

Statistique Descriptive

Introduction:

Définition: Ensemble des méthodes (des techniques) permettant d'analyser (traiter) des ensembles d'observations (des données).

Les outils utilisés en statistique :

- Mathématiques et outil informatique
- La statistique (science) \neq une statistique (ensemble de données chiffrées sur un sujet précis)
- Vocabulaire issu de la démographie (Exemple: population formée d'individu étudié selon un/plusieurs caractère(s))

Deux branches en statistique:

- **Statistique descriptive/exploratoire:** méthode dont l'objectif est la description des données étudiées.
Description en fonction de la présentation, la représentation graphique, calcul de résumé/caractéristique numérique.
La statistique descriptive ne prend pas en compte de probabilité
 - Classement
 - Organisation
 - Présentation
 - Analyse
- **Statistique inférentielle/mathématiques:** méthode dont l'objectif principal est de préciser un phénomène sur une population globale, à partir de son observation sur une partie restreinte de cette population.
 - Raisonnement par inférence (du particulier vers le général)
 - Décisions sur une population à partir d'un échantillon

Etapas:

1. Recueil, Organisation et traitement sommaire des informations
 - Objectif de l'étude
 - Sujet de collecte
 - Méthode de la collecte
2. Organisation des données sous forme exploitable.
 - ➔ Tableaux, graphiques, diagrammes...
3. Calcul des principales caractéristiques

Objectifs :

1. Acquérir une culture de base en statistique
2. Présenter un ensemble de méthode
3. Maîtriser les outils et techniques de base
4. Savoir choisir les outils adéquats pour pouvoir analyser et d'exploiter des données

Plan:

Chapitre 1 : Terminologie et concepts de base

Chapitre 2 : Distributions statistiques à un caractère

Chapitre 3 : Indices et taux de croissance

Chapitre 4 : Les distributions statistiques à deux caractères

Chapitre 1: Terminologie et concepts de base

I) Les données descriptives :

Elle cherche à décrire un ensemble d'éléments à partir de certaines de ses caractéristiques:

- Ensemble = **population**
- Éléments = **unités statistiques** ou **individus**
- Caractéristiques de ses éléments = **caractères**
- Caractères = **modalités**

1. Population et unité statistique

Au sens large du terme

- Population française: **personnes physiques**
- Ménages français: **ménages**
- Parc automobile française: **véhicules**
- Précipitations journalières dans une région donnée: **jours où il pleut**

2. Caractères

On cherche à révéler certaines caractéristiques de cette population.

- Age des individus de la population
- Revenu des ménages ou nombre de personnes qui compose un ménage
- Couleur ou puissance des véhicules
- Quantité de précipitations

3. Modalités

- Age: valeurs de 0 à 100 +
- Revenu des ménages: 500€ à 10000 ++ €
- Couleur des véhicules: rouge, vert...
- Puissance des véhicules: nombre de chevaux

→ Il existe 3 conditions :

1. **Incompatibles**: un même individu doit appartenir qu'à une seule modalité
Exemple: Une voiture ne pourra pas être rouge et verte à la fois
2. **Exhaustives**: un individu doit toujours pouvoir être classé
Modalités du type « Autres » ou « Divers »
3. **Sans ambiguïté**: éviter les erreurs de classement
A laquelle des modalités du caractère couleur appartiendraient une automobile blanche aux portières bleues ?
→ Les caractères peuvent être non mesurables

Caractère qualitatif (variable non mesurable) :

1. Un caractère est dit qualitatif si ses modalités ne sont pas mesurables.
 - Profession des individus
 - Etat matrimonial
 - Marque de leur voiture, etc.
2. Nomenclature: différentes rubriques qui représentent les différentes modalités du caractère.

Exemple de nomenclature:

- Agriculteurs, exploitants
- Artisans, commerçants et chef d'entreprise
- Cadres et professions intellectuelles supérieures
- Professions intermédiaires
- Employé
- Ouvriers
- Retraités
- Autres personnes sans activité professionnelle

Caractère quantitatif (variables mesurables) :

1. Un caractère est dit quantitatif si ses modalités sont mesurables.
2. Le caractère est appelé **variable statistique**:
 - **Une variable discrète** ne peut prendre qu'un nombre fini de valeurs dans un intervalle donné.
Exemple: le nombre de personnes composant un ménage.
 - **Une variable continue** peut prendre une infinité de valeurs dans un intervalle donné.
Exemple: l'âge d'un individu, la température extérieure, la vitesse.

