphiques

• Les **polygones des effectifs** : il s'agit de joindre les sommets des bâtons pour obtenir des polygones :



Représentations graphiques

Cas d'une variable quantitative continue

Comme les caractères continus ont plusieurs valeurs, on les regroupe en classes et on applique les formules concernant les caractères discrets aux centres des classes.

La représentation graphique se fait alors sous forme d'histogramme, graphique dans lequel chaque classe est représentée par un rectangle dont la surface est proportionnelle à l'importance de cette classe dans la population. Ainsi les bases de chaque rectangle sont les classes de la variable étudiée et les hauteurs sont les effectifs corrigés.

Remarque :

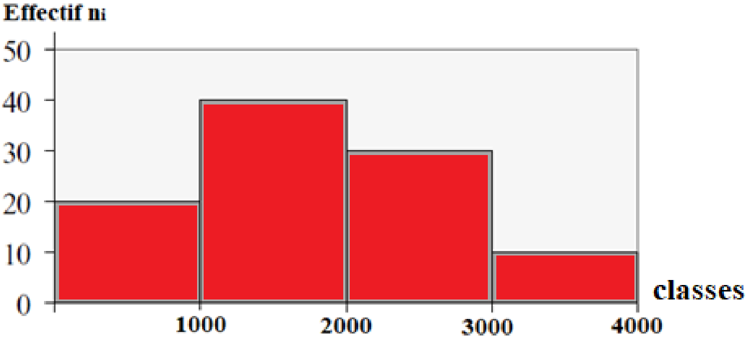
Si l'amplitude est la même pour toutes les classes, les hauteurs les rectangles correspondent simplement aux effectifs n_i .

Exemple:

On considère les tranches de revenus suivantes dans une population de $N = 100$ salariés :

Classes (en €)	Effectif n_i	Effectif cumulé N_i
$[0, 1000[$	20	20
$[1000, 2000[$	40	60
$[2000, 3000[$	30	90
$[3000, 4000[$	10	100
Total	**100**	–

• **L'histogramme** est donné comme suit :



Indicateurs de Position

Pour caractériser une série statistique quantitative, on peut construire plusieurs indicateurs comme:

- La médiane,
- Le mode,
- La moyenne,
- Les quantiles,...

La médiane

La médiane, notée Me, est la valeur de la variable qui partage la population statistique étudiée en deux effectifs égaux, les individus étant ordonnés selon les valeurs de la variable.

La médiane sera donc la valeur de la variable telle que 50% de la population se situe au-dessus et 50% se situe en dessous.

Indicateurs de Position

Dans le cas d'une variable discrète, le calcul de la médiane se fait comme suit:

• Si la taille N (N est l'effectif total) de la série statistique est un nombre impair, c-à-d $N = 2k + 1$, alors la médiane correspond à la $(k+1)^{\text{ème}}$ valeur de **série ordonnée**.

• Si la taille N de la série statistique est un nombre pair, c-à-d $N = 2k$, alors la valeur de la médiane est calculée, sur la **série ordonnée**, par la formule suivante :

$$Me = \frac{k^{\text{ème}} \text{ élement} + (k+1)^{\text{ème}} \text{ élement}}{2}$$

Indicateurs de Position

Exemple:

• **Cas où la taille N est impaire :**

La série suivante représente les notes obtenues par $N = 11$ élèves d'une classe :

$S = \{12; 9; 14; 15; 13; 8; 9; 17; 11; 13; 10\}$

Dans ce cas la taille de série est $N = 11 = 2 * 5 + 1$ (impaire)

Pour calculer la médiane on doit d'abord **ordonner** la série comme suit :

$S = \{8; 9; 9; 10; 11; 12; 13; 13; 14; 15; 17\}$

Donc la médiane correspond à la $(5 + 1)^{\text{ème}} = 6^{\text{ème}}$ valeur de la série **ordonnée**, d'où $Me = 12$.

Indicateurs de Position

• **Cas où la taille N est paire :**

Soit la série suivante (avec $N = 10$) :

$S = \{12; 9; 14; 15; 13; 8; 9; 17; 11; 13\}$

La série **ordonnée** est :

$S = \{8; 9; 9; 11; 12; 13; 13; 14; 15; 17\}$

Puisque $N = 10 = 2 \times 5$ est pair, alors la médiane est donnée par:

$$Me = \frac{(5^{\text{ème}} \text{ valeur} + 6^{\text{ème}} \text{ valeur})}{2} = \frac{12 + 13}{2} = 12.5$$

Le mode

Le mode, noté Mo, d'une série statistique est la modalité de la variable correspondant à l'effectif **le plus élevé**.

Remarque:

Une série statistique peut avoir plusieurs modes.

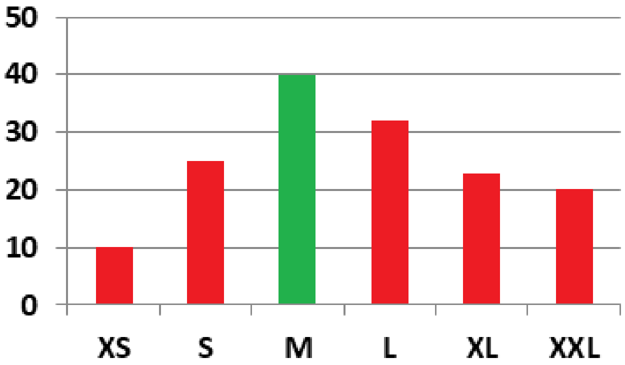
Exemple:

Le tableau suivant représente le nombre des chemises disponible dans une boutique selon la taille :

La taille de la chemise	Le nombre disponible
XS	10
S	25
M	40
L	31
XL	22
XXL	20
Total	**148**

Indicateurs de Position

- Le diagramme en barres associé est le suivant :



Le mode de cette série statistique est la modalité de la variable correspondant à l'effectif **le plus élevé** qui est dans ce cas la taille **M** (la modalité M).

Remarque:

Dans le cas de distributions groupées, on parle de classe modale, ainsi la classe modale est la classe ayant **l'effectif corrigé le plus élevé**.

La moyenne

La moyenne d'une série statistique quantitative X est donnée par:

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

Où les x_i sont les valeurs observées de la variable X et n_i représente l'effectif de chaque valeur x_i.

Indicateurs de Position

Remarque:

Pour les données groupées par des classes, les valeurs x_i sont remplacées par les centres des classes c_i. Dans ce cas la moyenne est :

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

La moyenne présentée ci-dessus s'appelle la moyenne arithmétique.

