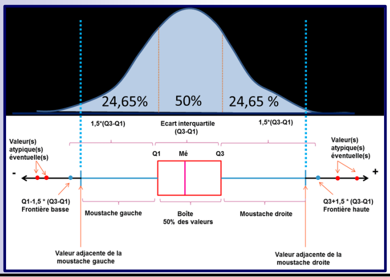
l'Université Mohammed V
École Nationale Supérieure d'Informatique et d'Analyse des Systèmes.

Statistique descriptive pour l'ingénieur (8h)

1ère année, 1er semestre, 2ème période



Prof. S.L. Aouragh
aouragh@hotmail.com
https://www.aouragh.ma
https://www.facebook.com/groups/2054755894753681/
https://www.youtube.com/channel/UCfmH7UvbkdI58EEJy7hASnw

Année universitaire : 2022/2023

1

Bibliographie

Des ouvrages :

- *Statistiques descriptives* de **Bernard Py** (Edition 2007-Economica)
- *Exercices corrigés de statistique descriptive* de **Bernard Py** (Edition 1999-Economica)
- *Statistiques descriptives* de **Bernard Grais** (Edition 2003-Dunod)
- *Méthodes statistiques* de **Bernard Grais** (Edition 2003-Dunod)
- *Statistiques pour l'économie et la gestion* de Anderson, Sweeney, Williams, traduit par Claire Borsenberger (3e édition 2010-De Boeck, éditeur)

Plan

- **Partie 1**: Séries simples *(deux séances)*
  - Terminologie et concepts de base.
  - Tableaux statistiques et représentations graphiques,
  - Paramètres de position (ou de tendance centrale)
  - Paramètres de dispersion
  - Paramètres de forme (asymétrie, aplatissement)
  - Paramètres de concentration.
- **Partie 2**: Les séries doubles *(une séance)*
  - Distribution conjointe, marginale, liaison entre 2 variables…
- **Partie 3**: Les séries chronologiques *(une séance)*
  - Décomposition d'une série chronologique
  - Modélisation d'une série chronologique

Partie 1
Les séries simples

Statistique et statistiques

Définitions:

La Statistique est l'ensemble des méthodes et techniques permettant de **recueillir, traiter et interpréter** un ensemble de données (informations chiffrées) associées à une situation ou un phénomène.

Elle permet d'obtenir de l'information à partir des données, et de prendre les meilleurs décisions.

Les Statistiques est l'ensemble de données ou d'informations relatives à un phénomène ou un processus donné,

Domaine d'application

La statistique est utilisée en plusieurs domaines:

• **Comptabilité, finance**

Les bilans ou comptes de résultats, gestion du capital, trésorerie, opérations avec les banques,

• **Biologie**

L'évolution d'une maladie,

•**Production**

Gestion des stocks ou du matériel, contrôle de la qualité

• **Achats, ventes**

Statistiques des ventes, études de marché.

Étude statistique

L'étude statistique concerne soit:

1. **Une seule variable** : statistique à une dimension, ou statistique **univarié**,
2. **Deux variables** à la fois : statistique à **deux dimensions**,
3. **Plus de deux variables** à la fois: statistique **multidimensionnelle**.

Deux directions en statistique

Statistique descriptive:

Organisation, **présentation** et **analyse** des données en mettant les points importants en évidence, en utilisant des tableaux et des graphes.

Statistique inférentielle:

Elle s'appelle aussi **statistique mathématique**, dont l'objet est de **formuler des lois de comportement** à partir d'observation souvent incomplètes.

Recueil des données statistiques

Pour **recueillir** des informations sur une population statistique, on dispose de deux méthodes :

- **La méthode exhaustive** ou **recensement** où chaque individu de la population est étudié selon le ou les caractères étudiés.

  Exemple: Recensement générale de la population marocaine.

- **La méthode de sondage** ou **échantillonnage** qui consiste à n'examiner qu'une partie de la population, appelée un échantillon.

  Exemple: Choix de 30 étudiants parmi 400 inscrits dans une filière .

Exemple 3 de données statistiques

L'exploitation des bases de données:

Une société possède environ 3 millions de clients. Pour chaque client elle dispose d'environ 30 données: *nom, adresse, date de début, quantité de produit, mode d'achat, ….*

En vue **d'identifier les clientes** qui sont le **plus susceptible d'acheter**, la société doit **exploiter les bases de données** qui vont lui renseigner sur le comportement d'achat des clients.

Par exemple: les plus anciens, la moyenne de la durée de payement minimale.

Vocabulaire 1/3

Population: La population est un ensemble d'individus: personnes, objets ou éléments sur lesquels on veut effectuer l'étude statistique.

Ces individus sont définis par une propriété commune donnée,

La taille d'une population est le nombre d'individus qui la composent.

Individu (ou unité statistique): Un élément de la population.

Echantillon: Sous-ensemble de la population.

Vocabulaire 2/3

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

Modalité:

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

Exemples:

1. Étude de la taille des étudiants
2. Étude du nombre d'enfants dans une famille
3. Étude de la taille des chemises dans un magasin
4. Étude de la couleur des voitures

Vocabulaire 3/3

Types de caractères:

- Variables
  - Qualitative
    - Nominale
    - Ordinale
  - Quantitative
    - Discrète
    - Continue

Exemples:

- Grade d'un fonctionnaire
- La mention du baccalauréat
- Nombre de buts marqués chaque mois
- Le poids d'un nouveau né

Variable qualitative nominale

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

Exemple: Nationalité d'un étranger, état matrimonial,…

Modalité:

Les valeurs prises par la variable s'appellent les **modalités**.

Exemple de modalité:

- Pour la variable: état matrimonial, les modalités sont: *célibataire, marié, veuf, divorcé*.
- Pour la couleur des cheveux : *blanc, brun, noir….*

Variable qualitative ordinale

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

Exemple:

- Un commerçant a fait un recensement des chemises dans son magasin, pour les classer selon la leur taille.
- Un questionnaire de satisfaction demande aux consommateurs d'évaluer une prestation en cochant l'une des six catégories suivantes:
  (a) nulle, (b) médiocre, (c) moyenne, (d) assez bonne, (e) très bonne, (f) excellente.

Variable quantitative discrète

Une variable est **quantitative discrète** si elle ne peut prendre que des **valeurs numériques isolées** d'un intervalle quelconque.

L'ensemble de ses modalités est un ensemble discrète.

Exemple:

Le nombre d'enfants par famille,

Le nombre d'accidents par mois,

Effectifs des étudiants inscrits à l'ENSIAS par ans durant les 10 dernières années.

Variable quantitative continue

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

Exemple:

- Salaire d'un fonctionnaire,
- Âge d'un étudiant,
- Taille ou poids d'un bébé

Tableau de données

Dans le cas d'une variable **qualitative ou quantitative discrète** on utilise souvent le tableau des effectifs et des fréquences suivant:

Modalités de X	x_1	…..	x_i	…..	x_k
Effectifs n_i	n_1	…..	n_i	…..	n_k
Fréquences f_i	f_1	…..	f_i	…..	f_k

n_i le nombre d'individu correspondant à la modalité x_i de la variable X,

$n = \sum_{i=1}^{k} n_i$ désigne l'effectif total,

$f_i = \frac{n_i}{n}$ désigne la fréquence de la valeur x_i , avec: $\sum_{i=1}^{k} f_i = 1$

Cas d'une variable quantitative discrète

Noms	Nombre d'enfants
M. Azim	2
M. Farid	3
Mme Latifi	0
Melle Fatiha	0
M. Ahmed	1
M. Salih	0
M. Berrada	1
Mme Réda	0
Melle Fatiha	2
M. Halim	4
M. Chadi	1
Mme Faouzi	3
M. Ali	2
Melle Loubna	0
M. Fatih	0
M. Said	1
M. Radi	2
Mme Faraj	2

On ne s'intéresse pas au nombre d'enfants de M. Azim ou de M. Farid par exemple, mais à la répartition du caractère « Nombre d'enfants » dans la population des 18 employés.

Nombres d'enfants	Effectifs	fréquence
0	6	33,33%
1	4	22,22%
2	5	27,78%
3	2	11,11%
4	1	5,56%
Total	18	100%

Variable quantitative continue

Afin de simplifier la présentation dans le cas des variables quantitatives continues on regroupe les effectifs proches dans une classe,
Par exemple:
Les valeurs: 175 d'effectif **1**, 176 d'effectif **2** et 177 d'effectif **1** peuvent être regrouper dans la classe [175;180[d'effectif **4**.

Classes des tailles en cm	[155 - 160 [	[160 - 165 [	[165 - 170 [	[170 - 175 [	[175 - 180 [
Effectifs	1	5	21	29	4

On définit l'amplitude $\mathbf{a_i}$ d'une classe $\mathbf{[x_{i-1},x_i[}$ par: $\mathbf{a_i=x_i-x_{i-1}}$

On définit la densité $\mathbf{d_i}$ d'une classe $\mathbf{[x_{i-1},x_i[}$ d'effectif $\mathbf{n_i}$ par: $d_i = \frac{n_i}{a_i}$

Rappel

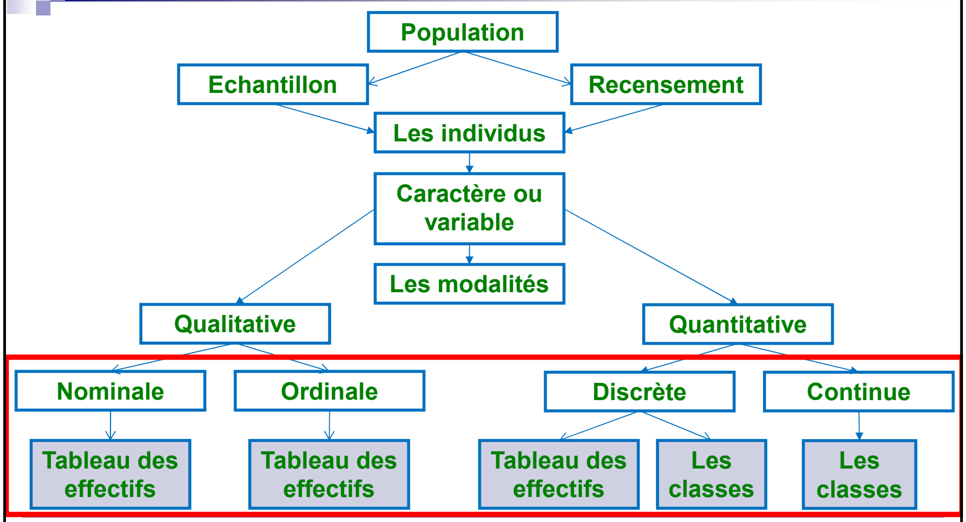


Tableau récapitulatif

Calcul des **effectifs cumulés** croissants et décroissants et des **fréquences cumulées** croissantes et décroissantes

X_i	n_i	Effectifs cumulés croissants	Effectifs cumulés décroissants	f_i	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
x_1	n_1	n_1	n	f_1	f_1	1
x_2	n_2	n_1+n_2	$n-(n_1)$	f_2	f_1+f_2	$1-(f_1)$
…	…	$n_1+n_2+n_3$	$n-(n_1+n_2)$		$f_1+f_2+f_3$	$1-(f_1+f_2)$
x_{k-1}	n_{k-1}	$n_1+\ldots+n_{k-1}$	$n-(n_1+\ldots+n_{k-2})$	f_{k-1}	$f_1+f_2+\ldots+f_{k-1}$	$1-(f_1+\ldots+f_{k-2})$
x_k	n_k	$n_1+\ldots+n_k=n$	n_k	f_k	$f_1+f_2+\ldots f_k=1$	f_k

Tableau récapitulatif

Pour une variable **quantitative continue**:
Les formules de calcul de l'amplitude, centre de classe, densité et effectifs corrigés.

Classe	n_i	Amplitude $a_i = x_i - x_{i-1}$	Centre de classe $c_i = \frac{x_i + x_{i-1}}{2}$	Densité $d_i = \frac{n_i}{a_i}$	Effectifs corrigés $n_{ic} = d_i \times \underset{i}{ppmc}(a_i)$
$[x_0;x_1[$	n_1	a_1	c_1	d_1	n_{1C}
$[x_1;x_2[$	n_2	a_2	c_2	d_2	n_{2C}
…	…	…	…	…	
$[x_{k-1};x_k[$	n_k	a_k	c_k	d_k	n_{kc}

23

Exemple du tableau récapitulatif

Classes	n_i	a_i	ci	fi	ficc	ficd	nicc	nicd	di	nic
[20;25[	9	*5*	*22,5*	*0,06*	*0,06*	*1*	*9*	*140*	*1,8*	*18*
[25;30[	17	*5*	*27,5*	*0,12*	*0,19*	*0,94*	*26*	*131*	*3,4*	*34*
[30;35[	36	*5*	*32,5*	*0,26*	*0,44*	*0,81*	*62*	*114*	*7,2*	*72*
[35;40[	27	*5*	*37,5*	*0,19*	*0,64*	*0,56*	*89*	*78*	*5,4*	*54*
[40;50[	45	*10*	*45*	*0,32*	*0,96*	*0,36*	*134*	*51*	*4,5*	*45*
[50;60[	6	*10*	*55*	*0,04*	*1*	*0,04*	*140*	*6*	*0,6*	*6*

24

Rappel

Compléter le tableau suivant.
Quel est le pourcentage des valeurs au moins égales à 5
Quel est le pourcentage des valeurs moins de 15

Classes	n_i	a_i	c_i	f_i	f_{icc}	f_{icd}	n_{icc}	n_{icd}	d_i	n_{ic}
[0; 5[										
[5;15[										
[15;30[										

Fin de la première séance

Les graphiques

Pour visualiser la **distribution statistique** d'une variable, on utilise des **graphiques**.
Il existe plusieurs types, selon le type de données.

Exemple:
Dans le cas d'une **variable qualitative**, les modalités ne peuvent pas être représentées sur un axe, selon une échelle donnée, car elles ne sont pas numériques.
On utilise surtout dans ce cas des **diagrammes circulaires** ou des **diagramme en tuyaux d'orgues** .

Représentations Graphiques usuelles

On distingue entre plusieurs types de graphiques :

Caractère qualitatif
- Diagramme en tuyaux d'orgues
- Diagramme circulaire / demi-circulaire

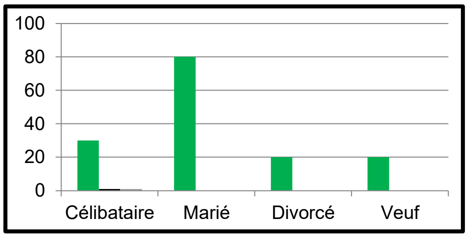
Caractère quantitatif discret
- Diagramme en bâtons
- Courbe cumulative des fréquences

Caractère quantitatif continu
- Histogramme
- Polygone de fréquences
- Courbe cumulative de fréquences

Graphe d'une variable qualitative

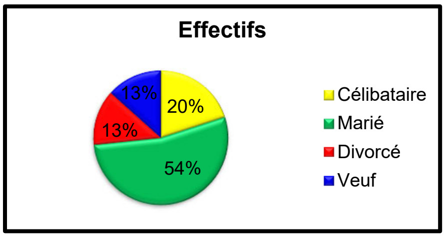
État matrimonial	Célibataire	Marié	Divorcé	Veuf	Total
Effectifs	30	80	20	20	**150**

Diagramme des tuyaux d'orgue



La longueur des tuyaux =n_i

Le graphique à secteurs



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

Graphe d'une variable quantitative discrète

xi	ni
0	4
1	7
2	8
3	8
4	6
5	4
6	3

Nombre d'enfant par famille

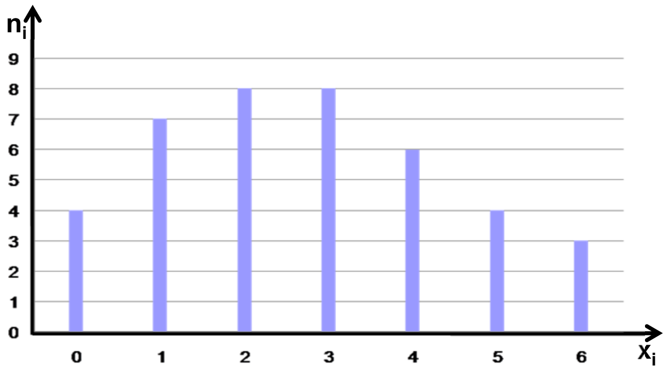
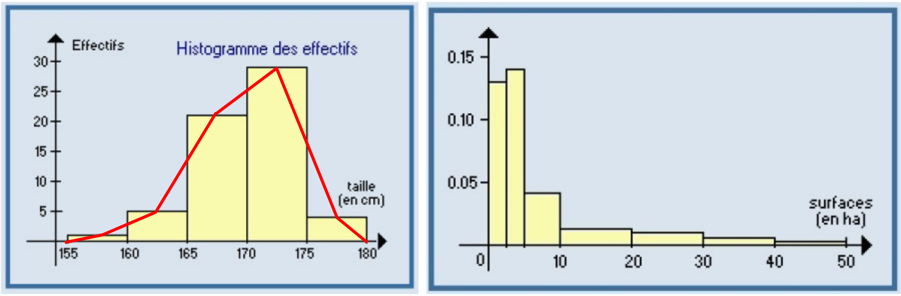


Diagramme des tuyaux d'orgue
La longueur des tuyaux = n_i

Graphe d'une variable quantitative continue



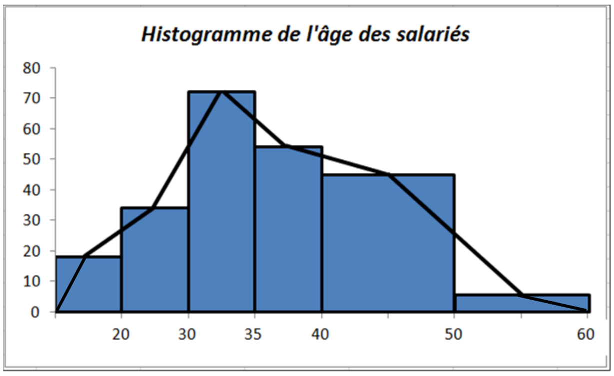
Exemples d'histogrammes.
A gauche : des classes de même amplitude
A droite : des classes de différentes amplitudes.

Graphe d'une variable quantitative

Classes	n_i	a_i	c_i	f_i	d_i	n_{ic}
[20;25[	9	*5*	*22,5*	*0,06*	1,8	*18*
[25;30[	17	*5*	*27,5*	*0,12*	3,4	*34*
[30;35[	36	*5*	*32,5*	*0,26*	7,2	*72*
[35;40[	27	*5*	*37,5*	*0,19*	5,4	*54*
[40;50[	45	*10*	*45*	*0,32*	4,5	*45*
[50;60[	6	*10*	*55*	*0,04*	0,6	*6*

Représenter graphiquement cette variable

Graphe d’une variable quantitative



Polygone des effectifs dans le cas des classes d’amplitudes différents

Types de graphiques

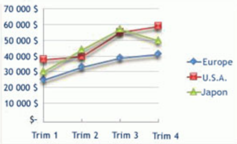
IL existe plusieurs type de graphiques parmi:

- ✓Histogrammes
- ✓Diagramme en bâtons
- ✓Graphiques en courbes
- ✓Graphiques en secteurs
- ✓Graphiques en barres
- ✓Graphiques en aires
- ✓Graphiques en nuages de points (XY)
- ✓Graphiques en surface
- ✓Graphiques en anneaux
- ✓Graphiques en bulles
- ✓Graphique en radar
- ✓…….

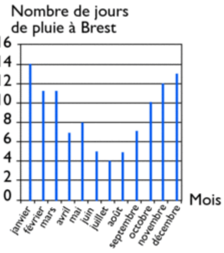
En barres

En barres doubles

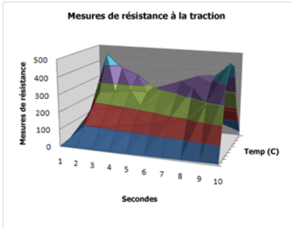
En secteurs



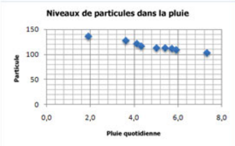
En courbes



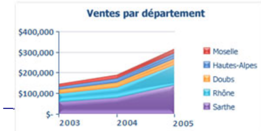
En bâtons



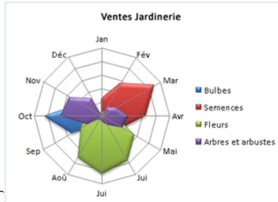
En surface



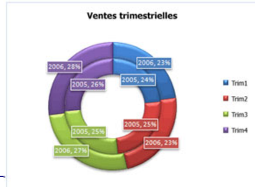
Nuages de points (XY)



En aires



En radar



En anneaux

Notations

X : une variable statistique (caractère)

x_i : (modalités) valeurs prises par la variable statistique X,

n : taille de l'échantillon

n_i : l'effectif de la modalité x_i

$f_i = n_i/n$: la fréquence de la modalité x_i

f_{icc}: la fréquence cumulée des valeurs prises par la variable X qui sont inférieures ou égales à x_i :

n_{icc} : l'effectif cumulé des valeurs prises par la variable X qui sont inférieures ou égales à x_i : $N_i = \sum_{k=1}^{i} n_k$

Le signe Σ désigne somme et Π désigne le produit

Fonction de répartition

La fonction de **répartition** ou fonction **cumulative** est la fonction F(x) qui: à tout réel associe la **proportion** des individus dont le caractère est **strictement inférieur** à x.

- F(x) définie pour toute valeur réel de x,
- F(x) = 0 pour tout x inférieur à la plus petite valeur prise par la variable,
- F(x) = 1 pour tout x supérieur à la plus grande valeur prise par la variable,

Fonction de répartition (variable discrète)

La fonction de répartition est donnée par:

F_i = fréquence de nombre de famille qui ont moins de x_i enfants n_{icc}/n

$$F(x)=\begin{cases} 0 & si \quad x < x_1 \\ F_i & si \quad x_i \leq x < x_{i+1} \\ 1 & si \quad x \geq x_k \end{cases}$$

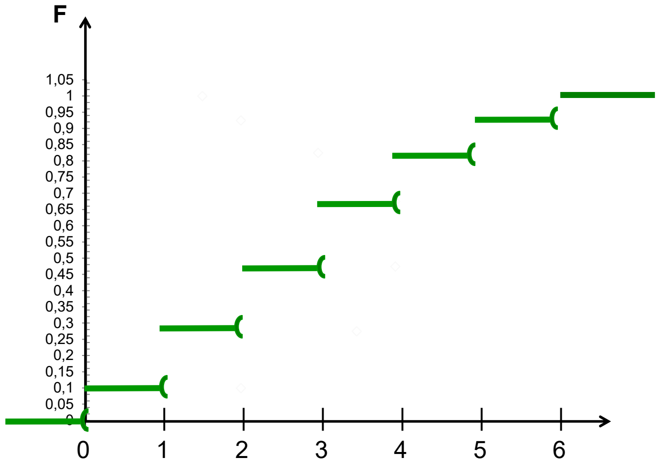
Exemple: le nombre d'enfants par famille,

xi	ni	Ni	Fi
0	4	4	0,1
1	7	11	0,275
2	8	19	0,475
3	8	27	0,675
4	6	33	0,825
5	4	37	0,925
6	3	40	1

F(0)= fréquence de nombre de famille moins de 0 enfants = 0/40
F(1) = fréquence de nombre de famille moins de 1 enfants = 4/40
F(2) = fréquence de nombre de famille moins de 2 enfants = 11/40
F(3) = fréquence de nombre de famille moins de 3 enfants = 19/40
F(4) = fréquence de nombre de famille moins de 3 enfants = 27/40
F(5) = fréquence de nombre de famille moins de 3 enfants = 33/40
F(6) = fréquence de nombre de famille moins de 6 enfants = 37/40
F(7) = fréquence de nombre de famille moins de 7 enfants = 40/40

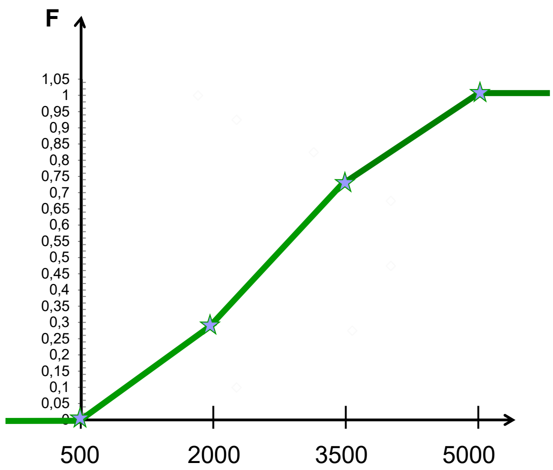
Représentation graphique de la fonction de répartition d’une variable discrète

x_i	n_i	f_i	F_i
0	4	0,1	0,1
1	7	0,175	0,275
2	8	0,2	0,475
3	8	0,2	0,675
4	6	0,15	0,825
5	4	0,1	0,925
6	3	0,075	1



Représentation graphique de la fonction de répartition d’une variable classée

x_i	n_i	f_i	F_i
[500;2000[	30	0,3	0,3
[2000;3500[	45	0,45	0,75
[3500;5000[	25	0,25	1



Les paramètres statistiques

Le but de l'étude statistique est aussi de résumer des données par des paramètres ou synthétiseurs.

Il existe 3 types de paramètres :

- **Paramètres de position** (ou de tendance centrale)
- **Paramètres de dispersion**
- **Paramètres de forme** (asymétrie, aplatissement)
- **Paramètres de concentration.**

Les paramètres de position

Le mode

Le mode (**Mo**) d'une série statistique est la modalité de la variable correspondant à l'effectif le plus élevé. Une série peut avoir plusieurs modes.

La médiane

La médiane (**Me**) d'une série est la valeur qui partage cette série, préalablement classée, en deux séries aux effectifs égaux.

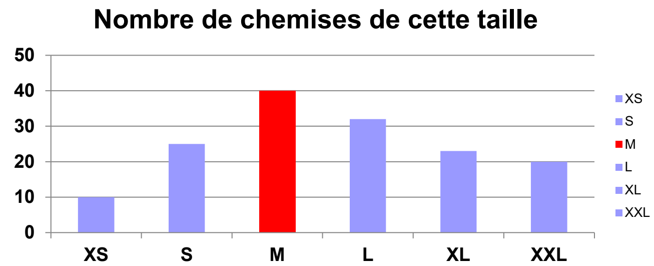
La moyenne arithmétique

Si les valeurs de X ne sont pas regroupées, la moyenne arithmétique d'une série quantitative est définie par:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Si les valeurs de X sont regroupées: $\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} \frac{n_i}{n} x_i = \sum_{i=1}^{k} f_i x_i$

Le mode



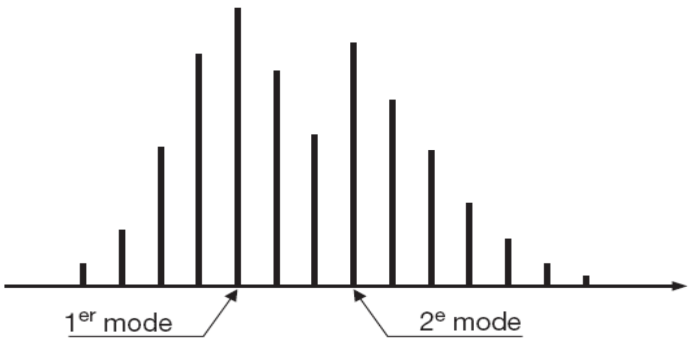
Le mode de cette série statistique est la modalité de la variable correspondant à **l'effectif** le **plus élevé** qui est dans ce cas la taille M.

Remarque:

De même pour toutes les variables de type qualitatif ou quantitatif non classé,

Cas particulier de mode: bimodale

Diagramme en bâtons d'une variable discrète



Exemple de distribution bimodale d'une variable discrète

Le mode d'une variable classée

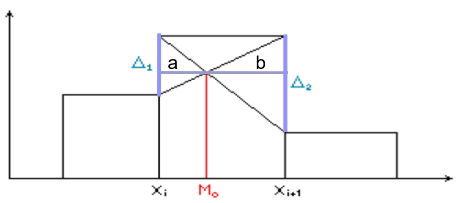
Pour les **variables quantitatives classées**, on parle d'abord de la **classe modale**:

- Si les classes sont d'**égales amplitudes**, la classe modale sera la classe où **l'effectif** est le plus élevé.
- Si les classes sont d'**inégales amplitudes**, la classe modale sera la classe où:

$$\left.\begin{array}{l} \text{➢ La } \textbf{densité} \quad d_i = \dfrac{n_i}{x_i - x_{i-1}} \\ \text{Ou} \\ \text{➢ La } \textbf{densité de fréquence} \quad d_{fi} = \dfrac{f_i}{x_i - x_{i-1}} \end{array}\right\} \text{est la plus élevée}$$

Le mode d'une variable classée

La classe modale est la classe pour laquelle l'effectif, la fréquence ou la densité de fréquence est la plus élevée. Pour déterminer sa valeur on utilise le schéma suivant:



La classe modale $[x_i,x_{i+1}[$ étant déterminée, le Mo vérifie:

Théorème de Thalès:

$$\frac{\Delta_1}{\Delta_2} = \frac{a}{b} = \frac{Mo - x_i}{x_{i+1} - Mo} \Longrightarrow Mo = \frac{x_i\Delta_2 + x_{i+1}\Delta_1}{\Delta_1 + \Delta_2} = x_i + \frac{\Delta_1(x_{i+1} - x_i)}{\Delta_1 + \Delta_2}$$

La médiane: Cas d'une variable non classée

Dans le cas d'une variable quantitative non classée, on détermine la médiane par:

- La série statistique doit être rangée par ordre croissant,

$$x_1<x_2<\ldots<x_p<x_{p+1}<\ldots<x_n$$

- On a deux cas:
  - ✓ Si n est impair et égal 2p+1 la médiane sera: x_{p+1}.
  - ✓ Si n est pair et égal 2p, la médiane est: $\dfrac{n_p + n_{P+1}}{2}$

Exemple:

Déterminer la médiane des deux séries suivantes:

1) 8 5 10 4 13 12 7 5 9.
2) 8 5 10 4 13 12 7 5.

La médiane

La médiane ne se calcule que pour les variables quantitatives et son calcul dépend du type de données. On distingue **quatre cas** :

- Les séries non groupées dont l'effectif est impair et où aucune valeur n'est répétée. **Exemple**: **{8 ,9, 5, 13, 25}**
- Les séries non groupées dont l'effectif est pair et où aucune valeur n'est répétée. **Exemple**: **{13,1,9,10,2,4,12,7}**
- Les séries groupées par valeurs. **Exemple**:

x_i	5	8	9	10
n_i	2	3	4	3

x_i	0	1	2	3
n_i	3	7	5	5

- Les séries groupées par classes de valeurs. **Exemple**:

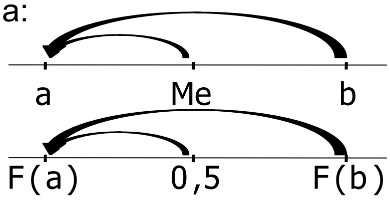
x_i	[0;5[	[5;10[	[10;15[	[15;20[
n_i	2	7	18	3

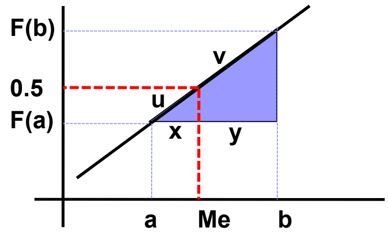
x_i	[0;5[	[5;10[	[10;20[
n_i	20	30	50

La médiane: Cas d’une variable classée

Soient a et b les bornes inférieurs et supérieurs de la classe contenant la médiane, F(a) et F(b) les valeurs des fréquences cumulées croissantes en a et b, alors:

On a:





$$\frac{x}{y} = \frac{u}{v}$$

$$\frac{Me - a}{b - a} = \frac{0{,}5 - F(a)}{F(b) - F(a)} \Longrightarrow Me = a + (b - a) \times \frac{0{,}5 - F(a)}{F(b) - F(a)}$$

Exemple de calcul du mode et médiane

La série suivante représente l’âge des salariés d’une société:

Classes	[20;25[	[25;30[	[30;35[	[35;40[	[40;50[	[50;60[
n_i	9	17	36	27	45	6

1. Définir la population étudiée, l’unité statistique, taille de la population, le caractère étudié, sa nature et ses modalités.
2. Les effectifs cumulés, les amplitudes, la densité, les fréquences, les fréquences cumules croissantes et décroissantes, sont déjà calculer dans le diapo 16
3. Déterminer la classe modale et déterminer le mode,
4. Déterminer la classe contenant la médiane et déterminer sa valeur

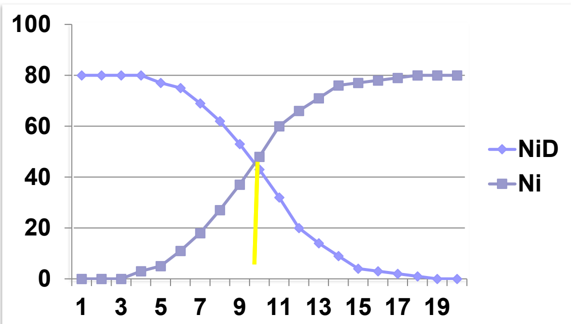
Mode et médiane pour une variable classée

Classes	n_i	a_i	$d_i=n_i/A_i$	$n_{ic}=d_i x ppcm(a_i)$	f_i	F_i
[20;25[	9	5	1,8	18	0,06	0,06
[25;30[	17	5	3,4	34	0,12	0,19
[30;35[	36	5	7,2	72	0,26	0,44
[35;40[	27	5	5,4	54	0,19	0,64
[40;50[	45	10	4,5	45	0,32	0,96
[50;60[	6	10	0,6	6	0,04	1

0,5

Même si la classe [40;50[a l'effectif le plus élevé mais la classe modale est [30;35[car elle a la densité ou l'effectif corrigé la plus élevée.
La médiane qui est équivalent à 0.5 pour les F_i se trouve dans l'intervalle [35;40[

La médiane



Représentation graphique des effectifs cumulés croissants et décroissants.

La médiane de la série correspond au point d'intersection de ces 2 courbes

La médiane

Propriété de la médiane:

La médiane donne des indications utiles sur la tendance centrale d'une distribution statistique. Elle n'est pas influencée par les valeurs extrêmes de la variable.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_7

La médiane

La valeur de la médiane ne change pas même si la valeur X_7 prend des valeurs différentes.

Exemple de la médiane

Compléter le tableau suivant et calculer la médiane

Classes	n_i	A_i	d_i	f_i	F_i
[1;2[	7				
[2;4[	8				
[4;5[	10				
[5;6[	3				
Total					

Les moyennes

On peut réduire un ensemble d'observations en une seule observation constante appelée **moyenne**.

La moyenne est donc une valeur qui présente comme si toutes les observations lui étaient égales.

On distingue plusieurs types de moyennes:

- La moyenne **arithmétique**,
- La moyenne **géométrique**,
- La moyenne **harmonique**,
- La moyenne **quadratique**.

La moyenne arithmétique

On distingue deux types: moyenne pour les variables quantitatives **non classées** et **classées**

La moyenne arithmétique (cas non classées)

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

La moyenne arithmétique (cas classées)

Le centre de la classe $[x_i, x_{i+1}[$, est : $c_i = \dfrac{x_i + x_{i+1}}{2}$

La moyenne est: $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} \dfrac{n_i}{n} c_i = \sum_{i=1}^{k} f_i c_i$

La moyenne géométrique

On distingue deux types: moyenne pour les variables quantitatives **non groupées** et **groupées**.

$$\overline{X}_g = \sqrt[n]{x_1 \times \ldots \times x_n} \quad \text{et} \quad \overline{X}_g = \sqrt[n]{x_1^{n_1} \times \ldots \times x_n^{n_k}} = \prod_{i=1}^{k} x_i^{n_i \frac{1}{n}} = \prod_{i=1}^{k} x_i^{f_i}$$

Dans le cas d'une variable classée on utilise c_i à la place de x_i.

Domaines d'application:

On utilise la moyenne géométrique dans:

- Le calcul du taux d'accroissement moyen,
- Le calcul des pourcentages moyens.

Exemple: calculer la moyenne géométrique de: 1; 2; 2; 4

La moyenne harmonique

On distingue deux types: moyenne pour les variables quantitatives **non groupées** et **groupées**.

$$\overline{X}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \quad ou \quad \overline{X}_h = \frac{n}{\sum_{i=1}^{k} n_i \frac{1}{x_i}} = \frac{1}{\sum_{i=1}^{k} f_i \frac{1}{x_i}}$$

Domaine d'application

Les calculs des durées moyennes,

Elle intervient lorsqu'on demande une moyenne de valeurs se présentant sous forme de quotient de deux variables x/y (km/h, kg/litre,...).

La moyenne quadratique

On distingue deux types: moyenne pour les variables quantitatives **non groupées** et **groupées**.

$$\overline{X}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \quad ou \quad \overline{X}_q = \sqrt{\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2} = \sqrt{\sum_{i=1}^{k} f_i x_i^2}$$

Domaines d'utilisation

La moyenne quadratique intervient dans le calcul de certains paramètres de dispersion

Exemple

Calculer la moyenne quadratique de : 2; 12; 2; 50. =25,749

Exemple

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

Classe de notes	[0;5[	[5;10[	[10;15[	[15;20[
Nombre d'étudiants	2	7	18	3

Calculer les quatre moyennes suivantes:

Les moyennes			
Arithmétique	Géométrique	Harmonique	Quadratique
$\overline{X} = \sum_{i=1}^{k} f_i x_i$	$\overline{X}_g = \prod_{i=1}^{k} x_i^{f_i}$	$\overline{X}_h = \frac{1}{\sum_{i=1}^{k} f_i \frac{1}{x_i}}$	$\overline{X}_q = \sqrt{\sum_{i=1}^{k} f_i x_i^2}$

Comparaison des moyennes

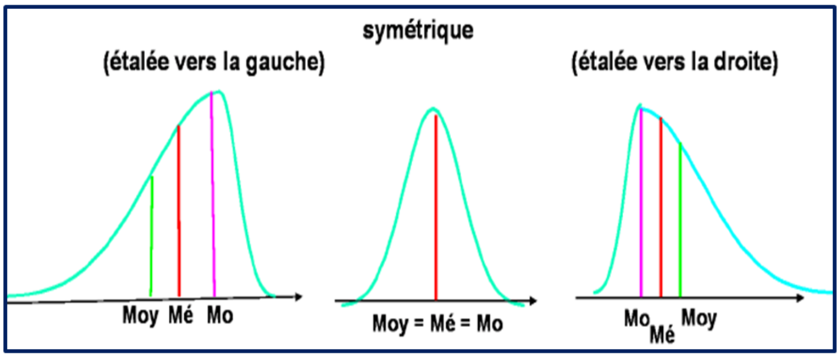
Pour la même série statistique, les quatre moyennes vérifient toujours la relation d'ordre suivante:

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

Conclusions :

1. Un inconvénient de la moyenne arithmétique est qu'elle est très sensible aux valeurs extrêmes de la série.
2. La moyenne géométrique est peu sensible aux valeurs extrêmes de la série.
3. La moyenne harmonique est plus sensible aux plus petites valeurs de la série qu'aux plus grandes.

Relation entre les paramètres de position



Asymétrie d'une distribution

Moyenne, mode, médiane et forme d'une distribution
La moyenne est influencée par les valeurs extrêmes de la distributions

Fin de la deuxième séance

Les paramètres de dispersion

La variance V et l'écart-type σ:

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

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

Etendu:

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

Quartiles :

Un quartile est chacune des 3 valeurs qui divisent les données triées en 4 parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population.

Les intervalles interquartiles :

L'intervalle interquartile d'une série statistique est égal à la différence **:**
Q3 – Q1.

Ecart interdécile:

On appelle **premier décile** d'une série la plus petite valeur D_1 des termes de la série pour laquelle au moins un dixième (10%) des données sont inferieures ou égales à D_1.
On appelle **écart interdécile** le nombre **D_9-D_1**.

L'interprétation des paramètres de dispersion

Etendu:
Les valeurs de la série sont réparties sur un intervalle d'amplitude égale à la valeur de l'étendu.

Les intervalles interquartiles :
50% des valeurs de la série sont dans l'intervalle **$[Q_1 ; Q_3]$**.
25% des valeurs de la série sont inférieures à **Q_1** (resp supérieures à **Q_3**)

Ecart interdécile:
80% des valeurs de la série sont dans l'intervalle **$[D_1 ; D_9]$**.
10% des valeurs de la série sont inférieures à **D_1** (resp supérieures à **D_9**)
La valeur à **D_1** est **D_9/D_1** fois plus élevé que à **D_9**,

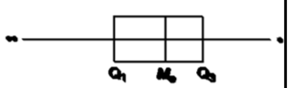
La variance V et l'écart-type σ:
La variance exprime la dispersion des valeurs autour de la moyenne, mais avec une unité de mesure différente de celle de la variable étudiée, on a introduit l'écart-type qui la même unité que la variable, mais son interprétation dépend de l'échelle de la variable.

Application des quantiles

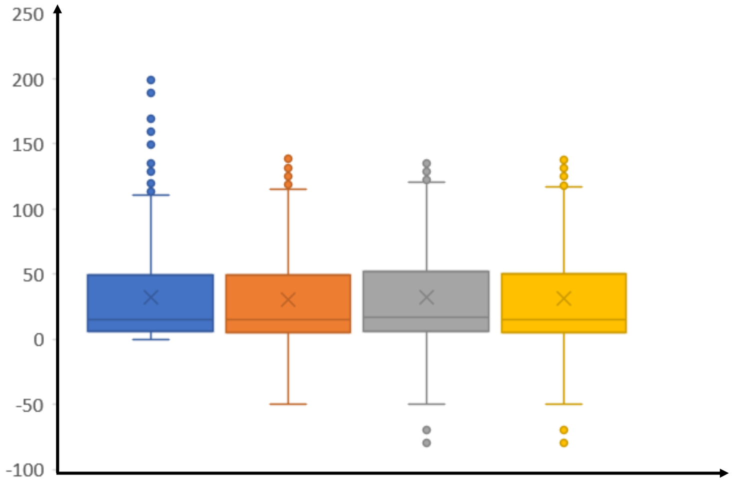
Le Quantile
Les quantiles sont des caractéristiques de position partageant la série statistique ordonnée en k parties égales.
Pour k = 4, les quantiles, appelés quartiles,
Pour k =10, les quantiles sont appelés déciles,

1	2	3	4	5	6	7	8	9
		Q_1		Q_2		Q_3		
		3		5		7		

Application
Le diagramme en boîte à **moustaches** ou **box-plot** permet de représenter schématiquement les principales caractéristiques d'une distribution en utilisant les quartiles.
La partie centrale de la distribution est représentée par une boîte de largeur arbitraire et de **longueur la distance interquartile**, la médiane est tracée à l'intérieur.
La boîte rectangle est complétée par des moustaches correspondant aux valeurs suivantes:
– Valeur supérieure : Min(la plus grande la valeur; $Q_3 + 1,5(Q_3 - Q_1)$)
– Valeur inférieure : Max(la plus petite valeur; $Q_1 - 1,5(Q_3 - Q_1)$)
Les valeurs extérieures « aux moustaches » sont représentées par des étoiles et peuvent être considérées comme **aberrantes**.



La boite à Moustache



Coefficient de variation

Pour une variable statistique réelle X, on appelle **coefficient de variation** le rapport:

$$C_v = \frac{\sigma_X}{\overline{X}},$$

$$où : \quad \sigma_X \text{ est l'écart-type de } X \text{ et } \overline{X} \text{ sa moyenne.}$$

- C'est un nombre sans unité, qui permet de comparer la distribution autour de la moyenne de deux variables statistiques de natures différentes:
  Exemple: la variabilité du poids des éléphants et des souris.
- Plus la valeur du coefficient de variation est **élevée**, plus la dispersion autour de la moyenne est **grande**.
- Il est généralement exprimé en pourcentage.
- Il permet d'apprécier **l'homogénéité** de la distribution, une valeur du coefficient de variation inférieure à 15 % traduit une bonne homogénéité de la distribution.

Exemple: les deux séries: 1, 10, 19 et 1000001, 1000010, 1000019

$\sigma_1=\sigma_2=4{,}24$. Mais les moyennes sont: $m_1=10$, $m_2=1000010$

Paramètres de dispersion

Pour $r \in \mathbb{N}^*$ et un caractère quantitatif X on définit:

- ✓ **Le moment d'ordre r** par:

$$m_r = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^r = \sum_{i=1}^{k} f_i x_i^r$$

- ✓ **Le moment centré d'ordre r** par:

$$\mu_r = \frac{1}{n}\sum_{i=1}^{k} n_i \left(x_i - \bar{x}\right)^r = \sum_{i=1}^{k} f_i \left(x_i - \bar{x}\right)^r$$

Paramètres de forme (asymétrie)

Coefficient d'asymétrie de Fisher γ_1

Il est défini par: $\gamma_1 = \frac{\mu_3}{\sigma^3}$

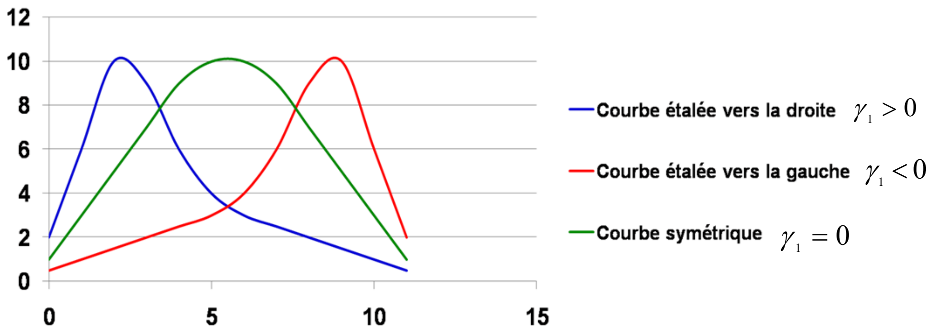
Si $\gamma_1 = 0$, la distribution est **symétrique** autour de la moyenne.

Si $\gamma_1 < 0$, la distribution est plus étalée **vers la gauche**.

Si $\gamma_1 > 0$, la distribution est plus étalée **vers la droite**.

Paramètres de forme (asymétrie)

Représentation graphique de trois séries statistiques de différents types d'asymétrie.



Paramètres de forme (aplatissement)

Coefficient d'aplatissement de Fisher γ_2:
Il est défini par:

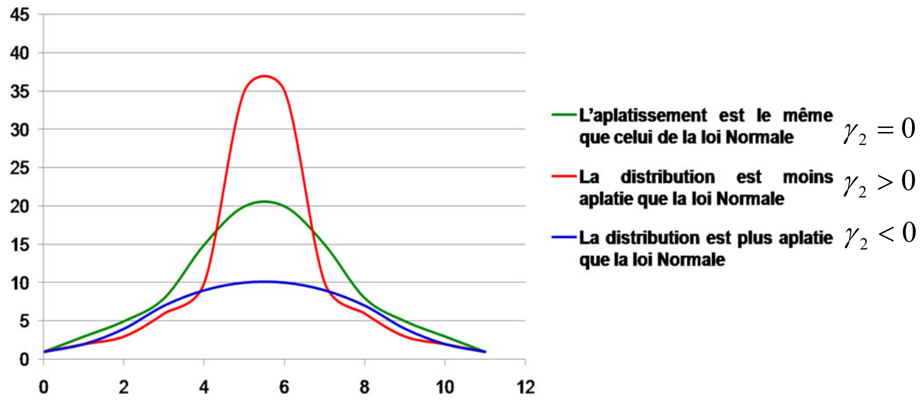
$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Si $\gamma_2 = 0$, L'aplatissement est le même que celui de la **loi Normale** (de Gauss).

Si $\gamma_2 < 0$, la concentration des valeurs autour de la moyenne est faible: la distribution est **plus aplatie** que la loi Normale.

Si $\gamma_2 > 0$, la concentration des valeurs autour de la moyenne est forte: la distribution est **moins aplatie** que la loi Normale.

Paramètres de forme (aplatissement)



Coefficient de Fisher γ_2 (aplatissement d'une distribution):

Paramètres de forme

Pratiquement, pour qu'une variable puisse être considérée comme suivant une **loi normale** ou de Gauss il faut que:

- Le coefficient **d'asymétrie** (en anglais Skewness) doit être inférieur à **|1|**
- Le coefficient **d'aplatissement** (en anglais Kurtosis) ou encore de concentration doit être inferieur à **|1,5|**

Exercice

Une enquête sur la consommation annuelle d'électricité a été effectuée sur une population de 2600 ménages. Les résultats figurent dans le tableau suivant:

Consommation annuelle en (kwh)	Nombre de ménages
[0,200[	455
[200,300[	614
[300,400[	532
[400,600[	385
[600,800[	422
[800,1000[	164
[1000,2000[	28

Exercice à la maison

Le tableau suivant donne le niveau de scolarité, en nombre d'années passées à l'école, d'un échantillon de 200 personnes

Niveau de scolarité	Effectif
[0,6[	40
[6,12[	80
[12,14 [	50
[14,16[	30

Exercice (suite)

1. Donner la population étudiée, et la taille de l'échantillon,
2. Donner le caractère X étudié, ses modalités et sa nature,
3. Calculer la moyenne, la médiane et le mode, comparer leurs valeurs, que peut-on dire ?
4. Calculer la variance, l'écart type et le coefficient de variation de X,
5. Calculer l'étendu de X, que peut-on dire de sa valeur ?
6. Calculer les quartiles Q_1, Q_3 , et en déduire l'écart inter-quartiles,
7. Calculer les déciles D_1, D_9 , et en déduire l'écart inter-décile,
8. Calculer μ_3 : le moment centré d'ordre 3, en déduire le coefficient d'asymétrie de Fisher γ_1 et interpréter le résultat,
9. Calculer μ_4 : le moment centré d'ordre 4, en déduire le coefficient d'aplatissement de Fisher γ_2 et interpréter le résultat.

Paramètre de concentration

Introduction:

D'après le rapport annuel de Bank Al-Maghrib (2005), le total cumulé des situations comptables des 16 banques agréées s'est chiffré à **461,5** milliards DH, tel que la part des **3 grandes** banques est **63,8%**, tandis que celle des **8 petites** banques est **4,2%**.

C'est le phénomène de la **concentration** de l'activité bancaire.

On dit que l'activité bancaire en 2005 est caractérisée par une **forte concentration**.

Paramètre de concentration

Exemple:

Soit la distribution suivante relative à la répartition de 80 salariés selon leur salaire horaire en DHS,

Salaire horaire en DHS	Nombre de salariés n_i
[10,20[	20
[20,40[	32
[40,80[	16
[80,100[	8
[100,160[	4

Indice de concentration de Gini

Soit **X** une variable divisée en **k** classes.

La **ième** classe **$[x_{i-1}, x_i[$** a, pour centre, **c_i** et, pour effectif, **n_i**.

- $s_i = n_i c_i$ la masse de caractère X dans la classe $[x_{i-1}, x_i[$.
- $S = \sum_{i=1}^{k} s_i$ la masse globale de X
- $g_i = \frac{s_i}{S}$ la fréquence de la masse de X possédée par les individus dans la classe $[x_{i-1}, x_i[$.
- $G_i = \sum_{j=1}^{i} g_j$ La masse cumulée relative à la classe $[x_{i-1}, x_i[$.

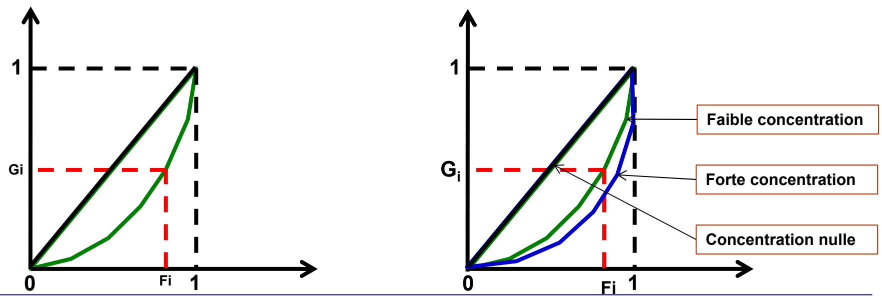
Paramètre de concentration

Salaire horaire en DHS	Nombre de salariés n_i	ci	$si=n_ic_i$	gi	Gi	fi	Fi	Gi-1+Gi	(Gi-1+Gi)*fi
[10,20[	20	15	300	0,09	0,09	0,25	0,25	0,09	0,02
[20,40[	32	30	960	0,28	0,36	0,40	0,65	0,45	0,18
[40,80[	16	60	960	0,28	0,64	0,20	0,85	1,01	0,20
[80,100[	8	90	720	0,21	0,85	0,10	0,95	1,49	0,15
[100,160[	4	130	520	0,15	1,00	0,05	1,00	1,85	0,09
			3460						0,64

Indice de concentration de Gini

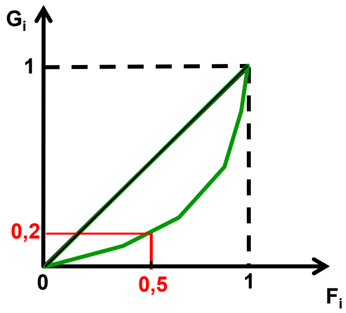
On appelle **courbe de concentration** (ou courbe de **Lorenz**) la ligne polygonale joignant les points de cordonnées (F_i,G_i).

Où: $G_i = \sum_{j=1}^{i} g_j$ et $F_i = \sum_{j=1}^{i} f_j$ $g_i = \frac{n_i c_i}{S}$ et $f_i = \frac{n_i}{n}$



Indice de concentration de Gini

Interprétation de la courbe de Lorenz:



On voit que 50% des salaires se partagent 20% de la masse salariale. Donc, on peut dire que la concentration est forte.

Indice de concentration de Gini

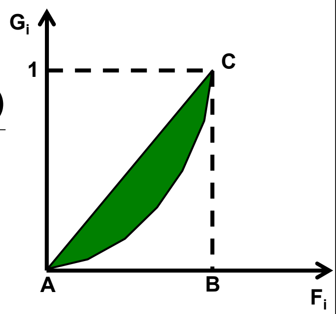
Définition:

L'indice de Gini est égal à:

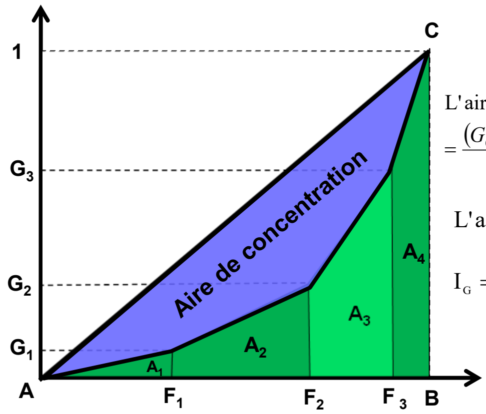
$$I_G = \frac{aire\ de\ concentration\ (en\ vert)}{aire\ du\ triangle\ ABC}$$

$$L'aire\ du\ triangle\ ABC = \frac{1 \times 1}{2} = 0{,}5$$

$$Donc\ I_G = 2 \times (aire\ de\ concentration)$$



Indice de concentration de Gini



$$\text{L'aire } A_3\ (Trapèse) = \frac{(G_3+G_2)(F_3-F_2)}{2} = \frac{(G_3+G_2)f_3}{2}$$

$$\text{L'aire en vert} = A_1 + A_2 + A_3 + A_4$$

$$= \frac{(G_0+G_1)f_1}{2} + \frac{(G_1+G_2)f_2}{2} + \frac{(G_2+G_3)f_3}{2} + \frac{(G_3+G_4)f_4}{2}$$

$$\text{L'aire en vert} = \frac{1}{2}\sum_{i=1}^{4}(G_{i-1}+G_i)f_i$$

$$I_G = 2\times(\text{l'aire en bleu}) = 2\times(\text{l'aire de concentration})$$
$$= 2\times(\text{l'aire(ABC)} - \text{L'aire en bleu})$$
$$= 2\times\left(\frac{1}{2} - \frac{1}{2}\sum_{1}^{4}(G_{i-1}+G_i)f_i\right)$$

$$\boxed{\text{Donc} \quad I_G = 1 - \sum_{1}^{4}(G_{i-1}+G_i)f_i}$$

Indice de concentration de Gini

Calcul de l'indice de Gini:

L'indice de concentration ou indice de Gini que l'on note par I_G est donné par:

$$\boxed{I_G = 1 - \sum_{i=1}^{k} f_i\,(G_{i-1} + G_i)} \qquad Avec\ G_0 = 0$$

Interprétation:

- On a toujours, $0 \leq I_G \leq 1$
- $I_G = 0$ concentration nulle,
- $I_G = 1$ concentration maximale,
- Plus la valeur de I_C est grande plus la concentration est forte.

Rappel (analogie)

Distribution $\{(x_i, n_i)_{1\le i\le k}\}$	Distribution $\{(x_i, n_i x_i)_{1\le i\le k}\}$
n_i	$s_i = n_i x_i$
$n = \sum_{i=1}^{k} n_i$	$S = \sum_{i=1}^{k} s_i$
$f_i = \frac{n_i}{n}$	$g_i = \frac{s_i}{S}$
$F_i = \sum_{j=1}^{i} f_j$	$G_i = \sum_{j=1}^{i} g_j$
M_e	M_l

Où M_l est la valeur de caractère X qui partage la masse globale en deux parties égales, il est calculé de la même façon que M_e

Partie 3
Les séries doubles

Exemple de série double

Construction de tableau de contingence:

Soit la distribution de 17500 jeunes salariés selon l'âge et le salaire net, en milliers de dirhams:

Âge (x_i) \ Salaire net(y_i)	[5,6[	[6,7[	[7,8[	Total
[20,22[	1200	500	100	1800
[22,24[	2500	3500	600	6600
[24,26[	1800	5000	2300	9100
Total	5500	9000	3000	17500

- **Âge**: en année
- **Salaire**: mensuels

Les cellules vertes représentent **la distribution marginale de caractère X,**

En bleu **la distribution marginale de caractère Y.**

Tableau de contingence

Modalités de Y

X \ Y	y_1	…	y_j	…	y_q	**Total**
x_1	n_{11}				n_{1q}	$n_{1.}$
…						…
x_i			n_{ij}			$n_{i.}$
…						…
x_p	n_{p1}				n_{pq}	$n_{p.}$
Total	$n_{.1}$	…	$n_{.j}$	…	$n_{.q}$	$n_{..}$

Modalités de X

Distribution marginale de Y

Distribution marginale de X

$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij}$$

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

Eléments d'un tableau de contingence

Les effectifs:

Les effectifs partiels: n_{ij} égale au nombre d'individus présentant la modalité x_i de la variable X et la modalité y_j de la variable Y

Les effectifs marginaux: ce sont les effectifs lus dans les 2 marges du tableau.

Pour X :

Variable X	x_1	x_2	$\cdots$	x_i	$\cdots$	x_p	Total
Effectifs	$n_{1\bullet}$	$n_{2\bullet}$	$\cdots$	$n_{i\bullet}$	$\cdots$	$n_{p\bullet}$	$n_{\bullet\bullet}$

Pour Y :

Variable Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_q	Total
Effectifs	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet q}$	$n_{\bullet\bullet}$

Eléments d'un tableau de contingence

Les fréquences:

Les fréquences partielles: $f_{ij} = \dfrac{n_{ij}}{n_{\bullet\bullet}}$

Les fréquences conditionnelles:

1. **La fréquence conditionnelle de X selon Y:** $f_{i/j} = \dfrac{n_{ij}}{n_{\bullet j}}$
2. **La fréquence conditionnelle de Y selon X:** $f_{j/i} = \dfrac{n_{ij}}{n_{i\bullet}}$

Les fréquences marginales: $f_{i\bullet} = \dfrac{n_{i\bullet}}{n_{\bullet\bullet}}$ et $f_{\bullet j} = \dfrac{n_{\bullet j}}{n_{\bullet\bullet}}$

Les relations entre les fréquences marginales et conditionnelles:

$$f_{i\bullet} \times f_{j/i} = f_{\bullet j} \times f_{i/j} = f_{ij}$$

Indépendance de deux variables

Définition :

Deux variables sont indépendantes si les variations de l'une n'entrainent pas de variations de l'autre.

Autrement:

Deux variables X et Y sont totalement indépendantes si les fréquences conditionnelles $f_{i/j}$ ne dépendent plus de j.

$$f_{i/j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i1}}{n_{\bullet 1}} = \frac{n_{i2}}{n_{\bullet 2}} = \ldots\ldots = \frac{n_{iq}}{n_{\bullet q}} = \frac{\sum_{j=1}^{q} n_{ij}}{\sum_{j=1}^{q} n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} = f_{i\bullet}$$

$$\Rightarrow \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \Rightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n_{\bullet\bullet}} \Rightarrow \frac{n_{ij}}{n_{\bullet\bullet}} = \frac{n_{i\bullet}}{n_{\bullet\bullet}} \times \frac{n_{\bullet j}}{n_{\bullet\bullet}} \quad \text{donc} \quad \boxed{f_{ij} = f_{i\bullet} \times f_{\bullet j}}$$

Exemple d'indépendance de 2 variables

Indépendance des variables

Le tableau suivant représente la distribution statistique de deux variables X et Y :

X \ Y	y_1	y_2	Total
x_1	3	5	8
x_2	6	10	16
Total	9	15	24

Calculer les fréquences partielles et les fréquences marginales.

Montrer que les caractères X et Y sont indépendants.

Les caractéristiques des séries à 2 caractères

La moyenne marginale de X:

$$\bar{x} = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i \qquad \text{avec} \qquad n_{\bullet\bullet} = \sum_{i=1}^{p} n_{i\bullet} = \sum_{j=1}^{q} n_{\bullet j} = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij}$$

La variance marginale de X:

$$V(x) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} (x_i - \bar{x})^2 \qquad \text{ou} \qquad V(x) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} n_{i\bullet} x_i^2 - \bar{x}^2$$

L'écart-type de X: $\sigma_X = \sqrt{V(x)}$

La covariance:

$$\operatorname{cov}(x, y) = \frac{1}{n_{\bullet\bullet}} \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

X	Y
i	j
p	q
$n_{i.}$	$n_{.j}$

Exercice

Pour 25 ménages, les âges de l'époux et de l'épouse, relevés sur le registres d'un état civil sont les suivants:

(22,17); (23,18); (24,17); (24,18); (24,20); (24,21); (25,18); (25,19); (25,20); (26,18); (26,19); (26,21); (26,23); (27,19); (27,21); (28,21); (28,22); (30,22); (30,23); (31,24); (31,25); (34,24); (35,24); (35,25); (36,25);

Sachant que chaque couple (x_i,y_j) représente respectivement l'âge de l'époux et l'âge de l'épouse au moment de mariage.

1. Ranger les données en classes de même amplitude 5, qui commencent par 20 pour X et par 15 pour Y.
2. Calculer l'âge moyenne des époux et des épouses.
3. Calculer la variance de l'âge d'épouse, et son écart-type.
4. Calculer la covariance des deux variables

Ajustement linéaire

Cadre, rappels et objectifs

On dispose de deux caractères X et Y quantitatifs.

On distingue trois objectifs :

- On cherche à savoir s'il existe un **lien** entre **X et Y**,
- On construit un modèle qui permet **d'exprimer Y** en fonction de **X**.
- On calcule les **prévisions** et on donne leurs **incertitudes**.

Etude de liaison entre deux variables

Lorsqu'on observe deux variables sur les mêmes individus, on peut s'intéresser à une liaison entre ces deux variables.

Trois types de liaison peuvent être envisagés:

1. **La liaison nulle:** lorsque il n'y a aucune influence d'un caractère sur l'autre.
   Exemple: salaire et la taille;
2. **La liaison totale:** lorsque il y a une liaison totale.
   Exemple: le périmètre et le rayon d'un cercle;
3. **La liaison relative:** est le cas général, les caractères sont dépendants l'un de l'autre dans un certaine mesure.
   Exemple: la consommation et le revenu,

Notion de corrélation

On dit qu'il y a une **corrélation** entre deux variables lorsqu'elles ont tendance à varier:

1. Soit dans le **même sens** (Exemple, si X augmente, Y augmente aussi),
2. Soit dans le **sens inverse** (Exemple, si X augmente, Y diminue).

Coefficient de corrélation:

On dispose de n individus dont on calcule leurs valeurs pour deux variables quantitatives X et Y : (x_1,y_1), …, (x_n, y_n).

Le **coefficient de corrélation linéaire** entre X et Y est :

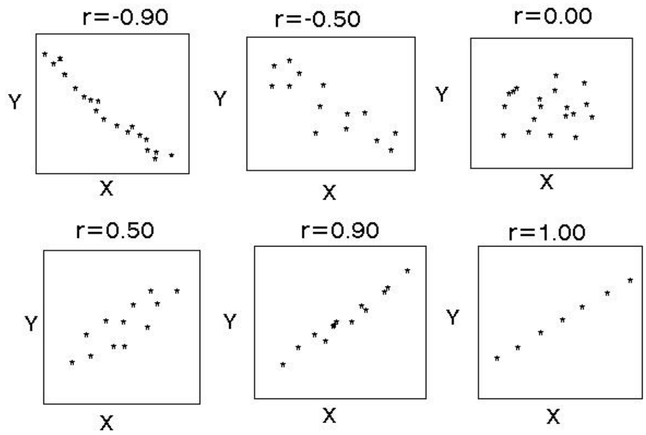
$$\boxed{r = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}}$$ avec $$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Notion de corrélation

Interprétation de coefficient de corrélation:

- Si **r est proche de 1**, il y a une forte corrélation positive entre X et Y (même sens de variation)
- Si **r est proche de -1**, il y a une forte corrélation négative entre X et Y (différence du sens de variation).
- Si **r = 0**, X et Y sont non corrélées : il n'y a pas d'association linéaire entre X et Y.
- Si **r = ±1**, alors chacune de ces deux variables peut définir l'autre d'une façon exacte.

Interprétation graphique



Ajustement linéaire

On dispose de deux caractères X et Y quantitatifs, on a:

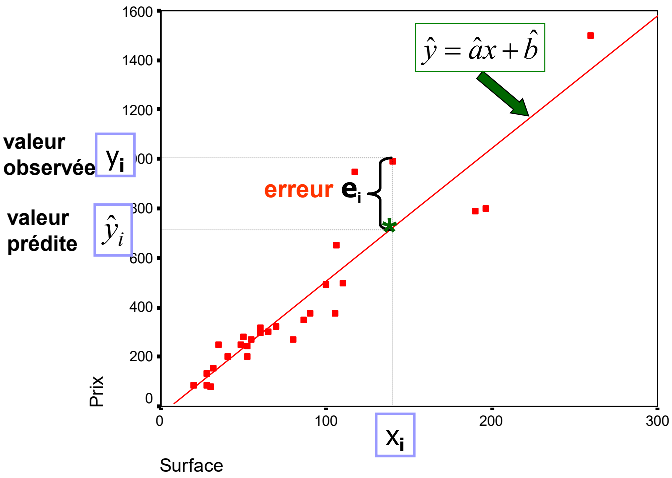
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad V(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\operatorname{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Donc le coefficient de corrélation sera:

$$\boxed{r = \frac{\operatorname{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

Ajustement par la méthode des moindres carrés



On cherche

$\hat{a}$ et $\hat{b}$

En minimisant

$$\sum_{i=1}^{n} e_i^2$$

Ajustement par la méthode des moindres carrés

Détermination de la droite de régression:

La méthode des moindre carrés consiste à trouver les coefficients **a** et **b** de la droite de régression **y=ax+b**, qui minimisent la distance quadratique entre $\hat{y}$ et y_i qui revient à minimiser:

$$S(a,b) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (ax_i + b))^2$$

Après un calcul, on obtient:

$$\hat{a} = \frac{\mathrm{cov}(X,Y)}{V(X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{xy}}{\frac{1}{n}\sum_{i=1}^{n} x_i^n - \bar{x}^2}; \quad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

Exercice

Soit X la note des mathématiques sur 20 points et Y la note de statistique sur 20 points pour 10 étudiants:

X	2	4	6	6	9	10	11	12	13	18
Y	3	6	6	7	9	10	10	11	14	14

1. Donner la droite des moindres carrés de Y en X ,
2. Donner la droite des moindres carrés de X en Y ,

Partie 2
Les séries chronologiques

Rappel

Les ventes trimestrielles de jus de fruits dans un grand magasin ont été, en milliers de litres, les suivantes:

Calculer les prévisions de semestre 1 et 2 de l'année 2001 ?

	1er	2ème	3ème	4ème
1996	170	300	610	120
1997	250	410	790	190
1998	290	460	890	250
1999	450	550	1100	270
2000	320	600	1260	280

Définition d'une série chronologique

Une **série chronologique** ou **chronique**, ou **série temporelle**, est une suite d'observations, échelonnées dans le temps, d'une variable quelconque.

Elle s'intéresse à l'évolution au cours du temps d'un phénomène, dans le but de **décrire**, **expliquer** puis **prévoir** ce phénomène dans le futur.

Exemple:

Les ventes d'une librairie en fonction de temps

Exemples des séries chronologiques

Exemple:

En économie:

- L'évolution des indices boursiers, des prix, des données économiques des entreprises, des ventes et achats de biens, des productions agricoles ou industrielles,
- L'état fait des prévisions sur le niveau de croissance de la production à court et à moyen terme.

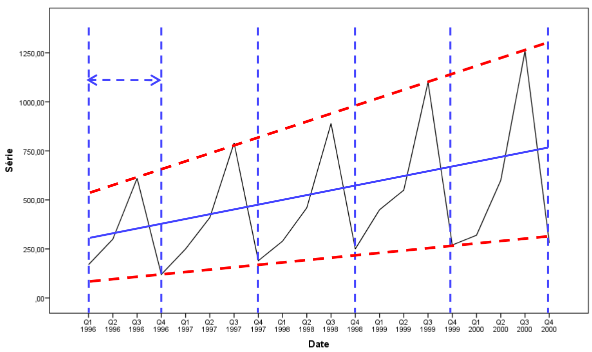
D'autres domaines:

- L'évolution du nombre du personnes atteintes d'une maladie.
- l'´evolution du nombre de voyageurs utilisant le train
- Nombre de clients qui visitent un supermarché par jour.
- La consommation d'électricité par mois.
- …

Représentation graphique d'une série chronologique

Les ventes trimestrielles de jus de fruits dans un grand magasin ont été, en milliers de litres, les suivantes:

	1er	2ème	3ème	4ème
1996	170	300	610	120
1997	250	410	790	190
1998	290	460	890	250
1999	450	550	1100	270
2000	320	600	1260	280



Description d'une série chronologique

Les composantes fondamentales d'une série chronologique $\mathbf{Y_t}$ sont: T_t, S_t, C_t, R_t

1. La **tendance** (ou trend) ($\mathbf{T_t}$) représente l'évolution à long terme de la série étudiée. Elle traduit le comportement « moyen » de la série.
2. La composante **saisonnière** (ou saisonnalité) ($\mathbf{S_t}$) correspond à un phénomène qui se répète aux intervalles de temps réguliers (périodiques). En général, c'est un phénomène saisonnier d'où le terme de variations saisonnières.

Description d'une série chronologique

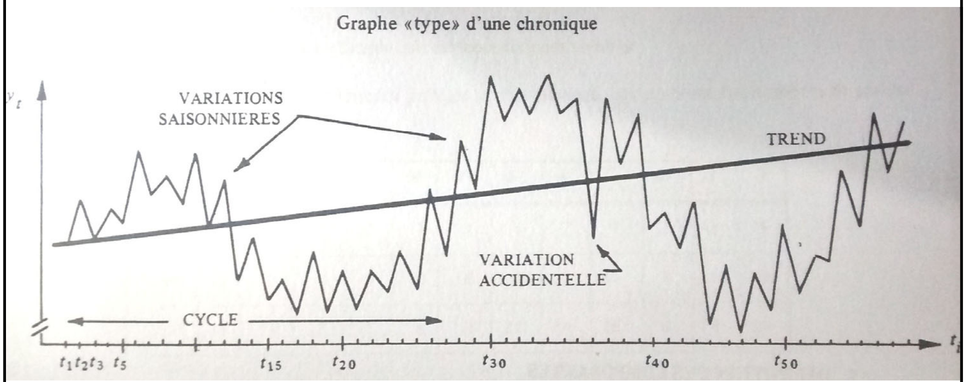
3. Un phénomène **cyclique** ($\mathbf{C_t}$): c'est souvent le cas en climatologie et en économie, mais souvent il n'est pas pris en compte dans les séries,

   Exemple : récession et expansion économique,...

4. La composante **résiduelle** (ou bruit ou résidu) ($\mathbf{R_t}$) correspond à des fluctuations irrégulières et imprévisibles, en général de faible intensité mais de nature aléatoire.

   Exemple : grève, guerre, sécheresse...

Les parties d'une série chronologique



Modèles des séries chronologiques

Deux modèles sont possibles:

Additif : $Y_t = T_t + S_t + C_t + R_t$

Multiplicatif: $Y_t = T_t \times S_t \times C_t \times R_t$

Pour faire cette détermination (modèle additif, multiplicatif) graphiquement, on trace les deux **droites** passant respectivement par les **minimum** et par les **maximum** de chaque **saison**.

Si ces deux droites sont **parallèles**, nous sommes en présence d'un modèle **additif**. Dans le **cas contraire**, c'est un modèle **multiplicatif**.

Remarque:

On peut toujours se ramener à partir du modèle multiplicatif au modèle additif en ajoutant le Logarithme:

$$\log(Y_t) = \log(T_t) + \log(S_t) + \log(C_t) + \log(R_t)$$

Choix de modèle: Additif ou multiplicatif

Méthode de Buys et Ballot :

On calcule, pour chacune des années, la moyenne et l'écart type.

On trace les points d'abscisse la moyenne et d'ordonnée l'écart type de la même année.

On trace la droite des moindres carrés de ces points.

- Si l'écart type est **indépendant** de la moyenne alors:

  Le modèle est additif.

  La pente (a) de la droite des moindres carrés est très proche de 0.
- Si l'écart type est **fonction** de la moyenne alors:

  Le modèle est multiplicatif.

  La pente (a) de la droite des moindres carrés n'est pas nulle.

Estimation des paramètres de la tendance

A- Méthode des moindres carrés:

La tendance peut prendre des formes fonctionnelles assez diverses citant:

Linéaire: $T_t = at + b + \varepsilon$

Polynomiale: [illegible]

Logarithmique: $T_t = a_0 \times a_1^t \times \varepsilon$

Cas linéaire:

$$a = \frac{\operatorname{cov}(X,Y)}{V(X)} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}; \quad b = \bar{y} - a\bar{x}$$

Et donc: $\boxed{T_t = at + b + \varepsilon}$

Estimation des paramètres de la tendance

Exemple :

Déterminer la tendance par la méthode des **moindres carrés** pour la série des ventes trimestrielles exprimées en millions de DH:

t	1	2	3	4	5	6	7	8	9	10	11	12
Y_t	3,6	3,9	4,3	3,4	3,8	4,1	5	3,9	4,7	5,1	5,8	4,7

Estimation des paramètres de la tendance

t	Y_t	t-moy(t)	y-moy(y)	(t-moy(t))²	(t-moy(t))(y-moy(y))
1	3,6	-5,5	-0,76	30,25	4,17
2	3,9	-4,5	-0,46	20,25	2,06
3	4,3	-3,5	-0,06	12,25	0,20
4	3,4	-2,5	-0,96	6,25	2,40
5	3,8	-1,5	-0,56	2,25	0,84
6	4,1	-0,5	-0,26	0,25	0,13
7	5	0,5	0,64	0,25	0,32
8	3,9	1,5	-0,46	2,25	-0,69
9	4,7	2,5	0,34	6,25	0,85
10	5,1	3,5	0,74	12,25	2,60
11	5,8	4,5	1,44	20,25	6,49
12	4,7	5,5	0,34	30,25	1,88
				143	21,25

Moyenne t 6,5
Moyenne Yt 4,36

Estimation des paramètres de la tendance

Exemple : On détermine le trend pour la série des ventes trimestrielle dans le cas d'un modèle additif

t	1	2	3	4	5	6	7	8	9	10	11	12
y_t	3,6	3,9	4,3	3,4	3,8	4,1	5	3,9	4,7	5,1	5,8	4,7

$f_t = at+b$, avec $a = \frac{Cov\,(t,y)}{V\,(t)}$ et $b = \bar{y} - a\bar{t}$

où : $V(t) = \frac{(12+1)(12-1)}{12} = \frac{13 \times 11}{12} = 11{,}917$

$$\bar{t} = \frac{1}{12}\sum_{i=1}^{12} i = \frac{12+1}{2} = 6{,}5 \qquad \bar{y} = \frac{1}{12}\sum_{i=1}^{12} y_i = \frac{52{,}3}{12} = 4{,}358 \text{ mDH}$$

$$Cov(t,y) = \frac{1}{12}\sum_{i=1}^{12} i.y_i - 6{,}5 \times 4{,}358 = \frac{361{,}2}{12} - 28{,}327 = 1{,}773$$

D'où, $a = 0{,}149$ et $b = 3{,}39$, donc $f_t = 0{,}149t + 3{,}39$.

Estimation des paramètres de la tendance

B. Méthode des moyennes mobiles:

Une moyenne mobile pour une période de temps est une moyenne arithmétique simple des valeurs de cette période et de celles avoisinantes,

Exemple:

Moyennes mobiles d'ordre 3:

Soit la série $y_1, y_2, \ldots, y_T$ on aura

$$\hat{y}_1 = \frac{y_1 + y_2 + y_3}{3}, \hat{y}_2 = \frac{y_2 + y_3 + y_4}{3}, \ldots, \hat{y}_{T-2} = \frac{y_{T-2} + y_{T-1} + y_T}{3},$$

La série $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{T-2}$ réduit les fluctuations aléatoires.

Estimation des paramètres de la tendance

Cas particulier de moyenne mobile:

Si l'ordre des moyennes mobiles est pair:

1. On calcule les MM d'ordre pair,
2. On calcule les MM d'ordre 2 de la nouvelle série (MM corrigées).

Exemple:

On calcule les MM d'ordre 4

Période	série	MM4	MMC4
1	15		
2	27		
		19,0	
3	20		20,25
		21,5	
4	14		19,50
		17,5	
5	25		
6	11		

Estimation des paramètres de la tendance

Exemple :

Déterminer la tendance par la méthode des **moyennes mobiles d'ordre 4** pour la série des ventes trimestrielles exprimées en millions de DH,

t	1	2	3	4	5	6	7	8	9	10	11	12
Y_t	3,6	3,9	4,3	3,4	3,8	4,1	5	3,9	4,7	5,1	5,8	4,7

Remarque:

Pour une meilleure estimation de la tendance, on choisit l'ordre des moyennes mobiles égal au nombre de saisons.

Calcul des coefficients saisonnières

Soit une série observée sur p périodes, de k valeurs pour chacune, et ne contient pas de composante cyclique, alors:

La série brute s'écrit, pour $1 \le i \le p$ et $1 \le j \le k$:

Le modèle additif: $Y_t = Y_{ij} = T_t + S_{ij} + R_t$

Le modèle additif: $Y_t = Y_{ij} = T_t \times S_{ij} \times R_t$

$\left.\right\} t = (i-1)k + j$

Exemple: Les ventes trimestrielle exprimées en millions de DH

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4
1988	3,6	3,9	4,3	3,4
1989	3,8	4,1	5	3,9
1990	4,7	**5,1**	5,8	4,7

Dans ce cas: p=3, k= 4, par exemple Y_{32}=5,1

Le modèle additif

Les coefficients saisonnières:

1. On détermine la tendance T_t;
2. On calcule les coefficients saisonniers $S_{ij} = Y_{ij} - T_t$

Les composantes saisonnières:

1. On calcule les composantes saisonnières bruts S'_j pour chaque saison j, $S'_j = \frac{1}{p}\sum_{i=1}^{p} S_{ij}$
2. On calcule la moyenne des S'_j : $s = \frac{1}{j}\sum_{j=1}^{k} S'_j$
3. Les composantes saisonnières sont: $\boxed{S_j = S'_j - s}$

Modèle multiplicatif

Les coefficients saisonnières

1. On détermine la tendance T_t;
2. On calcule les coefficients saisonniers $S_{ij} = \frac{Y_{ij}}{T_t}$

Les composantes saisonnières:

1. On calcule les composantes saisonnières bruts S'_j pour chaque saison j, $\Big\}\ S'_j = \frac{1}{p}\sum_{i=1}^{p} S_{ij}$
2. On calcule la moyenne des S'_j : $s = \frac{1}{j}\sum_{j=1}^{k} S'_j$
3. Les composantes saisonnières sont: $\boxed{S_j = \frac{S'_j}{s}}$

Exemple des coefficients saisonnières

Reprenons l'exemple des ventes trimestrielles de jus de fruits dans un grand magasin ont été, en milliers de litres, les suivantes:

	1er	2ème	3ème	4ème
1996	170	300	610	120
1997	250	410	790	190
1998	290	460	890	250
1999	450	550	1100	270
2000	320	600	1260	280

Calculer les coefficients saisonnières de cette série,

Les prévisions d'une série chronologiques

Les étapes à suivre:

- Détermination de la tendance Y_t,
- Détermination de composantes saisonnières S'_t,
- Les prévisions sont calculées par:
  - Pour le modèle additif: $\hat{Y}_t = T_t + S'_j$
  - Pour le modèle multiplicatif: $\hat{Y}_t = T_t \times S'_j$

Ici j est le reste de la division euclidienne de t par nombre de saisons dans la série,

Exemple:

Pour les trimètres, le nombre de saisons est 4, si on veut les prévisions à l'instants t=23. On a: 23=5*4+2 donc j=2

Exemple d'une série chronologiques

Les importations en produits maraîchers (Y_t), en milliers de tonnes, d'une région du Nord, sont données, en stock au premier jour de chaque trimestre, dans le tableau ci-dessous:

	Lundi	Mardi	Mercredi	Jeudi	Vendredi
Semaine 1	1	2	7	9	8
Semaine 2	2	3	11	12	9
Semaine 3	5	6	11	14	12

Calculer les prévisions pour t=13 et t=23

Exemple d’une série chronologique

Les importations en produits maraîchers (Y_t), en milliers de tonnes, d’une région du Nord, sont données, en stock au premier jour de chaque trimestre, dans le tableau ci-dessous:

Année	Tri 1	Tri 2	Tri 3	Tri 4
1986	1	2	7	9
1987	1	3	11	12
1988	5	6	11	14

Calculer les prévisions pour t=13 et t=23