**Insitut de Recherche sur l'Enseignement de la Mathématique,
de la Physique et de la Technologie(IREMPT)
Université Cheikh Anta Diop de Dakar**

LA STATISTIQUE AU LYCÉE

Mouhamed Bounama Sow, Babacar Diakhaté, Lamine Mbodj

Table des matières

1 PARAMETRES DE DISPERSION **7**

1.1 Notation indicielle 7
1.2 Etendue 9
1.2.1 Cas d'une variable discrète 9
1.2.2 Cas d'une variable continue 9
1.3 Ecart interquartile 10
1.3.1 Quartiles 10
1.3.2 Définition 11
1.4 Ecart moyen 13
1.4.1 Cas de série discrète 13
1.4.2 Cas de série continue 14
1.5 Variance et écart-type 14
1.5.1 Cas d'une série discrète 15
1.5.2 Cas d'une série continue 15
1.5.3 Méthode de calcul par le théorème de KOENIG 15
1.5.4 Propriétés 15
1.6 Exercices 16
1.7 Solutions 19

2 Séries statistiques à deux variables **25**

2.1 Séries à données individuelles 25
2.1.1 Définition 25
2.1.2 Nuage de points et point moyen 26
2.1.3 Utilisation du nuage 26
2.1.4 Ajustement linéaire 27
2.1.5 Coefficient de corrélation linéaire 34
2.2 Séries à données groupées 35
2.2.1 Exemple préliminaire 35
2.2.2 Définition 36
2.2.3 Nuage de points 37
2.2.4 Caractéristiques marginales 37

2.2.5 Caractéristiques conditionnelles 38

3 Utilisation d' une calculatrice en statistique dans le casd'une série double 41
3.1 Introduction 41
3.2 Mise en marche de la calculatrice 42
3.3 Entrée des données 42
3.4 Obtention des paramètres cités dansl'introduction 42
3.5 **Suppression des données** 43
3.6 Retour en mode normal 43

INTRODUCTION

A l'origine la Statistique rassemble et étudie des renseignements susceptibles d'intéresser l'État (status). C'est ainsi qu'initialement son étude portait sur les domaines économique et social.
Essentiellement descriptive à ses débuts, ce n'est qu'à partir du $16^{\text{ème}}$ siècle qu'elle a évolué vers l'analyse des données grâce notamment à l'astronomie. L'impact de la statistique dans les autres domaines de la vie (médecine, agronomie, démographie, sociologie, industrie,...) lui confère de plus en plus une importance dans l'enseignement de la Mathématique.
L'enseignement de la statistique au premier cycle concerne surtout le recueil de données et l'exploitation de celles-ci par le calcul des paramètres de position.
Cet enseignement se poursuit au second cycle avec l' introduction des paramètres de dispersion puis des séries à deux caractères quantitatifs.

Chapitre 1

PARAMETRES DE DISPERSION

Position du problème

-Activité préliminaire :
Deux éleves E_1 et E_2 d'une même classe ont obtenu les notes suivantes.

$$E_1 : 8; 10; 11; 8; 8; 10; 13$$

$$E_2 : 10; 8; 6; 11; 8; 13; 12$$

Vérifie que ces deux séries ont les mêmes paramètres de position.
Comment choisir alors l'éleve le plus régulier ?

-La connaissance des paramètres de position (mode, médiane, moyenne) ne suffit pas pour étudier une série statistique. L'étude de la structure de la série nécessite la connaissance de la distribution des valeurs et leur régularité. Pour cela le calcul d'autres paramètres, appelés **paramètres de dispersion**, s'impose. Ils ne concernent que les séries à caractère quantitatif. Ces caractéristiques de dispersion permettent de voir si les valeurs prises par le caractère sont suffisamment proches ou non des paramètres de position.

1.1 Notation indicielle

On considère la série statistique suivante suivante :

valeurs de caractère	x_1	x_2	x_3	$\ldots$	x_p
effectifs	n_1	n_2	n_3	$\ldots$	n_p

L'effectif total est :

$$N = n_1 + n_2 + n_3 + \ldots + n_p$$

cette écriture peut être simplifiée en utilisant le symbole $\sum$ de la façon suivante :

$$N = \sum_{i=1}^{p} n_i$$

La moyenne de cette série est

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + n_3x_3 + \ldots + n_px_p}{n_1 + n_2 + n_3 + \ldots + n_p}$$

qui s'écrit avec la notation indicielle $\sum$:

$$\overline{x} = \frac{\sum_{i=1}^{p} n_ix_i}{\sum_{i=1}^{p} n_i} = \frac{1}{N}\sum_{i=1}^{p} n_ix_i$$

Remarque 1

$*$*Soit $k \in R$, on a :*

$$k(\sum_{i=1}^{p} x_i) = k(x_1 + x_2 + x_3 + \ldots + x_p) = kx_1 + kx_2 + kx_3 + \ldots + kx_p = \sum_{i=1}^{p} kx_i$$

$*$ *Soit $a \in R$, on a :*

$$\sum_{i=1}^{p}(x_i + a) = (x_1 + a) + (x_2 + a) + (x_3 + a) + \ldots + (x_p + a)$$

$$= x_1 + x_2 + x_3 + \ldots + x_p + \underbrace{a + a + a + \ldots + a}_{p \text{ termes}}$$

$$= \sum_{i=1}^{p} x_i + pa$$

Donc

$$\sum_{i=1}^{p}(x_i + a) = \sum_{i=1}^{p} x_i + pa$$

Attention :

$$\sum_{i=1}^{p} x_i + a = a + \sum_{i=1}^{p} x_i$$

Exercice d'application

On considére la série statistique suivant

$valeurs\ \ x_i$	3	5	7	8
fréquences f_i	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{7}{20}$	$\frac{1}{5}$

Calcule la moyenne de cette série en utilisant le symbole $\sum$.
Calcule le nombre

$$\sum_{i=1}^{4} x_i|1 - f_i|$$

.

1.2 Etendue

Elle mesure la dispersion d'une série donnée.

1.2.1 Cas d'une variable discrète

L'étendue e d'une série statistique est la différence entre la plus grande et la plus petite valeur de la série.
Soient x_1 la plus petite valeur de la série et $x_n$la plus grande.L'étendue de cette série est

$$e = x_n - x_1$$

Exemple 1 *Considérons le "nombre de pièces" de chacun des appartements d'un immeuble donné :*

$Nombre\ de\ pieces$	1	2	3	4	5	6
$effectifs$	5	4	4	2	3	2

Il y'a 5 *appartements à* 1 *pièce,* 4 *à* 2 *pièces,* 4 *à* 3 *pièces,*2 *à* 4 *pièces,*3 *à* 5 *pièces et* 2 *à* 6 *pièces. L'étendue de cette série est :* $e = 6 - 1$.*Il y'a* 5 *pièces d'écart entre le plus petit et le lus grand appartement.*

1.2.2 Cas d'une variable continue

Soit une série continue dont la première classe est $[x_1,\ x_2[$ et la dernière $[x_{n-1},\ x_n[$. L'étendue e cette série est la différence entre la borne supérieure de la dernière classe et la borne inférieure de la première. Donc $e = x_n - x_1$.

Exemple 2 *Dans un immeuble les loyers payés selon le nombre de pièces de chaque appartement sont répartis de la façon suivante :*

Prix de la location	$[20.000; 40.000[$	$[40.000; 60.000[$	$[60.000; 80.000[$	$[80.000; 100.000[$
effectifs	14	26	18	12

Il y'a 14 *locataires qui paient entre* $20.000F$ *et* $40.000F$*,* 26 *locataires entre* $40.000F$ *et* $60.000F$*;* 18 *locataires entre* $60.000F$ *et* $80.000F$ *et* 12 *locataires entre* $80.000F$ *et* $100.000F$*.*
L'étendue de cette série est :

$$e = 100000F - 20000F = 80000F.$$

Il y'a un écart de $80000F$ *entre le loyer le plus cher et le loyer le moins cher.*

Remarque 2
Plus l'étendue est grande, plus la série est dispersée.
- *L'étendue a l'avantage d'etre facile à calculer mais elle présente l'inconvenient de ne prendre en compte que les valeurs extrêmes ; elle ne renseigne pas sur les valeurs intermédiaires.*

1.3 Ecart interquartile

1.3.1 Quartiles

Soit une série statistique à caractère quantitatif et dont les valeurs sont rangées dans l'ordre croissant.

Cas discret

Dans la cas d'une série discrète, la médiane **Mé** partage la population en deux groupes de même effectif. Appelons série inférieure la série des valeurs du caractère strictement inférieures à **Mé** et série supérieure celle des valeurs qui sont strictement supérieures à **Mé** . On appelle premier quartile, noté q_1, la médiane de la série inférieure et le troisième quartile, noté q_3, la médiane de la série supérieure.

Cas continu

Dans la cas d'une série continue d'effectif total N, le premier quartile q_1 appartient à la première classe dont l'effectif cumulé croissant (ECC) est supérieur ou égal à $\frac{N}{4}$ et le troisième quartile q_3 appartient à la première classe dont l'ECC est supérieur ou égal à $\frac{3N}{4}$.