Indicateurs de Position

Il existe d'autres types de moyennes telles que : **la moyenne géométrique**, **la moyenne harmonique** et **la moyenne quadratique**.

- La moyenne géométrique

$$\overline{X}_g = (x_1^{n_1} \times x_2^{n_2} \times \cdots \times x_k^{n_k})^{\frac{1}{N}} = \left[\Pi_{i=1}^{k} x_i^{n_i}\right]^{\frac{1}{N}}$$

- La moyenne harmonique

$$\overline{X}_h = \frac{N}{\sum_{i=1}^{k} n_i \frac{1}{x_i}}$$

- La moyenne quadratique

$$\overline{X}_q = \sqrt{\frac{1}{N} \sum_{i=1}^{k} n_i x_i^2}$$

Remarque:

Pour la même série statistique, les quatre moyennes vérifient toujours la relation d'ordre suivante:

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

Exemple (moyenne arithmétique):

Le tableau suivant représente la répartition des notes d'un échantillon de $N = 30$ étudiants.

Classe de notes	Nombre d'étudiants n_i	Centres c_i
$[0; 5[$	2	2.5
$[5; 10[$	7	7.5
$[10; 15[$	18	12.5
$[15; 20[$	3	17.5
Total	**30**	–

La moyenne arithmétique est :

$$\overline{X} = \frac{2 * 2.5 + 7 * 7.5 + 18 * 12.5 + 3 * 17.5}{30} = 11.16$$

Les quantiles

Soit $\alpha \in]0, 1[$. On appelle le quantile d'ordre α la valeur x_α de la variable telle que au moins $100 \times \alpha\%$ des observations sont inférieures ou égales à x_α et $100 \times (1 - \alpha)\%$ des observations qui sont supérieures ou égales à x_α.

Remarque:

La médiane Me est le quantile d'ordre $\alpha = 0.5$

Le tableau suivant résume quelques quantiles et leurs ordres.

Quantile	Ordre	Notation
Quartile	(0.25 , 0.5 , 0.75)	$(Q_1, Q_2 = Me, Q_3)$
Décile	(0.1 , 0.2 , ... , 0.9)	$(D_1, ..., D_9)$
Centile	(0.01, 0.02, ... , 0.99)	$(C_1, C_2, ..., C_{99})$

Indicateurs de Position

• **Méthode de calcul des quartiles Q_1 et Q_3:**

Premier cas : variables quantitatives discrètes

On considère d'abord une série statistique ordonnée de taille N,

– **Le premier quartile** Q_1 d'une série statistique est la plus petite valeur telle qu'au moins 25% des valeurs sont inférieures ou égales à Q1.

On distingue deux situations :

◦ Si $0.25 \times N$ est un nombre entier naturel alors le premier quartile Q_1 correspond à la valeur de rang $0.25 \times N$.

◦ Si $0.25 \times N$ n'est pas un entier, alors le rang de premier quartile Q_1 est le premier entier naturel supérieur à $0.25 \times N$.

Indicateurs de Position

– **Le troisième quartile** d'une série statistique est la plus petite valeur Q_3 telle qu'au moins 75% des valeurs sont inférieures ou égales à Q_3.

On distingue deux situations :

∘ Si $0.75 \times N$ est un nombre entier naturel alors Q_3 correspond à la valeur de rang $0.75 \times N$.

∘ Si $0.75 \times N$ n'est pas un entier, alors le rang de Q_3 est le premier entier naturel supérieur à $0.75 \times N$.

Exemple 1:

Soit la série statistique suivante :

$S = \{40, 30, 41, 50, 42, 10, 25, 111, 101, 110, 70, 55\}$

On ordonne la série et on obtient :

$S = \{10, 25, 30, 40, 41, 42, 50, 55, 70, 101, 110, 111\}$

Alors le premier quartile est $Q_1 = 30$. En effet, il y a $N = 12$ nombres dans cette série, et $0.25 \times 12 = 3$ (c'est un nombre entier). Le premier quartile est donc la $3^{\text{ème}}$ valeur de la série ordonnée, soit $Q_1 = 30$.

De même, on a : $0.75 \times 12 = 9$ (nombre entier), alors le troisième quantile Q_3 correspond à la $9^{\text{ème}}$ valeur, soit $Q_3 = 70$.

Exemple 2:

Si on considère la série précédente avec $N = 13$ nombres, c-à-d,

$S = \{10, 25, 30, 40, 41, 42, 50, 55, 70, 101, 110, 111, 120\}$

Dans ce cas on a : $0.25 \times 13 = 3.25$ (n'est pas un entier), alors le premier quartile correspond à la $4^{\text{ème}}$ valeur de la série ordonnée, soit $Q_1 = 40$.

Pour Q_3, on a : $0.75 \times 13 = 9.75$ (n'est pas un entier), alors le troisième quantile Q_3 correspond à la $10^{\text{ème}}$ valeur, soit $Q_3 = 101$.

Deuxième cas : variables quantitatives continues

On détermine la classe $c_q = [a_{i-1}; a_i[$ (respectivement $c_l = [b_{i-1}; b_i[$) qui contient Q_1 (respectivement Q_3) : c'est la première classe dont l'effectif cumulé dépasse $0.25 \times N$ (respectivement $0.75 \times N$).

Q_1 et Q_3 s'obtiennent ensuite par les formules suivantes :

$$Q_1 = a_{i-1} + (a_i - a_{i-1}) \times \frac{0.25N - N_{q-1}}{N_q - N_{q-1}}$$

$$Q_3 = b_{i-1} + (b_i - b_{i-1}) \times \frac{0.75N - N_{l-1}}{N_l - N_{l-1}}$$

avec N_q est l'effectif cumulé de la classe c_q.

