I - Séries statistiques à une variable:

1- Définition:

la statistique est une méthode scientifique dont l'objet est de recueillir, d'organiser, de résumer et d'analyser les données d'une enquête, d'une étude ou d'une expérience, aussi bien que de tirer les conclusions logiques et de prendre les décisions qui s'imposent à partir des analyses effectuées.

2 - Notions de Base:

2.1 - La population: est une ensemble d'éléments définis par une propriété commune donnée.

Exemple: Si l'on veut étudier la durée de vie des ampoules électriques fabriquées par une compagnie, la population considérée est l'ensemble de toutes les ampoules fabriquées par cette compagnie.

2.2 - L'échantillon: est une sous ensemble de la population.

Exemple: pour établir la durée de vie des ampoules produites par une machine; on peut prélever au hasard un certain nombre d'ampoules (un échantillon) parmi toutes les celles produites par cette machine.

2.3 - L'individu (ou unité statistique):

Chaque élément de la population ou de l'échantion.

Exemple: dans l'exemple précédant, Chaque ampoule constitue un individu ou une unité statistique.

2.4 - Le caractère (ou la variable):

c'est l'aspect particulier que l'on désire étudier, les variables sont désignées par X;

Exemple:

concernant un groupe de personnes, on peut s'intéresser à leur âge, leur taille, ...

- 2.5 - Les modalités: sont les différentes manières d'être que peut présenter un caractère.

Exemple: quant au nombre d'enfants par famille, les modalités de ce caractère peuvent etre 0, 1, 2, ..., 20.

2-6- caractère qualitatif:
Ses modalités ne s'expriment pas par un nombre.
Exemple: la religion, l'opinion, la nationalité, ...
2-7- caractère quantitatif:
Ses modalités sont numériques.
Exemple: l'age, la taille, le poids, ...
2-8- caractère quantitatif discret:
l'ensemble des valeurs que peut prendre le caractère est fini ou dénombrable, le plus souvent, ces valeurs sont entières.
Exemple le nombre d'enfant dans une famille.
2-9- caractère quantitatif continu:
le caractère peut prendre théoriquement n'importe quelle valeur dans un intervalle donnée de nombres réels.
Exemple: la taille d'un individu, le poids, ...
2-10- Série statistique:
l'ensemble des différentes données associées à un certain nombre d'individus.
Exemple: la série suivante résulte d'une courte enquête auprès de quelques personnes pour connaître leur age.
18. 21 19 19 19 17 22 27 18 18 17 20 20 23.
2-11- l'effectif (ou la fréquence absolue) associée à une valeur d'un caractère X_i est le nombre de fois où cette valeur du caractère X_i a été observée, noté par n_i.
2-12- l'effectif total:
Représente le nombre d'individus d'un échantillon, il est symbolisé par N

$$N = n_1 + \dots + n_K, \quad (K : \text{nombre des modalités})$$

2-13- La fréquence relative associée à X_i: notée par f_i

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

2-14- l'effectif cumulé croissant $n_{ic}^{\uparrow}$: est donnée par la valeur suivante:

$$n_{ic}^{\uparrow} = n_1 + n_2 + \dots + n_i$$

2-15- la effectif relative cumulée croissant $f_{ic}^{\uparrow}$:

$$f_{ic}^{\uparrow} = f_1 + f_2 + \dots + f_i$$

2-16 - L'effectif cumulé décroissant $n_{ic}^{\downarrow}$.

$$n_{ic}^{\downarrow} = N - (n_1 + \dots + n_{i-1}) \; ; \; n_0 = 0. \qquad i \neq 1$$

2-17 - L'effectif relatif cumulé décroissant $f_{ic}^{\downarrow}$:

$$f_{ic}^{\downarrow} = 1 - (f_1 + \dots + f_{i-1}) \; ; \; f_0 = 0$$

2-18 - L'étendue de la série est l'écart qui sépare la plus grande et la plus petite valeur du caractère, est noté par e, tel que :

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

- 2-19 - Tableau statistique : Ce tableau établit la correspondance entre deux séries de nombres, l'une constituée par les valeurs du caractère étudié, l'autre par les effectifs correspondants (ou relatifs ou cumulés).