On peut déterminer q_1 et q_3 par une interpolation linéaire.
Si $[a_1; b_1[$ est la classe contenant q_1, n_1 son ECC et n'_1 celui de la classe précédente alors q_1 s'obtient de la manière suivante :

$$\frac{q_1 - a_1}{\frac{N}{4} - n'_1} = \frac{b_1 - a_1}{n_1 - n'_1}$$

d'où

$$q_1 = \frac{(b_1 - a_1)(\frac{N}{4} - n'_1)}{n_1 - n'_1} + a_1$$

Si $[a_3; b_3[$ est la classe contenant q_3, n_3 son ECC et n'_3 celui de la classe précédente alors q_3 s'obtient de la manière suivante :

$$\frac{q_3 - a_3}{\frac{3N}{4} - n'_3} = \frac{b_3 - a_3}{n_3 - n'_3}$$

d'où

$$q_3 = \frac{(b_3 - a_3)(\frac{3N}{4} - n'_3)}{n_3 - n'_3} + a_3$$

Remarque 3 *Le deuxième quartile, noté q_2, est la médiane **Mé** de la série.*

1.3.2 Définition

On appelle intervalle interquartile l'intervalle $[q_1; q_3]$. Son amplitude $q_3 - q_1$ est appelée **écart interquartile**.

Exemple 3 *Le tableau suivant donne l'âge de chacune des onze personnes d'une assemblée de notables d'un quartier.*

Age	50	57	60	62	65	66
effectif	3	2	1	2	1	2

$$\underbrace{50 - 50 - \underbrace{50}_{q1} - 57 - 57}_{\text{série inférieure}} - \underbrace{60}_{Mé} - \underbrace{62 - 62 - \underbrace{65}_{q3} - 66 - 66}_{\text{série supérieure}}$$

Interprétation :
- *Le quart des personnes a au maximum* 50 *ans.*
- *La moitié des personnes a au maximum* 60 *ans.*
- *Les trois quarts des personnes ont au maximum* 65 *ans.*

L'intervalle interquartile est $[50; 65]$ *et l'écart interquartile est :* $65 - 50 = 15$

Exemple 4 *Au cours d'une dictée le relevé du nombre de fautes commises par douze élèves a donné le tableau suivant :*

nombre de fautes	0	1	2	3	5
nombre d'élèves	2	4	1	2	3

$$\underbrace{0-0-1\underbrace{\ }_{q1}1-1-1}_{\text{série inférieure}}\underbrace{\ }_{Mé}\underbrace{2-3-3\underbrace{\ }_{q3}5-5-5}_{\text{série supérieure}}$$

On a : $q_1 = 1$ *;* $Mé = 1.5$ *et* $q_3 = 4$.
Interprétation :
- 25 *pour* 100 *d'élèves, soit* 3 *élèves ont commis au plus une faute.*
- 50 *pour* 100 *d'élèves, soit* 6 *élèves ont commis au plus une faute et demie.*
- 75 *pour* 100 *d'élèves, soit* 9 *élèves ont commis au plus quatre fautes.*

L'intervalle interquartile est $[1; 4]$ *et l'écart interquartile est* $4 - 1 = 3$.

Remarque 4 *Il y'a un écart de* 3 *fautes entre les "élèves moyens".*

Exemple 5 *On considère la série suivante :*

Taille (en cm)	*Effectifs*	*ECC*
$[155; 160[$	25	25
$[160; 165[$	80	105
$[165; 170[$	45	150
$[170; 175[$	30	180
$[175; 180[$	20	200
$Total$	200	

On a : $q_1 \in [160; 165[$, *d'où*

$$\frac{q_1 - 160}{50 - 25} = \frac{165 - 160}{105 - 25},$$

$$q_1 = \frac{5 \times 25}{80} + 160,$$

donc $q_1 = 161,5625$.
25% *des effectifs, soit* 50 *individus ont une taille comprise entre* 155 *et* $161,5625$ *cm.*

On a : $q_3 \in [165; 170[$, *d'où*

$$\frac{q_3 - 165}{150 - 105} = \frac{170 - 165}{150 - 105},$$

$$q_3 = \frac{5 \times 45}{45} + 165,$$

donc $q_3 = 170$.
75% *des effectifs, soit* 150 *individus ont une taille comprise entre* 155 *et* 170 *cm.*

$[161,5625; 170]$ *est l'intervalle interquartile.*
50% *des effectifs, soit* 100 *individus ont une taille comprise entre* $161,5625$ *et* 170 *cm*

L'écart interquartile est alors : $170 - 161,5625 = 8,4375$.
Il y'a un écart de $8,4375cm$ *entre les* 100 *individus de "taille moyenne".*

1.4 Ecart moyen

L'écart moyen est un nombre positif ou nul qui permet d'apprécier l'éloignement des valeurs de caractère par rapport à la moyenne de la série.

1.4.1 Cas de série discrète

Soit une série de valeurs $x_1, \cdots, x_n$, d'effectif total N et de moyenne $\bar{x}$.
On appelle **écart moyen** de la série la moyenne arithmétique des écarts entre les différentes valeurs x_i et la moyenne $\bar{x}$ de la série.

$$e_m = \frac{\sum_{i=1}^{n} n_i |x_i - \bar{x}|}{N}$$

N.B : On appelle écart entre deux valeurs a et b la valeur absolue de la différence : $|a-b|$.

Exemple 6 *Reprenons l'exemple* 4 *du paragraphe* 1.3.2. *Calculons la moyenne* $\bar{x}$ *de la série et complétons le tableau suivant :*

x_i	*Effectifs* x_i	$\lvert x_i - \bar{x}\rvert$	$n_i\lvert x_i - \bar{x}\rvert$
0	*2*	*2.25*	*4.5*
1	*4*	*1.25*	*5*
2	*1*	*0.25*	*0.25*
3	*2*	*0.75*	*1.5*
5	*3*	*2.75*	*8.25*
Totaux	*12*	.	*19.5*

On a alors

$$\bar{x} = \frac{0+4+2+6+15}{12} = 2.25$$

et

$$e_m = \frac{19.5}{12} = 1.625.$$

1.4.2 Cas de série continue

Soit une série statistique de classes

$$[a_1, b_1[, \quad [a_2, b_2[, \quad \cdots \quad , [a_n, b_n[,$$

d'effectif total N et de moyenne $\bar{x}$. On appelle **écart moyen** de la série la moyenne arithmétique des écarts entre les différents centres c_i des classes et la moyenne $\bar{x}$ de la série.

$$e_m = \frac{\sum_{i=1}^{n} n_i |c_i - \bar{x}|}{N}.$$

Exemple 7 *Reprenons l'exemple* 5 *du paragraphe* 1.3.2. *Calculons la moyenne* $\bar{x}$ *de la série et complétons le tableau suivant :*

Taille (cm)	*Centre des classes*	*Effectifs* x_i	$\lvert c_i - \bar{x}\rvert$	$n_i \lvert c_i - \bar{x}\rvert$
[155 ; 160[	*157.5*	*25*	*8.5*	*212.5*
[160 ; 165[	*162.5*	*80*	*3.5*	*280*
[165 ; 170 [	*167.5*	*45*	*1.5*	*67.5*
[170 ; 175 [	*172.5*	*30*	*6.5*	*195*
[175 ; 180 [	*177.5*	*20*	*11.5*	*230*
Totaux	.	*200*	.	*985*

$$\bar{x} = \frac{3937.5 + 13000 + 7537.5 + 5175 + 3550}{200} = \frac{33200}{200} = 166,$$

d'où

$$e_m = \frac{985}{200} = 4.925.$$

1.5 Variance et écart-type

Pour éviter l'utilisation souvent fastidieuse des valeurs absolues, il est préférable de calculer les carrés des écarts que de calculer la valeur absolue de ces écarts employée dans la définition de l'écart-moyen.

1.5.1 Cas d'une série discrète

Soit une série de valeurs $x_1, \cdots, x_n$, d'effectif total N et de moyenne $\bar{x}$.
On appelle **variance** de la série la moyenne arithmétique des carrés des différences entre les valeurs x_i et la moyenne $\bar{x}$ de la série.

$$V = \frac{\sum_{i=1}^{n} n_i(x_i - \bar{x})^2}{N}.$$

On appelle écart-type de la série la racine carrée de la variance.

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

1.5.2 Cas d'une série continue

Soit une série statistique de classes

$$[a_1, b_1[, \quad [a_2, b_2[, \quad \cdots \quad , [a_n, b_n[,$$

d'effectif total N et de moyenne $\bar{x}$. On appelle **variance** de la série la moyenne arithmétique des carrés des différences entre les centres c_i des classes et la moyenne $\bar{x}$ de la série.

$$V = \frac{\sum_{i=1}^{n} n_i(c_i - \bar{x})^2}{N}.$$

Dans ce cas aussi l'écart-type est la racine carré de la variance.
N.B : La variance est aussi appelée **fluctuation**.

1.5.3 Méthode de calcul par le théorème de KOENIG

Pour calculer la variance, on peut utiliser la formule de Koenig suivante :

$$V = \frac{1}{N}\sum_{i=1}^{n} n_i x_i^2 - \bar{x}^2.$$

Ainsi, la variance est égale à la différence entre la moyenne des carrés des valeurs du caractère et le carré de la moyenne arithmétique de la série.

1.5.4 Propriétés

Soit X le caractère de la série, $V(X)$ sa variance et $\sigma(X)$ son écart-type. On a :

$$V(X) \geq 0;$$
$$V(aX + b) = a^2 V(X);$$
$$\sigma(aX + b) = |a|\sigma(X);$$

avec a un réel non nul et b réel fixé.

1.6 Exercices

Exercice 1 *Chaque élève d'une classe de troisième a pesé à l'aide d'une balance électronique, une pièce de* 1 *euro. Voici les masse (en grammes) :*

$$7.49; 7.49; 7.51; 7.55; 7.51; 7.52; 7.51; 7.54; 7.51; 7.51; 7.52; 7.51; 7.48;$$

$$7.58; 7.50; 7.54; 7.51; 7.60; 7.59; 7.56; 7.49; 7.53; 7.51; 7.54; 7.53.$$

1. *Définis la série.*
2. *Calcule la médiane et la moyenne de la série.*
3. *Calcule la médiane de la série inférieure et celle de la série supérieure. Déduis-en l'écart interquartile.*
4. *Calcule l'étendue, l'écart moyen et l'écart-type de la série.*

Exercice 2 *Dans une écurie, on a sélectionné* 20 *lutteurs dont les poids (en kg) sont :*

$$95; 90; 100; 105; 90; 110; 82; 95; 92; 90;$$

$$105; 100; 105; 102; 90; 100; 85; 92; 90; 82.$$

1. *a). Définis la série statistique puis calcule l'étendue et l'écart interquartile de cette série.*
   b). Interprète les résultats trouvés.
2. *Calcule l'écart moyen de la série.*
3. *Calcule l'écart-type de la série.*

Exercice 3 *Dans un établissement scolaire, le relevé du nombre d'heures effectuées durant un mois par chaque professeur est le suivant :*

Nombre d'heures	*Nombre de professeurs*
[20 ; 30[	*8*
[30 ; 40[	*14*
[40 ; 50 [	*15*
[50 ; 60 [	*25*
[60 ; 70 [	*20*
[70 ; 80 [	*16*
[80 ; 90 [	*12*

1. *Quelle est l'étendue de la série.*
2. *Calcule l'écart interquartile de la série.*
3. *Calcule l'écart moyen de la série.*
4. *Calcule l'écart-type de la série.*

Exercice 4 *Dans le registre d'une maternité, nous constatons que le poids de chaque nouveau-né (exprimé en (kg)) est compris entre* $2kg$ *et* $5kg$*. Voici le tableau des effectifs :*

Poids	*[2 ; 2.6[*	*[2.6 ; 3.2[*	*[3.2 ; 3.8[*	*[3.8 ; 4.4[*	*[4.4 ; 5[*
Effectifs	*6*	*32*	*42*	*19*	*1*

1. *Calcule l'étendue de la série.*
2. *Calcule les quartiles et interprète les résultats. Déduis-en l'écart interquartile de la série.*
3. *Calcule l'écart moyen et l'écart-type de la série.*

Exercice 5 *Voici, dans l'ordre croissant les notes obtenues par les élèves de troisième à un devoir de latin :*

$$3\,;5\,;7\,;8\,;9\,;10\,;10\,;13\,;13\,;15\,;16\,;19\,;19.$$

1. *Détermine la médiane de la série.*
2. *Calcule les quartiles et interprète les résultats obtenu.*
3. *Calcule l'écart interquartile et l'écart moyen.*

Exercice 6 *La hauteur (en mm) de pluie recueillie dans chacune des* 25 *localités d'une région a donné le tableau suivant :*

Hauteur de pluie	*28*	*30*	*32*	*35*	*38*	*40*	*43*	*50*
Nombre de localités	*3*	*2*	*3*	*4*	*6*	*3*	*3*	*1*

1. *Quelle est la hauteur moyenne de pluie dans la région ?*
2. • *Calcule l'écart de hauteur entre la localité la plus arrosée et celle la moins arrosée.*
   • *Calcule l'écart interquartile.*
3. *Calcule l'écart moyen puis la variance de la série.*

Exercice 7 *Le tableau suivant donne la répartition des salaires mensuels (en euro) des employés d'une entreprise.*

Grille de salaire	*Effectifs*
Moins de 1000	*12*
[1000 ; 1200[	*30*
[1200 ; 1400 [	*42*
[1400 ; 1600[	*27*
[1600 ; 1800 [	*30*
[1800 ; 2000[	*23*
[2000 ; 2200[	*40*
[2200 ; 2400[	*26*
Plus de 2400	*20*

1. *A quelle classe appartient la médiane Mé de la série ?*
2. *Pour chacun des paramètres suivants, précise la classe à laquelle il appartient : quartile Q_1, quartile Q_3, décile D_1 et décile D_9.*
3. *Représente la série des effectifs cumulés croissants à l'aide d'histogrammes. Déduis-en le tracé du polygone des fréquences cumulées croissantes.*
4. *On suppose que la répartion est uniforme dans chacune des classes. Donne une valeur approchée de chacun des cinq paramètres ci-dessus.*
5. *Calcule la valeur exacte de la médiane Mé et celle du premier décile D_1.*

Exercice 8 *La pesée de* 100 *sacs d'arachide d'une fabrique d'huile a fourni le tableau suivant :*

Poids (en kg)	*Effectifs*
[10 ; 20[	*12*
[20 ; 30[	*8*
[30 ; 40[	*10*
[40 ; 45[	*20*
[45 ; 50[	*15*
[50 ; 60[	*15*
[60 ; 80[	*11*
[80 ; 90[	*9*

1. *Calcule l'écart interquartile.*
2. *Calcule les $5^{ème}$ et $6^{ème}$ déciles. Donnes-en une interprétation.*
3. *Calcule l'écart-type de la série.*

1.7 Solutions

Exercice 2

1. *a*) Calculs :

Poids : x_i	82	85	90	92	95	100	102	105	110
Nombre de lutteurs : n_i	2	1	5	2	2	3	1	3	1

• Étendue : $e = 110 - 82 = 28$
• Le 1^{er} quartile est $q_1 = 90$. Le $3^{\text{ème}}$ quartile est $q_3 = \frac{100+102}{2} = 101$.
L'écart interquartile est : $q_3 - q_1 = 101 - 90 = 11$.
b) Interprétation :
• L'étendue est de 28 kg : la différence de poids entre 2 lutteurs est toujours ≤ 28 kg.
• L'écart interquartile est de 11 kg : la différence de poids entre les lutteurs de la catégorie moyenne est ≤ 11 kg.

2. La moyenne de la série est :

$$\overline{x} = \frac{\sum_{i=1}^{i=9} n_i x_i}{N} = \frac{1900}{20} = 95$$

.L'écart moyen est :
$e_m = \frac{(2\times13)+(1\times10)+(5\times5)+(2\times3)+(2\times0)+(3\times5)+(1\times7)+(3\times10)+(1\times15)}{20} = \frac{134}{20}$.

$$e_m = 6,7.$$

3. L'écart type σ est la racine carrée de la variance

$$V = \frac{1}{N}\sum_{i=1}^{9} n_i(x_i - \overline{x}^2) = 61,5.$$

L'écart-type $\sigma = \sqrt{61,5} \simeq 7,84$.

Exercice 4

1. L'étendue de la série est la différence entre la plus grande et la plus petite des bornes. $e = x_{max} - x_{min}$

$$e = 5 - 2 = 3$$

.

2. Dressons le tableau statistique des effectifs cumulés croissants

[2 ; 2,6[	[2,6 ; 3,2[	[3,2 ; 3,8[	[3,8 ; 4,4[	[4,4 ; 5[
6	38	80	99	100

• Calcul du 1er quartile q_1 :

$$\frac{q_1 - 2,6}{3,2 - 2,6} = \frac{25 - 6}{38 - 6} \Rightarrow q_1 \simeq 2,96.$$

Interprétation : 25 % des nouveau-nés ont un poids $\leq 2,96\ kg$.
• Calcul du second quartile q_2 = Mé :

$$\frac{q_2 - 3,2}{3,8 - 3,2} = \frac{50 - 38}{80 - 38} \Rightarrow q_2 \simeq 3,37.$$

Interprétation : 50 % des nouveau-nés ont un poids $\leq 3,37\ kg$.
• Calcul du 3er quartile q_3 :

$$\frac{q_3 - 3,2}{3,8 - 3,2} = \frac{75 - 38}{80 - 38} \Rightarrow q_3 \simeq 3,73.$$

Interprétation : 75 % des nouveau-nés ont un poids $\leq 3,73\ kg$.
L'écart interquartile est $q_3 - q_1 = 0,77$.

3. • Calcul de la moyenne de la série :

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{5} n_i c_i = \frac{336,2}{100} = 3,362.$$

• Calcul de l'écart- moyen e :

$$e = \frac{\sum_{i=1}^{5} n_i |c_i - \overline{x}|}{N} \simeq 0,42.$$

• Calcul de la variance V de la série :

$$V = \frac{\sum_{i=1}^{5} n_i (c_i - \overline{x})^2}{N} \simeq 0,27.$$

• Calcul de l'écart- type σ :

$$\sigma = \sqrt{V} \simeq 0,51.$$

Exercice 6

1. La moyenne de la série est :
   $\overline{x} = \frac{(3\times28)+(2\times230)+(3\times32)+(4\times35)+(6\times38)+(3\times40)+(3\times43)+(1\times50)}{25} = \frac{907}{25} = 36,28$. La hauteur moyenne de pluie dans la région est égale à $36,28mm$.
2. • L'étendue de la série est : $e = 50 - 28 = 22$.
   Donc l'écart de hauteur est égal à $22mm$.
   • Le 1^{er} quartile est $q_1 = 32$ et le $3^{\text{ème}}$ quartile $q_3 = 40$.
   Donc l'écart interquartile est :$q_3 - q_1 = 40 - 32 = 8$.
3. L'écart-moyen est :

$$e_m = \frac{1}{N}\sum_{i=1}^{8} n_i |x_i - \overline{x}| = \frac{110,72}{25} = 4,428.$$

La variance est :

$$V = \frac{1}{N}\sum_{i=1}^{8} n_i \left(x_i - \overline{x}\right)^2,$$

ou

$$V = \frac{1}{N}\sum_{i=1}^{8} n_i x_i^2 - \overline{x}^2,$$

$$V = \frac{33635}{25} - (36,28)^2 = 29,161.$$

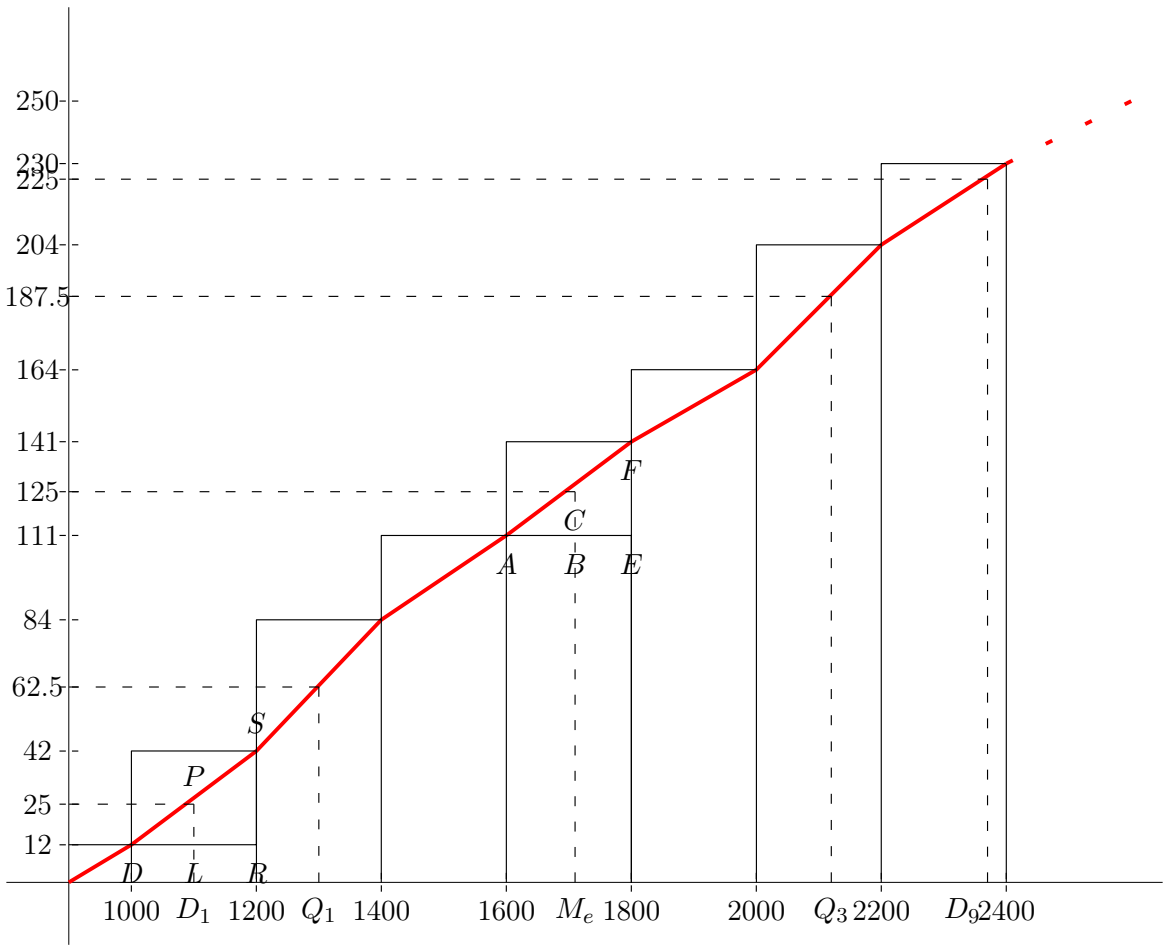


FIGURE 1.1 – Exercice 7

Exercice 7

1. Mé $\in [1600; 1800[$.
2. $q_1 \in [1200; 1400[$; $q_3 \in [2000; 2200[$; $D_1 \in [1000; 1200[$; $D_9 \in [2200; 2400[$.
3. (graphe histogramme).ici1
4. A partir du graphique précèdent on obtient par projection sur l'axe des modalités la valeur approchée de chacun des cinq paramètres demandés :

$$q_1 \simeq 1300, \quad q_3 \simeq 2120, \quad \text{Mé} \simeq 1710, \quad D_1 \simeq 1100, \quad D_9 \simeq 2370.$$

5. Calcul de la médiane :

$$\frac{\text{Mé} - 1600}{1800 - 1600} = \frac{125 - 111}{141 - 111},$$

$$\text{Mé} = 1693,3\mathbf{3}.$$

Calcul de D_1 :

$$\frac{D_1 - 1000}{1200 - 1000} = \frac{25 - 12}{42 - 12},$$

$D_1 = 1086,6\mathbf{6}$.

Chapitre 2

Séries statistiques à deux variables

Introduction

Il s'agit, dans ce chapitre, d'étudier simultanément, sur une même population d'effectif N, deux caractères de même type (quantitatifs ou qualitatifs) ou de types différents (l'un qualitatif et l'autre quantitatif). Chaque individu E_i de la population fournit une valeur x_i de l'un des caractères X et une valeur y_i de l'autre caractère Y. Il est commode de représenter l'observation de E_i par le point $M_i(x_i; y_i)$ dans un repère du plan ou à l'aide d'un diagramme. Les séries à étudier seront de deux sortes : les séries à données individuelles ou séries injectives et les séries à données groupées ou séries non injectives. Une série est dite **injective** si deux points distincts ne peuvent pas avoir le même couple de coordonnés (x, y).

Remarque 5 *Dans la suite, les séries à étudier seront des séries à caractères quantitatifs.*

2.1 Séries à données individuelles

2.1.1 Définition

Une série statistique est dite à données individuelles s'il n'existe pas deux individus distincts de la population qui prennent le même couple de valeurs de caractères.

Remarque 6 *Le tableau statistique (des effectifs) associé à une telle série est appelé* ***tableau linéaire****.*

Exemple 8 *Voici un tableau donnant les moyennes de mathématiques et de français d'un même élève lors de cinq contrôles consécutifs :*

Français : X	*7*	*10*	*13*	*11*	*12*
Maths : Y	*8*	*8*	*10*	*9*	*10*

Ici la population compte 5 *individus que sont les* 5 *contrôles ordonnés.*
Le tableau est un tableau linéaire.
la série est injective et pourtant nous avons 2 *valeurs égales à* 8 *dans la* $2^{ième}$ *ligne ce qui signifie que l'injection n'est pas celle d'une application qui fait passer de* X *à* Y *mais de l'application qui, à* E_i*, associe* $M_i(x_i; y_i)$*.*

2.1.2 Nuage de points et point moyen

Définition 1 *On appelle nuage de points associé à une série statistique à deux caractères, l'ensemble des points* M_i *de coordonnés* (x_i, y_i) *représentant les observations faites sur les individus* E_i *de la population.*

Représenter le nuage de points revient à placer tous les points $M_i(x_i, y_i)$ dans le plan rapporté à un repère.

Définition 2 *On appelle point moyen ou barycentre du nuage de points, le point* $G(\overline{x}; \overline{y})$*.*

Exemple 9 *Le tableau suivant indique l'évolution de* 1971 *à* 1979 *du prix moyen au kg (en francs) d'une denrée*

Année : X	*1971*	*1972*	*1973*	*1975*	*1976*	*1977*	*1978*	*1979*
Prix : Y	*1,20*	*1,70*	*1,80*	*2,60*	*2,75*	*3,25*	*3,30*	*3,65*

Le nuage est l'ensemble des 8 *points suivant :*

$$\Big\{M_1(1971; 1,20); M_2(1972; 1,70); \ldots; M_8(1979; 3,65)\Big\}.$$

Graphique
ici2

Attention ! Ne pas entourer la nuage si sa direction n'est pas demandée.

2.1.3 Utilisation du nuage

Le nuage de points permet, dans sa représentation graphique de déceler ou pas une tendance de l'une des variable en fonction de l'autre.
Graphiques
ici3

2.1.4 Ajustement linéaire

Un ajustement linéaire vise à trouver une droite "qui passe le plus prés possible" de tous les points du nuage. On peut le faire de plusieurs façons dont les 3 que voici :

a. **Ajustement à main levée**
On cherche une droite qui **semble** "passer le plus prés possible" de tous les points du nuage, le choix de cette droite étant arbitraire.
Exemple 10 *Dans l'exemple* 9*, pour tracer la droite d'ajustement D, Monsieur A a choisi les points :* $M_3(1973; 1,80)$ *et* $M_7(1978; 3,3)$ *et a trouvé :* $D : y = 0,3x - 590,1$ *comme équation.*
Quant à Madame B son choix a porté sur les points : $M_4(1975; 2,6)$ *et* $M_6(1977; 3,25)$ *et a trouvé :* $D : y = 0,325x - 639,275$ *comme équation.* ***Remarques :***

• *Si cette méthode a l'avantage d'être simple, par contre elle présente l'inconvénient d'être subjectif car chaque individu peut obtenir sa propre droite d'ajustement.*

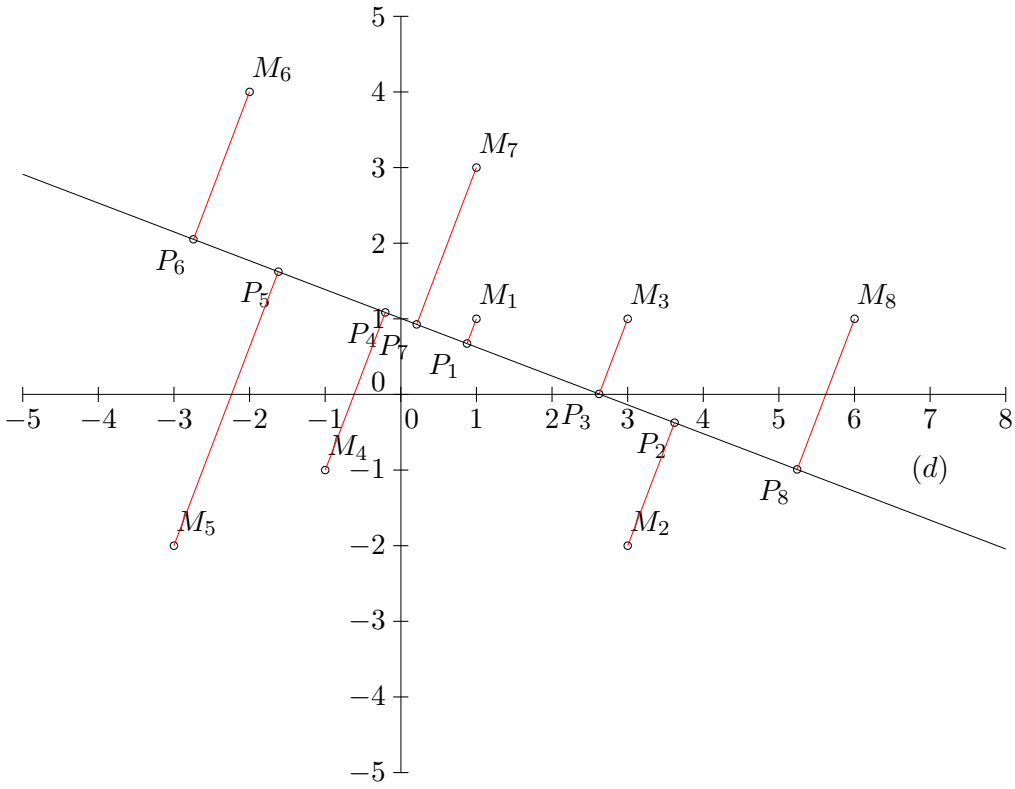
• *Les points choisis pour obtenir la droite D ne sont pas nécessairement des points du nuage.*

b. **Ajustement par la méthode de Mayer**
La méthode d'ajustement de Mayer consiste à partager la série en deux séries s_1 et s_2 de même effectif ou d'effectifs différents d'une unité. On détermine ensuite les points moyens G_1 et G_2 respectivement des séries s_1 et s_2. La droite (G_1G_2) est appelée droite d'ajustement par la méthode de Mayer (ou droite de Mayer).

Exemple 11 *Reprenons le tableau de l'exemple* 9*. La série est ordonnée et a un effectif pair. On va donc la partager en* 2 *séries d'effectifs* 4 *chacun.*
La série s_1 *est la série des points* $M_i, 1 \leq i \leq 4$ *associés aux années* $1971, 1972, 1973, 1975$*, la série* s_2 *celle des* 4 *points restants.*
On obtient : $G_1(1972,75; 1,823)$*,* $G_2(1977,5; 3,2375)$ *et l'équation de la droite de Mayer suivante :* $y = 0,297x - 586,913$
Remarques :
• *La méthode de Mayer est toujours associée à un partage à préciser*
• *La méthode de Mayer n'est pas toujours la meilleure méthode d'ajustement.*

c. **Ajustement par la méthode des moindres carrés**

Le nuage de points M_i représentant une série statistique double (X, Y) est tel que ses points soient le plus proche possible d'une droite D, c'est à dire la moyenne des distances M_iP_i est la plus petite possible ; P_i étant le projeté de M_i sur D parallèlement l'axe des ordonnées.



L'équation de D est donnée par $y = ax + b$

$M_iP_i = |y_i - (ax_i + b)|$.
Le but est de minimiser la somme

$$S = \sum_{i=1}^{N} M_iP_i^2 = \sum_{i=1}^{N} \Big(y_i - (ax_i + b)\Big)^2, N \text{ étant l'effectif total.}$$

Pour minimiser S, il suffit de déterminer les valeurs $\widehat{a}$ de a et $\widehat{b}$ de b nécessaires à cela.

• Calcul du coefficient $\widehat{b}$

$$S = \sum_{i=1}^{N} \Big(y_i - (ax_i + b)\Big)^2 = \sum_{i=1}^{N} \Big((y_i - ax_i) - b\Big)^2$$

$$= \sum_{i=1}^{N} (y_i - ax_i)^2 - 2b \sum_{i=1}^{N} (y_i - ax_i) + N.b^2.$$

En fixant a, S est un trinôme du second degré en b :

$$S = N.b^2 - 2K.b + T,$$

où $K = \sum_{i=1}^{N}(y_i - ax_i)$ et $T = \sum_{i=1}^{N}(y_i - ax_i)^2$.
Puisque $N > 0$, S admet un minimum pour $b = \widehat{b} = \frac{K}{N}$.
Le trinôme est minimal pour

$$\widehat{b} = \frac{\sum_{i=1}^{N}(y_i - ax_i)}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i - a\frac{1}{N}\sum_{i=1}^{N} x_i,$$

donc

$$\widehat{b} = \overline{y} - a\overline{x}.$$

• Calcul du coefficient $\widehat{a}$

$$S = \sum_{i=1}^{N}\Big(y_i - (ax_i + b)\Big)^2 = \sum_{i=1}^{N}\Big((y_i - b) - ax_i\Big)^2$$

$$S = \sum_{i=1}^{N}(y_i - b)^2 - 2a\sum_{i=1}^{N} x_i(y_i - b) + a^2\sum_{i=1}^{N} x_i^2.$$

En fixant b, S est un trinôme du second degré en a :

$$S = \alpha.a^2 - 2\beta.a + \gamma,$$

où $\alpha = \sum_{i=1}^{N} x_i^2$, $\beta = \sum_{i=1}^{N} x_i(y_i - b)$ et $\gamma = \sum_{i=1}^{N}(y_i - b)^2$.
Puisque $\alpha > 0$, S admet un minimum pour $a = \widehat{a} = \frac{\beta}{\alpha}$.
Le trinôme est minimal pour

$$\begin{aligned}
\widehat{a} &= \frac{\sum_{i=1}^{N} x_i(y_i - b)}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N}(x_i y_i - bx_i)}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N} x_i y_i - b\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N} x_i y_i - Nb\overline{x}}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N} x_i y_i - N(\overline{y} - \widehat{a}\overline{x})\overline{x}}{\sum_{i=1}^{N} x_i^2}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y} + \widehat{a}N\overline{x}^2}{\sum_{i=1}^{N} x_i^2} \\
&= \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2} + \frac{\widehat{a}N\overline{x}^2}{\sum_{i=1}^{N} x_i^2},
\end{aligned}
$$

donc

$$
\widehat{a} - \frac{\widehat{a}N\overline{x}^2}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2}.
$$

$$
\frac{\widehat{a}\sum_{i=1}^{N} x_i^2 - \widehat{a}N\overline{x}^2}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2}
$$

$$
\widehat{a}\Big(\sum_{i=1}^{N} x_i^2 - N\overline{x}^2\Big) = \sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y},
$$

donc

$$
\begin{aligned}
\widehat{a} &= \frac{\sum_{i=1}^{N} x_i y_i - N\overline{x}\cdot\overline{y}}{\sum_{i=1}^{N} x_i^2 - N\overline{x}^2} \\
&= \frac{N\Big(\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\cdot\overline{y}\Big)}{N\Big(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2\Big)} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\cdot\overline{y}}{\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \overline{x}^2}.
\end{aligned}
$$

L'équation de la droite de régression de Y en X, obtenue lorsque S est minimum, est alors :

$$
D: \quad y = \widehat{a}x + \widehat{b}, \quad \text{avec} \quad \widehat{b} = \overline{y} - \widehat{a}\cdot\overline{x}.
$$

L'équation de la droite D devient $y = \widehat{a}x + \overline{y} - \widehat{a}\overline{x}$, donc

$$
D: \quad y - \overline{y} = \widehat{a}(x - \overline{x}).
$$

Remarque 7 *La covariance du couple de caractères (X, Y), notée $cov(X, Y)$ [1], est :*

$$
cov(X, Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}),
$$

1. $cov(X^2) = var(X)$

ou bien

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \overline{x}\cdot\overline{y}$$

De cette remarque on tire :

$$\widehat{a} = \frac{cov(X,Y)}{var(X)}.$$

De façon analogue, une équation de la droite de régression de X en Y est donnée par

$$D' : \qquad x - \overline{x} = \widehat{a}'(y - \overline{y}),$$

avec $\widehat{a}' = \frac{cov(X,Y)}{var(Y)}$.

Remarque 8 *Les deux droites de régression de Y en X et de X en Y passent toutes deux par le point moyen $G(\overline{x};\overline{y})$.*

Exemple 12 *La taille (X) et le poids (Y) de 15 enfants âges de 9 ans sont 2 caractères d'une série statistique à 2 variables. Les résultats des observations faites sur chaque enfant de la population sont donnés dans le tableau suivant*

Enfants i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Taille en (cm) (X)	*131*	*133*	*140*	*136*	*124*	*136*	*127*	*130*	*139*	*142*
Poids en kg (Y)	*33.0*	*28.5*	*28.5*	*33.0*	*28.0*	*29.0*	*24.0*	*23.5*	*31.5*	*34.0*

11	*12*	*13*	*14*	*15*
147	*134*	*131*	*129*	*135*
36.5	*27.0*	*27.5*	*31.0*	*34.5*

1.) Détermine une équation de la droite de régression de Y en X par la méthode des moindres carrés.
2.) Détermine une équation de la droite de régression de X en Y par la méthode des moindres carrés.

SOLUTION

1.)

$$\overline{x} = \frac{\sum_{i=1}^{i=15} x_i}{N} = \frac{2014}{15} = 134.27cm$$

$$\overline{y} = \frac{\sum_{i=1}^{i=15} y_i}{N}\frac{449.5}{15} = 29.97kg,$$

d'où

$$\hat{a} = \frac{\sum_{i=1}^{i=15} x_i y_i - 15\bar{x}\bar{y}}{\sum_{i=1}^{i=15} x_i^2 - 15\bar{x}^2} \approx \frac{206.63}{510.93} \approx 0.404 kg/cm.$$

$$y - \bar{y} = \hat{a}(x - \bar{x}) \Rightarrow y = \hat{a}x + \bar{y} - \hat{a}\bar{x},$$

$$y = (0.404)x - 24.28.$$

2.)

$$\hat{b} = \frac{\sum_{i=1}^{i=15} x_i y_i - 15\bar{x}\bar{y}}{\sum_{i=1}^{i=15} y_i^2 - 15\bar{y}^2} \approx \frac{206.63}{199.74} \approx 1.04 cm/kg.$$

$$x - \bar{x} = \hat{b}(y - \bar{y}) \Rightarrow x = \hat{b}y + \bar{x} - \hat{b}\bar{y},$$

$$x = (1.04)x - 3.10 \Leftrightarrow y = (0.96)x - 2.98.$$

Exemple 13 *Voici trois points $A_1(-2;-2)$, $A_2(0;1)$, $A_3(2;1)$. On donne la somme*

$$S = \sum_{i=1}^{i=3} [y_i - (ax_i + b)]^2.$$

1.) Détermine deux polynômes g et h de degré 2 chacun tels que $S = g(a) + h(b)$
2.) détermine la valeur de a pour que g soit minimal et la valeur de b pour que h soit minimal.
3.) En admettant que ces deux valeurs trouvées de a et b minimisent la somme S, donne une équation de la droite Δ correspondante.
4.) Représente la droite Δ et les points $A_i(2;1), i = 1, 2, 3$.

SOLUTION

$$1.) S = [-2 - (-2a + b)]^2 + [1 - (a \times 0 + b)]^2 + [1 - (-2a + b)]^2$$

$$(2a - b - 2)^2 + (1 - b)^2 + (1 - 2a - b)^2 = 8a^2 - 12a + 3b^{2+6} = g(a) + h(b),$$

on a donc

$$g(a) = 8a^2 - 12a + 3b^2 + 6$$

et

$$h(b) = 3b^2 + 6.$$

2.) $g'(a) = 16a - 12 \Rightarrow \hat{a} = \frac{12}{16} = \frac{3}{4}$
$h'(b) = 6b \Rightarrow \hat{b} = 0$

3.) $\Delta: \quad y = \hat{a}x + \hat{b} \qquad \Delta: \quad y = \frac{3}{4}x.$
(NB : $\overline{x} = 0, \overline{y} = 0$, $\sum_{i=1}^{i=3} x_i y_i = 6$, $\sum_{i=1}^{i=3} x_i^2 = 8$, $\hat{a} = \frac{6}{8} = \frac{3}{4}$, $\hat{b} = 0$, $y = \frac{3}{4}x$*)*
4.)

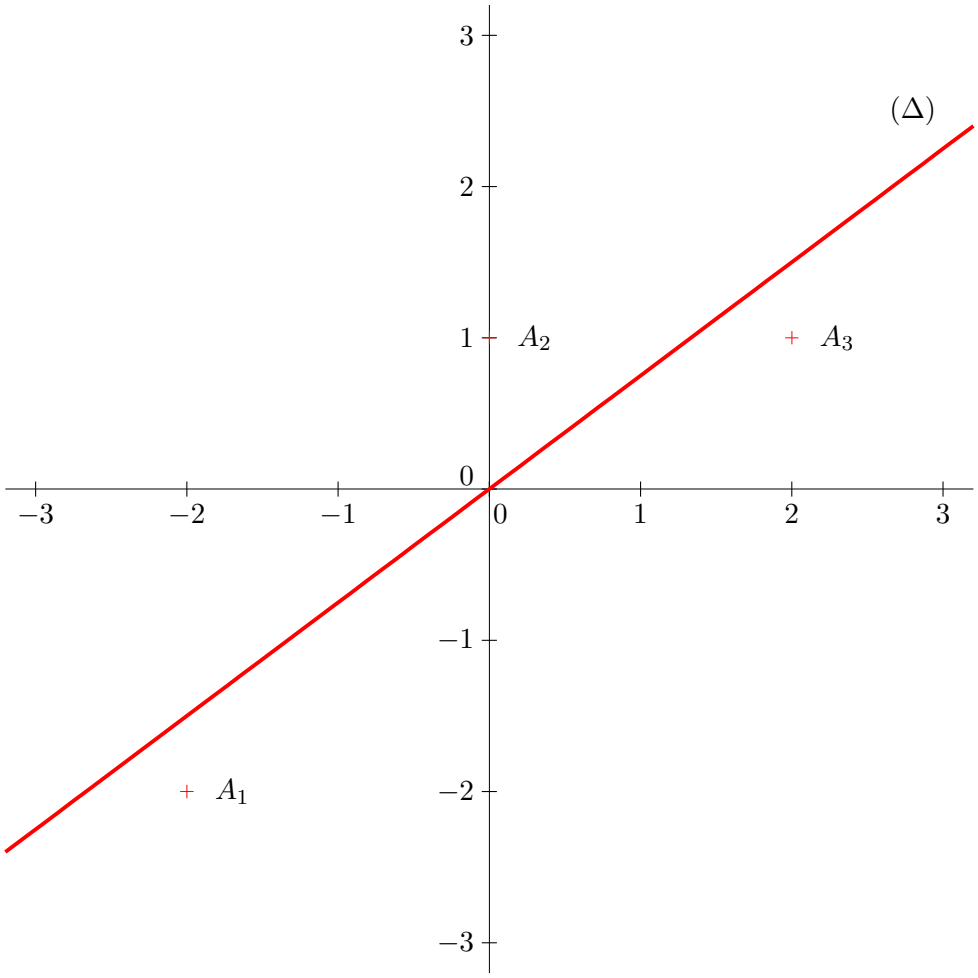


FIGURE 2.1 – Exercice

2.1.5 Coefficient de corrélation linéaire

a. **Définition 3** *Soit (X,Y) un couple de caractères. On appelle coefficient de corrélation linéaire entre X et Y le nombre réel $r = \frac{cov(X,Y)}{\sigma(X)\sigma(Y)}$, où $cov(X,Y)$ est la covariance de la série double (X,Y), $\sigma(X)$ est l'écart type de X et $\sigma(Y)$ celui de Y.*

b. **Propriétés**.

b1. Le coefficient de corrélation linéaire entre X et Y est du signe de $cov(X,Y)$.

b2. On a $-1 \leq r \leq 1$

Démonstration 1 *Reprenons la fonction S de deux variables a et b définie par*

$$S = \sum_{i=1}^{i=N} \Big[y_i - (ax_i + b)\Big]^2,$$

$\hat{a}$ et $\hat{b}$ sont les valeurs d a et b qui minimisent S :

$$S = \sum_{i=1}^{i=N} \Big[y_i - (\hat{a}x_i + \hat{y} - \hat{a}\bar{x})\Big]^2, \ car\ \hat{b} = \hat{y} - \hat{a}\bar{x}.$$

$$\begin{aligned} S = & \sum_{i=1}^{i=N}(y_i - \bar{y})^2 - 2\hat{a}\sum_{i=1}^{i=N}(y_i - \bar{y})(x_i - \bar{x}) + \hat{a}^2\sum_{i=1}^{i=N}(x_i - \bar{x})^2 \\ = & \ n\sigma_Y^2 - 2\hat{a}n \cdot cov(XY) + \hat{a}^2 n\sigma_X^2 \\ = & \ n\Big[\sigma_Y^2 - 2\hat{a} \cdot cov(XY) + \hat{a}^2\sigma_X^2\Big]. \end{aligned}$$

La somme S étant positive ou nulle (car somme de carré), l'équation en $\hat{a}$:

$$\sigma_Y^2 - 2\hat{a} \cdot cov(XY) + \hat{a}^2\sigma_X^2 = 0,$$

admet un discriminant réduit négatif ou nul :

$$\begin{aligned} \Delta' = & \ \Big[cov(XY)\Big]^2 - \sigma(X)^2\sigma(Y)^2 \leq 0 \\ = & \ \sigma(X)^2\sigma(Y)^2\Big[\Big(\frac{cov(XY)}{\sigma(X)\sigma(Y)}\Big)^2 - 1\Big] \leq 0, \end{aligned}$$

ce qui implique que

$$r^2 - 1 \leq 0 \qquad où \qquad r = \frac{cov(XY)}{\sigma(X)\sigma(Y)},$$

soit

$$-1 \leq r \leq 1.$$

b3. Si les deux droites de régression ont pour équations respectives

$$y = \hat{a}(x - \bar{x}) + \bar{y} \qquad \text{et} \qquad x = \hat{a'}(y - \bar{y}) + \bar{x},$$

alors $r^2 = \hat{a} \times \hat{a'}$

Démonstration 2 *On sait que* $\hat{a} = \frac{cov(X,Y)}{V(X)}$ *et* $\hat{a'} = \frac{cov(X,Y)}{V(Y)}$
d'où

$$\hat{a} \times \hat{a'} = \frac{cov^2(X,Y)}{V(X)V(Y)} = \frac{cov^2(X,Y)}{\sigma(X)^2\sigma(Y)^2} = \Big(\frac{cov(X,Y)}{\sigma(X)\sigma(Y)}\Big)^2 = r^2.$$

c. **Interprétation**
Le coefficient de corrélation linéaire du couple (X, Y) permet d'apprécier le degré de dépendance entre les deux caractères X et Y.

1. Lorsque Le coefficient de corrélation linéaire r est proche de 1 ou de -1, soit $r \geq \frac{3}{4}$, alors on dit que la corrélation est **forte** .
   Dans ce cas les deux droites de régression sont proches.
2. Lorsque Le coefficient de corrélation linéaire r est proche de 0, soit $r < \frac{3}{4}$, alors la corrélation est **faible** .
3. Lorsque Le coefficient de corrélation linéaire $r = 1$ ou $r = -1$, alors la corrélation est **parfaite** .
   Dans ce cas les deux droites de régression sont confondues.

2.2 Séries à données groupées

2.2.1 Exemple préliminaire

Des études statistiques en français (X) faites dans une classe ont donné le tableau suivant

Notes de français	05	07	09	11	12	14
Nombre d'élèves	1	4	8	3	2	2

Elles ont donné en mathématiques (Y), dans la même classe, le tableau suivant

Notes de français	06	07	09	10	12
Nombre d'élèves	2	4	8	1	5

Le surveillant a confectionné pour chaque élève e_{ij} de la classe une fiche sur laquelle figurent ses notes de français x_i et et de mathématiques y_j. Ainsi, la répartition des élèves de la classe suivant le couple de notes (x_i, y_j) permet de savoir le nombre n_{ij} d'élèves ayant x_i en français et y_j en mathématiques. Pour cela on a établi le tableau à double entrée suivant.

$\frac{Francais\,X}{Maths\,Y}$	05	07	09	11	12	14	Totaux
06	1	0	0	1	0	0	2
07	0	4	0	0	0	0	4
09	0	0	6	0	2	0	8
10	0	0	0	0	0	1	1
12	0	0	2	2	0	1	5
Totaux	1	4	8	3	2	2	20

Remarque sur la lecture du tableaux
On note que :

. 2 élèves ont 12 en Français et 09 en Maths,
. 4 élèves ont 07 en Français et 07 en Maths,
. Aucun élève n'a 11 en Français et 10 en Maths.

2.2.2 Définition

Considérons une population de N individus sur laquelle on étudie simultanément les deux caractères quantitatifs X et Y.
Désignons par $x_1, x_2, \ldots, x_p$ les p modalités du caractère X et $y_1, y_2, \ldots, y_q$ les q modalités du caractère Y avec $1 \leq p \leq N$ et $1 \leq q \leq N$.
Soit n_{ij} le nombre d'individus de la population présentant à la fois les modalités x_i de X et y_j de Y.
L'ensemble des triplets (x_i, y_j, n_{ij}) est appelé série statistique double. Ces résultants sont consignés dans le tableau suivant appelé tableau de statistique à double entrée ou tableau de contingence.

Y\ X	x_1	x_2	$\ldots$	x_i	$\ldots$	x_p	Total
y_1	n_{11}	n_{21}	$\ldots$	n_{i1}	$\ldots$	n_{p1}	$n_{.1}$
y_2	n_{12}	n_{22}	$\ldots$	n_{i2}	$\ldots$	n_{p2}	$n_{.2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_j	n_{1j}	n_{2j}	$\ldots$	n_{ij}	$\ldots$	n_{pj}	$n_{.j}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_q	n_{1q}	n_{2q}	$\ldots$	n_{iq}	$\ldots$	n_{pq}	$n_{.q}$
Total	$n_{1.}$	$n_{2.}$	$\ldots$	$n_{i.}$	$\ldots$	$n_{p.}$	N

Où

* $n_{i.} = \sum_{i=1}^{i=q} n_{ij}$, $i = 1, \ldots, p$,
  $n_{.j} = \sum_{i=1}^{i=p} n_{ij}$, $j = 1, \ldots, q$.
* n_{2j} individus présentent les modalités x_p et y_2,
  n_{p2} individus présentent les modalités x_2 et y_j,
  $n_{1.}$ individus présentent la modalité x_1,

$n_{.p}$ individus présentent la modalité y_p.

2.2.3 Nuage de points

On désigne par M_{ij} le point de coordonnées (x_i, y_j), $1 \leq i \leq p$, $1 \leq j \leq q$.
On convient de représenter la série statistique double par l'ensemble des points $M_{ij}(x_i, y_j)$ dans le plan rapporté à un repère orthogonal.
NB Si la série n'est pas injective, alors les points sont soit pondérés, soit représentés par des disques de rayons différents
Exemple Graphe

2.2.4 Caractéristiques marginales

a) Série marginale
A partir du tableau de contingence établi dans le paragraphe 2.2.2, on définie la série marginale de la variable X comme étant l'ensemble des couples $(x_i, n_{i.})$, $1 \leq i \leq p$.
De même la série marginale de la variable Y comme étant l'ensemble des couples $(y_i, n_{.j})$, $1 \leq j \leq q$.
Ainsi obtient-on les tableaux linéaires suivants

X	x_1	x_2	...	x_i	...	x_p
Effectif	$n_{1.}$	$n_{2.}$	...	$n_{i.}$	...	$n_{p.}$

Y	y_1	y_2	...	y_i	...	y_q
Effectif	$n_{.1}$	$n_{.2}$	...	$n_{.i}$	...	$n_{.q}$

b) Moyennes marginales Soit la série statistique double (x_i, y_j, n_{ij}), x_i étant les valeurs du caractère X et y_j celles du caractère Y.

∗ On appelle **moyenne marginale** par rapport au caractère au caractère X, le réel $\bar{x} = \frac{1}{N}\sum_{i=1}^{p} n_{i.} \cdot x_i$ N étant l'effectif total.

∗ On appelle **moyenne marginale** par rapport au caractère au caractère Y, le réel $\bar{y} = \frac{1}{N}\sum_{j=1}^{q} n_{.j} \cdot y_j$ N étant l'effectif total.
NB

$$\bar{X} = \sum_{i=1}^{p} f_{i.} \cdot x_i \,; \qquad \bar{Y} = \sum_{i=1}^{q} f_{j.} \cdot y_j$$

c) Variance et écart type marginaux

∗ On appelle **variance marginale** par rapport au caractère au caractère X, le réel positif

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

$*$ On appelle **variance marginale** par rapport au caractère au caractère Y, le réel positif

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

$*$ **L'écart type marginal** est la racine carré de la variance marginale :

$$\sigma_X = \sqrt{V(X)}\,; \quad \sigma_Y = \sqrt{V(Y)}$$

2.2.5 Caractéristiques conditionnelles

Considérons le tableau de contingence établi dans la paragraphe 2.2.5.

a) **Séries conditionnelles**

En fixant $i = 2$ par exemple, on obtient le tableau suivant

$Y/X = x_2$	y_1	y_2	$\ldots$	y_j	$\ldots$	y_p
$n_{ij}/X = x_2$	n_{21}	n_{22}	$\ldots$	n_{2j}	$\ldots$	n_{2q}

où $Y/X = x_2$ représente les modalités de Y lorsque X prend la valeur x_2 et $n_{ij}/X = x_2$ représente les effectifs partiels des couples (x_2, y_j). ce tableau détermine la série conditionnelle par rapport à x_2.

$*$ On appelle **série conditionnelle** par rapport à x_{ln} l'ensemble des couples (y_j, n_{lnj}), ln étant un entier fixé entre 1 et p.

$*$ De même on appelle **série conditionnelle** par rapport à y_m l'ensemble des couples (x_i, n_{im}), m étant un entier fixé entre 1 et q.

Exemple

Reprenons l'exemple préliminaire du paragraphe 2.2.1

- La série conditionnelle par rapport à $x_3 = 09$ est donnée dans le tableau suivant

$Y/X = 09$	0.6	0.7	0.9	10	12
$n_{ij}/X = 09$	0	0	6	0	2

- La série conditionnelle par rapport à $y_1 = 06$ est donnée dans le tableau suivant

$X/Y = 06$	0.5	0.7	0.9	11	12	14
$n_{ij}/Y = 06$	1	0	0	1	0	0

b) **Fréquences conditionnelles**

On appelle **fréquence conditionnelle** de la modalité x_i par rapport à y_j, le rapport noté $f_{x_i/y_j} = \frac{n_{ij}}{n_{.j}}$ et **fréquence conditionnelle** de y_j par rapport à x_i, le rapport

$f_{y_j/x_i} = \frac{n_{ij}}{n_{i.}}$.
Par exemple, la fréquence conditionnelle de x_1 par rapport à y_2 est égale à $f_{x_1/y_2} = \frac{n_{12}}{n_{.2}}$
la fréquence conditionnelle de y_q par rapport à x_1 est égale à
$f_{y_q/x_1} = \frac{n_{1q}}{n_{1.}}$.

c) **Moyennes conditionnelles**

On appelle **moyenne conditionnelle** de X par rapport à y_m de Y, le réel noté

$$\bar{X_m} = \frac{1}{n_{.m}} \sum_{i=1}^{p} n_{im} x_i,$$

avec m un entier fixé entre 1et q.
De même, on appelle **moyenne conditionnelle** de Y par rapport à x_{ln} de X, le réel

$$\bar{Y_{ln}} = \frac{1}{n_{ln.}} \sum_{i=1}^{q} n_{lnj} y_j,$$

avec ln un entier fixé entre 1et p.
Par exemple,
la moyenne conditionnelle de X par rapport à y_1 est égale à

$$\bar{X_1} = \frac{1}{n_{.1}}(n_{11}x_1 + n_{21}x_2 + \cdots + n_{i1}x_i + \cdots + n_{p1}x_p).$$

la moyenne conditionnelle de Y par rapport à x_2 est égale à

$$\bar{Y_2} = \frac{1}{n_{2.}}(n_{21}y_1 + n_{22}y_2 + \cdots + n_{2j}y_j + \cdots + n_{2q}x_q).$$

d) **Variance et écart type conditionnels**

On appelle **variance conditionnelle** de X par rapport à $Y = y_m$ le réel positif noté

$$V_m\bar{(X)} = \frac{1}{n_{.m}} \sum_{i=1}^{p} n_{im}(x_i - \bar{X_m})^2 = \frac{1}{n_{.m}} \sum_{i=1}^{p} n_{im} x_i^2 - \bar{X_m}^2,$$

La **variance conditionnelle** de Y par rapport à $X = x_{ln}$ le réel positif noté

$$\bar{V_{ln}}(X) = \frac{1}{n_{ln.}} \sum_{j=1}^{q} n_{lnj}(y_j - \bar{Y_{ln}})^2 = \frac{1}{n_{ln.}} \sum_{j=1}^{q} n_{lnj} j_j^2 - \bar{Y_{ln}}^2,$$

L'écart type conditionnel de X est la racine carré de la variance conditionnelle de X par rapport à $Y = y_m$

$$\sigma_m(X) = \sqrt{V_m(X)}.$$

De même **L'écart type conditionnel** de Y par rapport à $X = x_{ln}$ est la racine carré de la variance conditionnelle de Y par rapport à $X = x_{ln}$

$$\sigma_{ln}(Y) = \sqrt{V_{ln}(Y)}.$$

Chapitre 3

Utilisation d' une calculatrice en statistique dans le cas d'une série double

3.1 Introduction

Dans le cadre de ce travail, la calculatrice utilisée est de marque charp El-531 VMB. Son utilisation se limite au cas d'une série double (X, Y) à données individuelles oú les variables x_i du caractère X et y_i du caractère Y peuvent être consignées dans un tableau linéaire de la forme

X	x_1	x_2	$\ldots$	x_i	$\ldots$	x_n
Y	y_1	y_2	$\ldots$	y_i	$\ldots$	y_n

Cette calculatrice permet d'obtenir entre autres :

a- L'effectif total N (noté n dans la calculatrice),

b- les moyennes $\overline{x}$ et $\overline{y}$ des deux séries.

c- la somme des valeurs $\sum_{i=1}^{n} x_i$ et $\sum_{i=1}^{n} y_i$,

d- la somme des carrés des valeurs et la somme des produits des composantes des couples de valeurs :

$$\sum_{i=1}^{n} x_i^2;\ \sum_{i=1}^{n} y_i^2;\ \sum_{i=1}^{n} x_i y_i,$$

e- les écarts types $\sigma(X)$ et $\sigma(X)$ et les variances $V(X)$ et $V(Y)$ des deux caractères,

f- le coefficient directeur b et l'ordonnée à l'origine a de la droite de régression D de Y en X

$$D : y = a + bx,$$

g- le coefficient de corrélation linéaire r du couple (X, Y).

3.2 Mise en marche de la calculatrice

⊛ Pour allumer la calculatrice, il suffit d'appuyer sur la touche [on/c]

⊛ Pour mettre la calculatrice en mode statistique double, il faut appuyer successivement sur les touches [2nd F] et [MODE] puis sur la touche [2]. Il s'affiche alors sur l'écran **Stat xy**

3.3 Entrée des données

Chaque donnée est un couple (x_i, y_i), $1 \leq i \leq n$.
Pour faire entrer des données (x_1, y_1), (x_2, y_2), (x_3, y_3), ..., (x_n, y_n), on appuie successivement sur les touches :

x_1	STO	y_1	M_+
x_2	STO	y_2	M_+
x_3	STO	y_3	M_+
⋮	⋮	⋮	⋮
x_n	STO	y_n	M_+

3.4 Obtention des paramètres cités dans l'introduction

a). Après l'entrée de la dernière donnée, la calculatrice affiche automatiquement l'effectif total n. On peut retrouver ultérieurement cet effectif total n en appuyant sur les touches [RCL] et n (touche [0]).

b). On obtient la moyenne $\bar{x}$ de la première série en appuyant sur les touches [RCL] et $\bar{x}$ (touche [4]) et la moyenne $\bar{y}$ de la deuxième série en appuyant sur les touches [RCL] et $\bar{y}$ (touche [7]).

c). La somme $\sum_{i=1}^{n} x_i$ s'obtient en appuyant sur les touches [RCL] et $\sum x$ (touche [·]) et la somme $\sum_{i=1}^{n} y_i$ en appuyant sur les touches [RCL] et $\sum y$ (touche [2]).
N.B : $\sum_{i=1}^{n} x_i = n\bar{x}$; $\sum_{i=1}^{n} y_i = n\bar{y}$.

d). Les sommes $\sum_{i=1}^{n} x_i^2$; $\sum_{i=1}^{n} y_i^2$ et $\sum_{i=1}^{n} x_i y_i$ s'obtiennent en appuyant respectivement sur [RCL] et $\sum x^2$ (touche [+/−]) ; sur [RCL] et $\sum_{i=1}^{n} y^2$ (touche [3]) et sur [RCL] et $\sum xy$ (touche [1]).
N.B : On peut calculer les variances $V(X)$ et $V(Y)$ à partir de $\sum_{i=1}^{n} x_i^2$ et $\sum_{i=1}^{n} y_i^2$

et la covariance à partir de $\sum_{i=1}^{n} x_i y_i$.

e). L'écart-type $\sigma(X)$ s'obtient en appuyant sur $\boxed{\text{RCL}}$ et $\boxed{\sigma x}$ $\Big($touche $\boxed{6}$ $\Big)$ et l'écart-type $\sigma(Y)$ en appuyant sur $\boxed{\text{RCL}}$ et $\boxed{\sigma y}$ $\Big($touche $\boxed{9}$ $\Big)$.

N.B : On peut calculer les variances en élevant les écart-types au carré.

f). Soit $\Delta :\ y = a + bx$ la droite de régression de Y en X. Le coefficient directeur b de Δ s'obtient en appuyant sur $\boxed{\text{RCL}}$ et $\boxed{b}$ $\Big($touche $\boxed{\supset}$ $\Big)$ et l'ordonnée à l'origine a en appuyant sur $\boxed{\text{RCL}}$ et $\boxed{\text{a}}$ $\Big($touche $\boxed{\subset}$ $\Big)$.

g). Le coefficient de corrélation linéaire r s'obtient en appuyant sur $\boxed{\text{RCL}}$ et $\boxed{r}$ $\Big($touche $\boxed{\div}$ $\Big)$.

Remarque : On peut remarquer que pour obtenir un paramètre quelconque, il suffit d'abord d'appuyer sur $\boxed{\text{RCL}}$ puis sur la touche appropriée.

3.5 Suppression des données

* Pour effacer une donnée non encore mémorisée, il suffit d'appuyer sur $\boxed{\text{on/c}}$ (les données mémorisées antérieurement restent toujours).
* Pour effacer la dernière donnée mémorisée (après appui sur la touche $\boxed{M_+}$), on appuie sur $\boxed{\text{2nd F}}$ et sur $\boxed{M_-}$ (touche $\boxed{M_+}$)
* Tant que les données ne sont pas effacées, elles demeurent toujours mémorisées dans la calculatrice même si celle-ci est éteinte. Pour effacer toutes les données mémorisées, il suffit d'appuyer sur $\boxed{\text{2nd F}}$ et sur $\boxed{\text{CA}}$ (touche $\boxed{\text{DEL}}$) ou bien sur $\boxed{\text{2nd F}}$ et sur $\boxed{\text{MODE}}$ puis sur $\boxed{0}$ (ou $\boxed{1}$ ou $\boxed{2}$).

3.6 Retour en mode normal

* Pour quitter le mode statistique et revenir en mode normal, il suffit d'appuyer sur $\boxed{\text{2nd F}}$ et sur $\boxed{\text{MODE}}$ puis sur la touche $\boxed{0}$.
* On éteint la calculatrice en appuyant sur les touches $\boxed{\text{2nd F}}$ et $\boxed{\text{OFF}}$.