Exemple:

Considérons le tableau statistique suivant :

Classe	Effectif n_i	Effectif cumulé N_i
$[2;4[$	16	$N_1 = 16$
$[4;5[$	25	$N_2 = 41$
$[5;7[$	29	$N_3 = 70$
$[7;11[$	30	$N_4 = 100$
Total	**100**	–

Dans cet exemple on a,

$Q_1 \in c_q = [4, 5[$ (car le premier effectif cumulé qui dépasse $0.25 \times 100 = 25$ est $N_2 = 41$)

et $Q_3 \in c_l = [7, 11[$ (car le premier effectif cumulé qui dépasse $0.75 \times 100 = 75$ est $N_4 = 100$)

Donc Q_1 et Q_3 sont donnés par:

$$Q_1 = 4 + (5 - 4)\frac{0.25 \times 100 - 16}{41 - 16} = 4.36$$

$$Q_3 = 7 + (11 - 7)\frac{0.75 \times 100 - 70}{100 - 70} = 7.67$$

Indicateurs de Position

Remarque:

On peut déterminer la médiane Me de la même manière:

On a $Me = Q_2$, la classe médiane est $c_m = [a_{i-1}; a_i[$: c'est la première classe dont l'effectif cumulé dépasse la valeur $\frac{N}{2} = \frac{100}{2} = 50$, c-à-d $c_m = [5; 7[$, donc

$$Q_2 = 5 + (7 - 5)\frac{0.5 \times 100 - 41}{70 - 41} = 5.62$$

Indicateurs de dispersion

La variance et l'écart-type

• **La variance** d'une série statistique quantitative X est définie par la formule suivante :

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

Avec les x_i représentent les valeurs observées de la variable étudiée X.

Remarque:

Dans le cas d'une variable continue les valeurs x_i sont remplacées par les centres des classes c_i.

• **L'écart-type** est définie comme la racine carrée de la variance :

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

Indicateurs de dispersion

Etendu

L'étendu est la différence entre la valeur maximale et la valeur minimale d'une variable.

Les intervalles interquartiles

L'intervalle interquartile I d'une série statistique est égal à la différence entre le troisième et le premier quartile :

$$I = Q_3 - Q_1$$

Exemple:

On reprend la série statistique suivante :

$$S = \{40, 30, 41, 50, 42, 10, 25, 111, 101, 110, 70, 55\}$$

• La valeur minimale est **10** et la valeur maximale est **111**, alors l'étendu est égal à **111-10=101**.

• On a :

$Q_1 = 30$ et $Q_3 = 70$

donc l'intervalle interquartile est $I = Q_3 - Q_1 = 70 - 30 = 40$

• **La variance :**

On calcul d'abord la moyenne arithmétique, on a :

$\overline{X} = (10+25+30+40+41+42+50+55+70+101+110+111)/12 = 57.08$

La variance est :

$V(X) = 1051.624$

L'écart type est : $\sigma_X = \sqrt{V(X)} = 32.42$

Paramètres de forme

Les paramètres de forme sont utilisés pour connaître la forme de la distribution d'une série statistique X et de la comparer avec celle d'une série statistique suivant **la loi normale** (ou loi de Gauss)

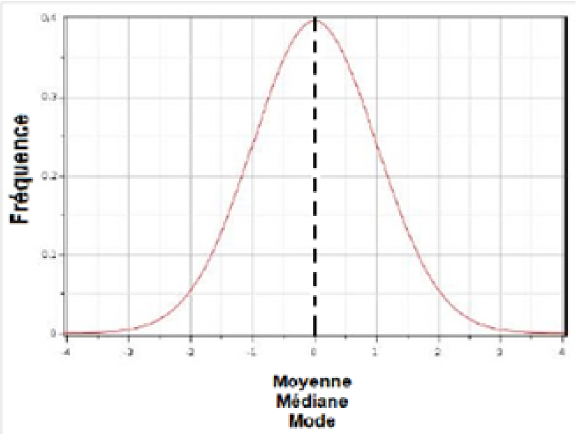
• Caractéristiques de la distribution normale :

La loi normale est considérée une des lois les plus utilisées en statistique pour modéliser des phénomènes issus de plusieurs événements aléatoires.

La loi normale est centrée autour de sa moyenne. De plus, la **moyenne**, la **médiane** et le **mode** sont confondus.

Paramètres de forme

La figure suivante montre la distribution d'une série statistique suivant la loi normale :



Paramètres de forme

Coefficient d'asymétrie (skewness)

On considère une variable statistique X, le coefficient d'asymétrie de Fisher, noté γ_1 est donné par la formule suivante :

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

Avec $\mu_3 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^3$ est le moment centré d'ordre 3 de la variable X, et σ représente d'écart-type de X.

Paramètres de forme

On distingue trois cas :

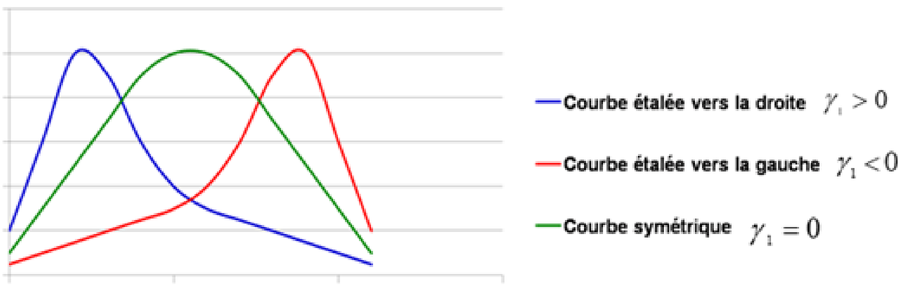
• Si $\gamma_1 = 0$, alors la distribution est symétrique, le mode, la moyenne et la médiane sont confondus.

• Si $\gamma_1 < 0$, alors la distribution présente une asymétrie à droite de la médiane et donc la queue de distribution est plus étalée vers la gauche..

•Si $\gamma_1 > 0$, alors la distribution présente une asymétrie à gauche de la médiane et donc la queue de distribution est plus étalée vers la droite..

Paramètres de forme

La figure suivante montre la représentation graphique de trois séries statistiques de différents types d'asymétrie :



Paramètres de forme

Le coefficient d'aplatissement (kurtosis)

Coefficient d'aplatissement de Fisher, noté γ_2, est défini par :

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Avec $\mu_4 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^4$ est le moment centré d'ordre 4 de la variable X, et σ représente d'écart-type de X.

Paramètres de forme

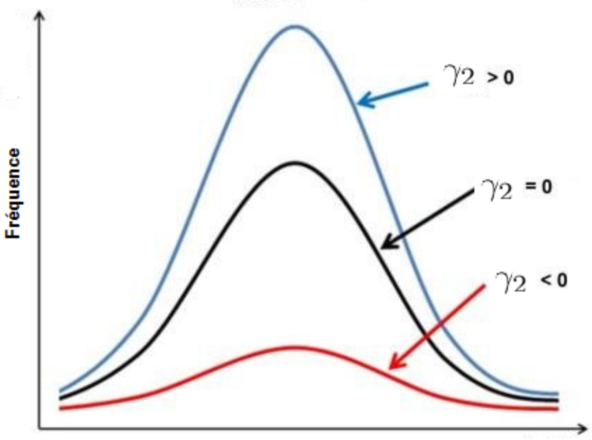
On distingue trois cas :

• Si $\gamma_2 = 0$, alors l'aplatissement est le même que celui d'une distribution normale.

• Si $\gamma_2 < 0$, alors la courbe représentant la distribution est aplatie (plus que celle d'une normale).

• Si $\gamma_2 > 0$, lalors la courbe représentant la distribution est dite concentrée ou affilée (moins aplatie que celle d'une normale).

Le graphe ci-après illustre ces trois cas :



Rappel:

Statistique descriptive bivariée

Une série statistique à deux variables est une série statistique pour laquelle deux caractères X et Y sont relevés pour chaque individu. On souhaite déterminer, essentiellement, les liens existants entre ces deux caractères.

Exemple:

Considérons un échantillon composée de 300 clients d'une banque ayant emprunté une même somme. Nous nous intéressons au nombre de mois nécessaires pour le remboursement des emprunts (variable X) et aux revenus mensuels de ces clients (variable Y).

Tableau de contingence et distributions

Définition

Soient X et Y deux variables statistiques définies sur un échantillon de taille n, avec :

$\{x_1, x_2, \cdots, x_p\}$ sont les p valeurs prises par X,

$\{y_1, y_2, \cdots, y_q\}$ sont les q valeurs prises par Y.

(X, Y) sera appelée série statistique double ou bivariée et nous disposons de $p \times q$ couples (x_i, y_j) de valeurs observées,

avec :$i = 1, 2, \cdots, p$ et $j = 1, 2, \cdots, q$

Exemple

Revenons à l'exemple précédent : On suppose que les clients vont rembourser la somme soit en **3 mois**, **6 mois**, **8 mois** ou **12 mois**.

Leurs revenus sont soit de **1000Dh**, **3000Dh**, **5000Dh**, **7000Dh** ou de **8000Dh** par mois.

Ainsi, on peut écrire :

$\{x_1 = 3,\ x_2 = 6,\ x_3 = 8,\ x_4 = 12\}$ sont les valeurs prises par X,

$\{y_1 = 1000,\ y_2 = 3000,\ y_3 = 5000,\ y_4 = 7000,\ y_5 = 8000\}$ sont les valeurs prises par Y.

Ici : $p = 4,\ q = 5$. Ainsi, pour la série statistique double (x, y) étudiée, nous avons $p \times q = 20$ couples qui peuvent être observés : (3,1000), (3,3000), (3,5000),...,(6,1000), (6,3000)...

Effectifs et fréquences partiels

• **L'effectif partiel:**

L'effectif partiel d'un couple (x_i, y_j) est le nombre n_{ij} de couples observés égaux à (x_i, y_j).

• **Tableau de contingence:** est un tableau à double entrée contenant la distribution des effectifs partiels:

Statistique descriptive bivariée

$x \setminus y$	y_1	y_2	$\cdots$	y_q
x_1	n_{11}	n_{12}	$\cdots$	n_{1q}
x_2	n_{21}	n_{22}	$\cdots$	n_{2q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\cdots$	n_{pq}

• **La fréquence partielle:** la fréquence partielle du couple (x_i, y_j) est le rapport :

$$f_{ij} = \frac{n_{ij}}{n} \quad , \quad i = 1, 2, \cdots, p \ ; \ j = 1, 2, \cdots, q$$

Exemple:

On reprend l'exemple précédent. On considère le tableau de contingence correspondant suivant :

$x \setminus y$	$y_1 = 1000$	$y_2 = 3000$	$y_3 = 5000$	$y_4 = 7000$	$y_5 = 8000$
$x_1 = 3$	0	3	9	24	27
$x_2 = 6$	9	12	18	21	24
$x_3 = 8$	15	15	15	18	24
$x_4 = 12$	24	18	12	9	3

• **L'effectif partiel :**

- L'effectif partiel du couple (x_1, y_1) est le nombre $n_{11} = 0$ correspondant au couple observé (3,1000) : aucun individu, parmi ceux qui ont un revenu de 1000Dh par mois, ne rembourse l'emprunt en 3 mois.

- L'effectif partiel du couple (x_3, y_2) est le nombre $n_{32} = 15$ correspondant au couple observé (8,3000).

- L'effectif partiel du couple (x_2, y_4) est le nombre $n_{24} = 21$ correspondant au couple observé (6,7000).

• **La fréquence partielle :**

- La fréquence partielle du couple (x_1, y_1) est $f_{11} = \frac{n_{11}}{n} = \frac{n_0}{300} = 0.$

- La fréquence partielle du couple (x_3, y_4) est $f_{34} = \frac{n_{34}}{n} = \frac{18}{300} = 0.06.$

- La fréquence partielle du couple (x_4, y_5) est $f_{45} = \frac{n_{45}}{n} = \frac{3}{300} = 0.01:$

c-à-d le pourcentage des clients qui remboursent en **12 mois** et ont un revenu mensuel de **8000Dh** est 1%.

Distributions marginales

L'étude marginale d'une série statistique double (x, y) est l'étude de **la distribution de** X sans tenir compte du caractère Y ou celle de Y sans tenir compte de X.

Les distributions marginales peuvent être traitées comme une série univariée.

Définition

L'effectif marginal : l'effectif marginal de la valeur x_i (resp. y_j) est la somme des effectifs partiels des couples contenant x_i (resp. y_j).

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \ , \ i = 1, 2, \cdots, p$$

$$n_{.j} = \sum_{i=1}^{p} n_{ij} \ , \ j = 1, 2, \cdots, q$$

Définition

La fréquence marginale : la fréquence marginale de la valeur x_i(resp. y_j) est la somme des fréquences partielles des couples contenant x_i (resp. y_j).

$$f_{i.} = \sum_{j=1}^{q} f_{ij} = \sum_{j=1}^{q} \frac{n_{ij}}{n} = \frac{n_{i.}}{n} \ , \ i = 1, 2, \cdots , p$$

$$f_{.j} = \sum_{i=1}^{p} f_{ij} = \sum_{i=1}^{p} \frac{n_{ij}}{n} = \frac{n_{.j}}{n} \ , \ j = 1, 2, \cdots , q$$

On peut compléter le tableau de contingence ci-dessus en ajoutant une colonne et une ligne contenant les effectifs marginaux des modalités x_i et y_j.

$x \setminus y$	y_1	y_2	$\cdots$	y_q	**Total**
x_1	n_{11}	n_{12}	$\cdots$	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	$\cdots$	n_{2q}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\cdots$	n_{pq}	$n_{p.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.q}$	$\sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} = n$

Remarque:

On a:

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

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = \sum_{i=1}^{p} f_{i.} = \sum_{j=1}^{q} f_{.j} = 1$$

Exemple:

On reprend l'exemple précédent. On complète le tableau de contingence correspondant en ajoutant les effectifs marginaux en x et en y:

$x \setminus y$	$y_1 = 1000$	$y_2 = 3000$	$y_3 = 5000$	$y_4 = 7000$	$y_5 = 8000$	$n_{i.}$
$x_1 = 3$	0	3	9	24	27	$n_{1.} = 63$
$x_2 = 6$	9	12	18	21	24	$n_{2.} = 84$
$x_3 = 8$	15	15	15	18	24	$n_{3.} = 87$
$x_4 = 12$	24	18	12	9	3	$n_{4.} = 66$
$n_{.j}$	$n_{.1} = 48$	$n_{.2} = 48$	$n_{.3} = 54$	$n_{.4} = 72$	$n_{.5} = 78$	$\mathbf{n = 300}$

• L'effectif marginal :

- l'effectif marginal de la valeur x_1 est la somme des effectifs partiels des couples contenant x_1 :

$$n_{1.} = \sum_{j=1}^{5} n_{1j} = n_{11} + n_{12} + n_{13} + n_{14} + n_{15} = 0 + 3 + 9 + 24 + 27 = 63$$

- l'effectif marginal de la valeur y_2 est la somme des effectifs partiels des couples contenant y_2 :

$$n_{.2} = \sum_{i=1}^{4} n_{i2} = n_{12} + n_{22} + n_{32} + n_{42} = 3 + 12 + 15 + 18 = 48$$

• La fréquence marginale :

- la fréquence marginale de la valeur x_3 est la somme des fréquences partielles des couples contenant x_3 :

$$
\begin{aligned}
f_{3.} &= \sum_{j=1}^{5} f_{3j} = \sum_{j=1}^{5} \frac{n_{3j}}{n} \\
&= \frac{n_{31}}{n} + \frac{n_{32}}{n} + \frac{n_{33}}{n} + \frac{n_{33}}{n} + \frac{n_{35}}{n} \\
&= \frac{n_{3.}}{n} \\
&= \frac{15}{300} + \frac{15}{300} + \frac{15}{300} + \frac{18}{300} + \frac{24}{300} = \frac{87}{300} \\
&= 0.29
\end{aligned}
$$

- la fréquence marginale de la valeur y_1 est la somme des fréquences partielles des couples contenant y_1 :

$$
\begin{aligned}
f_{.1} &= \sum_{i=1}^{4} f_{i1} \\
&= \frac{0}{300} + \frac{9}{300} + \frac{15}{300} + \frac{24}{300} \\
&= \frac{48}{300} \\
&= 0.16
\end{aligned}
$$

Moyennes marginales de X et de Y

La moyenne marginale de x est le nombre $\overline{x}$ défini par:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_i$$

De même, la moyenne marginale de y est le nombre $\overline{y}$ défini par:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}y_j$$

Variances marginales de X et de Y

La variance marginale de X est définie par :

$$S_x = Var(x) = \frac{1}{n}\sum_{i=1}^{p} n_{i.}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2 - (\overline{x})^2$$

De même, la variance marginale de Y est :

$$S_y = Var(y) = \frac{1}{n}\sum_{j=1}^{q} n_{.j}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2 - (\overline{y})^2$$

Exemple:

Dans le tableau précédent on a:

• La moyenne marginale de X :

$$\overline{x} = \frac{1}{300}\sum_{i=1}^{4} n_{i.}x_i = \frac{1}{300}(n_{1.}\ x_1 + n_{2.}\ x_2 + n_{3.}\ x_3 + n_{4.}\ x_4)$$

$$= \frac{1}{300}((63 \times 3) + (84 \times 6) + (87 \times 8) + (66 \times 12)) = 7,27$$

- La moyenne marginale de Y :

$$\overline{y} = \frac{1}{300}\sum_{j=1}^{5} n_{.j} y_j = \frac{1}{300}(n_{.1}\ y_1 + n_{.2}\ y_2 + n_{.3}\ y_3 + n_{.4}\ y_4 + n_{.5}\ y_5)$$

$$= \frac{(1000 \times 48) + (3000 \times 48) + (5000 \times 54) + (7000 \times 72) + (8000 \times 78)}{300}$$

$$= 5300$$

• La Variance marginale de X :

$$S_x^2 = Var(x) = \frac{1}{n}\sum_{i=1}^{p} n_{i.}x_i^2 - (\overline{x})^2 = \frac{1}{300}\sum_{i=1}^{4} n_{i.}x_i^2 - (7,27)^2$$

$$= \frac{1}{300}[n_{1.}x_1^2 + n_{2.}x_2^2 + n_{3.}x_3^2 + n_{4.}x_4^2] - (7,27)^2$$

$$= \frac{1}{300}[63 \times 3^2 + 84 \times 6^2 + 87 \times 8^2 + 66 \times 12^2] - (7,27)^2$$

$$= 9,357$$

• La Variance marginale de Y :

$$S_y^2 = Var(y) = \frac{1}{n}\sum_{j=1}^{q} n_{.j}y_j^2 - (\overline{y})^2 \;= \frac{1}{300}\sum_{i=1}^{5} n_{.j}y_j^2 - (5300)^2$$

$$= \frac{1}{300}[n_{.1}y_1^2 + n_{.2}y_2^2 + n_{.3}y_3^2 + n_{.4}y_4^2 + n_{.5}y_5^2] - (5300)^2$$

$$= \frac{1}{300}\,[48\times 1000^2 + 48\times 3000^2 + 54\times 5000^2 + 72\times 7000^2 + 78\times 8000^2] - (5300)^2$$

$$= 6410000$$

Indépendance statistique

Définition

On dit que les variables statistiques (caractères) X et Y sont indépendantes si et seulement si :

$$f_{ij} = f_{i.} \times f_{.j} \Leftrightarrow n_{ij} = \frac{n_{i.} n_{.j}}{n} \quad \forall i = 1, 2, \cdots, p \quad et \quad \forall j = 1, 2, \cdots, q$$

Remarque:

Pour montrer que deux variables X et Y ne sont pas indépendantes (liées), il suffit de trouver un couple (i, j) pour lequel

$$f_{ij} \neq f_{i.} \times f_{.j}$$

Statistique descriptive bivariée

Définition

On appelle la covariance de x et y la quantité :

$$
\begin{aligned}
Cov(x,y) &= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(x_i-\overline{x})(y_j-\overline{y}) \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}(x_i-\overline{x})(y_j-\overline{y}) \\
&= \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_iy_j - \overline{x}\,\overline{y} \\
&= \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}x_iy_j - \overline{x}\,\overline{y}
\end{aligned}
$$

Interprétation :

La covariance indique si les variables x et y varient dans le même sens ou dans deux sens opposés.