→ Même si la variable n'est présentée qu'avec une ou deux décimales, elle reste une variable continue.

Répartition des communes en France métropolitaine

Tailles des communes	Effectifs (n_i)
Moins de 200	9649
200 à 499	10578
500 à 999	6959
1000 à 1999	4454
2000 à 4999	2992
5000 à 9999	1045
10000 ou plus	887
Ensemble	36564

Source : Insee-Recensement de la population en 2008

→ Caractère: taille des communes → quantitatif discret.

Tableau de données:

1. Série dite groupée: données regroupées par classes / Série non groupée: données présentées individuellement.
2. Amplitude d'une classe (a_i): écart entre ses deux bornes.
3. 1^{ère} classe: amplitude de 199
4. Extrémités de classe (e_i^-, e_i^+): min et max de la classe.
5. 2^{ème} classe: [200,500[ou [200,499[.
6. Centre de classe (c_i): amplitude de classe divisé par 2
- 7.

$$C_2 = \frac{499 - 200}{2} = 149,5$$

- Les extrémités de classe peuvent poser problème si elles ne sont pas définies correctement.
- Il ne s'agit que d'une moyenne théorique, les moyennes réelles seront calculées à partir des effectifs.
- Le centre de classe n'est que rarement utilisé comme moyenne.
- En général, on essaye de déterminer la borne supérieure et le centre de classe de cette dernière classe de façon a priori et arbitraire.

II) L'opérateur somme

- Somme de n éléments, x_1, x_2, \dots, x_n :

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

Exemple: $x_1=10, x_2=5, x_3=20$

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 = 10 + 20 + 5 = 35$$

Propriétés 1:

$$\sum_{i=1}^n a \cdot x_i = a \sum_{i=1}^n x_i$$

Exemple:

$$\sum_{i=1}^n a \cdot x_i = a \cdot x_1 + a \cdot x_2 + \dots + a \cdot x_n = a \cdot (x_1 + x_2 + \dots + x_n) = a \sum_{i=1}^n x_i$$

$$\sum_{i=1}^3 5 = 5 \cdot 10 + 5 \cdot 20 + 5 \cdot 5 = 75$$

Propriétés 2:

$$\sum_{i=1}^n xi^2 = x1^2 + x2^2 + \dots + xn^2$$

Exemple:

$$\sum_{i=1}^3 xi^2 = 10^2 + 5^2 + 20^2 = 525$$

Propriétés 3:

$$\sum_{i=1}^n xi \cdot yi = \sum_{i=1}^n xi \cdot \sum_{i=1}^n yi$$

Propriétés 4:

$$\sum_{i=1}^n a = a + a + \dots + a = n \cdot a$$

Exemples:

$$\sum_{i=1}^4 6 = 6 + 6 + 6 + 6 = 4 \cdot 6 = 24$$

Propriétés 5:

$$\sum_{i=k}^n a = (n - k + 1) \cdot a$$

Exemples:

$$\sum_{i=2}^4 6 = (4 - 2 + 1) \cdot 6 = 3 \cdot 6 = 18$$

Propriétés 6:

$$\sum_{i=k}^l xi + \sum_{i=l+1}^n xi = \sum_{i=k}^n xi$$

Exemple: $x_1=10, x_2=5, x_3=20, x_4=15$ et $x_5=2$

$$\sum_{i=1}^2 xi + \sum_{i=3}^5 xi = \sum_{i=1}^5 xi$$
$$(10+5) + (20+15+2) = (10+5+20+15+2)$$

➤ **Doubles sommes:**

$$\sum_{i=1}^n \sum_{j=1}^m xi . yj$$

Exemple:

Pour $n=3$ et $m=2$

$$\sum_{i=1}^3 \sum_{j=1}^2 xi . yj = x1 . y1 + x1 . y2 + x2 . y1 + x2 . y2 + x3 . y1 + x3 . y2$$

Exemple: $x_1=10, x_2=5, x_3=20, y_1=3$ et $y_2=4$

$$\sum_{i=1}^3 \sum_{j=1}^2 xi . yj = x1 . y1 + x1 . y2 + x2 . y1 + x2 . y2 + x3 . y1 + x3 . y2$$
$$= 10.3+10.4+5.3+20.3+20.4$$

Propriétés 1:

$$\sum_{i=1}^n \sum_{j=1}^m xi . yj = \sum_{i=1}^n \sum_{j=1}^m xi . yj \text{ ou } \left(\sum_{i=1}^n \sum_{j=1}^m yj . xi \right)$$

Exemple :

$$= \sum_{i=1}^2 \sum_{j=1}^3 xi . yi = x1 . y1 + x1 . y2 + x2 . y1 + x2 . y2 + x3 . y1 + x3 . y2$$

$$= \sum_{i=1}^3 \sum_{j=1}^2 xi . yi$$

Propriétés 2:

$$\sum_{i=1}^n \sum_{j=1}^m xi . yj = \sum_{i=1}^n xi . \sum_{j=1}^m yj$$

Exemple:

$$\sum_{i=1}^3 xi \sum_{j=1}^2 yj = (x1 + x2 + x3) . (y1 + y2)$$
$$= x1 . y1 + x1 . y2 + x2 . y1 + x2 . y2 + x3 . y1 + x3 . y2$$

$$= \sum_{i=1}^3 \sum_{j=1}^2 xi . yi$$

III) Effectif et fréquence

A partir des effectifs d'une population, on va calculer des fréquences pour lesquelles on ramène les effectifs en pourcentages.

2 types de fréquences:

- Fréquences relatives
- Fréquences cumulées/absolues

1. Effectifs et Fréquences Relatives/simples

A. Série statistique

1. n observations, p modalités
2. n_i : effectif de la modalité i
3. $n = n_1 + n_2 + \dots + n_p = \sum_{i=1}^p n_i$

B. Fréquence relative d'une modalité i: proportion d'unités de la population qui appartient à cette modalité.

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

C. $\sum_{i=1}^p f_i = \sum_{i=1}^p \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^p n_i (= n) = \frac{n}{n} = 1$

Exemple: répartition par tranches d'âge de la population française en 2009

Age	n_i en millions	f_i
0 à 19 ans	15,9	0,2465
20 à 39 ans	16,9	0,2620
40 à 59 ans	17,5	0,2713
60 à 74 ans	8,7	0,1348
75 ans et plus	5,5	0,0854
Total	64,5	1,0000

Source : Insee- Estimation de population au 1^{er} Janvier 2011

2. Effectifs et Fréquences Cumulées/croissantes/absolues

Définition: Fréquence cumulée (ou fréquence cumulée croissante ou ascendante) de la modalité i: pourcentage d'unités de la population qui vérifie la modalité i ou moins

$$F = \sum_{i=1}^i f_i = f_1 + f_2 + \dots + f_i$$

Exemple:

$$F_4 = \sum_{i=1}^4 f_i = f_1 + f_2 + f_3 + f_4$$

Age	n _i en millier	f _i	Fi
0 à 19 ans	15,9	0,2465	0,2465
20 à 39 ans	16,9	0,2620	0,5085
40 à 59 ans	17,5	0,2713	0,7798
60 à 74 ans	8,7	0,1348	0,9146
75 ans et plus	5,5	0,0854	1,0000
Total	64,5	1,0000	

Remarque sur les tableaux statistiques :

1. Le titre
2. Les intitulés des lignes et des colonnes
3. L'unité utilisée
4. La source

IV) Représentation graphique

Définition: Représenter un caractère qualitatif ou un caractère quantitatif qui ne comporte pas trop de modalités

Exemple: Effectifs des candidats inscrits au baccalauréat général par série 2011

Série	Effectif	f _i
Littéraire	55600	0,17
SES	104957	0,32
Scientifique	167228	0,51
Total	331575	1,00

Source : Ministère de l'Education Nationale, DEPP

1^{ère} modalité: les candidats ont passé un bac littéraire

2^{ème} modalité: les candidats ont passé un bac SES

3^{ème} modalité: les candidats ont passé un bac scientifique

Diagramme en tuyaux d'orgue

Tracer pour chaque modalité du caractère un rectangle:

1. Hauteur proportionnelle à la fréquence
2. Largeur identique pour chaque modalité

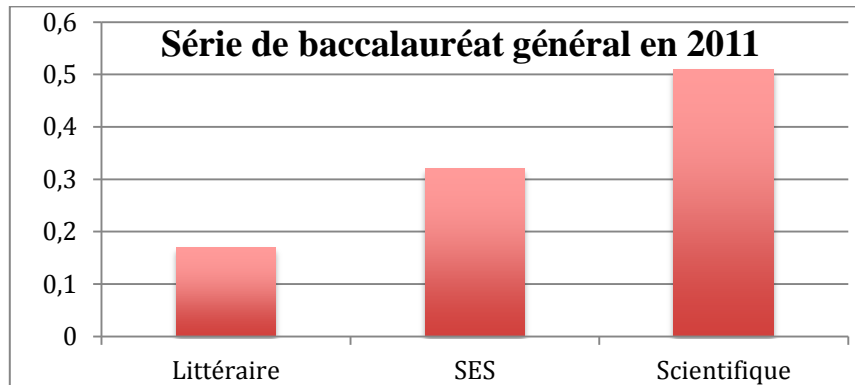
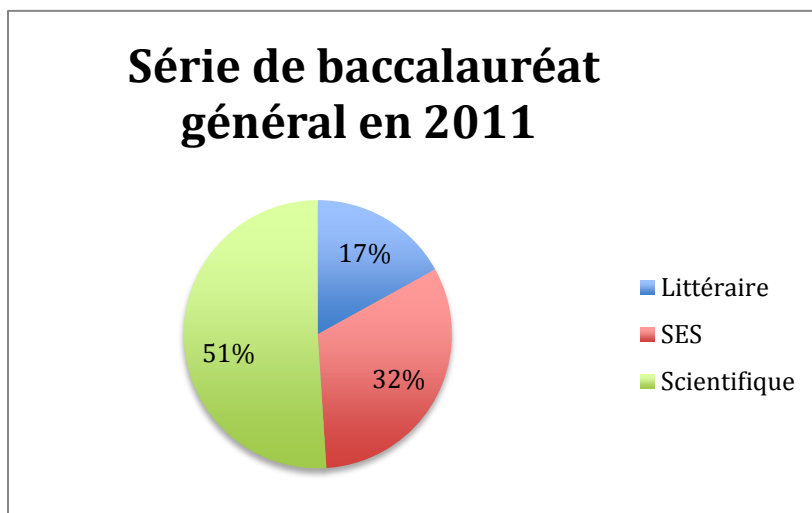


Diagramme circulaire:

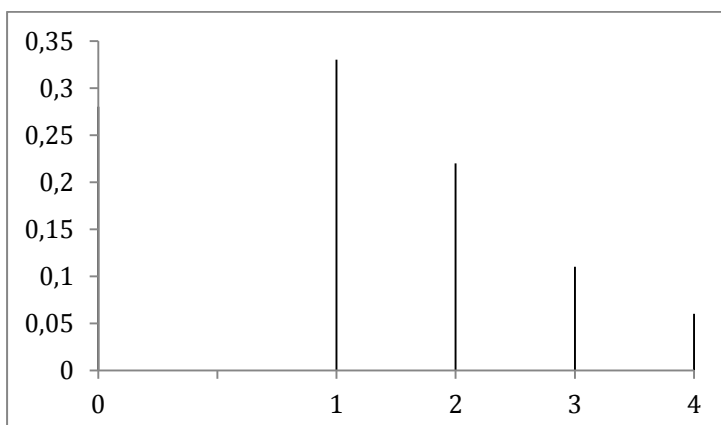
Aire de chaque secteur sera proportionnelle à la fréquence de la modalité qu'il représente

Diagramme en bâtons:

1. Représenter les fréquences d'un caractère quantitatif avec des variables discrètes
2. À chaque x_i , correspond un segment vertical de longueur proportionnelle aux effectifs ou fréquences



Série	Effectif	f_i
x_i	n_i	
0	50	0,28
1	60	0,33
2	40	0,22
3	20	0,11
4	10	0,06
Total	180	1,00



Les histogrammes:

1. Représenter les fréquences d'un caractère quantitatif dont les valeurs sont présentées par classe.
2. Variable continue mais aussi variable discrète s'étalant sur un intervalle large
3. Rectangles dont la surface est proportionnelle à la fréquence (et non pas à la hauteur comme avec un diagramme à tuyaux d'orgue)
4. Corriger la fréquence en fonction de l'amplitude de la classe

Pour construire un histogramme:

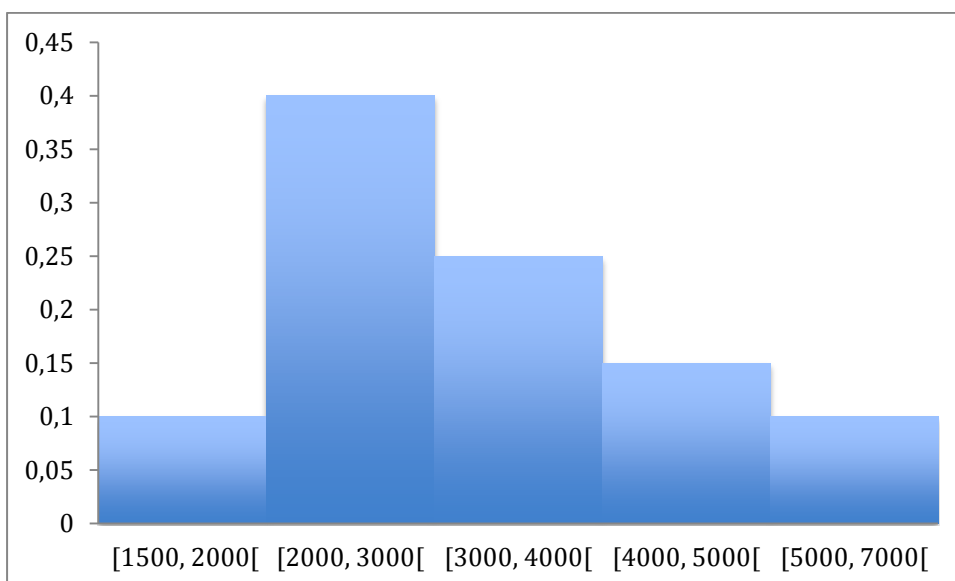
- On regarde si l'amplitude des classes est constante
- Si l'amplitude des classes n'est pas constante, on doit corriger cette amplitude
- Pour corriger ces fréquences, il faut choisir une unité d'amplitude (généralement on prendra comme unité d'amplitude, l'amplitude la plus fréquente mais également l'amplitude la plus petite)
- Une fois que cette amplitude est choisie, on va calculer le nombre d'unités d'amplitude par classe
- On va obtenir les fréquences corrigées en divisant les fréquences par le nombre d'unités d'amplitude par classe, on obtient des fréquences moyennes notées h_i

Exemple: Distribution des salaires mensuels nets dans une entreprise de 400 personnes

Salaires en euros	n_i	f_i	d_i	h_i	F(x)
[1500, 2000[40	0,1	0,5	0,2	0,1
[2000, 3000[160	0,4	1	0,4	0,5
[3000, 4000[100	0,25	1	0,25	0,75
[4000, 5000[60	0,15	1	0,15	0,9
[5000, 7000[40	0,1	2	0,05	1
Total	400	1,0			

Exemple: Unité de 1000 euros → $d_5 = 2$ unités

$h_5 = 0,1/2 = 0,05$



Courbe cumulative:

1. Fonctions de répartition: représente les fréquences cumulées
2. Caractères quantitatifs
 - Variable discrète: courbe en escalier

Exemple 1: Sondage sur le nombre de livres lus dans un mois

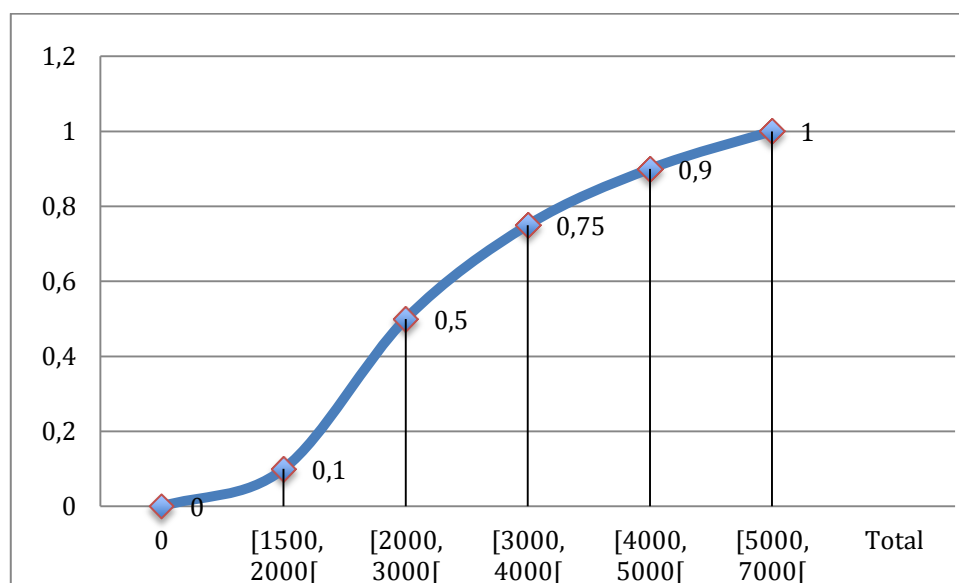
x_i	f_i	F_i
0	0,1	0,1
1	0,3	0,4
2	0,25	0,65
3	0,2	0,85
4	0,1	0,95
5	0,05	1,00

(Faire courbe cumulative en escalier)

➔ Puisque la variable x est continu, on peut lire la fonction pour n'importe quelle valeur/modalité.

Exemple 2: Quel pourcentage de personne ont un salaire inférieur à 3500€ ? 62,5% des salariés de mon entreprise ont un salaire inférieur à 3500€ par lecture graphique. Trouver comment faire une coordonné

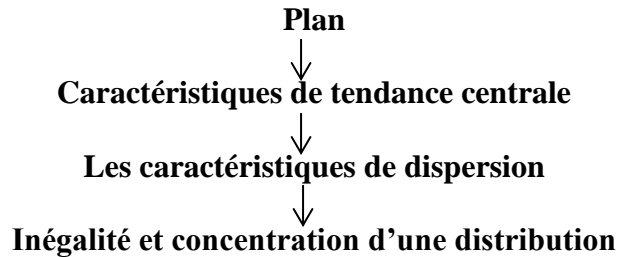
Salaire en Euros	n_i	f_i	$F(x)$
[1500,2000[40	0.1	0.1
[2000,3000[160	0.4	0.5
[3000,4000[100	0.25	0.75
[4000,5000[60	0.15	0.9
[5000,7000[40	0.1	1.00



Chapitre 2: Distributions statistiques à un caractère

Le premier chapitre nous a permis de comprendre comment repérer, ranger et présenter les données.

Grâce aux tableaux et aux graphiques, on obtient une première image de la distribution des données, nous allons maintenant aller plus loin en calculant les indicateurs.



Introduction:

- Une distribution caractéristique à un caractère est une série de données portant sur un seul caractère d'une population
Exemple: la distribution des revenus en France.
- Pour écrire la série et éventuellement pouvoir la comparer à une autre, on utilise des indicateurs :
 - **Indicateurs de tendance centrale** tentent de donner une indication du niveau de la variable
Exemple : moyenne
 - **Indicateurs de dispersion** tentent de mesurer l'étendue de la série et plus précisément la dispersion des valeurs autour de la tendance centrale
Exemple : Variance

I) Caractéristique de tendance centrale

Ces indicateurs sont utilisés pour donner une indication du niveau de la variable, on va calculer trois types d'indicateurs :

- le mode
- la médiane
- la moyenne

1. Le mode

- Le mode correspond à la modalité du caractère qui est la plus fréquente : M_0 , on parle également de valeurs dominantes.
On doit faire la distinction entre les caractères qualitatifs ou quantitatifs discret et des caractères quantitatifs continus

- **Caractère qualitatif**

Exemple: nombre de médailles de la France aux J.O. de Pékin, le caractère étudié est la médaille, il s'agit donc d'un caractère qualitatif avec trois modalités.

x_i (métal)	n_i	f_i
Or	7	0,175
Argent	16	0,4
Bronze	17	0,425
Total	40	1

$$M_0 = \{\text{Bronze}\}$$

➔ **Dans le cas d'une série à plusieurs modes, on parle d'une série plurimodale.**

Exemple: l'Italie a remporté 8 médailles d'Or, 10 d'argent et 10 de bronze.
On a donc une série plurimodale avec le bronze et l'argent.

- **Caractère quantitatif continu**

➔ Il faut prendre en compte l'amplitude de chaque classe pour éviter qu'une classe ne soit surreprésentée du fait de sa taille.

Exemple: Revenu annuel exprimé en milliers d'euros :

x_i	n_i	a_i	$\frac{a_i}{a}$	h_i
[10, 20[10	10	1	10
[20, 30[40	10	1	40
[30, 50[200	20	2	100
[50, 90[240	40	4	60
[90,100]	10	10	1	10
Total	500			

➔ Lorsque l'on a des variables continues, le mode correspond à un intervalle, le mode n'est pas égal à la modalité [50, 90[car cette classe est d'amplitude plus grande que les autres classes.

Pour pouvoir trouver le mode, on procède comme pour l'histogramme, on va corriger les effectifs en fonction des amplitudes de la classe.

➔ $a=10$, soit 10000€: calcul a_i , puis les effectifs corrigés h_i

$$M_0 = [30, 50[$$

➔ Le mode correspond à la modalité/à la classe ayant le plus grand nombre en termes d'effectifs corrigés

➔ Le mode est un indicateur simple, qui s'interprète facilement, il ne fournit que les informations partielles en ne s'intéressant qu'à la modalité la plus courante

2. La médiane

- La médiane est la valeur du caractère qui sépare en deux groupes de même effectif les observations classées par ordre croissant : M_e
- Il y'a donc la moitié de l'effectif de la population en dessous de la médiane et l'autre moitié au-dessus de la médiane
- La médiane ne se calcule uniquement que dans le cadre des variables quantitatives
- Il faut faire une distinction entre les variables quantitatives discrètes et variables quantitatives continus.

▪ Variables discrètes

- n impair: observation de rang $\frac{(n+1)}{2}$

Exemple: 9 personnes selon leur taille:

153 157 159 165 **167** 172 173 178 182

→ $M_e = 167$, 50% des individus ont une taille inférieure à 167 et vice versa

- n pair: modalités de rang $\frac{(n)}{2}$ et $\frac{(n)}{2} + 1$

Exemple: 10 personnes selon leur taille:

153 157 159 165 **167** **172** 173 178 182 184

→ $M_e = [167, 172[$, on parle d'intervalles médians

➔ Dans le cadre de données groupées, on va calculer les fréquences cumulées ou les effectifs cumulés

Exemple: Ménages français selon le nombre de personnes en 2009 :

Nombre de personnes	n_i	f_i	F_i
1	9080	0,333	0,333
2	8945	0,328	0,661
3	4036	0,148	0,809
4	3409	0,125	0,934
5 ou plus	1800	0,066	1,000
Total	27270	1,000	

Source : INSEE- Estimation de population au 1^{er} Janvier 2009

➔ On va regarder la modalité pour laquelle la fréquence cumulée dépasse les 50%

➔ $M_e =$ « 2 personnes par ménages »

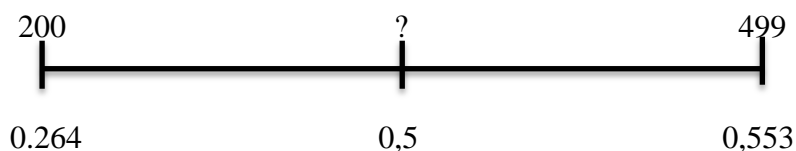
➔ Dans le cadre de variables continus, on peut déterminer la médiane de façon précise (ce qui n'est pas le cas pour les variables discrètes) par la méthode d'interpolation linéaire.

➔ Pour la calculer: on va déterminer dans un premier temps la classe médiane, c'est à dire la classe où passe la fréquence cumulée de 50% puis on va déterminer par interpolation linéaire la valeur de cette médiane.

Exemple: Répartition des communes suivant leur population en 2008
Cf. exemple 1, $M_e = [200, 499[$

Tailles des communes	Effectifs (n_i)	F_i	F_i
Moins de 200	9649	0,264	0,264
200 à 499	10578	0,284	0,553
500 à 999	6959	0,190	0,743
1000 à 1999	4454	0,132	0,845
2000 à 4999	2992	0,082	0,947
5000 à 9999	1045	0,028	0,975
10000 ou plus	887	0,025	1,000
Ensemble	36564	1,000	

Interpolation linéaire:



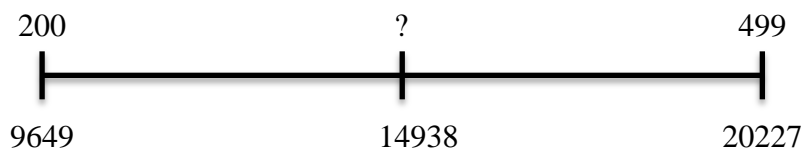
- En supposant que les effectifs soient uniformément répartis dans la classe [200, 499[, la médiane se situe proportionnellement à la même distance que la valeur 0,5 sur l'axe [0,264, 0,553].
- La valeur 0,5 se situe à une distance qui représente 81,66% de l'axe [0,264, 0,553]
- $M_e = 200 + \frac{0,5-0,264}{0,553-0,264} \times (499 - 200) = 444,1634$

→ $M_e = 445$ personnes

→ 50% des communes françaises ont moins de 445 habitants et 50% ont plus de 445 habitants.

➔ On peut aussi calculer la médiane à partir des effectifs cumulés:

- $n = 36564$ communes, $\frac{n}{2} = 18282$
- 9649 communes comptent moins de 200 habitants (borne inférieure de la classe)
- 20227 plus de 499 habitants (9649+10578)



$$\rightarrow M_e = 200 + \frac{18282-9649}{20227-9649} \times (499 - 200) = 444,0222$$

➔ La médiane n'est vraiment pertinente que dans le cas d'une variable continue ou alors d'une variable quantitative discrète avec un très grand nombre d'observations.

On ne peut pas calculer la médiane pour un caractère qualitatif et son interprétation pose problème dans le cas d'une variable discrète.

3. La moyenne arithmétique

- La moyenne se calcule uniquement pour des variables quantitatives.
- La moyenne arithmétique simple du caractère X, notée \bar{x} est égale à la somme des observations rapportée à l'effectif total :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Exemple: Notes obtenues par 5 étudiants lors d'un examen : 18 9 14 6 13

$$\bar{x} = \frac{1}{5} (18 + 9 + 14 + 6 + 13) = \frac{60}{5} = 12$$

➔ Dans le cas de données groupées, il faut calculer une moyenne arithmétique pondérée du caractère X, notée \bar{x} est égale à la somme pondérée des observations rapportées à l'effectif total :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \cdot x_i$$

Avec k le nombre de groupes et $n = n_1 + n_2 + \dots + n_i$

Exemple 1: Notes obtenues à un contrôle par un groupe de 50 étudiants :

I	x_i	n_i	$n_i \cdot x_i$
1	6	8	48
2	9	12	108
3	12	21	252
4	16	8	128
5	20	1	20
	Total	50	556

$$\bar{x} = \frac{556}{50} = 11,12$$

Exemple 2:

- Groupe de 10 personnes dont 2 ont un salaire mensuel de 2000€ et 8 un salaire mensuel de 3000€

- Revenu total de ces 10 individus: $2 \times 2000 + 8 \times 3000 = 28000€$

- Revenu moyen:

$$\bar{x} = \frac{1}{10} (2 \times 2000 + 8 \times 3000) = 2800$$

➔ L'unité que l'on retrouve derrière la valeur de la moyenne est la même que celle de la variable

➤ **Fréquences:**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \cdot x_i = \sum_{i=1}^k \frac{n_i}{n} \cdot x_i$$
$$\bar{x} = \sum_{i=1}^k f_i \cdot x_i$$

Exemple: $f_1 = 2/10 = 0,2$ et $f_2 = 8/10 = 0,8$

$$\bar{x} = 0,2 \times 2000 + 0,8 \times 3000 = 2800$$

➔ Dans le cas de variables continues, les valeurs sont regroupées par classe, on remplace des x_i par le centre de classe

➤ **Moyennes arithmétique pondérée:**

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \cdot c_i$$

Avec $c_i = \frac{\text{borne supérieure} + \text{borne inférieure}}{2}$

→ En fréquence:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i \cdot c_i$$

Exemple 1: Salaire dans une entreprise

Salaires en euros	n_i	c_i	$n_i \cdot c_i$
[1500, 2000[40	1750	70000
[2000, 3000[160	2500	400000
[3000, 4000[100	3500	350000
[4000, 5000[60	4500	270000
[5000, 7000[40	6000	240000
Total	400		1330000

$$\bar{x} = \frac{1330000}{400} = 3325$$

➔ La méthode pour calculer les moyennes arithmétiques peut conduire à des erreurs si les centres de classe ne correspondent pas aux moyennes réelles de chaque classe.

Exemple 2: Commune en France métropolitaine suivant leur population en 2008

Population	c_i	$n_i \cdot c_i$
1128543	100	964900
3444420	249,5	2639211
4907236	749,5	5215770,5
6211790	1499,5	66808723
9173250	3499,5	10470504
7188472	7499,5	7836977,5
30081252	50000	44350000
62134963		138286086

Source: Insee – