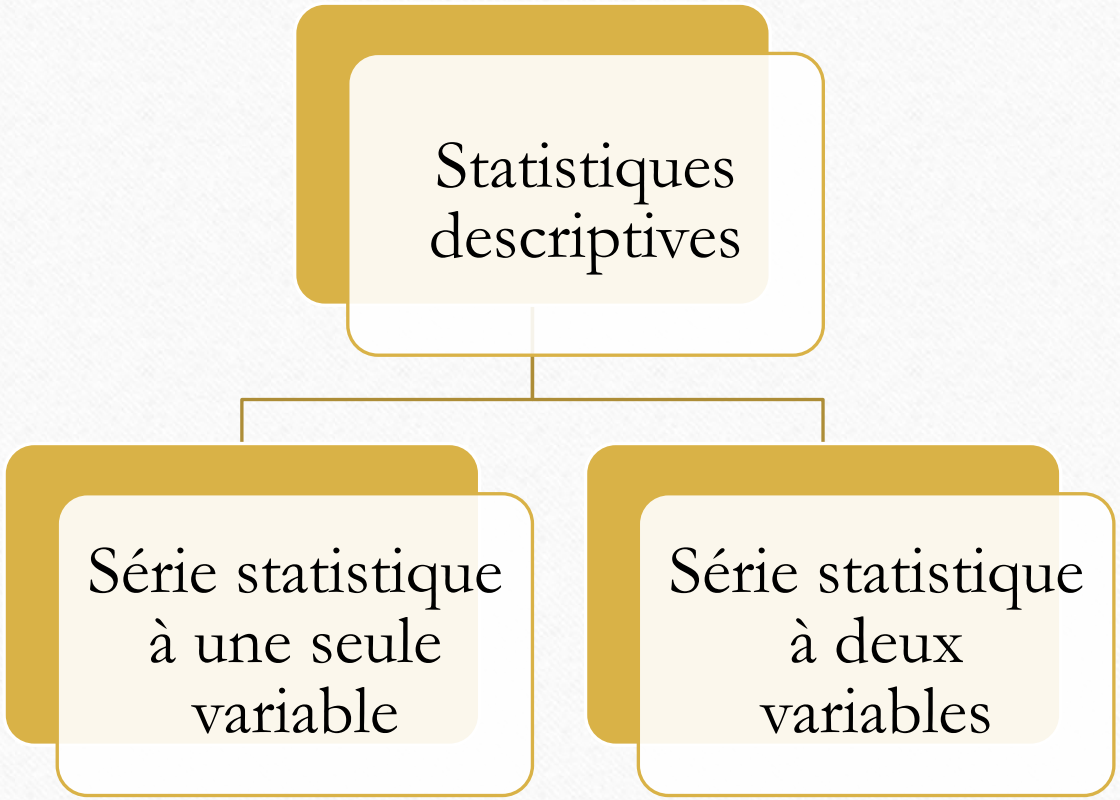
Statistiques descriptives

Résumé du cours

Présenté par: Mme BENAZZOU Salma

Le plan


Statistiques descriptives
Série statistique à une seule variable
Série statistique à deux variables

Vocabulaire statistique

- ✓ **Population ou univers** : L'ensemble de référence, l'ensemble des unités étudiées ou observées.
- ✓ **Individu ou unité statistique :** Tout élément de la population cible
- ✓ **Caractère ou variable statistique:** C'est l'aspect particulier auquel on s'intéresse, la statistique se réfère à deux grandes catégories de caractères :

- Qualitatif : couleur des yeux, nationalité,….
- Quantitatif : nombre d'étudiant, nombre de pièces fabriquées,…….. (discret ou continue)

- ✓ **Modalité (x_i) :** les différentes rubriques associées à un caractère, le nombre de modalité est généralement noté k .Exemple : pour le caractère état matrimonial, on pourra avoir 4 modalités (k=4) qui sont : célibataire, marié, divorcé, et veuf.

- ✓ **Effectif ou fréquence absolue (n_i)** : C'est le nombre de fois que la modalité xi est observée $\sum_{i=1}^{k} ni = N$
- ✓ **Effectif relative (fi) :** C'est le pourcentage des individus ayant la modalité i dans la population étudiée on a : fi=$\frac{ni}{N}$ et $\sum_{i=1}^{k} fi = 1$

Exemple 1

Les 33 élèves d'une classe ont obtenu les notes suivantes lors d'un devoir :

Note (xi)	5	8	10	11	12	14	15	18	20
Effectif (ni)	1	4	3	8	7	3	4	2	1

- ✓ **Population ou univers** : Les élèves d'une classe
- ✓ **Individu ou unité statistique :** Chaque étudiant représente un individu de la série statistique
- ✓ **Caractère ou variable statistique:** La note , son type est quantitatif discret
- ✓ **Modalité (x_i) :** Ici on a 9 modalités
- ✓ **Effectif relative (fi) :** C'est le pourcentage des individus ayant la modalité i dans la population étudiée on a : fi=$\frac{ni}{N}$ et $\sum_{i=1}^{k} fi = 1$

Exemple 2

La répartition de 6920 supermarchés en France suivant la surface en m^2 :

Surface (xi)	**[400 ;800[**	**[800 ;1000[**	**[1000 ;2500]**
Effectif (ni)	2613	928	3379

- ✓ **Population ou univers** : Les supermarchés en France
- ✓ **Individu ou unité statistique :** Chaque supermarché représente un individu de la série statistique
- ✓ **Caractère ou variable statistique:** La surface , son type est quantitatif continue
- ✓ **Modalité (x_i) :** Ici on a 3 modalités
- ✓ **Effectif relative (fi) :** C'est le pourcentage des individus ayant la modalité i dans la population étudiée on a : fi=$\frac{ni}{N}$ et $\sum_{i=1}^{k} fi = 1$

Exemple 3

La répartition de 1000 candidats convoqués pour participer au test d'admissibilité à la formation en management pour l'accession à l'ENCG d'Agadir, selon la série de bac se présente comme suit :

La série de Bac (xi)	Le nombre des candidats (ni)
Science économique	250
Science mathématique	200
Science expérimentale	400
T.G.A	50
T.G.C	100

:

- ✓ **Population ou univers** : Les candidats convoqués pour participer au test d'admissibilité à la formation en management
- ✓ **Individu ou unité statistique :** Chaque candidat représente un individu de la série statistique
- ✓ **Caractère ou variable statistique:** La série de bac , son type est qualitatif
- ✓ **Modalité (x_i) :** Ici on a 5 modalités
- ✓ **Effectif relative (fi) :** C'est le pourcentage des individus ayant la modalité i dans la population étudiée on a : $fi=\frac{ni}{N}$ et $\sum_{i=1}^{k} fi = 1$

Série statistique à une seule variable

Série statistique à deux variables

La représentation graphique

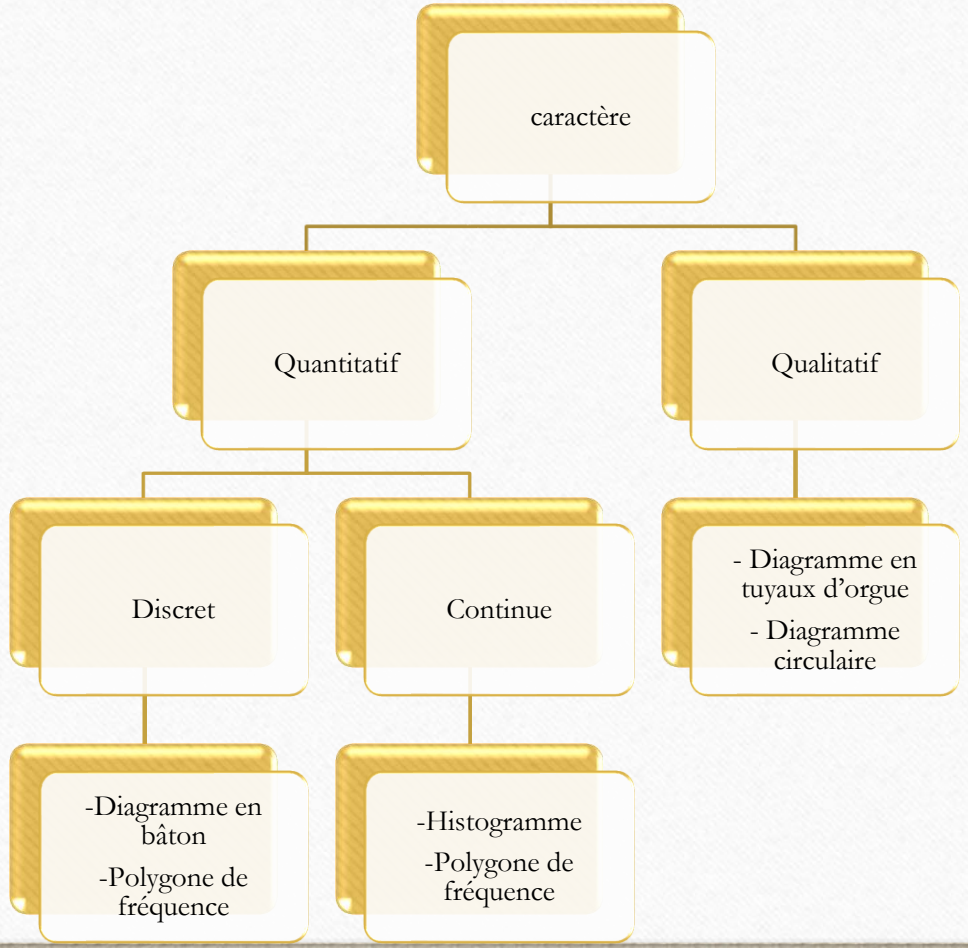
Les caractéristiques de position

Les caractéristiques de dispersion

Série statistique à une seule variable

Série statistique à deux variables

I- La représentation graphique



Série statistique à une seule variable

Série statistique à deux variables

Caractère qualitatif : diagramme en tuyaux d'orgue et diagramme circulaire

Série statistique

xi	ni
x1	n1
x2	n2
x3	n3
x4	n4

Diagramme en tuyaux d'orgue

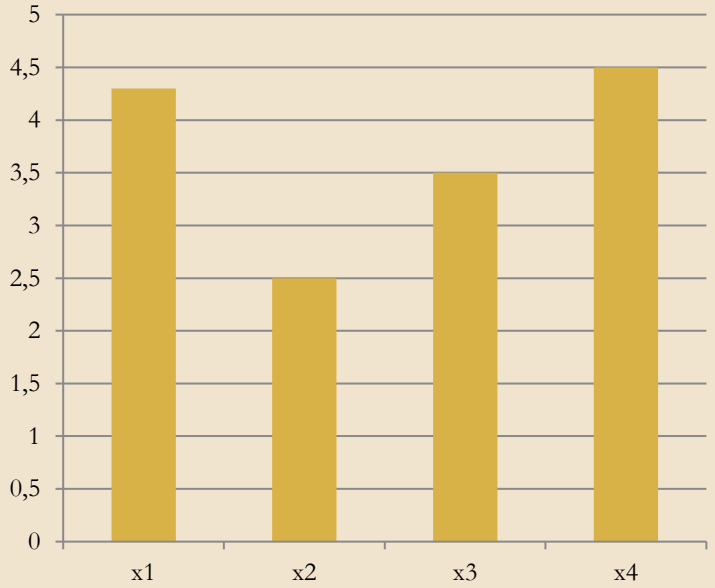
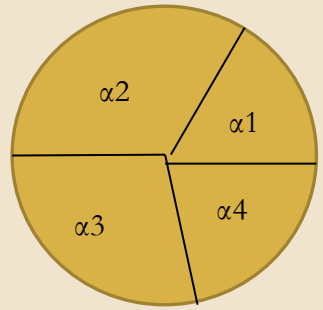


Diagramme circulaire



$$\begin{cases} N \to 360° \\ n_i \to \alpha_i \end{cases} \Rightarrow \alpha_i = 360 \times \frac{n_i}{N} = 360 \times f_i$$

Caractère qualitatif : diagramme en tuyaux d'orgue et diagramme circulaire

Exemple

La répartition des candidats convoqués pour participer au test d'admissibilité à la formation en management pour l'accession à l'ENCG d'Agadir, selon la série de bac se présente comme suit :

série de Bac (xi)	nbr des candidats (ni)	αi=ni/N*360
Sc. eco	250	90°
Sc. math	200	72°
Sc. exp	400	144°
T.G.A	50	18°
T.G.C	100	36°

Diagramme en tuyaux d'orgue

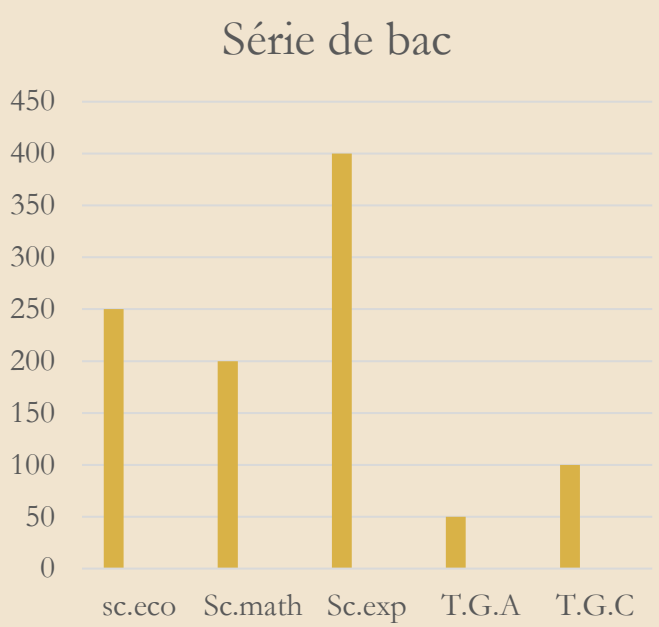
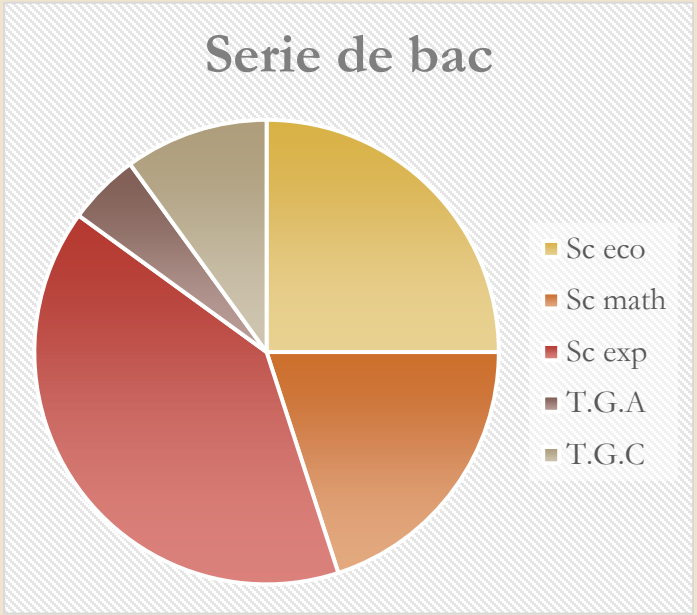


Diagramme circulaire



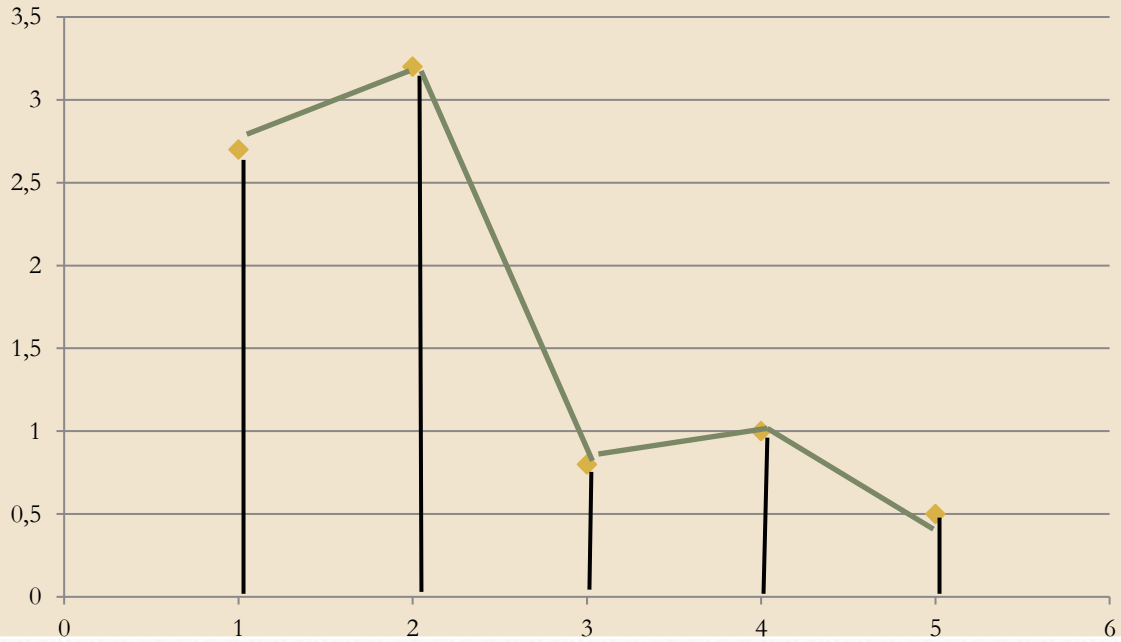
Série statistique à une seule variable

Série statistique à deux variables

Caractère quantitatif discret : diagramme en bâton et polygone de fréquence

Exemple	Diagramme en bâton et polygone de fréquence

xi	ni
1	2,7
2	3,2
3	0,8
4	1
5	0,5



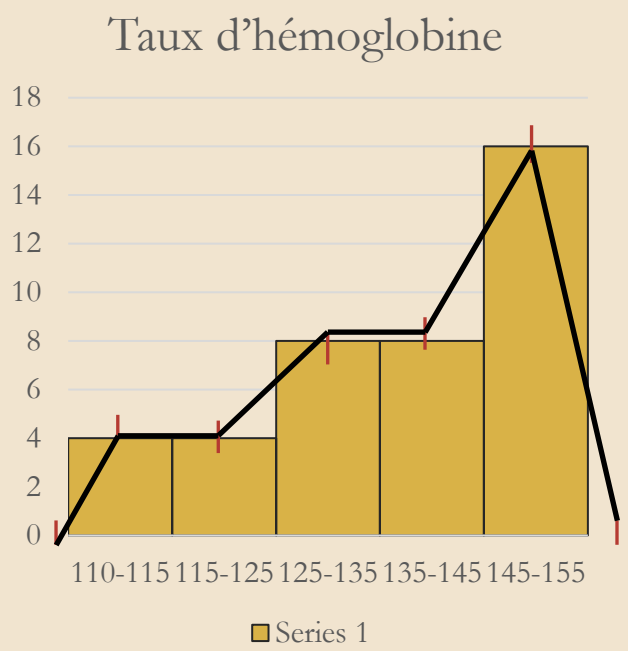
Caractère quantitatif continue : Histogramme et polygone de fréquence

1 er cas : les amplitudes sont égales

On a relevé le taux d'hémoglobine chez 40 personnes adultes présumées en bonne santé. On obtient le tableau suivant:

Le taux: xi	Nbr de personnes : ni	ai=BS-BI
105-115	4	10
115-125	4	10
125-135	8	10
135-145	8	10
145-155	16	10

Histogramme



Polygone de fréquence

Le polygone de fréquence joint les points : **(c_i, n_i)**

Le polygone de fréquence pour une variable continue, doit être toujours fermé avec l'axe des abscisses en prenant deux points aux deux extrémités de l'histogramme, ces deux points sont : (Borne inf_1 **-a/2**, 0) et (Borne supk+a/2**,** , 0)

-Le centre d'un intervalle est ci=(BS+BI)/2

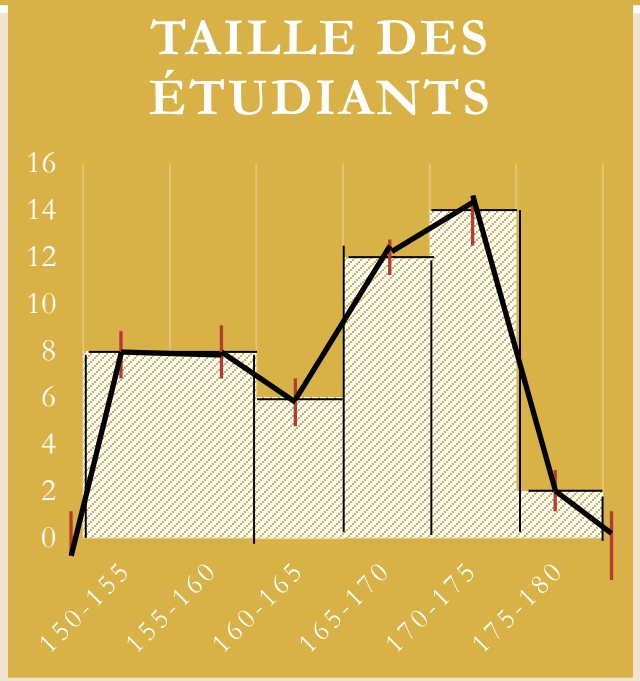
Caractère quantitatif continue : Histogramme et polygone de fréquence

2 Emme cas : les amplitudes changent

Le tableau suivant représente la distribution de 50 étudiants en fonction de leurs tailles.

Taille en cm	Nbr d'étudiant	ai= BSi-BIi	nicorr=ni*aN/ai Ici aN=5
150-160	16	10 > 5	8
160-165	6	5 ≤ 5	6
165-170	12	5 ≤ 5	12
170-175	14	5 ≤ 5	14
175-180	2	5 ≤ 5	2

Histogramme



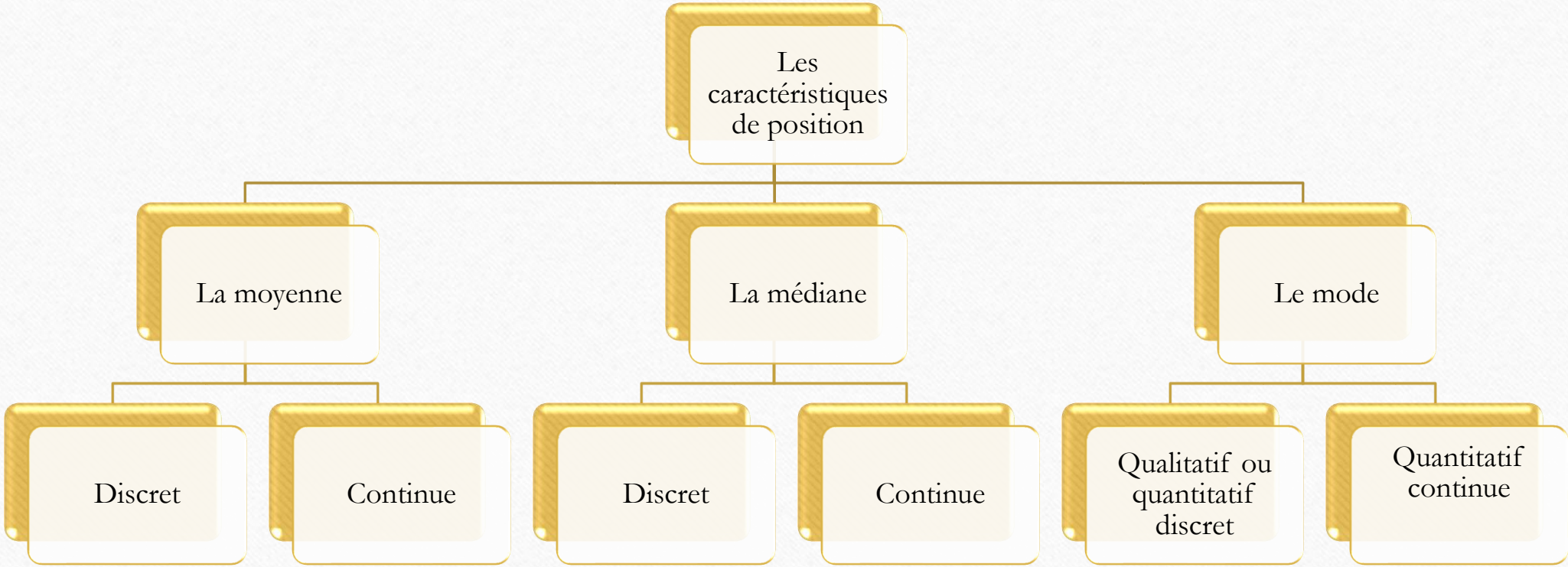
Polygone de fréquence

Le polygone de fréquence joint les points : (c_i, n_{icor}) pour les classes ayant une amplitude $a_i \leq a_N$
$(\text{Borne inf}_i + aN/2 , n_{icor})$ et $(\text{Borne sup}_i - aN/2 , n_{icor})$ pour toutes les classes ayant une amplitude $a_i > a_N$
Le polygone de fréquence pour une variable continue, doit être toujours fermé avec l'axe des abscisses en prenant deux points aux deux extrémités de l'histogramme, ces deux points sont : $(\text{Borne inf}_1 - aN/2 , 0)$ et $(\text{Borne sup}_k + aN/2 , 0)$

Série statistique à une seule variable

Série statistique à deux variables

II- Les caractéristiques de position et interprétations



Série statistique à une seule variable

Série statistique à deux variables

I- La moyenne : C'est la valeur que devrait avoir chaque individu de façon équitable

Caractère quantitatif discret

La moyenne est $\bar{X} = \frac{1}{N}\sum x_i n_i$

Exemple:

Les 33 élèves d'une classe ont obtenu les notes suivantes lors d'un devoir :

Note (xi)	5	8	10	11	12	14	15	18	20
Effectif (ni)	1	4	3	8	7	3	4	2	1

On a $\bar{X} = \frac{1}{N}\sum x_i n_i = \frac{1}{33}.397 = 12{,}03$

Interprétation: Si tous les étudiants ont eu la même note , ca sera 12,03

Caractère quantitatif continue

La moyenne est $\bar{X} = \frac{1}{N}\sum c_i n_i$

Exemple:

Taille en cm	Nbr d'étudiant	Ci= (Bsi+Bii)/2
150-160	16	155
160-165	6	162,5
165-170	12	167,5
170-175	14	172,5
175-180	2	177,5

On a $\bar{X} = \frac{1}{N}\sum c_i n_i = \frac{1}{50}.8235 = 164{,}7\ cm$

Interprétation: Si tous les étudiants avait la même taille , ca sera 164,7 cm

II- La médiane : Partage la population en 2

Caractère quantitatif discret

Avant de déterminer la valeur de la médiane, il faut classer la série statistique par ordre croissant. Deux cas de figure peuvent se présenter :

- N=2p — $Me=\frac{x_p+x_{p+1}}{2}$
- N=2p+1 — $Me=x_{p+1}$

Note (xi)	5	8	10	11	12	14	15	18	20
Effectif (ni)	1	4	3	8	7	3	4	2	1
Effectif cumulé	1	5	8	16	23	26	30	32	33

N=33=2*16+1 alors Me=Me=x_{16+1} Donc =x_{17}=12

5 8 8 8 8 10 10 10 11 11 …. 11 12 …..12 14 14 14 …..

Caractère quantitatif continue

Etape 1 : La détermination de la classe médiane
La classe médiane est la première classe dont l'effectif cumulé croissant est supérieur ou égale à N/2
Etape 2 : La détermination de la médiane
Soit i l'indice de la classe médiane, on a alors :

$Me= BIi+(BSi-BIi)\ \frac{\frac{N}{2}-N_{i-1}}{N_i-N_{i-1}}$

Taille en cm	Nbr d'étudiant	ai	N
150-160	16	10	16
160-165	6	5	22
165-170	12	5	34
170-175	14	5	48
175-180	2	5	50

On a N=50 donc
N/2=25
Donc la classe médiane est :165-170
Alors
$Me= 165+5\ \frac{25-22}{34-22}$
Me=166,25 cm

I- Le mode : La modalité la plus fréquente

Caractère qualitatif ou quantitatif discret

La modalité xi dont l'effectif ni est le plus grand est le mode Mo=xi

Note (xi)	5	8	10	11	12	14	15	18	20
Effectif (ni)	1	4	3	8	7	3	4	2	1

L'effectif le plus grand est 7 , donc le mode Mo=12

La série de Bac (xi)	Le nombre des candidats (ni)
Science économique	250
Science mathématique	200
Science expérimentale	250
T.G.A	50
T.G.C	100

Ici on a 2 modes : sc économique et sc expérimentale . C 'est une série statistique bimodale

Caractère quantitatif continue

La classe modale i est celle dont l'effectif ni est le plus grand et on a : Mo= BIi+ai $\frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$

Remarque : Si les classes ont des amplitudes différentes, on travaillera avec les n_{icor} au lieu des n_i

Taille en cm	Nbr d'étudiant	ai= BSi-BIi	nicorr=ni*aN/ai Ici aN=5
150-160	16	10 > 5	8
160-165	6	5 ≤ 5	6
165-170	12	5 ≤ 5	12
170-175	14	5 ≤ 5	14
175-180	2	5 ≤ 5	2

La classe modale est 170-175

Donc :

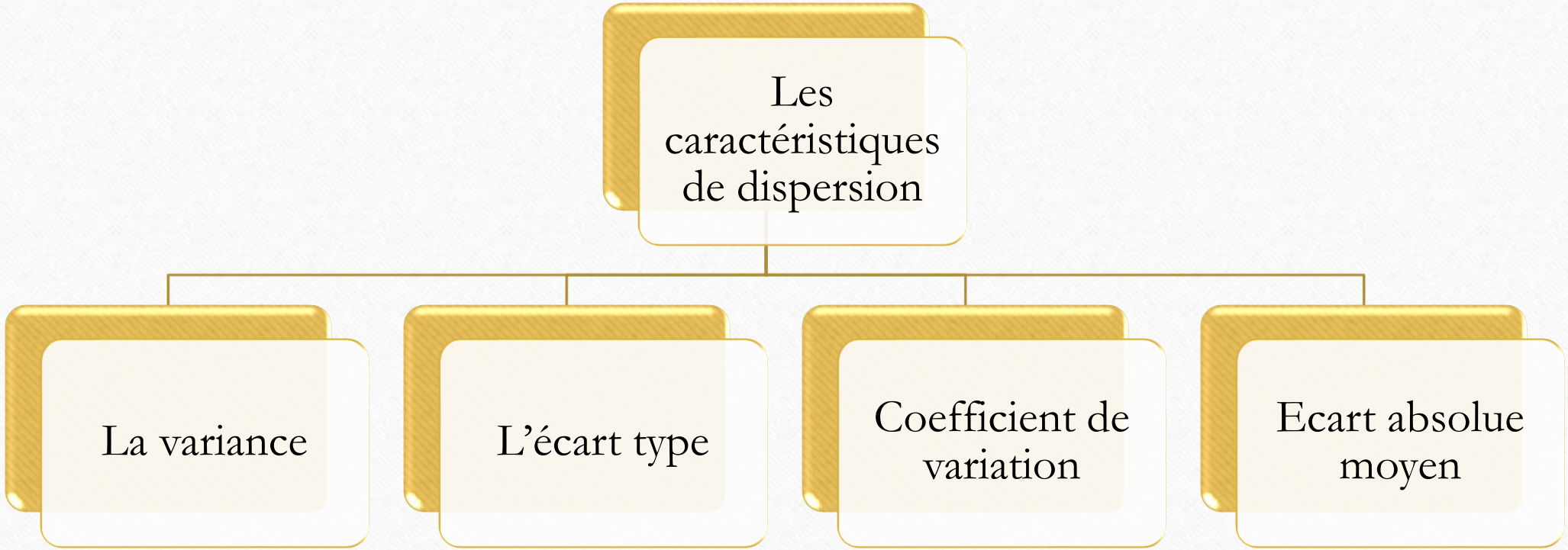
Mo=170+5 $\frac{14-12}{(14-12)+(14-12)}$

Mo= 170,71 cm

Série statistique à une seule variable

Série statistique à deux variables

II- Les caractéristiques de dispersion et interprétations



Série statistique à une seule variable

Série statistique à deux variables

Variance

- Var (X)= $\frac{1}{N}\sum n_i x_i^2 - (\bar{X})^2$
- Var (X)=$\frac{1}{N}\sum n_i c_i^2 - (\bar{X})^2$

Ecart type

- On a $\sigma_x = \sqrt{\text{var}(X)}$

Coefficient de variation

- C.V= $\frac{\sigma_x}{\bar{X}} \times 100$

Ecart absolue moyen

- On a : $E.A.M_{(\bar{X})} = \frac{1}{N}\sum_{i=1}^{k} n_i |c_i - \bar{X}|$

$$E.A.M_{(Me)} = \frac{1}{N}\sum_{i=1}^{k} n_i |c_i - Me|$$

$$E.A.M_{(Mo)} = \frac{1}{N}\sum_{i=1}^{k} n_i |c_i - Mo|$$

Série statistique à une seule variable

Série statistique à deux variables

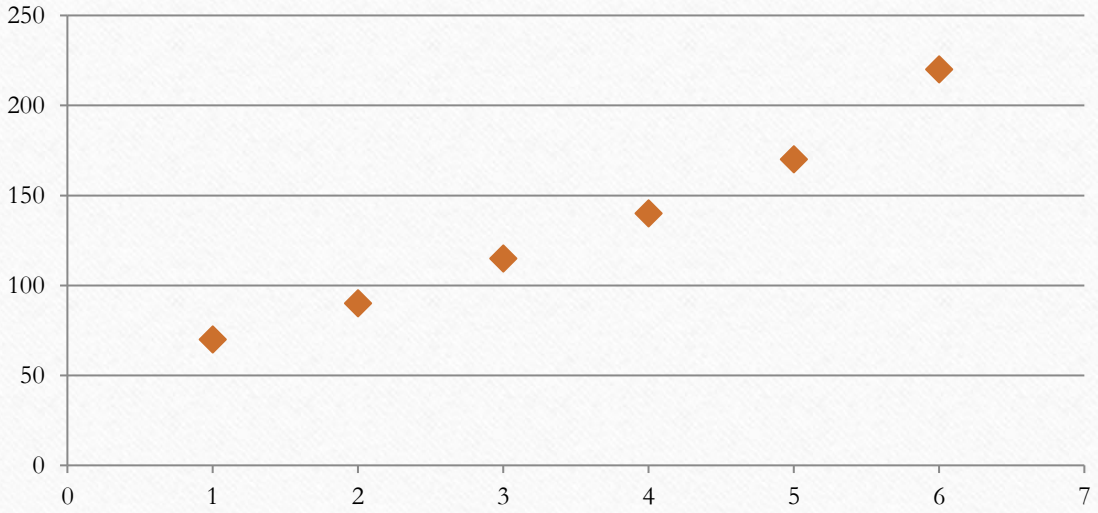
L'ajustement linéaire

- Le nuage de point
- Le modèle de Mayer
- Le modèle des moindre carrées

Le nuage de point:

Le problème qui se pose dans les séries statistiques à deux variables est principalement celui du lien qui existe ou non entre chacune des variables.
Exemple : Le tableau suivant donne l'évolution du nombre d'adhérents d'un club du rugby de 2001 à 2006.

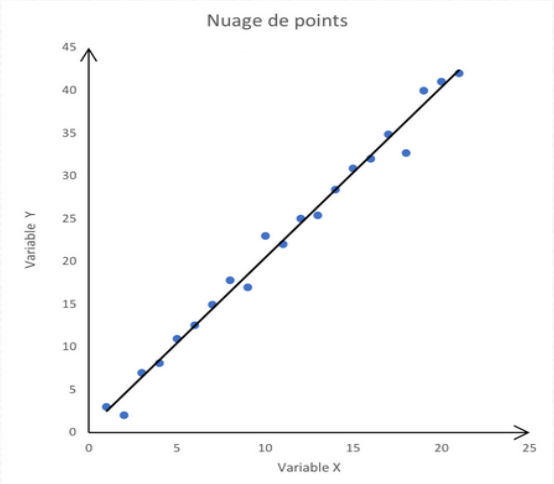
Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

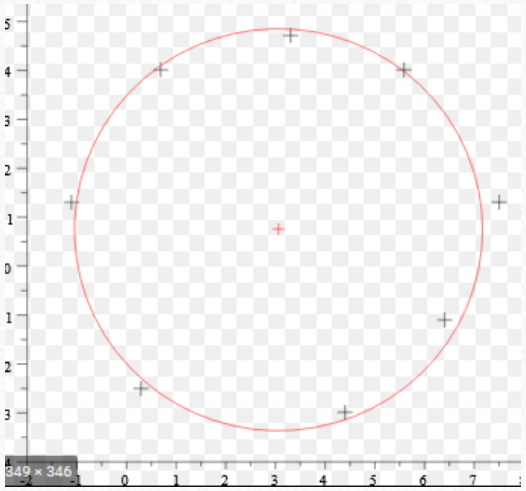


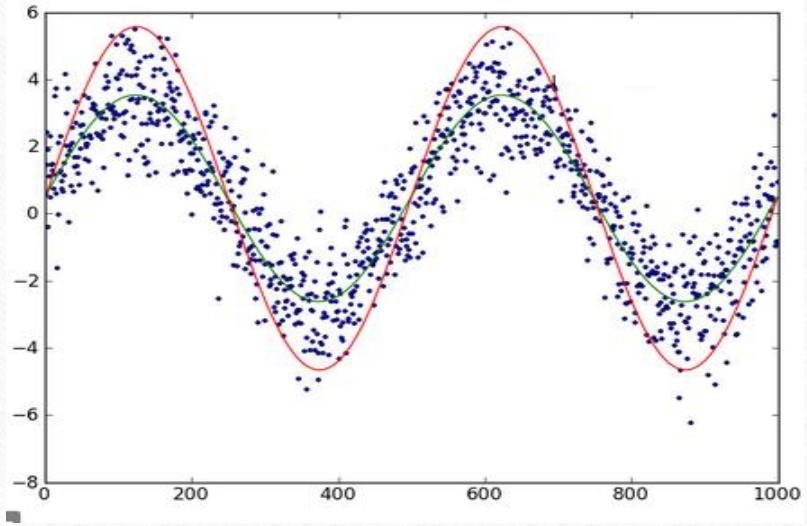
Série statistique à une seule variable

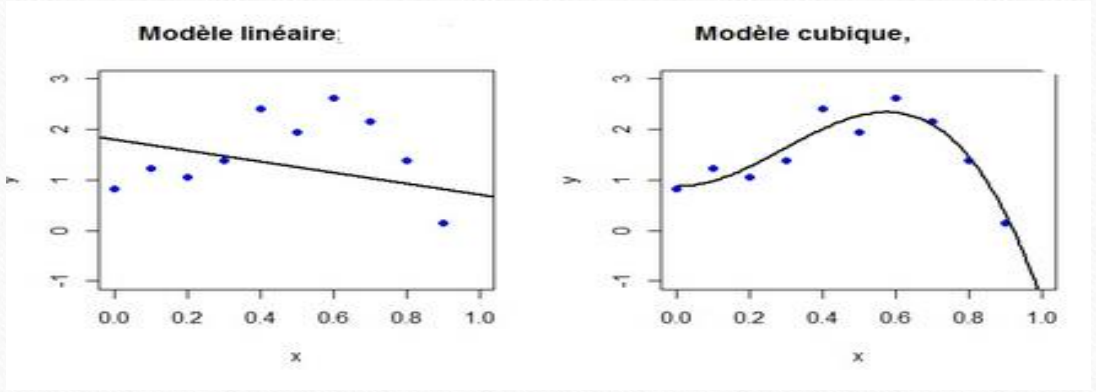
Série statistique à deux variables

La modélisation









Le point moyen:

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

Soit une série statistique à deux variables X et Y dont les valeurs sont les couples (xi, yi)
On appelle point moyen de la série, le point G de coordonnées :

$$x_G = \frac{x_1 + x_2 + ... + x_n}{n}$$

Et

$$y_G = \frac{y_1 + y_2 + ... + y_n}{n}$$

Ici Le point moyen et $G(\bar{x}, \bar{y})$

$$\bar{x} = \frac{1+2+3+4+5+6}{6} = 3,5$$

$$\bar{y} = \frac{70+90+115+140+170+220}{6} = 134,16$$

La méthode de Mayer: Y=aX+b

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

Consiste à déterminer la droite passant par 2 points moyen de nuage de points (le nuage est partagé suivant **les valeurs croissantes des xi** en 2 nuages d'égale importance).

G1 : des années allant de 2001 à 2003

$\overline{x_{G1}}=\frac{1+2+3}{3}=2$

$\overline{y_{G1}}=\frac{70+90+115}{3}=91,66$

G2 : des années allant de 2004 à 2006

$\overline{x_{G2}}=\frac{4+5+6}{3}=5$

$\overline{y_{G2}}=\frac{140+170+220}{3}=176,66$

G1$\in$ (D)donc $\overline{y_{G1}}=a\overline{x_{G1}}+b$

G2$\in$ (D)donc $\overline{y_{G2}}=a\overline{x_{G2}}+b$

Donc

91,66=2a+b

176,66=5a+b

Donc a = 28,33 et b = 35

Donc la droite de Mayer est Y=28,33 X + 35

La méthode de Mayer: Prévisions

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

Donc la droite de Mayer est Y=28,33 X + 35
Combien d'adhérents s'inscrirai t- il en 2024 ?
Pour X=24 on a Y = 28,33*24+35 = 714,92
En quel année s'inscrira t il 1000 adhérents

Pour Y=1000 on a X=(1000-35)/28,33 = 34

La méthode de moindre carrées: Droite de régression de Y en X est Y=aX+b

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

Le coefficient de corrélation linéaire

•Le coefficient de corrélation linéaire est définit par la relation suivante : $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$

$$\sigma_{xy} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$$
$$\sigma_x = \sqrt{\frac{1}{n}\sum {x_i}^2 - (\bar{x})^2}$$
$$\sigma_y = \sqrt{\frac{1}{n}\sum {y_i}^2 - (\bar{y})^2}$$

On a $\sigma_{xy} = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$ et $(\sigma_x)^2 = \frac{1}{n}\sum {x_i}^2 - (\bar{x})^2$ et $(\sigma_y)^2 = \frac{1}{n}\sum {y_i}^2 - (\bar{y})^2$

Alors $\sigma_{xy} = \frac{1}{6}(1\times 70 + 2\times 90 + 3\times 115 + 4\times 140 + 5\times 170 + 6\times 220) - 3{,}5 \times 134{,}16 = 84{,}6$

$\sigma_x = \sqrt{\frac{1}{n}\sum {x_i}^2 - (\bar{x})^2} = \sqrt{\frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - (3{,}5)^2} = 1{,}7$

$\sigma_y = \sqrt{\frac{1}{n}\sum {y_i}^2 - (\bar{y})^2} = \sqrt{\frac{1}{6}(70^2 + 90^2 + 115^2 + 140^2 + 170^2 + 220^2) - (134{,}16)^2} = 50{,}21$

Alors $r = \frac{84{,}6}{1{,}7\times 50{,}21} = 0{,}99$

La méthode de moindre carrées: Droite de régression de Y en X est Y=aX+b

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

Le coefficient de corrélation linéaire

•Le coefficient de corrélation linéaire est définit par la relation suivante : $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$

•Plus le coefficient est proche de 1 en valeur absolue, meilleur est l'ajustement linéaire
•Lorsque r=1 ou r=-1, la droite de régression passe par tous les points du nuage
•Lorsque la corrélation est forte, le nuage de point peut être approximer par la droite de régression
•Lorsque la corrélation est faible, le nuage de point ne peut pas être ajusté par une droite mais il se peut qu'une autre courbe permette un bon ajustement.

•Dans l'exemple , on a r=0,99 donc |r| est très proche de 1 , alors on peut envisager une relation linéaire entre X et Y

La méthode de moindre carrées: Droite de régression de Y en X est Y=aX+b

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

La droite de régression:

- La droite est Y=aX+b
- On a $a=\frac{\sigma_{xy}}{(\sigma_x)^2}=\frac{84,6}{(1,7)^2}=29,27$
- On a G $(\bar{x},\bar{y}) \in (D)$ alors $\bar{y}=a\,\bar{x}+b$ donc $b=\bar{y}-a\,\bar{x}$
- Alors b= 134,16-29,27*3,5 = 31,71
- Donc Y=29,27 X+31,71

La méthode de moindre carrées: Prévisions

Année	2001	2002	2003	2004	2005	2006
Rang (xi)	1	2	3	4	5	6
Nombre d'adhérents (yi)	70	90	115	140	170	220

La droite de régression:

La droite de régression est : Y=29,27 X+31,71

Combien d'adhérents s'inscrirai t- il en 2030 ?

Pour X=30 on a Y = 29,27*30+31,71 = 909,81

En quel année s'inscrira t il 2000 adhérents

Pour Y=2000 on a X=(2000-31,71)/29,27 = 67,24