Nous avons les cas suivants :

• Si $Cov(x, y) > 0 \;\Rightarrow$ les deux variables ne sont pas indépendantes, elles sont positivement liées et varient dans le même sens.

• Si $Cov(x, y) < 0 \;\Rightarrow$ les deux variables ne sont pas indépendantes, elles sont négativement liées et varient dans deux sens opposés.

• Si $Cov(x, y) = 0 \;\Rightarrow$ les variations de l'une des deux variables n'entraînent pas la variation de l'autre.

Exemple:

On revient à l'exemple précédent et on calcul de la covariance entre x et y :

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}x_iy_j - \overline{x}\,\overline{y}$$

$$= \frac{1}{n}[n_{11}x_1y_1 + n_{12}x_1y_2 + \cdots + n_{15}x_1y_5 + n_{21}x_2y_1 + \cdots + n_{25}x_2y_5$$

$$+ \cdots \quad + n_{41}x_4y_1 + \cdots + n_{45}x_4y_5] - (7,27 \times 5300)$$

$$= \frac{34940}{300} - 38531 = -38414,53 < 0$$

Une covariance négative. Par conséquent, dans notre exemple, x et y sont négativement liés (évoluent dans deux sens opposés).

Etude de la dépendance linéaire entre deux variables

Dans ce qui suit, nous considérons deux variables statistiques discrètes x et y. Pour chaque unité i d'une population, on observe deux valeurs x_i et y_i. La série statistique double observée peut être écrite comme :

$$\{(xi, y_i)\ , i = 1, \cdots, n\}$$

X	x_1	x_2	$\cdots$	x_n
Y	y_1	y_2	$\cdots$	y_n

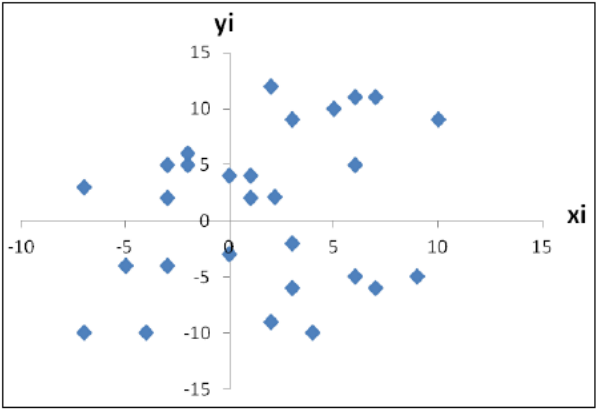
Nuage de points

Il s'agit de l'ensemble des points de coordonnées x_i et y_j d'une série statistique à deux variables. Ce nuage comporte n points.

Une représentation graphique du nuage de points s'effectue sur un repère en représentant les x_i en abscisses et les y_i en ordonnées.

Cette représentation permet de détecter visuellement l'existence ou non d'une **éventuelle liaison**.

Statistique descriptive bivariée

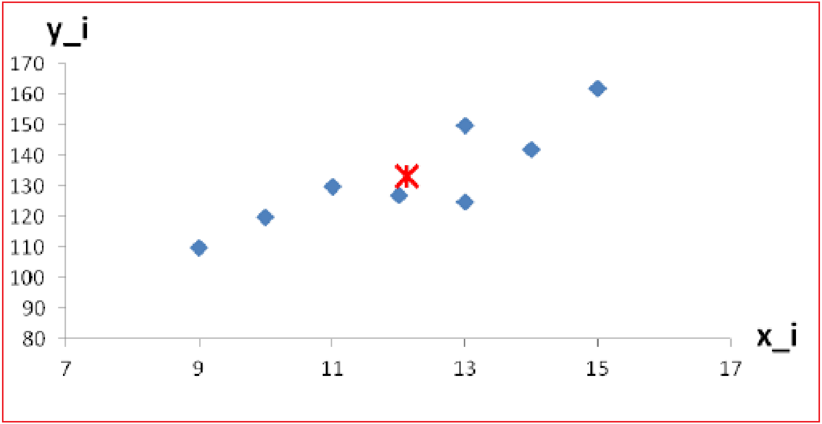


Exemple:

Une entreprise a étudié les montants alloués aux compagnes publicitaires avec le chiffre d'affaires enregistré au cours des dernières années. Le tableau ci-après résume les résultats de cette étude (en millions).

Dépenses publicitaires (X)	10	9	11	13	14	12	13	15
Chiffre d'affaires (Y)	120	110	130	125	142	127	150	162

Statistique descriptive bivariée



Remarque:

Le point en rouge représente le point moyen du nuage de points $G(\overline{x}, \overline{y})$ de coordonnées $\overline{x}$ et $\overline{y}$ où $\overline{x}$ est la moyenne arithmétique de X et $\overline{y}$ est la moyenne arithmétique de Y.
Dans cet exemple, les coordonnées du point moyen sont:

$$\overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = \frac{1}{8}[10+9+11+13+14+12+13+15] = \frac{97}{8} = 12,125$$

et

$$\overline{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{1}{8}[120+110+130+125+142+127+150+162] = \frac{1066}{8} = 133,25$$

Covariance des variables x et y

• On rappelle que :

$$S_x^2 = Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \overline{x}^2$$

et

$$S_y^2 = Var(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(\frac{1}{n}\sum_{i=1}^{n}y_i^2\right) - \overline{y}^2$$

• La covariance entre x et y est définie comme suit :

$$S_{xy} = Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i y_i\right) - (\overline{x}\,\overline{y})$$

Statistique descriptive bivariée

Exemple:

On revient à l'exemple précédent:

Dépenses publicitaires (X)	10	9	11	13	14	12	13	15
Chiffre d'affaires (Y)	120	110	130	125	142	127	150	162

Statistique descriptive bivariée

- La variance de x est :

$$S_x^2 = Var(x) = \left(\frac{1}{8}[10^2 + 9^2 + 11^2 + 13^2 + 14^2 + 12^2 + 13^2 + 15^2]\right) - (12,125)^2$$
$$\simeq 3,61$$

- Par conséquent, l'écart-type de x est :

$$S_x = \sqrt{Var(x)} = 1,9$$

- La variance de y est :

$$S_y^2 = Var(y) = \left(\frac{1}{8}[120^2 + 110^2 + 130^2 + 125^2 + 142^2 + 127^2 + 150^2 + 162^2]\right) - (133,25)^2$$

$$\simeq 252,188$$

- Par conséquent, l'écart-type de y est :

$$S_y = \sqrt{Var(y)} \simeq 15,88$$

Statistique descriptive bivariée

• La covariance de entre x et y est :

$$S_{xy} = Cov(x,y) = \left(\frac{1}{8}\sum_{i=1}^{8} x_i y_i\right) - (12,125 \times 133,25)$$

$$= \left(\frac{1}{8}[(10 \times 120) + (9 \times 110) + (11 \times 130) + (13 \times 125) + (14 \times 142)\right.$$

$$\left. + (12 \times 127) + (13 \times 150) + (15 \times 162)]\right) \; - \; (12,125 \times 133,25)$$

$$\boxed{\simeq 26,47}$$

La covariance est positive, ce qui peut être interprété par l'existence d'une relation positive entre les dépenses publicitaires et les chiffres d'affaires.

Coefficient de corrélation

Bien que la covariance mesure le sens et la force de liaison entre deux variables, cette quantité peut être influencée par les unités des variables X et Y. Par conséquent, la covariance reste un indicateur important, mais non suffisant.

Afin d'avoir une mesure plus fiable, nous avons recours au coefficient de corrélation. Ce coefficient mesure **le sens** de la relation linéaire entre deux variables X et Y ainsi que **l'intensité** de cette liaison.

Définition

On appelle coefficient de corrélation de la série statistique double, le nombre r (ou r_{xy}) défini par :

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)} \times \sqrt{Var(y)}} = \frac{Cov(x,y)}{S_x \times S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}}$$

Remarques:

• r est un nombre sans dimension (sans unité)

• Le coefficient r prend le même signe que celui de la covariance :

$$sign(r) = sign(Cov(x, y))$$

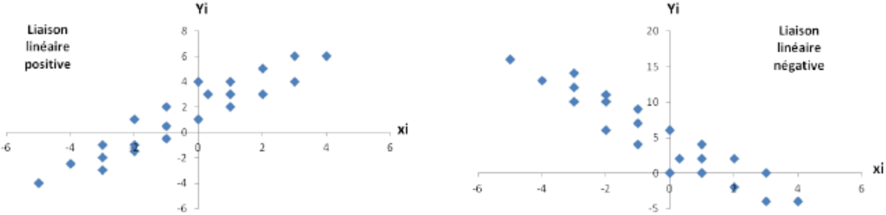
. • On a : $-1 \leq r \leq 1$. En outre :

- Plus le coefficient est proche de 1, plus la relation linéaire **positive** entre les variables est **forte**.

- Plus le coefficient est proche de -1, plus la relation linéaire **négative** entre les variables est **forte**.

- Plus le coefficient est proche de 0, plus la relation linéaire entre les variables est **faible**.

Statistique descriptive bivariée



Exemple

Calculons le coefficient de corrélation entre les dépenses dédiées à la publicité et le chiffre d'affaires dans l'exemple précédent:

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)} \times \sqrt{Var(y)}} = \frac{26,47}{1,9 \times 15,88} \simeq 0,88$$

- La covariance est positive. Par conséquent, le coefficient de corrélation est positif également.

- r est positif : la relation linéaire entre X et Y est **positive** (l'augmentation d'une variable entraînera l'augmentation de l'autre).

- r est proche de 1 : la **dépendance linéaire** entre X et Y est **forte**.

Ajustement linéaire et droite de régression

Dans le cas de l'existence d'une forte corrélation (liaison linéaire) entre deux variables x et y, on peut essayer d'ajuster le nuage de points par une droite d'équation :

$$y = ax + b$$

avec a et b sont des constantes appelées coefficients de régression et la droite représentée par cette équation est appelée droite d'ajustement (ou droite de régression).

y sera appelée la variable **à expliquer (ou dépendante)**

x est la variable **explicative (ou indépendante)**.

En particulier, on s'attend à ce que la droite soit la plus proche possible du nuage de points.

Pour cela, et à partir d'une série bivariée $\{(x_i, y_i)\ ,\ i = 1, \cdots, n\}$ on note :

$$\widehat{y_i} = \widehat{a}x_i + \widehat{b}$$

qui sont les valeurs **ajustées (ou prédites)** de la variable y par la variable x.

avec,

$$\begin{cases} \widehat{a} = \frac{Cov(x,y)}{Var(x)} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\ \widehat{b} = \overline{y} - \widehat{a}\overline{x} \end{cases}$$

Définition (Résidu)

On appelle résidu de l'observation i, la quantité :

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{b} - \widehat{a}x_i$$

C'est l'écart entre la $i^{\text{ème}}$ valeur observée et la $i^{\text{ème}}$ valeur prédite de y.

Qualité d'ajustement:

Pour mesurer la qualité d'ajustement, on calcule le coefficient de détermination noté R^2, qui est égal au coefficient de corrélation au carrée. C'est à dire :

$$R^2 = r^2 = \frac{S_{xy}}{S_x^2 S_y^2}$$

R^2 indique la proportion de la variance de y expliquée par x dans le modèle de régression, et on a :

- $0 \leq R^2 \leq 1$

- Si $R^2 = 0$, alors les variables x et y ne sont pas linéairement corrélées.

- Si $R^2 = 1$, alors la relation linéaire explique toute la variation (l'ajustement est parfait). On dit qu'il y a une corrélation linéaire parfaite entre x et y.

- Plus R^2 est proche de 1, plus **la qualité de l'ajustement linéaire est bonne** et plus la corrélation linéaire entre x et y est **forte**.

Exemple:

On revient à l'exemple précédent, l'équation de régression de y en x est :

$$y = \widehat{a}x + \widehat{b}$$

avec,

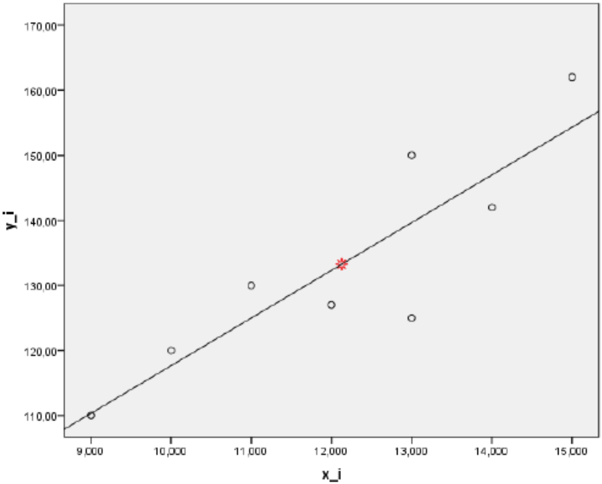
$$\begin{cases} \widehat{a} = \frac{Cov(x,y)}{Var(x)} = \frac{26.47}{3.61} = 7.33 \\ \widehat{b} = \overline{y} - \widehat{a\overline{x}} = 133.25 - (7.33 \times 12.125) = 44.37 \end{cases}$$

d'où, la droite de régression est:

$$y = 7.33x + 44.37$$

Statistique descriptive bivariée

Le graphe ci-dessous contient le nuage de points, la droite de régression d'équation $y = 7.33x + 44.37$ ainsi que le point moyen représenté en rouge.



Coefficient de détermination:

Le coefficient de détermination est égal à $R^2 = r^2 = (0.88)^2 = 0.77$. Ce qui signifie que 77% des variations de y sont expliquées par x.

De plus, la valeur de ce coefficient est proche de 1, ce qui signifie que la qualité de l'ajustement est **"bonne"**, c'est à dire que les points y_i sont assez proches des points $\widehat{y}_i$.

Prévision :

Avec une équation de droite de régression, on peut faire des prévisions. En effet, supposons que l'entreprise en question prévoit d'investir $17 millions$ pour les compagnes publicitaires et souhaite savoir le niveau de son chiffre d'affaires.

Dans ce cas, on peut dire que le chiffre d'affaire y attendu est:
$y = 7.33 \times 17 + 44.37 = 168.98$ millions.

Etude de l'indépendance entre deux variables qualitatives

Test d'indépendance de khi-deux

Ce test est utilisé pour tester d'indépendance entre deux variables qualitatives.

• **Problème de test:**

On cherche à tester les hypothèses suivantes:

$$\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendantes} \quad \text{(pas de correspondance)} \\ H_1 : X \text{ et } Y \text{ sont dépendantes} \quad \text{(il y a une correspondance)} \end{cases}$$

Pour tester ces hypothèses, on considère le tableau de contingence suivant des effectifs observés n_{ij}, ensuite on calcul le tableau des effectifs théoriques E_{ij}:

Test d'indépendance de khi-deux

X \ Y	y_1	...	y_j	...	y_J	Total
x_1						$n_{1.}$
.			.			.
.			.			.
.						.
x_i	.	...	n_{ij}	...	.	$n_{i.}$
.			.			.
.			.			.
.			.			.
x_I			.			$n_{I.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

Table 1: Tableau des effectifs observés

Test d'indépendance de khi-deux

Exemple:

X : Sexe (x_1 =Masculin ; x_2 =Féminin)

Y : Niveau d'étude (y_1 =Licence, y_2 =Master, y_3 =Doctorat)

X \ Y	L	M	D	Total
M	40	25	10	75
F	60	15	5	80
Total	100	40	15	155

Table 2: Tableau des effectifs observés

Test d'indépendance de khi-deux

X \ Y	y_1	...	y_j	...	y_J	Total
x_1						$n_{1.}$
.			.			.
.			.			.
.						.
x_i	.	...	$E_{ij} = \frac{n_{i.}\, n_{.j}}{n}$	...	.	$n_{i.}$
.			.			.
.			.			.
.			.			.
x_I			.			$n_{I.}$
Total	$n_{.1}$	...	$n_{.j}$	...	$n_{.J}$	n

Table 3: Tableau des effectifs théoriques

avec

$$E_{ij} = \frac{\text{Total de ligne } i \times \text{Total de collone } j}{\text{Total}} = \frac{n_{i.}\, n_{.j}}{n}$$

Test d’indépendance de khi-deux

Exemple:

X \ Y	L	M	D	Total
M	E_{11}	E_{12}	E_{13}	75
F	E_{21}	E_{22}	E_{23}	80
Total	100	40	15	155

Table 4: Tableau des effectifs théoriques

$$E_{11} = \frac{75 \times 100}{155} \ ; E_{12} = \frac{75 \times 40}{155} \ ; E_{13} = \frac{75 \times 15}{155}$$

$$E_{21} = \frac{80 \times 100}{155} \ ; E_{22} = \frac{80 \times 40}{155} \ ; E_{23} = \frac{80 \times 15}{155}$$

Exemple:

X \ Y	L	M	D	Total
M	48.39	19.35	7.26	75
F	51.61	20.64	7.74	80
Total	100	40	15	155

Table 5: Tableau des effectifs théoriques

Test d'indépendance de khi-deux

• **La statistique du test:**

La statistique du test est donnée par:

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

Où n_{ij} est l'effectif observé.

Plus la valeur de T est grande, plus le tableau observé est éloigné du tableau théorique (ie le tableau qui contient les effectifs E_{ij}).

Exemple:

$$T = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = \frac{(40 - 48.39)^2}{48.39} + \frac{(25 - 19.35)^2}{19.35} + \cdots + \frac{(5 - 7.74)^2}{7.74} = 8.014$$

Test d'indépendance de khi-deux

• **Règle de décision:**

• Si $T \geq T_c$, alors les deux variables X et Y sont dépendantes.

• Si $T < T_c$, alors les deux variables X et Y sont indépendantes.

Où $T_c = \chi^2_{1-\alpha}((I-1)\times(J-1))$ est le quantile d'ordre $1-\alpha$ d'une loi de χ^2 à $(I-1)(J-1)$ degrés de liberté.

Exemple:

Dans l'exemple précédent, on a $I = 2$ et $J = 3$ donc $T_c = 5.99$

On a $T = 8.014 > T_c = 5.99$ donc les deux variables "Sexe" et "Niveau d'étude" sont dépendantes.

Test d'indépendance de khi-deux

Exemple 2:

On considère deux variables qualitatives X et Y observées sur une population de taille n=592.

avec:

X : représente les couleurs des yeux
Y : représente les couleurs des cheveux

Le tableau de contingence obtenu est le suivant:

X \ Y	brun	châtain	roux	blond
marron	68	119	26	7
noisette	15	54	14	10
vert	5	29	14	16
bleu	20	84	17	94