2-20 - Pour la classe $[e_i ; e_{i+1}[$:

- Les nombres $e_i ; e_{i+1}$ s'appellent les limites (ou les bornes) de la classe.
- La demi somme $\frac{e_{i+1} + e_i}{2} = c_i$ s'appelle le centre de la classe.
- La différence $e_{i+1} - e_i$ s'appelle l'amplitude de la classe.

- 3 - Choix du nombre des classes :

Le nombre des classes ne doit pas être trop grand et ne doit pas être trop faible. Il y a des formules pour calculer ce nombre, comme la règle de Sthigerse tel que le nombre des classes k est définie comme suite :

$$k = 1 + 3{,}322 \log_{10} N$$

Exemple :

X est une variable statistique prenant les valeurs suivantes :

45 50 54 55 56 45 46 47 50 51 49 53 52 51 48 54
55 56 49 48 47 50 45 47 48 48 49 50 56 55

1 - Regrouper ces ~~classes~~ valeurs en classes dans tableau statistique.

On a $N = 30 \Rightarrow k = 1 + 3{,}322 \log_{10}(30) = 5.906 \simeq 6$.

$$a = \frac{e}{k} = \frac{56 - 45}{5{,}906} \simeq \frac{56 - 45}{6} = \frac{11}{6} = 1.83 \simeq 2$$

1) le tableau statistique :

classes	n_i (effectif)
[45 . 47[	.04
[47 . 49[	07
[49 . 51[	07
[51 , 53[	03
[53 . 55[	03
[55 . 57[	06
Σ	N=30

4 - Représentations graphiques :

Les représentations graphiques ont l'avantage d'offrir une meilleure vue d'ensemble de la série statistique que les tableaux. Elles permettent par simple lecture, de voir les caractéristiques essentielles de la série, et aussi de comparer des séries différentes.

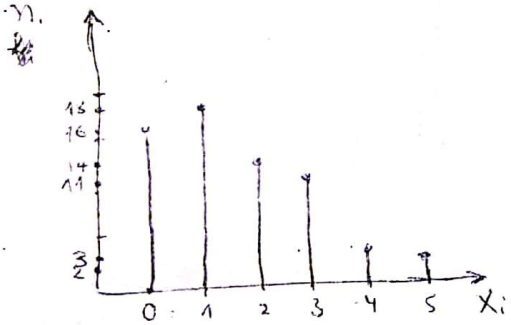
4 - 1 - Caractère discontinu :

lorsque le caractère est discontinu, on utilise diagramme en bâtons.

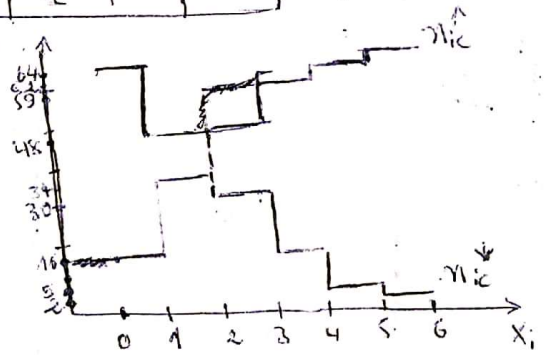
Exemple :

le nombre d'enfants par famille.

Nombre d'enfants X_i	Nombre de familles n_i (Effectif)	f_i	$n_{ic}^{\uparrow}$	$n_{ic}^{\downarrow}$
0	16	0,250	16	64
1	18	0,281	34	48
2	14	0,218	48	30
3	11	0,172	59	16
4	3	0,047	62	5
5	2	0,031	64	2
Σ	64	1	—	—



le diagramme en bâtons (n_i en fonction X_i)



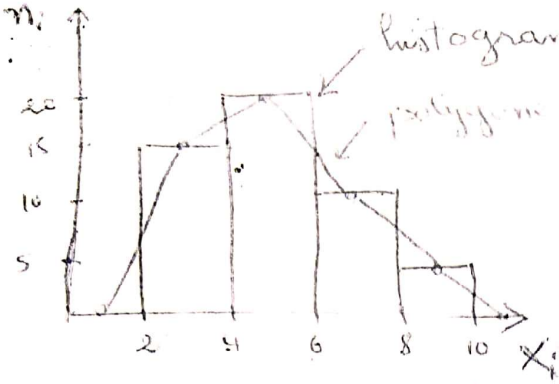
le graphe des fréquences cumulées (diagramme intégral)

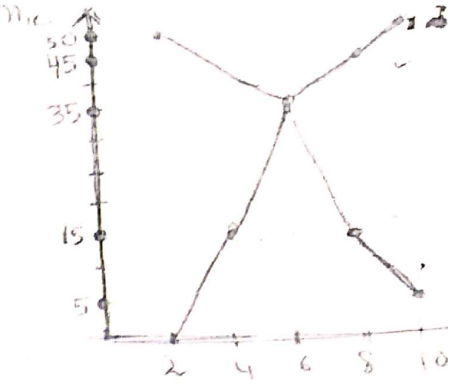
4 - 2 - Caractère continu :

Exemple

On considère le tableau statistique suivant :

classes	n_i	$n_{ic}^{\uparrow}$	$n_{ic}^{\downarrow}$
[2, 4[	15	15	50
[4, 6[	20	35	35
[6, 8[	10	45	15
[8, 10[	5	50	5



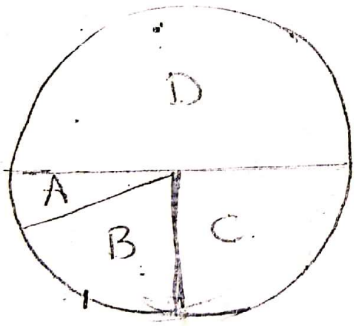


4-3 – Caractère qualitatif

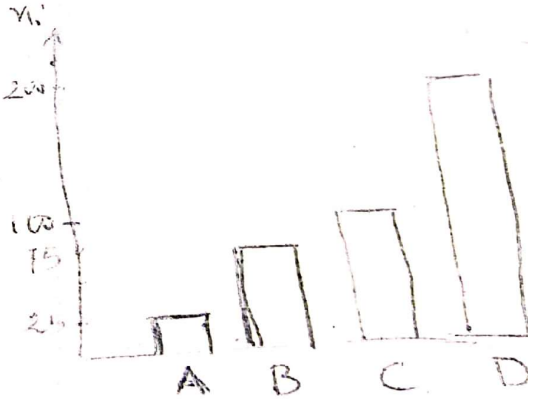
Exemple : le tableau suivant représente les parts des différentes

x_i	n_i	α_i
A	25	22,5
B	75	67,5
C	100	90°
D	200	180°
Σ	400	360°

$\alpha_i = \frac{n_i \times 360}{400}$



Représentation en secteur circulaire



bâtons de l'An

I- Paramètres de position, paramètres de dispersion:

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

I.1 - Paramètres de position:

1.1 - La moyenne arithmétique:

Soit $x_1, x_2, x_3, \dots, x_n$ une suite finie de nombres, La moyenne arithmétique est le rapport :

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Si chaque valeur x_i apparaît n_i fois dans la série, on peut encore écrire

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{k} n_i x_i$$ N : l'effectif total, k : nombre de modalités.

Remarque :

Pour caractère continue, On prend pour valeur de x_i les centres de classes.

Exemples:

- La moyenne arithmétique des valeurs 8, 4, 3, 6, 2 :

$$\bar{x} = \frac{8+4+3+6+2}{5} = \frac{23}{5} = 4,6$$

- La moyenne arithmétique des valeurs 8, 5, 5, 5, 5, 6, 6, 3, 3, 2

$$\bar{x} = \frac{8\times1 + 5\times3 + 6\times2 + 3\times2 + 2\times1}{10} = 4.8$$

- Soit le tableau statistique suivant:

x_i	n_i	$n_i x_i$
0	11	0
1	22	22
2	45	90
3	40	120
4	4	16
5	19	55
6	11	12
Σ	N=150	375

On a :

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{7} n_i x_i = \frac{375}{150} = 2,5$$

- Soit le tableau statistique :

classes	n_i	$C_i = X_i$	$n_i X_i$
$[8, 10[$	1	9	9
$[10, 12[$	2	11	22
$[12, 14[$	4	13	52
$[14, 16[$	6	15	90
$[16, 18[$	5	17	85
$[18, 20[$	2	19	38
Σ	N=20	——	296

On a :

$$\bar{X} = \frac{\sum_{i=1}^{6} n_i X_i}{N} = \frac{296}{20} = 14.8.$$

5-1-2- La moyenne harmonique:

Soit une suite finie de nombres $X_1, X_2, \dots, X_n$ ~~et l'ensemble des inverses de~~

On appelle moyenne harmonique des X_i l'inverse de la moyenne arithmétique des inverses $\frac{1}{X_i}$. Soit:

$$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

Exemple: une voiture parcourt un circuit fermé à une vitesse de 10 km/h durant le 1er tour, de 20 km/h durant le second, de 30 km/h durant le 3ème tour. Déterminer la vitesse moyenne de la voiture durant les trois tours.

Soit ℓ la longueur du circuit (en km). La durée totale des 3 parcours est: $t = \frac{\ell}{10} + \frac{\ell}{20} + \frac{\ell}{30}$

La vitesse moyenne v recherchée est donc:

$$V = \frac{3\ell}{t} = \frac{3\ell}{\frac{\ell}{10} + \frac{\ell}{20} + \frac{\ell}{30}} = \frac{3}{\frac{1}{10} + \frac{1}{20} + \frac{1}{30}} = 16.6 \text{ km/h}.$$

5-2-3- La moyenne géométrique:

Soit une suite finie de nombres $X_0, X_1, X_2, \dots X_n$ tels que $X_1 \dots X_n$

$X_1 = X_0 \cdot r$ $X_2 = X_0 \cdot r^2$, $X_3 = X_2 \cdot r = X_0 \cdot r^3$, par récurrence $X_n = X_0 r^n$ $X_2 = X_1 \cdot r$. $X_3 = X_1 \cdot r^2$ $X_n = X_1 \cdot r^{n-1}$

On appelle moyenne géométrique des (X_i) le terme:

$$g = (X_0 \times X_1 \times X_2 \times \dots \times X_n)^{1/n+1} = X_0 \cdot r^{n/2} \qquad g = (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{\frac{1}{n}} = X_1 \cdot r^{\frac{n-1}{2}}$$

5-1-4- La médiane(Me):

Définition: la médiane est la valeur du caractère telle qu'il y ait autant d'individus pour lesquels le caractère est inférieur à Me que d'individus pour lesquels le caractère est supérieur à Me.

Série statistique d'un caractère discret:

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

Exemple: Dans la série de 17 valeurs suivantes:

1, 12, 2, 11, 4, 10, 4, 9, 4, 9, 5, 8, 6, 8, 8, 7, 12, 12

La série ordonnée: 1, 2, 4, 4, 4, 5, 6, 7, 8, 8, 9, 9, 10, 11, 12, 12, 12

$17 = 2(8)+1$

$\boxed{Me = 8}$

- Si la série compte un nombre pair de valeurs soit $2k$ valeurs, la médiane sera la demi-somme de la $k^{ième}$ et de la $(k+1)^{ième}$.

Exemple: La médiane de la série statistique suivante:

19, 4, 17, 5, 14, 8, 12, 8, 11, 9

4, 5, 8, 8, 9, 11, 12, 14, 17, 19

$10 = 2 \times 5$

$$Me = \frac{9+11}{2} = 10$$

Série statistique d'un caractère continu:

La médiane obtenue par est donnée par:

$$Me = \ell_1 + \left[\frac{(\ell_2 - \ell_1)}{n_0}\left(\frac{N}{2} - n_{i-1}^{\uparrow c}\right)\right]$$

ℓ_1, ℓ_2 : les extrémités de la classe contenant la médiane

n_0 : l'effectif de cette classe.

$n_{i-1}^{\uparrow c}$: l'effectif cumulé jusqu'à ℓ_1

N : l'effectif total.

Exemple:

On considère la série statistique suivante:

Classes	n_i	$n_{ic}^{\uparrow}$
[38, 40[	11	11
[40, 42[	28	39
[42, 44[	16	55
[44, 46[	25	80
[46, 48[	15	95
[48, 50[	5	100
Σ	N = 100	—

On a $\frac{N}{2} = \frac{100}{2} = 50$ donc la classe correspondante qui contenant la médiane est $[42.44[$ (car $50 \leq 55$)

$l_1 = 42$. $l_2 = 44$. $n_0 = 16$, $n_{i-1}c = 39$

$$M_e = l_1 + \left[\frac{(l_2 - l_1)}{n_0}\left(\frac{N}{2} - n_{i-1}c\right)\right] = 42 + \left[\left(\frac{44-42}{16}\right)(50 - 39)\right] = 43,37.$$

2) Le mode (dominant valeur dominante) Mo

Définition:

Le mode d'une série statistique est la valeur la plus fréquente de cette série.

Remarque: Pour une série classés, la ~~définition précédente n'est plus valable~~. On définit la classe modale: c'est la classe dont l'effectif est relativement le plus élevé, et on attribue au mode la valeur centrale de cette classe.

Exemple:

Pour la série statistique: 13.14.13.15.16.13, 12.12.12.13.11.15; Mo = 13

Pour l'exemple précédent (caractère continue): La classe modale est $[40.42[$ et $Mo = \frac{40+42}{2} = 41$.

5 2_Paramètres de dispersion

Les caractéristiques quantifient les fluctuations des valeurs observées autour de la valeur centrale et permettent d'apprécier l'étalement de la série les principaux sont: l'écart type ou sa carré appelé variance, le coefficient de variation et l'étendue.

Variance et écart-type:

- La variance d'une série de valeurs du caractère est la moyenne arithmétique des carrés des écarts de ces valeurs par rapport à leur moyenne arithmétique:

$$V = \frac{1}{N} \sum_{i=1}^{K} n_i (X_i - \bar{X})^2$$

- L'écart-type est la racine carrée de la variance.

$$\sigma = \sqrt{V}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{K} n_i (X_i - \bar{X})^2}$$

$$\sigma = \sqrt{\left(\frac{1}{N} \sum_{i=1}^{K} n_i X_i^2\right) - \bar{X}^2}$$ (formule simplifiée)

Propriétés:

- L'écart-type σ caractérise la dispersion d'une série de valeurs. Plus σ est petit plus les donnée sont regroupées autour de la moyenne arithmétique $\bar{X}$ et plus la population est homogène.

6_Coefficient de variation

Il s'exprime par l'expression suivante:

$$CV = \frac{\sigma}{\bar{X}} \times 100$$ (sous la forme d'un pourcentage)

- Propriétés:
- Le coefficient de variation ne dépend pas des unités choisies.
- Il permet d'apprécier l'homogénéité de la distribution, une valeur du coefficient de variation $\leq 15\%$ traduit une bonne homogénéité de la distribution.
- Il permet de comparer deux distributions, même si les données ne sont pas exprimées avec la même unité ou si les moyennes arithmétiques des deux séries sont très différentes.

Si :

68% de observations comprise entre $\bar{x} - \sigma$ et $\bar{x} + \sigma$,

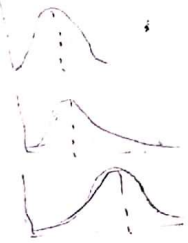
95% ... dans $\bar{x} \pm 2\sigma$

99% ... $\bar{x} \pm 3\sigma$

On dit que la distribution est normale (distribution de Gauss)

f- Paramètres de forme:

Distribution symétrique:



1) Une distribution est symétrique si : $Mo = Me = \bar{X}$

2) Si $Mo < Me < \bar{X}$ la distribution est étalée vers la droite:

3) Si $Mo > Me > \bar{X}$: " " " la gauche

II- Séries numériques à deux dimensions:

Soient X et Y les deux caractères étudiés, p le nombre de modalités prises par X, q le nombre de modalités prises par Y et N le nombre total d'observations.

On étudie par exemple, le poids et la taille d'un nombre N d'individus,

~~1- Le monde de Représentation graphique des données~~

1- Tableaux statistiques:

On suppose que les deux variable étudiées sont des caractères quantitatifs. Les tableaux statistiques portent le nom de tableaux de contingence

Dans chaque case du tableau, on écrit l'effectif n_{ij} de l'échantillon, c'est à dire le nombre de données tel que $X = X_i$ et $Y = Y_j$

On définit les fréquences absolues suivantes:

- les fréquences marginales:

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

X \ Y	Y_1	...	Y_j	...	Y_p	$n_{i.}$
X_1	n_{11}					$n_{1.}$
X_2						
⋮						⋮
X_i			n_{ij}			
⋮						⋮
X_q						$n_{q.}$
$n_{.j}$	$n_{.1}$				$n_{.p}$	N

tableau de contingence.

~~Les fréquences relatives marginaux.~~

- La fréquence marginale $n_{i.}$ est donc le nombre d'individus possédant la modalité i du caractère X quelle que soit la distribution du caractère Y

2 - Représentation graphique

Un nuage de points dans $\mathbb{R}^2$.

3 - Mesure de dépendance: ($\delta(\bar{w}, 1)$

L'étude de la distribution simultanée (conjointe) de deux variables a pour but de préciser le type de liaison pouvant exister entre ces deux variable, la nature et l'intensité (force) de cette liaison, à l'aide de différents coefficients.

Les propriétés ~~marginales~~ marginaux:

- Les moyennes ~~marginales~~:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{q} n_{i.} X_i$$

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{p} n_{.j} Y_j$$

- ~~Les moyennes marginales les écart-types marginaux~~

- Les variances:

$$V_M(X) = \left(\frac{1}{N} \sum_{i=1}^{q} n_{i.} X_i^2 \right) - \bar{X}^2$$

$$V_M(Y) = \left(\frac{1}{N} \sum_{j=1}^{p} n_{.j} Y_j^2 \right) - \bar{Y}^2$$

- Les écart-types:

$$\sigma_M(X) = \sqrt{V_M(X)}$$

$$\sigma_M(Y) = \sqrt{V_M(Y)}$$

Exemple:

Au cours des épreuves d'un examen, 45 étudiants ont obtenu les notes suivantes à deux matières (X: statistique et Y: ~~Anglais~~ Algèbre)

X \ Y	5	10	15	$n_{i.}$
5	1 (25)	2 (100)	3 (225)	6
10	4 (200)	5 (500)	6 (900)	15
15	7 (225)	8 (1200)	9 (2025)	24
$n_{.j}$	12	15	18	N = 45 (5450) (5400) 7510

X \ Y	5	10	15	$n_{i.}$	$n_{i.}X_i$	$n_{i.}X_i^2$
5	1 (25)	2 (100)	3 (225)	6	5×6=30	150
10	4 (200)	5 (500)	6 (900)	15	150	1500
15	7 (225)	8 (1200)	9 (2025)	24	360	5400
$n_{.j}$	12	15	18	N=45 (5450)	$\sum n_{i.}X_i$ = 540	$\sum n_{i.}X_i^2$ = 7050
$n_{.j}Y_j$	60	150	270	$\sum n_{.j}Y_j$ = 480		
$n_{.j}Y_j^2$	300	1500	4050 (5850)	$\sum n_{.j}Y_j^2$		

L'effectif total:
N = 45,

Les fréquences marginaux selon X: $n_{1.} = 6$, $n_{2.} = 15$, $n_{3.} = 24$

Les " " " Y: $n_{.1} = 12$, $n_{.2} = 15$, $n_{.3} = 18$

Les propriétés marginaux:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{3} n_{i.} X_i = \frac{1}{45}(6 \cdot 5 + 15 \cdot 10 + 24 \cdot 15) = 12 \qquad \bar{X} \Rightarrow \frac{540}{45}$$

$$\bar{Y} = \frac{1}{N}\sum_{j=1}^{3} n_{.j} Y_j = \frac{1}{45}(12 \cdot 5 + 15 \cdot 10 + 18 \cdot 15) = 10.66 = \frac{480}{45}$$

$$V_M(X) = \left(\frac{1}{N}\sum_{i=1}^{3} n_{i.} X_i^2\right) - \bar{X}^2 = \left[\left(\frac{1}{45}(6 \cdot 25 + 15 \cdot 100 + 24 \cdot 225)\right)\right] - (12)^2 = 12,66$$

$$\frac{7050}{45} - (12)^2$$

$$V_M(Y) = \left(\frac{1}{N}\sum_{j=1}^{3} n_{.j} Y_j^2\right) - \bar{Y}^2 = \left[\left(\frac{1}{45}(12 \cdot 25 + 15 \cdot 100 + 18 \cdot 225)\right)\right] - (10,66)^2 = 16,15$$

$$\frac{5850}{45} - (10,66)^2 = 16,15$$

$$\sigma_M(X) = 3,56$$

$$\sigma_M(Y) = 4,02$$

La covariance commune $V\,cov(X.Y)$:

donnée par:

$$cov(X.Y) = \left[\frac{1}{N}\sum_{i=1}^{q}\sum_{j=1}^{p} n_{ij} X_i Y_j\right] - (\bar{X} \cdot \bar{Y})$$

$$cov(X.Y) = \left(\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right) - (\bar{X} \cdot \bar{Y})$$

Exemple: Dans l'exemple précédent:

$$cov(X.Y) = \left(\frac{1}{45} \cdot 5450\right) - (12 \cdot 10,66) = -6,94$$

$$126,66 - 127,92 = -1,26$$

$$\alpha = \frac{cov(X.Y)}{V_M(X)} = \frac{-6,94}{(4,02)^2}$$

Le coefficient de corrélation:

donnée par: $$\rho = \frac{cov(X.Y)}{\sigma_M(X) \cdot \sigma_M(Y)} \quad \text{avec } -1 \le \rho \le 1.$$

- Si $|S| \to 1$ alors la relation entre x et X est linéaire
- Si $|S| \to 1$ et $S > 0$, la relation est positive.
- Si $|S| \to 1$ et $S < 0$, " " " négative
- Si $|S| \to 0$, alors il n'y a pas une relation linéaire entre X et Y.

Dans Exemple: Dans l'exemple précédent :

$$S = \frac{-6.94}{3.5 \times 4} = -0,49$$ $\begin{pmatrix} |S| < 0,5 & \text{faible} \\ |S| > 0,5 & \text{fort} \end{pmatrix}$.

Il y a une relation faible entre X et Y

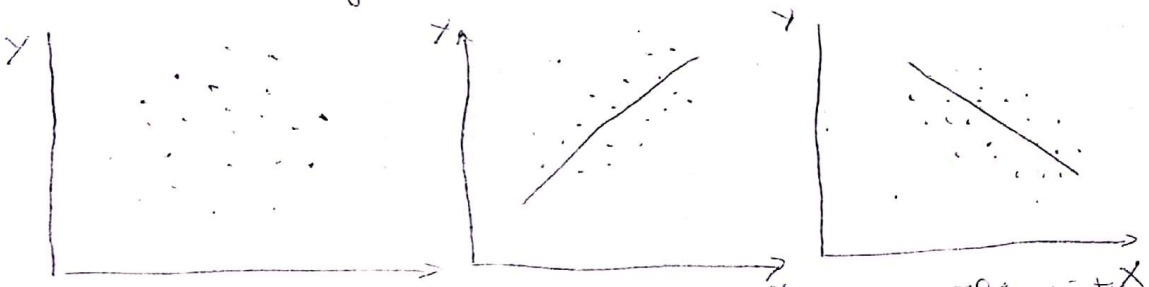
Ajustement linéaire

Si

$$S = \frac{-1,26}{3,5 \times 4} = \frac{-1,26}{14} = 0,09$$

Ajustement linéaire - corrélation:

Considérons une population de taille N et supposons que sur chaque élément de cette population, on effectue deux observations portant sur deux caractères différents associés aux deux variables aléatoires X et Y. Le problème de corrélation est celui qui consiste à rechercher s'il existe une relation entre les variables X et Y.
A chaque élément i de l'échantillon, on peut associer un couple de valeurs (X_i, Y_i) qui peut être représenté par un nuage de N points constituant un diagramme de dispersion.



a) variables non corrélées — corrélation positive — corrélation négative

- l'ajustement consiste à rechercher une fonction $f(x)$ dont le graphe se rapproche le plus possible des points du diagramme.

Il existe donc entre X_i et Y_i une relation de la forme:

$$Y_i = f(X_i) + \varepsilon_i \qquad (\varepsilon_i : \text{un écart résiduel})$$

- la méthode d'ajustement consiste à déterminer les paramètres de $f(x)$ qui minimisent ces écarts. ~~Cela revient à minimiser la somme des valeurs abso~~

Ou encore: (minimiser S):

$$S = \sum_{i=1}^{N} (Y_i - f(X_i))^2 .$$

C'est la méthode dite "des moindres carrés".

- Nous nous limiterons ici au cas l'ajustement linéaire

$f(x) = aX + b$ (la droite de régression)

Déterminer a et b :

Pour que la somme $S = \sum_{i=1}^{N} (Y_i - aX_i - b)^2$ soit minimum.

Il suffira que :

$$\frac{\partial S}{\partial a} = 0 \quad \text{et} \quad \frac{\partial S}{\partial b} = 0$$

$$\Rightarrow a = \frac{\sum_{i=1}^{N} (X_i Y_i) - N\bar{X}\bar{Y}}{\sum_{i=1}^{N} X_i^2 - N\bar{X}^2} = \frac{\left(\frac{1}{N}\sum_{i=1}^{n} X_i Y_i\right) - \bar{X}\bar{Y}}{\left(\frac{1}{N}\sum_{i=1}^{N} X_i^2\right) - \bar{X}^2} = \frac{cov(X,Y)}{\sigma_M(X)}$$

$$b = \bar{Y} - (a \cdot \bar{X})$$

Généralisation de la méthode des moindres carrés:

Par la méthode des moindres carrés, on peut ajuster une fonction qui dépend de plus de deux paramètres.

Exemple:

Soit les 5 points $(x,y) = (1,9)\ (2,22),\ (3,38)\ (4,82),\ (5,154)$

- Ajuster par les moindres carrés une courbe d'equation: $y = k.a^x$ (k et a deux constantes)

Solution:

Linéarisons l'equation en prenant les logarithmes népériens de chaque membre:

$y = k\,a^x \iff \ln y = x \ln a + \ln k \iff Y = Ax + B$ $\begin{pmatrix} \ln y = Y, \\ \ln a = A, \\ \ln k = B \end{pmatrix}$

x_i	1	2	3	4	5	
y_i	9	22	38	82	154	
Y_i	2,197	3,091	3,638	4,407	5,037	

- De ce tableau on déduit: $\overline{x} = 3$, $\overline{Y} = 3,674$

Puis par les moindres carrés:

$$A = \frac{\frac{1}{5}\sum_{i=1}^{5} x_i Y_i - \overline{x}\,\overline{Y}}{\frac{1}{5}\sum_{i=1}^{5} x_i^2 - \overline{x}^2} \simeq 0,710$$

$$\Rightarrow \quad a = e^{A} = 2,013$$

$$B = \overline{Y} - A\overline{X} = 1,575 \quad \Rightarrow K = e^{B} = 4,833$$

$$y = k\,a^x = 4,83 \times (2,013)^x$$

- Utilisation d'Excel pour calculer des grandeurs statistiques:
- Tapez les données (valeurs de X_i) dans une colonnes (par exemple A)
- Insérer une fonction de grandeur demandé sous ~~ses~~ cette colonne.

Exemple Transformer ce tableau dans excel:

X_i	n_i
4	3
5	2
6	4
7	1

→

	A
1	4
2	4
3	4
4	5
5	5
6	6
7	6
8	6
9	6
10	7
	=

- Selectionner une catégorie : statistiques

= AVERAGEA (A_1 ; A_{10}) = 5,3 moyenne arithmétique.

= ECARTTYPE (A_1 ; A_{10}) = 1,059 ecat type.

= Mode (A_1 : A_{10}) = 6

= Mediane (A_1 ; A_{10}) = 5,5

= Moyenne. geometrique (A_1 A_n) = 5,20

= Moyenne. harmonique

= VAR (A_1 ; A_{10}) = 1,12.

Deux Variables:

X \ Y	6	7	8
4	1	2	0
5	0	0	2
6	3	1	0
7	1	0	0

→

	A	B
1	4	6
2	4	7
3	4	7
4	5	8
5	5	8
6	6	6
7	6	6
8	6	6
9	6	7
10	7	6
	=	

COVAIANCE (A_1 : A_{10} ; B_1 : B_{10}) = -0,31

COFFICIENT. CORRELATION (A_1 : A_{10} ; B_1 : B_{10}) = -0,39