Cours de Statistique Descriptive

N. BELKHIR

Ecole Supérieure des sciences appliquées de Tlemcen
Formation préparatoire

Table des matières

1 Séries Statistiques à un caractère

1.1 Définition et vocabulaire

La statistique descriptive est un ensemble de méthodes scientifique qui consistent à collecter des données statistiques, puis à analyser, à commenter et à critiquer ces données.
Comme toute science, la statistique fait appel à un vocabulaire spécialisé .
Population : un ensemble sur lequel on effectue des observations.
Exemple : ensemble des étudiants, ensemble des malades ….
Unité statistique (individu) : est un élément de la population
Exemple : un étudiant, un malade….
Echantillon : est un sous ensemble de la population.
Variable statistique (caractère) : est ce qui est observé.
Exemple : note des étudiants, la guérison des malades….
Les variables statistique sont notées par des lettres majuscule X, Y, Z,…
Une variable statistique peut être :

- **Qualitative** : si ses valeurs sont des nombres

On distingue deux cas :

- ✓ Variable discrète : si elle ne prend que des valeurs isolées.
- ✓ Variable continue : si elle peut prendre n'importe quelle valeur intermédiaire entre deux valeurs données.

- **Qualitative** : si ses valeurs sont des mots (littérales).

On distingue trois cas :

- ✓ Ordinale : si on peut ordonner ses valeurs

Exemple : taille de vêtements : S<M<L<XL<XXL

- ✓ Nominale : si on ne peut pas ordonner ses valeurs
  Exemple : profession, lieu de naissance….
- ✓ Dichotomique : si elle admet seulement deux valeurs
  Exemple : sexe : Féminin, Masculin

1.2 Tableau statistique :

Un tableau de statistique comporte un titre indiquant l'objet du travail, l'unité de mesure, a référence de la source des données statistiques.
un tableau théorique est donné comme suit

Les valeurs de la variable X	Les effectifs n_i	Les fréquences f_i
x_1	n_1	f_1
x_2	n_2	f_2
…	…	…
x_k	n_k	f_k
Total	N	1

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

Remarque : les ensembles : $\{(x_i, n_i), i = 1,2, \dots, k\}, \{(x_i, f_i), i = 1,2, \dots, k\}$ sont appelés distribution statistique.

Si la variable est continue ses valeurs sont regroupés dans des classes de la forme : [e_1 , e_2 [, [e_2 , e_3 [, [e_3 ; e_4 [,... [e_k ; e_{k+1} [et on notera n_1 , n_2 , ... , n_k les effectifs associés.

n_i est le nombre d'individus appartenant à la classe [e_i ; e_{i+1} [.

L'amplitude de la classe i est donnée par $a_i = e_{i+1} - e_i$

Le centre de la classe i est donnée par $C_i = \frac{e_{i+1} + e_i}{2}$

Le nombre de classes k est donné par :

Règle de Sturge : $k = 1 + 3.3\log(N)$

Règle de Yule : $k = 2.5\sqrt[4]{N}$

L'amplitude $a = \frac{x_{max} - x_{min}}{k}$

Effectifs cumulé croissant : $N_i \uparrow = \sum_{l=1}^{i} n_k = n_1 + n_2 + \cdots + n_i$ c'est le nombre d'individus qui correspond à une valeur inférieure à x_i
Effectifs cumulés décroissant : $N_i \downarrow = \sum_{l=i}^{k} n_k = n_i + n_{i+1} + \cdots + n_k$ c'est le nombre d'individus qui correspond à une valeur supérieure ou égale à x_i
Fréquences cumulée croissante : $F_i \uparrow = \sum_{l=1}^{i} f_k = f_1 + f_2 + \cdots + f_i$ c'est la proportion des individus qui correspond à une valeur inférieure ou égale à x_i
Fréquences cumulée décroissante : $F_i \downarrow = \sum_{l=i}^{k} f_k = f_i + f_{i+1} + \cdots + f_k$ c'est la proportion des individus qui correspond à une valeur supérieure à x_i

Remarque : les ensembles : $\{(x_i, N_i), i = 1,2, \dots, k\}, \{(x_i, F_i), i = 1,2, \dots, k\}$ sont appelés répartition statistique.

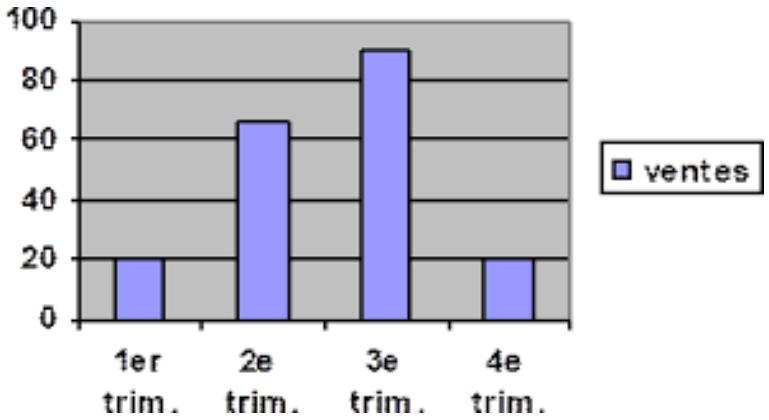
1.3 Représentation graphique

1.3.1 Graphique pour les variables qualitatives

Les variables qualitatives sont représentées par

- **Diagramme en barre**
  Chaque rectangle a une base constante et une hauteur proportionnelle à l'effectif n_i ou à la fréquence f_i

Exemple :

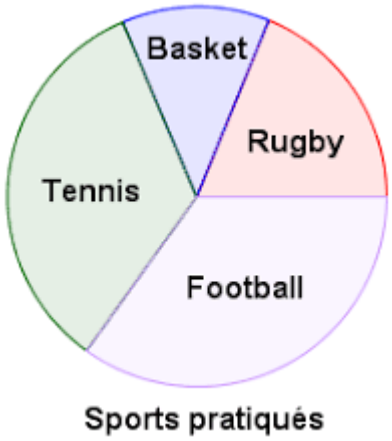


- **Diagramme circulaire**

  Chaque modalité est représentée par un secteur circulaire dont l'angle (et donc la surface) est proportionnel à son effectif. le à Le rayon du cercle est arbitraire.

Pour une modalité *i* l'angle $\theta_i = f_i \times 360$

Exemple :



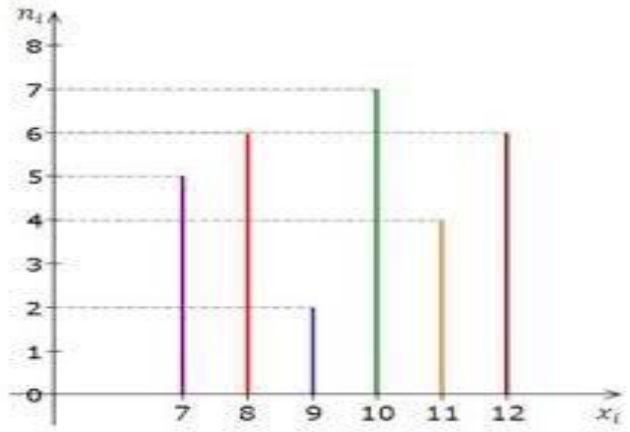
1.3.2 Graphique pour les variables quantitatives

Les variables quantitatives ont deux types de graphiques :

Graphiques de distribution

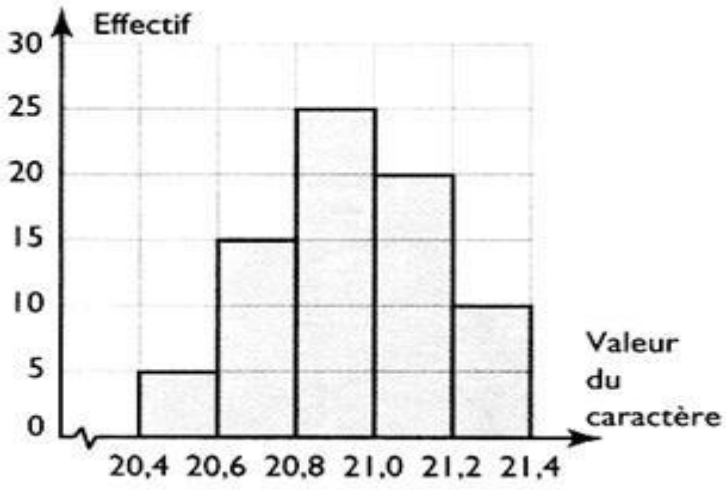
- Les variables discrètes sont représentées par des bâtons :
  Les x_i sont placés suivant une échelle sur l'axe des abscisses, et les effectifs n_i sont représentés par un "bâton" de longueur n_i (axe des ordonnées).

Exemple :



- Les variables continues sont représentées par des histogrammes
  L'histogramme consiste à représenter les effectifs (resp. les fréquences) des classes par des rectangles contigus dont la surface (et non la hauteur) représente l'effectif (resp. la fréquence).
  Pour un histogramme des effectifs, la hauteur du rectangle correspondant à la classe *i* est donc donnée par : $h_i = \frac{n_i}{a_i}$, h_i est appelé densité d'effectif.
  L'aire de l'histogramme est égale à l'effectif total.
  Pour un histogramme des fréquences on a : $d_i = \frac{f_i}{a_i}$
  On appelle d_i la densité de fréquence.

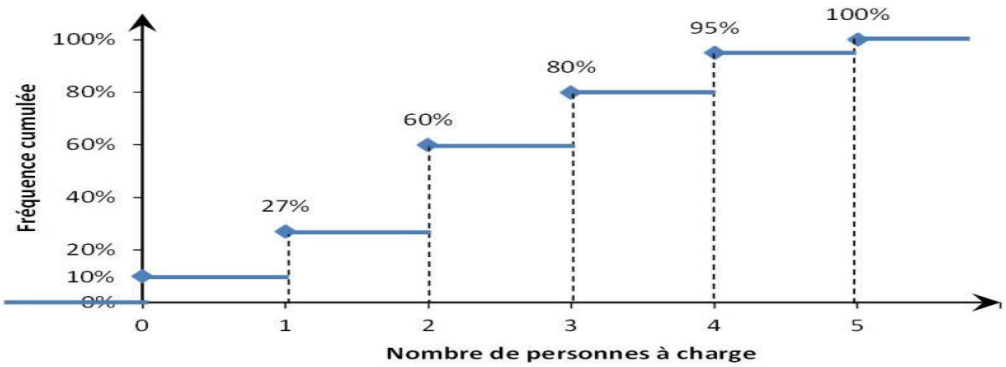
Exemple :



Graphiques de répartition

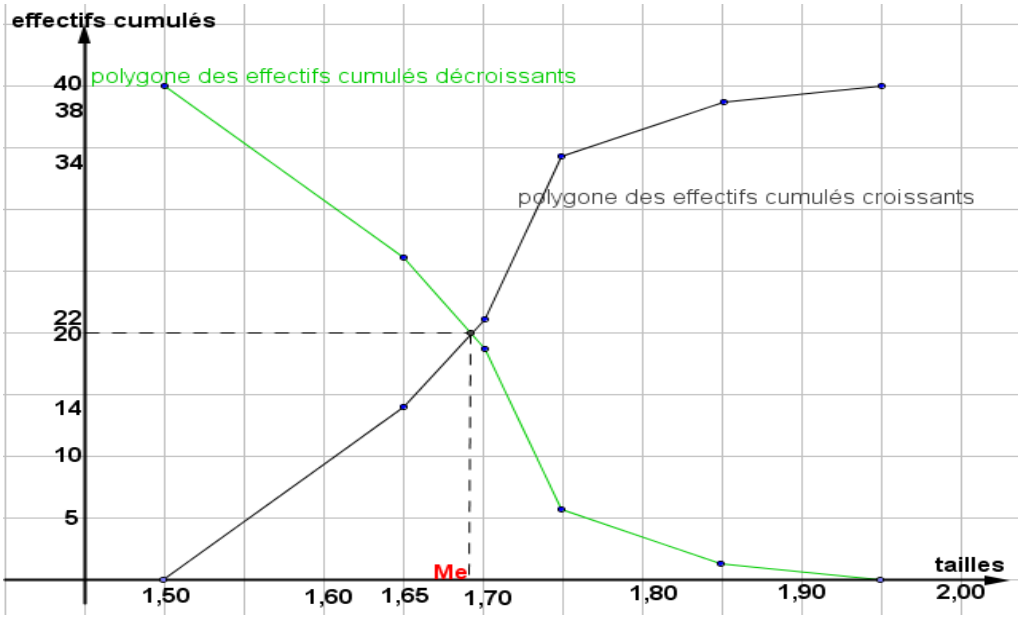
- Les variables discrètes sont représentées par des courbes des fonctions en Escaliers.

Exemple :



- Les variables continues sont représentées par les courbes cumulatives croissantes et décroissantes

  Exemple :



Pour tracer la courbe de $F^{\nearrow}$ (ou $N^{\nearrow}$) on utilise les bornes supérieure des classes et pour tracer la courbe de $F^{\searrow}$ (ou $N^{\searrow}$) on utilise les bornes inférieure des classes.

1.4 Paramètres Statistiques

Les paramètres statistiques ont pour but de résumer, à partir de quelques nombres clés, l'essentiel de l'information relative à l'observation d'une variable statistique.

On définira plusieurs sortes de paramètres :

1.4.1 Paramètres de tendance centrale

Ces paramètres représentent une valeur numérique autour de laquelle les observations sont réparties.

1.4.1.1 Moyenne arithmétique

La moyenne peut être calculée seulement pour une variable quantitative.
Pour une série brute numérique x_1 , x_2 , ... , x_N , la moyenne est donnée par

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

Pour une série groupée, où les x_i sont les différentes valeurs de la variable et les n_i les effectifs associés

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

Remarque

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

- Si la variable est continue

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

1.4.1.2 Mode

Le mode peut être calculé pour tous types de variables (quantitative et qualitative).
Le mode est la valeur de la variable qui le plus grand effectif (plus grande fréquence) , il est noté Mo.

- **Variable qualitative**

Si on reprend la variable 'situation de famille', dont le tableau statistique est le suivant:

modalités	célibataire	Marié	divorcé	veuf	Total
n_i	30	80	20	20	150

Le mode est Mo= 'Marié'

- **Variable quantitative discrète**
  Exemple : Nombre de pièces par logement

x_i	1	2	3	4	5	6	Total
n_i	5	10	25	13	8	4	65

Le nombre de pièces fréquent est Mo = 3

- **Variable quantitative continue**
  Dans le cas continu, le mode se trouve dans la classe ayant le plus grand effectif appelée classe modale.
  Il se calcule sur l'histogramme par la formule suivante :

$$\boldsymbol{M_o = L + \frac{E_1}{E_1 + E_2} a_i}$$

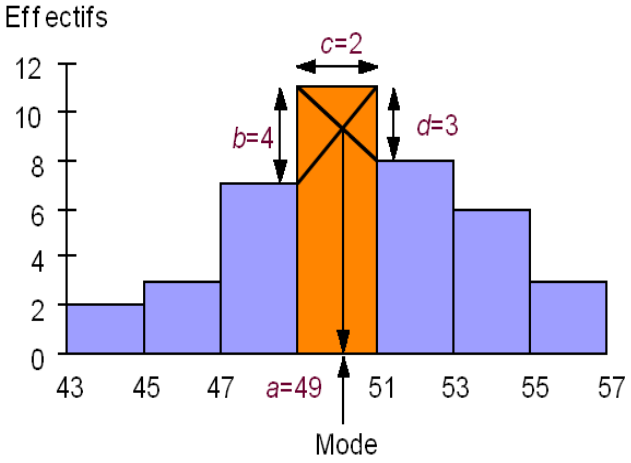
Où
L : est la borne inférieure de la classe modale.
$\boldsymbol{E_1}$ **:** est l'écart entre l'effectif de la classe modale et l'effectif de la classe précédente.
$\boldsymbol{E_2}$ **:** est l'écart entre l'effectif de la classe modale et l'effectif de la classe suivante.
$\boldsymbol{a_i}$: est l'amplitude de la classe modale.

Cas de classes de mêmes amplitudes
Lors d'une course de vitesse, les 40 participants ont mis les temps ci-contre pour effectuer le parcours.

Temps	c_i	n_i
$[43,45[$	44	2
$[45,47[$	46	3
$[47,49[$	48	7
$[49,51[$	50	11
$[51,53[$	52	8
$[53,55[$	54	6
$[55,57[$	56	3

La classe modale est $[49,51[$
Ainsi

$$Mo = 49 + \frac{4}{4+3} \times 2 = 50.14$$



Cas de classes d'amplitudes inégales

Si les amplitudes sont différentes, afin de constituer l'histogramme, il est nécessaire de :

Calculer, pour chaque classe, l'amplitude ai ;

Calculer la densité $h_i = \frac{n_i}{a_i}$ pour un histogramme des effectifs, et $d_i = \frac{f_i}{a_i}$ pour un histogramme de fréquences ;

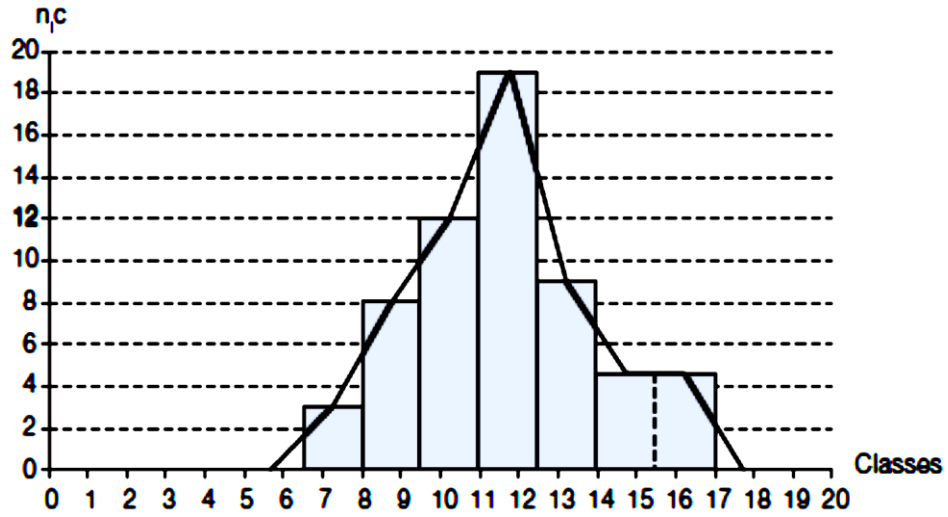
Affecter à chaque rectangle une hauteur proportionnelle à la densité d_i de la classe correspondante.

Exemple :

Le responsable des ressources humaines d'une entreprise a relevé la distribution statistique suivante correspondant à l'ancienneté du personnel cadre dans l'entreprise, exprimée en années :

Tracez l'histogramme des effectifs et estimez la proportion de cadres ayant une ancienneté comprise entre 10 et 13,25 années.

Classes	n_i	a_i	h_i	$n_{ic} = h_i \times pgcd(a_i)$
$[6.5,8[$	3	1.5	2	3
$[8,9.5[$	8	1.5	5.33	8
$[9.5,11[$	12	1.5	8	12
$[11,12.5[$	19	1.5	12.67	19
$[12.5,14[$	9	1.5	6	9
$[14,17[$	9	3	3	4.5
Total	60	/		



Pour estimer l'effectif des cadres de l'intervalle $[10,13.25[$, on décompose celui-ci en trois intervalles : $[10,11[, [11,12.5[\; et \; [12.5,13.25[$. Il suffit alors de multiplier, pour chacun de ces intervalles, l'amplitude par la densité pour obtenir l'effectif total recherché

Intervalles	a_i	h_i	$n_i \; estimés$
$[10,11[$	1	8	8
$[11,12.5[$	1.5	12.67	19.01
$[12.5,13.25$	0.75	6	4.5
Total	/	/	31.51

Ce qui donnera pour estimation de la proportion recherchée

$\frac{31.51}{60} = 0.5251 \; soit \; 52.51\%$

La classe modale est $[11,12.5[$

$$Mo = 11 + \frac{7}{7+10} \times 1.5 = 11.62$$

1.4.1.3 Médiane

La **médiane** est la valeur de la variable qui partage la population en deux sous populations égales.

Méthode de calcul

Pour une série statistique brute

On ordonne les données chiffrées par ordre croissant.

Soit *N* le nombre d'unités statistiques et *k* le rang d'une valeur dans la série.

- **Si *N* est pair :** Dans ce cas la médiane est égale à la moyenne arithmétique de $x_k \; et \; x_{k+1}$ où *k* est tel que *N = 2k [k=n/2]*

Exemple : si on prend la série S = {4, 0, 1, 1, 2, 2, 2, 3, 3, 4}. On classe les valeurs par ordre croissant:

0 1 1 2 2 2 3 3 4 4

Puisque $N = 10 = 2k \; alors \; k = 5$*n=10*, *n* est pair.

Donc :

$$M_e = \frac{x_5 + x_6}{2} = \frac{2+2}{2} = 2$$

- **Si *N* est impair :** Dans ce cas la médiane est égale à x_{k+1} où *k* est tel que $N = 2k + 1$

Exemple : si on prend la série précédente et on enlève le 10ème élément, on alors la série

0 1 1 2 2 2 3 3 4

$N = 9 = 2k + 1 \; et \; donc \; k = 4 \; et \; k + 1 = 5$

$$M_e = x_5 = 2$$

Pour une série statistique groupée

La médiane est aussi la valeur de la variable statistique qui a pour fréquence cumulée croissante la valeur 0.5.

Dans le cas d'une série statistique classée dans un tableau de statistique (série groupée), il n'existe pas toujours une valeur médiane Me pour laquelle la fréquence cumulée croissante vaut exactement 0.5. Il faut donc dans ce cas chercher la première valeur pour laquelle sa fréquence cumulée dépasse 0.5.

Exemple :

x_i	n_i	f_i	F_i ↗
0	1	0.05	0.05
1	3	0.16	0.21
2	5	0.26	0.47
3	5	0.26	0.73
4	4	0.21	0.94
5	1	0.05	0.99
Total	19	1	/

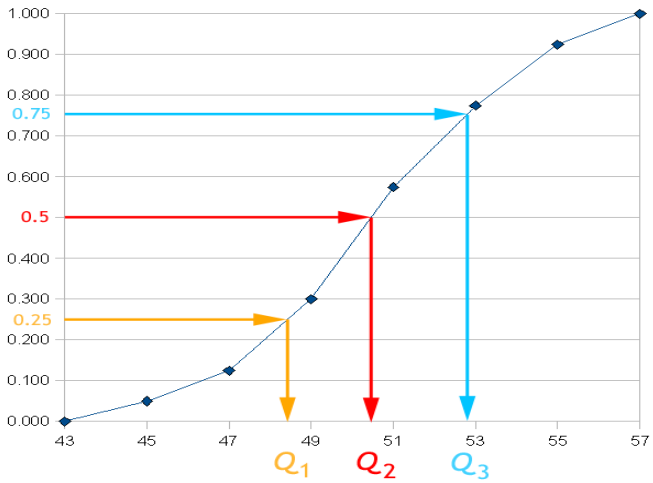
Me=3

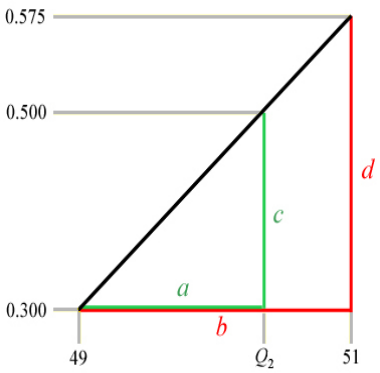
Dans le cas d'une variable continue, la médiane se calcule en utilisant le polygone des fréquences cumulées., on repère quel segment coupe la droite horizontale d'ordonnée 0.5, puis on calcule la médiane par proportionnalité grâce au théorème de *Thalès*.

Exemple :

Temps	c_i	n_i	fi	Fi ↗
$[43,45[$	44	2	0.05	0.05
$[45,47[$	46	3	0.075	0.125
$[47,49[$	48	7	0.175	0.3
$[49,51[$	50	11	0.275	0.575

[51,53[	52	8	0.2	0.775
[53,55[	54	6	0.15	0.925
[55,57[	56	3	0.075	1
Total	/	40	1	/





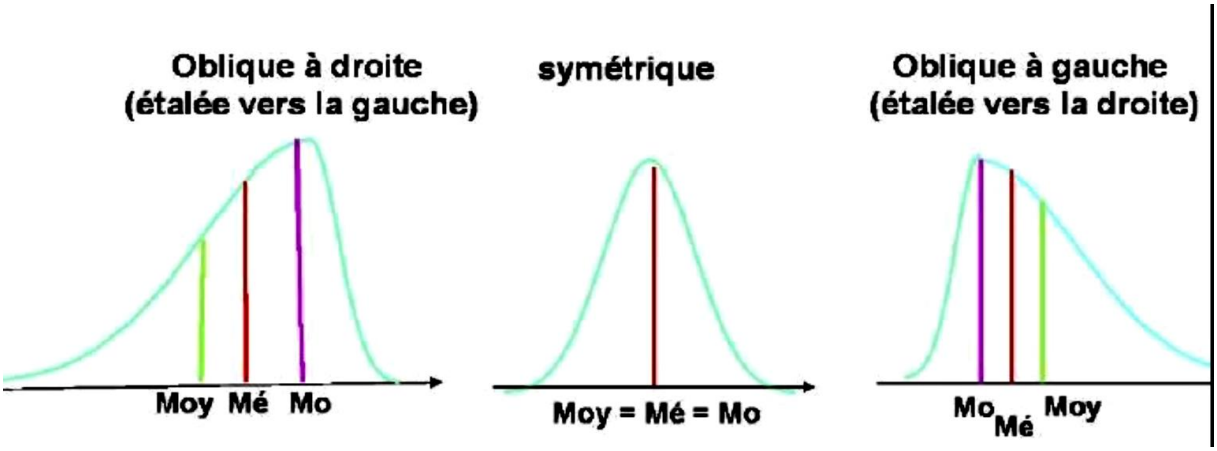
$$\frac{a}{b}=\frac{c}{d}\Rightarrow\frac{Me-49}{51-49}=\frac{0.5-0.3}{0.575-0.3}\Rightarrow Me=49+2\times\frac{0.2}{0.275}=50.45$$

Remarque sur la forme de la distribution statistique

Si $\bar{X}=Me=Mo$, la distribution statistique est symétrique.

Si $\bar{X}<Me<Mo$, la distribution statistique est asymétrique à gauche.

Si $\bar{X}>Me>Mo$, la distribution statistique est asymétrique à droite .



1.4.2 Paramètres de dispersions

1.4.2.1 Etendue

C'est l'écart entre la plus grande valeur et la plus petite valeur des données statistiques :

$$E=x_{max}-x_{min}$$

1.4.2.2 Variance

La variance est un indicateur de la dispersion d'une série par rapport à sa moyenne. Elle est donnée par :

Pour une série statistique brute

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2 = \frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \bar{X}^2$$

Pour une série groupée :

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

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

Remarques :
Plus la variance est élevée, plus la dispersion autour de la moyenne est élevée.
La variance est un nombre réel strictement positif.

Dans le cas d'une variable continue pour calculer la variance il suffit de remplacer les xi par les ci

Exemple 1

x_i	n_i	f_i	$f_i x_i$	$f_i {x_i}^2$
0	5	0.05	0	0
1	15	0.15	0.15	0.15
2	20	0.2	0.4	0.8
3	30	0.3	0.9	2.7
4	25	0.25	1	4
5	5	0.05	0.25	1.25
Total	100	1	1.7	8.9

$\bar{X} = 1.7$
$Var(X) = 8.9 - 1.7^2 = 6.01$

1.4.2.3 Ecart type

l'écart type est la mesure de dispersion la plus utilisée, il est donné par la racine carré de la variance

$$\sigma(x) = \sqrt{Var(X)}$$

Pour l'exemple précédent $\sigma(X) = \sqrt{Var(X)} = \sqrt{6.01} = 2.45$

Exemple 2

Temps	c_i	f_i	$f_i c_i$	$f_i {c_i}^2$
$[43,45[$	44	0.05	2.2	96.8
$[45,47[$	46	0.075	3.45	158.7
$[47,49[$	48	0.175	8.4	403.2
$[49,51[$	50	0.275	13.75	687.5
$[51,53[$	52	0.2	10.4	540.8
$[53,55[$	54	0.15	8.1	437.4
Total		1	46.3	2324.4

$Var(x) = 2324.4 - 46.3^2 = 180.71$

Et $\sigma(x) = \sqrt{V(x)} = \sqrt{180.71} = 13.44$

Remarques

• Si l'écart-type est faible, cela signifie que les valeurs sont assez concentrées autour de la moyenne.
• Si l'écart-type est élevé, cela veut dire au contraire que les valeurs sont plus dispersées autour de la moyenne.

2 Séries statistiques à deux caractères

2.1 Introduction

La mesure de deux variables, désignées par X et Y, sur n éléments (individus, objets, ...) donne lieu à une série statistique bivariée, $\{(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_N, y_N)\}$

L'analyse statistique d'une telle série s'effectue selon une démarche ayant de nombreux points communs avec celle suivie dans le cas uni varié. Cependant, l'objectif poursuivi ici est double : il consiste à explorer, organiser et décrire les données afin :

a) d'analyser les valeurs observées pour X d'une part et pour Y d'autre part;

b) d'analyser le lien éventuel entre les valeurs prises par X et celles prises par Y

2.2 Distribution conjointe

- On note $x_i, i = 1, \dots, k$ les k modalités ou valeurs de la variable X
- On note $y_j, j = 1, \dots, l$ les l modalités ou valeurs de la variable Y
- Les deux variables X et Y sont mesurées simultanément sur chacun des n individus de la population. On note , n_{ij} l'effectif correspondant au couple (x_i, y_j)

2.2.1 Distribution conjointe des effectifs

On appelle distribution conjointe des effectifs de X et Y l'ensemble des informations $(x_i, y_j, n_{ij}), i = 1,2, \dots, k, j = 1,2, \dots, l$

Cette distribution conjointe peut être présentée sous la forme d'un tableau à doubles entrées appelé **tableau de contingence** (tableau croisé) ci-dessous.

X \ Y	y_1	y_2	...	y_j	...	y_l	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	
...	...						
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	
...	...	...	...	...	...	...	
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	

Remarque:

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N$$

Exemple:

Une enquête réalisée auprès de 80 familles d'une ville comporte les deux questions suivantes :

- Quel est le nombre de pièces de votre appartement ? → variable X
- Combien avez-vous d'enfants jusqu'à ce jour ? → variable Y

Le tableau de contingence suivant résume les réponses données à ces deux questions.

X \ Y	0	1	2	3	4
1	4	4	2	0	0
2	9	16	4	0	0
3	4	12	9	2	0
4	1	6	1	1	2
5	0	1	0	1	1

$N = 80$

Ce tableau nous indique par exemple que, parmi les 80 familles prises en compte, 16 ont eu 2 enfants et n'ont qu'une seule pièce; 9 familles ont eu 3 enfants et ont 2 pièces.

2.2.2 Distribution conjointe des fréquences

On appelle distribution conjointe des fréquences de X et Y l'ensemble des informations $(x_i, y_j, f_{ij}), i = 1,2, \dots, k, j = 1,2, \dots, l$.

La fréquence f_{ij} du couple (x_i, y_j) est définie par :

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

On peut ainsi construire le tableau de fréquences ci-dessous.

Y / X	y_1	y_2	...	y_j	...	y_l	Total

x_1		f_{11}	f_{12}	…	f_{1j}	…	f_{1l}	
x_2		f_{21}	f_{22}	…	f_{2j}	…	f_{2l}	
…	…							
x_i		f_{i1}	f_{i2}	…	f_{ij}	…	f_{il}	
…	…							
x_k		f_{k1}	f_{k2}	…	f_{kj}	…	f_{kl}	
Total								1

Remarque

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1$$

Exemple

X Y	0	1	2	3	4	
1	0.05	0.05	0.025	0	0	
2	0.1125	0.2	0.05	0	0	
3	0.05	0.15	0.1125	0.025	0	
4	0.0125	0.075	0.0125	0.0125	0.025	
5	0	0.0125	0	0.0125	0.0125	
						1

Si les données sont brutes : elles sont représentées par N couples (x_i, y_i) $ou\ les\ x_i\ et\ les\ y_i$ sont les valeurs de X et Y pour l'individu i ; elles sont généralement présentées sous forme de tableau :

Individus	1	2	…	i	…	N
X	x_1	x_2	…	x_i	…	x_N
Y	y_1	y_2	…	y_i	…	y_N

Remarque

A partir des données brutes on peut toujours construire le tableau de contingence, alors qu'à partir du tableau de contingence on ne peut pas reconstituer la liste des données brutes

2.3 Distributions marginales :

On ajoute au tableau de contingence les totaux en ligne et en colonne

X \ Y	y_1	y_2	...	y_j	...	y_l	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1l}	$n_{1.}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2l}	$n_{2.}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{il}	$n_{i.}$
...	...	...	...	...	...	...	...
x_k	n_{k1}	n_{k2}	...	n_{kj}	...	n_{kl}	$n_{k.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.l}$	N

- En marge à droite (totaux en ligne) : la distribution de X : pour chaque indice i, l'effectif $n_{i.}$ est le nombre total d'observations de la modalité x_i de X quelque soit la modalité de Y

$$n_{i.} = \sum_{j=1}^{l} n_{ij} = total\ de\ la\ ligne\ i$$

Les k couples $(x_i, n_{i.})$ définissent la distribution marginale de la variable X et on a :

$$\sum_{i=1}^{k} n_{i.} = N$$

- En marge on bas (totaux en colonne) : la distribution de Y : pour chaque indice j, l'effectif $n_{.j}$ est le nombre total d'observations de la modalité y_j de Y quel que soit la modalité de X

$$n_{.j} = \sum_{i=1}^{k} n_{ij} = total\ de\ la\ colonne\ j$$

Les l couples $(y_j, n_{.j})$ définissent la distribution marginale de la variable Y .

on a :

$$\sum_{j=1}^{l} n_{.j} = N$$

Remarque:

On peut aussi introduire des fréquences marginales :

- à la valeur x_i est associée la fréquence marginale

  $f_{i.} = \frac{n_{i.}}{N}$, i=1,...,k

- à la valeur y_j est associée la fréquence marginale

  $f_{.j} = \frac{n_{.j}}{N}$, j=1,...,l

- On peut également considérer des effectifs cumulés marginaux**,** des fréquences cumulées marginales**,** etc.

Exemples

Cas de deux variables quantitatives

X \ Y	0	1	2	3	4	Total
1	4	4	2	0	0	10
2	9	16	4	0	0	29
3	4	12	9	2	0	27
4	1	6	1	1	2	11
5	0	1	0	1	1	3
Total	18	39	16	4	3	N=80

Population : les familles

Le nombre d'observations N=80

X : « le nombre de pièces par logement » : variable discrète

Y : « le nombre d'enfants» : variable discrète

Distribution de X

X	1	2	3	4	5	Total
$n_{i.}$	10	29	27	11	3	80

Distribution de Y

Y	0	1	2	3	4	Total
$n_{.j}$	18	39	16	4	3	80

Cas de deux variables qualitative

Pour étudier les liens entre le niveau scolaire et l'assiduité, on a fait une enquête en mesurant sur des élèves le niveau scolaire X et l'absentéisme en classe Y

Population : les élèves

Le nombre d'observations N=27

X : « le niveau scolaire » : variable ordinale A<B

Y : « l'assiduité » : variable ordinale Rare<Moyen<Fréquent

X / Y	Rare	Moyen	Fréquent	$n_{i.}$
A	7	4	4	15
B	8	2	2	12
$n_{.j}$	15	6	6	27=N

Distribution de X

X	A	B	Total
$n_{i.}$	15	12	27

Distribution de Y

Y	Rare	Moyen	Fréquent	Total
$n_{.j}$	15	6	6	27

2.4 Principales caractéristiques

2.4.1 Moyenne des distributions marginales

Moyenne de X :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\, x_i = \sum_{i=1}^{k} f_{i.}\, x_i$$

Moyenne de Y :

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\, y_j = \sum_{j=1}^{l} f_{.j}\, y_j$$

2.4.2 Variance des distributions marginale

Variance et écart type de X :

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i.}\, x_i{}^2 - \bar{X}^2$$

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

Variance et écart type de Y :

$Var(Y) = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\, y_j{}^2 - \bar{Y}^{\,2}$ $\qquad$ $\sigma(Y) = \sqrt{Var(Y)}$

2.5 Distribution conditionnelle :

Le but est d'analyser le comportement de l'une des deux variables quand l'autre a une valeur donnée.

La distribution des observations suivant les modalités de Y sachant que la variable X prend la modalité xi est appelée distribution conditionnelle de Y pour $X = x_i$

A la ligne i du tableau de contingence, on lit la distribution de Y sachant que $X = x_i$, notée $Y\backslash X = x_i$

Exemple :

$Y\backslash X = A$	Rare	Moyen	fréquent	Total
$n_{j\backslash i=1} = n_{1j}$	7	4	4	15

La fréquence conditionnelle de Y sachant que $X = x_i$ est

$$f_{j\backslash i} = \frac{n_{ij}}{n_{i.}}$$

Exemple

$Y\backslash X = A$	Rare	Moyen	fréquent	Total
$f_{j\backslash i=1}$	7/15	4/15	4/15	1

La distribution des observations suivant les modalités de la variable X sachant que la variable Y prend la modalité y_j, est appelée distribution conditionnelle de X pour $Y = y_j$

A la colonne j du tableau de cotingence on lit la distribution de la variable X sachant que $Y = y_j$

Exemple

$X\backslash Y = Moyen$	A	B	Total
$n_{i\backslash j=2} = n_{i2}$	4	2	6

La fréquence conditionnelle de X sachant que $Y = y_j$ est

$$f_{i\backslash j} = \frac{n_{ij}}{n_{.j}}$$

Exemple :

X\Y=Moyen	A	B	Total
$f_{i\backslash j=2}$	4/6	2/6	1

2.6 Lien entre les variables

Les deux variables X et Y sont indépendantes si et seulement si l'une des égalités est vérifiée ;

1) $f_{ij} = f_{i.}.f_{.j}$
2) $n_{ij} = \frac{1}{n} n_{i.}.n_{.j}$
3) $f_{i\backslash j} = f_{i.}$
4) $f_{j\backslash i} = f_{.j}$

2.7 Principales caractéristiques

2.7.1 Moyenne des distributions conditionnelles

Moyenne de X sachant que $Y = y_j$:

$$\mu(X\backslash Y = y_j) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij}\, x_i = \sum_{i=1}^{k} f_{i\backslash j}\, x_i$$

Moyenne de Ysachant que $X = x_i$:

$$\mu(Y\backslash X = x_i) = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij}\, y_j = \sum_{j=1}^{l} f_{j\backslash i}\, y_j$$

2.7.2 Variance des distributions conditionnelles :

Variance de X sachant que $Y = y_j$:

$$Var(X\backslash Y = y_j) = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij}\, {x_i}^2 - \ (\mu(X\backslash Y = y_j))^2$$

Variance de Y sachant que $X = x_i$:

$$Var(Y\backslash X = x_i) = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij}\, {y_j}^2 - \ (\mu(Y\backslash X = x_i))^2$$

Exemple : une entreprise employant 100 hommes relève pour chaque agent son âge ,noté X et le nombre de journées d'absence durant un mois noté Y

X / Y	0	1	2	3	$n_{i.}$
$[20,30[$	0	0	5	15	20
$[30,40[$	0	15	20	0	35
$[40,50[$	15	10	5	0	30
$[50,60[$	0	5	5	5	15
$n_{.j}$	15	30	35	20	100

Distribution conditionnelle de $X\backslash Y = 1$

$X\backslash Y = 1 = y_2$	[20,30[	[30,40[	[40,50[	[50,60[	Total
c_i	25	35	45	55	/
$n_{i\backslash j=2}$	0	15	10	5	30
$n_{i\backslash j=2} c_i$	0	525	450	275	1250
$n_{i\backslash j=2} {c_i}^2$	0	18375	20250	15125	53750

$$\mu(X\backslash Y = 1) = \frac{1}{n_{.2}} \sum_{i=1}^{4} n_{i2}\, c_i = \frac{1}{30} \times 1250 = 41.67$$

$$Var(X\backslash Y = 1) = \frac{1}{n_{.2}} \sum_{i=1}^{4} n_{i2}\, {c_i}^2 - \ (\mu(X\backslash Y = 1))^2 = \frac{1}{30} \times 53750 - (41.67)^2 = 55.28$$

$$\sigma(X\backslash Y = 1) = \sqrt{Var(X\backslash Y = 1)} = \sqrt{55.28} = 7.43$$

Distribution conditionnelle de $Y\backslash X = 55$

$Y\backslash X = 55 = c_4$	0	1	2	3	Total
$n_{j\backslash i=4}$	0	5	5	5	15
$n_{j\backslash i=4} y_j$	0	5	10	15	30
$n_{j\backslash i=4} {y_j}^2$	0	5	20	45	70

$$\mu(Y\backslash X = 55) = \frac{1}{n_{4.}} \sum_{j=1}^{4} n_{4j}\, y_j = \frac{1}{15} \times 30 = 2$$

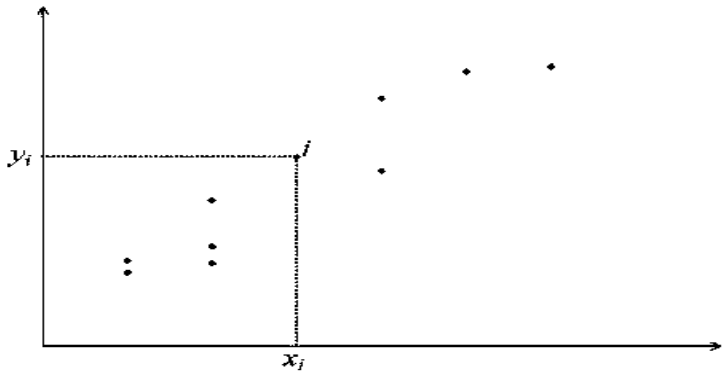
$$Var(Y\backslash X = 55) = \frac{1}{n_{4.}} \sum_{j=1}^{4} n_{4j}\, {y_j}^2 - \ (\mu(Y\backslash X = 55))^2 = \frac{1}{15} \times 70 - 4 = 0.66$$

$$\sigma(Y\backslash X = 55) = \sqrt{Var(Y\backslash X = 55)} = \sqrt{0.66} = 0.81$$

2.8 Nuage de points

Si les deux variables X et Y sont quantitatives, une manière simple de visualiser les données consiste à représenter chaque individu i par un point dans le plan $\mathbb{R}^2$ Cette représentation graphique des données porte le nom de nuage de points. Le centre de gravité du nuage de point est le point G de coordonnées $(\bar{X}, \bar{Y})$

L'examen du nuage de points permet de découvrir l'existence d'une structure d'association particulière entre les deux variables étudiées



La question qu'on peut se poser :

Peut-on quantifier l'intensité de cette association ?

2.9 Covariance, Coefficient de Corrélation

La covariance et le coefficient de corrélation sont des outils pour mesurer la dépendance linéaire entre deux variables quantitatives X et Y.

2.9.1 Covariance :

La covariance est le nombre réel défini par :

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}(x_i - \bar{X})(y_j - \bar{Y})$$

Formule pratique de calcul :

$$Cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j\right) - \bar{X}\bar{Y}$$

Propriétés :

- $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,X) = V(x)$

- $Cov(aX+b,cY+d) = acCov(X,Y)$ pour a, b, c dans $\mathbb{R}$
- Si X et Y sont indépendantes alors $Cov(X,Y) = 0$
- La réciproque est fausse : la covariance peut être nulle sans que les variables X et Y soient indépendantes.

2.9.2 Coefficient de corrélation :

Le coefficient de corrélation linéaire de X et Y est défini par :

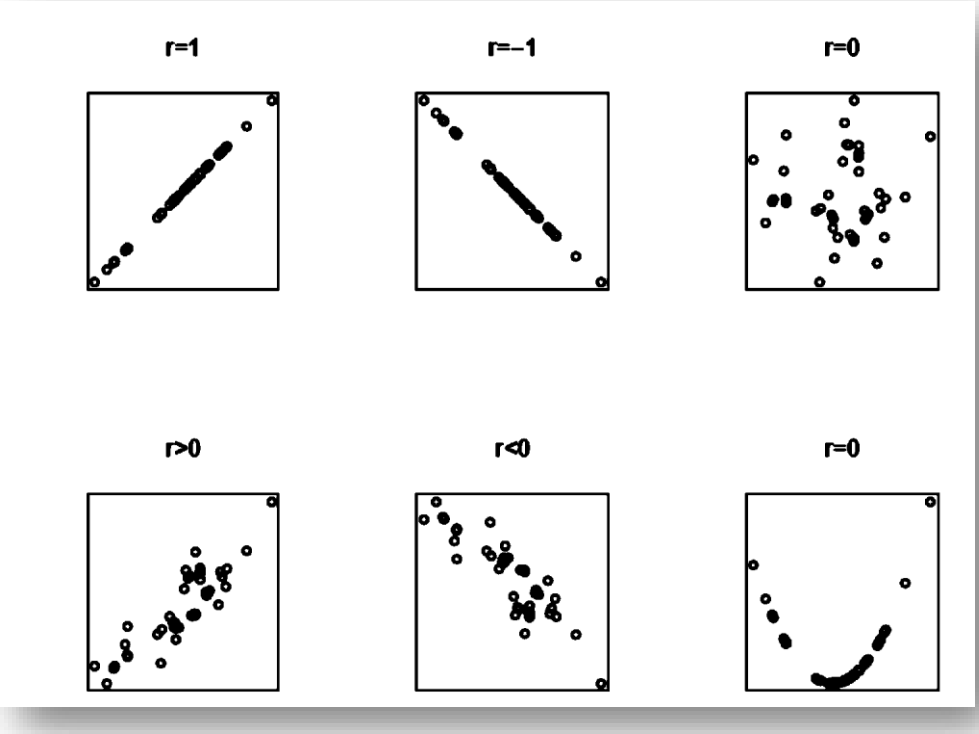
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

Propriétés :

- Le coefficient de corrélation mesure la présence et l'intensité de la liaison entre X et Y
- $\rho(X,Y) \in [-1,1]$

1. $\rho(X,Y) = 1$: liaison linéaire exacte $Y = aX + b \; avec \; a > 0$
2. $\rho(X,Y) = -1$: liaison linéaire exacte $Y = aX + b \; avec \; a < 0$
3. $\rho(X,Y) > 0$: liaison relative, X et Y ont tendance à varier dans le même sens.
4. $\rho(X,Y) < 0$: liaison relative, X et Y ont tendance à varier dans le sens contraire.
5. $\rho(X,Y) = 0$: il n'y a aucun lien linéaire (mais il peut exister un lien non linéaire)

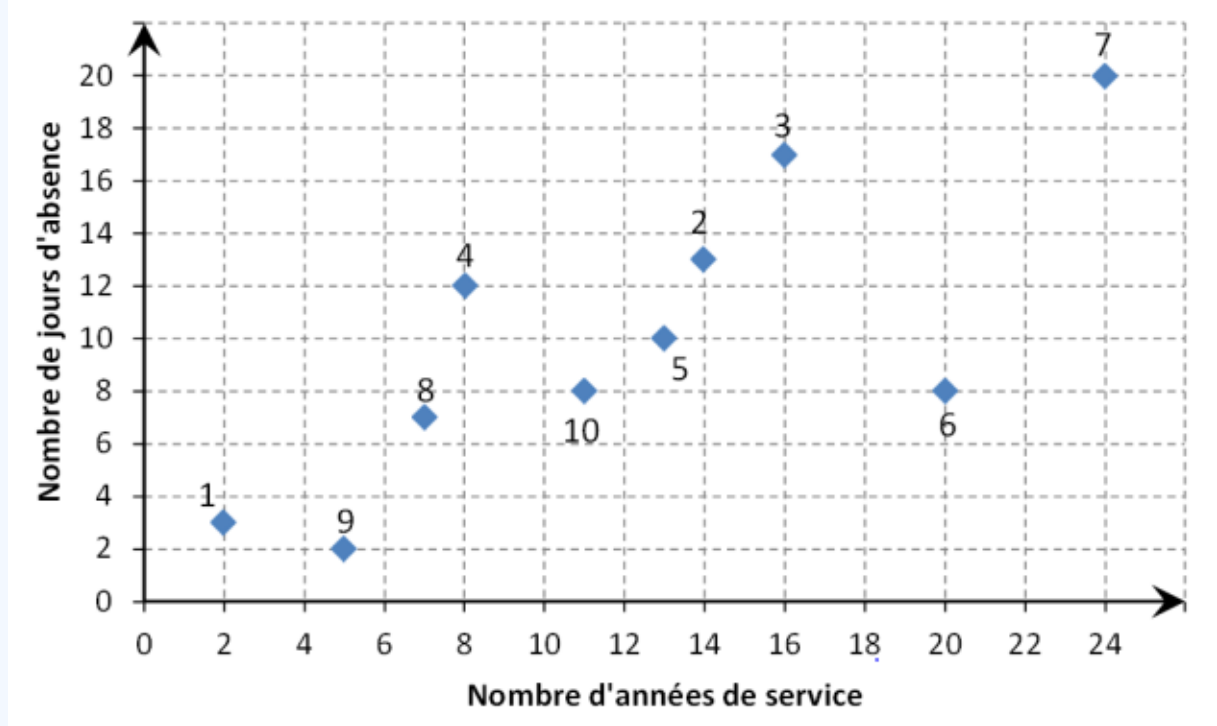
Exemples :



Exemple

Les données suivantes représentent le nombre d'années de service et le nombre de jours d'absence de 10 fonctionnaires.

Fonctionnaire : i	1	2	3	4	5	6	7	8	9	10
Nbre d'années de service : x_i	2	14	16	8	13	20	24	7	5	11
Nbre de jours d'absence : y_i	3	13	17	12	10	8	20	7	2	8



Calcul de la covariance et du coefficient de corrélation.

i	1	2	3	4	5	6	7	8	9	10	total
x_i	2	14	16	8	13	20	24	7	5	11	120
y_i	3	13	17	12	10	8	20	7	2	8	100
$x_i y_i$	3	182	272	96	130	160	480	49	10	88	1467
${x_i}^2$	4	196	256	64	169	400	576	49	25	121	1860
${y_i}^2$	9	169	289	144	100	64	400	49	4	64	1292

Dans cet exemple la série statistique est brute

$$Cov(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right) - \bar{X}\bar{Y}$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{10} \times 120 = 12$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{10} \times 100 = 10$$

Donc

$$Cov(X,Y) = \frac{1}{10} \times 1467 - (12 \times 10) = 27.5$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {x_i}^2 - \bar{X}^2} = \sqrt{\frac{1}{10} \times 1860 - (12)^2} = 6.48$$

$$\sigma(Y) = \sqrt{Var(Y)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} {y_i}^2 - \bar{Y}^2} = \sqrt{\frac{1}{10} \times 1292 - (10)^2} = 5.40$$

Donc

$$\rho(X,Y) = \frac{27.5}{6.48 \times 5.40} = 0.78$$

3 Ajustement Statistique

3.1 Ajustement linéaire

3.1.1 Introduction

Dans le cas où l'association entre X et Y correspond à une relation de dépendance statistique, les deux variables ne jouent pas un rôle symétrique : l'une est (approximativement) fonction de l'autre.

Si Y est fonction de X (c'est-à-dire que le comportement de Y est influencé par celui de X), on dira alors que

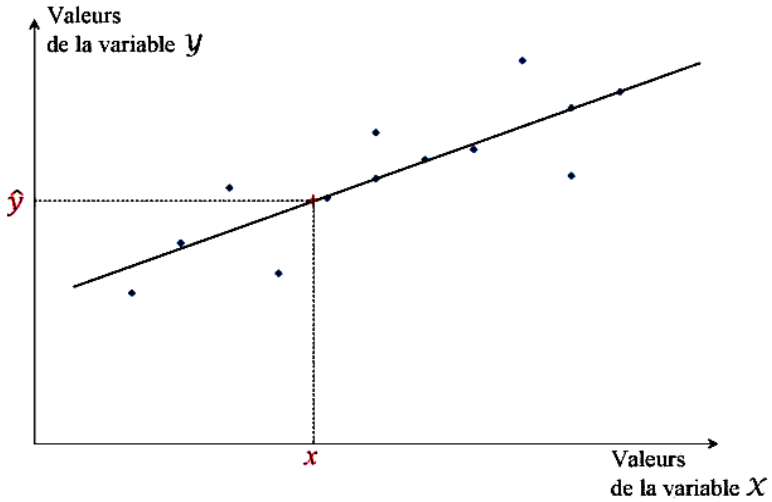
- Y est la variable dépendante (ou expliquée),
- X est la variable explicative.
- Ce type de dépendance statistique est appelé « régression linéaire ».

Le but est de représenter graphiquement cette relation particulière à l'aide d'une droite traversant le nuage de points.

3.1.2 Détermination de la droite de régression : le principe des moindres carrés

Comme toute droite, la droite de régression de Y en X peut être définie au moyen d'une équation du premier degré.

De manière générale, désignons par $(x, \hat{y})$ les coordonnées d'un point appartenant à la droite de régression : l'abscisse x de ce point est une valeur quelconque de la variable explicative X et son ordonnée $\hat{y}$ correspond à la valeur ajustée ou prédite pour Y lorsque X vaut x.

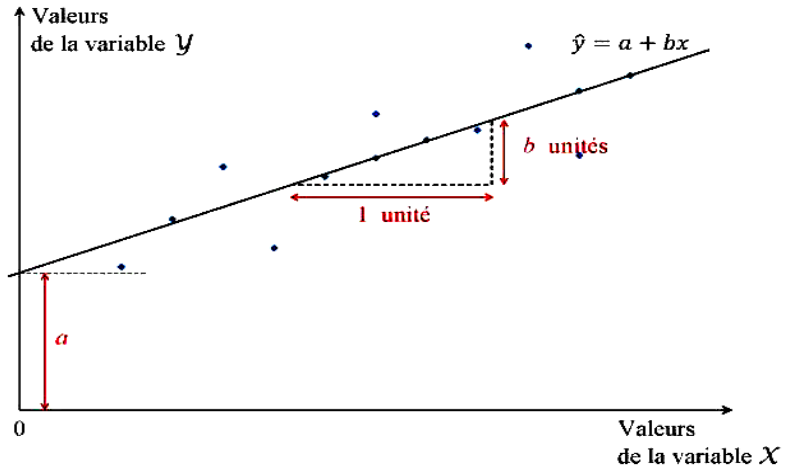


Dans ce cas, l'équation de la droite de régression de Y en X est de la forme :

$$\hat{y} = ax + b$$

où

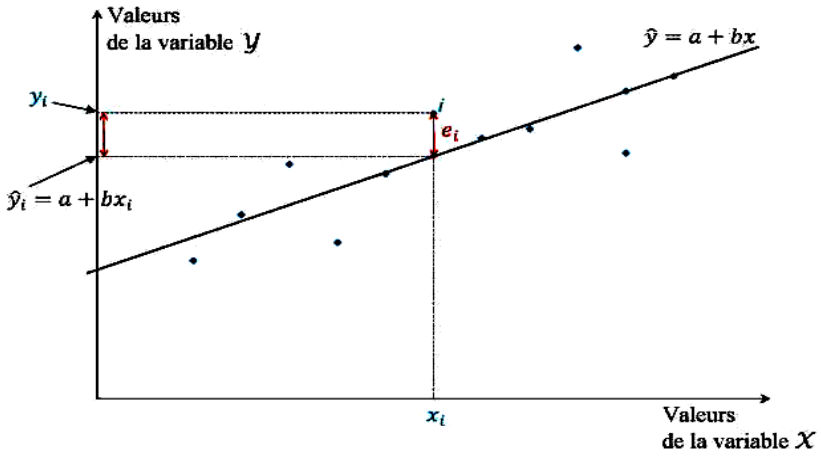
- a est la pente de la droite (aussi appelé le coefficient de régression de Y en X.
- b est la valeur de $\hat{y}$ lorsque $x = 0$.



Déterminer la droite de régression de Y en X revient alors à déterminer les valeurs des coefficients a et b qui permettent à la droite d'assurer un ajustement optimal du nuage de points.

3.1.3 Le principe des moindres carrés

Considérons l'individu i représenté par le point de coordonnées (x_i, y_i).



Pour cet individu i, la droite de régression donne $\hat{y}_i = ax_i + b$ comme valeur ajustée pour la variable Y. Il existe donc un écart appelé résidu ou erreur d'ajustement entre la valeur y_i réellement observée pour la variable dépendante et la valeur ajustée $\hat{y}_i$ fournie par la droite de régression :

$$e_i = y_i - \hat{y}_i = y_i - (ax_i - b)$$

Ainsi, à chaque point du nuage, est associé un résidu (une erreur d'ajustement) e_i le meilleur ajustement est fourni par la droite qui minimise globalement les erreurs d'ajustement $e_i, i = 1, \dots, N$

Le principe des moindres carrés consiste à minimiser la somme des carrés des résidus $\sum_{i=1}^{N} {e_i}^2$pour déterminer les constantes a et b

$$a = \frac{cov(X,Y)}{V(X)} \qquad b = \bar{Y} - a\bar{X}$$

Remarque:

Si nous remplaçons a et b par leurs expressions trouvées ci-dessus dans l'équation de la droite de régression de Y en X cette équation devient

$$\hat{y} = \bar{Y} + \frac{cov(X,Y)}{V(X)} \ (x - \bar{X})$$

Cette reformulation de l'équation nous indique que, si $x = \bar{X}$ alors $\hat{y} = \bar{Y}$, nous constatons ainsi que la droite de régression passe par le centre de gravité G, de coordonnées $(\bar{X}, \bar{Y})$ du nuage de points.

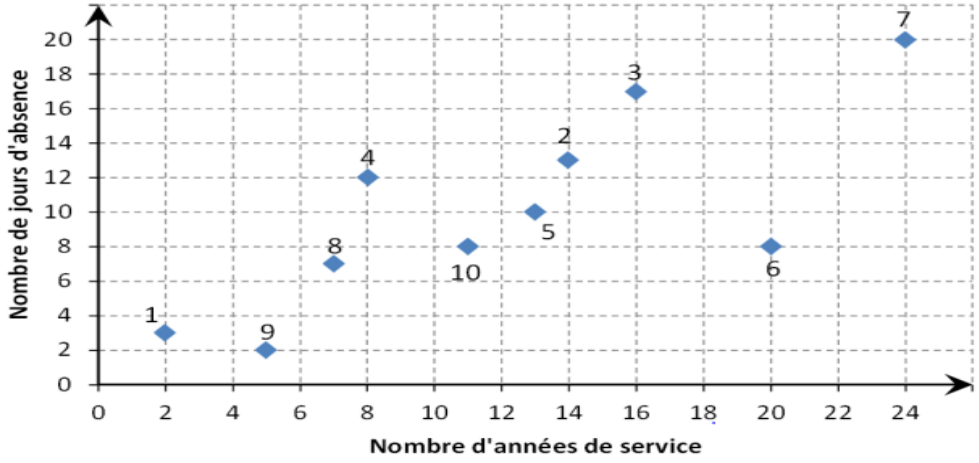
Exemple: Les données suivantes représentent le nombre d'années de service et le nombre de jours d'absence de 10 fonctionnaires

X : Nbre d'années de service

Y : Nbre de jours d'absence

i	1	2	3	4	5	6	7	8	9	10	total
xi	2	14	16	8	13	20	24	7	5	11	120
yi	3	13	17	12	10	8	20	7	2	8	100
xiyi	3	182	272	96	130	160	480	49	10	88	1467
Xi^2	4	196	256	64	169	400	576	49	25	121	1860

Nuage de points :



Déterminons la doite de régression

$$\hat{y} = ax + b$$

On a

$$a = \frac{cov(X,Y)}{V(X)} \quad et \quad b = \bar{Y} - a\bar{X}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{10} x_i = \frac{120}{10} = 12$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{10} y_i = \frac{100}{10} = 10$$

$$Covv(X,Y) = \frac{1}{n}\sum_{i=1}^{10} x_i y_i - \bar{X}\,\bar{Y} = \frac{1}{10} \times 1467 - 12 \times 10 = 27.3$$

$$V(X) = \frac{1}{n}\sum_{i=1}^{10} {x_i}^2 - \bar{X}^2 = \frac{1}{10} \times 1860 - (12)^2 = 42$$

Donc : $a = \frac{27.3}{42} = 0.65 \;\; et \;\; b = \; 10 - \; 0.65 \times 12 = 2.2$

La droite de régression a pour équation

$$\hat{y} = 0.65x - 2.2$$

Cette doite contient le point $G(\bar{X},\bar{Y}) = (12{,}10)$

Pour la dessiner il suffit d'ajouter un autre point par exemple pour x=20 donc le point $M(20{,}10.8)$

Prévision :

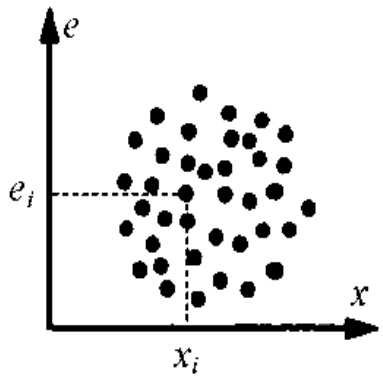
Si un fonctionnaire a 10 ans de service il aura :

$$0.65 \times 10 - 2.2 = 4.3 \cong 4 \; absences$$

Remarques

- la droite de régression nous permet de faire la meilleure prédiction possible en fonction des données connues
- la qualité de la prédiction va dépendre de la force de la correlation :
  plus la correlation est forte et plus la prédiction est fiable.
  plus la correlation est faible et moins la prédiction est fiable

- les prédictions doivent être limitées à des individus qui se trouvent qans le nuage de points (les valeurs connues doivent être comprises entre le minimum et le maximum des valeurs de l'échantillon)
- l'usage d'une droite de régression ne doit pas être automatique dès que le coefficient de correlation est proche de 1 ou -1. Un examen du nuage de points nous informe sur le caractère linéaire ou non entre les deux variables.(voir T.D)
- L'ajustement linéaire réalisé ne peut être considéré comme valable que si la distribution des résidus est bien symétrique et régulière, et si le graphique des résidus ne fait apparaître aucune structure particulière.



3.2 Ajustement par une fonction puissance

Considérons un ajustement de type : $y = bx^a$….(1)

On cherche à déterminer $a\ et\ b$ de sorte que cette fonction réalise un bon ajustement pour le couple (X, Y) c'est-à-dire que l'on ait :

$$\forall\ 1 \leq i \leq N,\ y_i \cong b{x_i}^a$$

Si on compose les deux côtes dans (1) par la fonction *ln*, (1) devient :

$$lny = \ln(\ bx^a) = lnb + alnx$$

Posons :$v_i = ln\ y_i\ ,\ u_i = lnx_i\ \ et\ B = lnb$

Alors les propositions suivantes sont équivalentes :

$$\forall\ 1 \leq i \leq N,\ y_i \cong b{x_i}^a\ \Leftrightarrow\ \forall\ 1 \leq i \leq N, v \cong au_i + B$$

Si on réalise un ajustement linéaire de V en fonction de U que l'on note $v = au + B$ la courbe d'équation $y = bx^a$ réalise un ajustement puissance de Y en fonction de X.

Pour mesurer la validité de l'ajustement puissance du couple (X, Y) on calcule le coefficient de corrélation du couple $(U, V) = (lnX, lnY)$

Pour réaliser un ajustement puissance :

- On calcule $u_i = lnx_i\ et\ v_i = ln\ y_i$

- On détermine l'équation de la droite de régression de v en u par la méthode des moindres carrés : $v = au + B$ avec

$$a = \frac{Cov(U,V)}{Var(U)} \ \ et\ B = \bar{V} - a\bar{U}$$

- De l'équation obtenue on déduit l'équation de la fonction puissance $y = bx^a$ avec $b = e^B$

Exemple

Ajuster les points suivants par une fonction puissance

											Totale
x_i	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	/
y_i	0.1	0.5	1.4	2.7	5.1	7.6	11.2	15.9	22.3	28.1	/
u_i	-0.69	0	0.40	0.69	0.91	1.1	1.25	1.38	1.50	1.61	8.15
v_i	-2.30	-0.69	0.33	0.99	1.63	2.03	2.41	2.76	3.10	3.33	13.59
$u_i v_i$	1.587	0	0.132	0.1831	1.4833	2.233	2.675	3.8088	4.65	5.3613	22.6135
${u_i}^2$	0.4761	0	0.16	0.4761	0.8281	1.21	1.5625	1.9044	2.25	2.5921	11.4593
${v_i}^2$	5.29	0.4761	1.1089	0.9801	2.6569	4.1209	5.8081	7.6176	9.61	11.089	47.7575

La droite de régréssion de *v* en *u* est donnée par :

$v = au + B$ avec

$a = \frac{Cov(U,V)}{Var(U)} \ \ et\ B = \bar{V} - a\bar{U}$

$\bar{U} = \frac{1}{10}\sum_{i=1}^{10} u_i = \frac{1}{10} \times 8.15 = 0.815$

$\bar{V} = \frac{1}{10}\sum_{i=1}^{10} v_i = \frac{1}{10} \times 13.59 = 1.359$

$Cov(U,V) = \frac{1}{10}\sum_{i=1}^{10} u_i v_i - \bar{U}\bar{V} = \frac{1}{10} \times 22.6135 - 0.815 \times 1.359 \cong 1.15$

$Var(U) = \frac{1}{10}\sum_{i=1}^{10} {u_i}^2 - \bar{U}^2 = \frac{1}{10} \times 11.4593 - (0.815)^2 \cong 0.48$

Donc $a = \frac{1.15}{0.48} = 2.39$

$$B = \bar{V} - a\bar{U} = 1.359 - 2.39 \times 0.815 \cong -0.58$$

La droite de régression de *v* en *u* est :

$$v = 2.39u - 0.58$$

Calculons $\rho(U,V)$

$$Var(V) = \frac{1}{10}\sum_{i=1}^{10} {v_i}^2 - \bar{V}^2 = \frac{1}{10} \times 47.7575 - (1.359)^2 \cong 2.93$$

$$\rho(U,V) = \frac{Cov(U,V)}{\sigma(U)\sigma(V)} = \frac{1.15}{\sqrt{0.48 \times 2.93}} = 0.97$$

Donc il y a une très forte corrélation entre le deux variables $U\ et\ V$

L'équation de la fonction puissance est :

$$y = bx^a \quad b = e^B = e^{-0.58} = 0.56$$

D'où l'équation : $y = 0.56x^{2.39}$

Représentation graphique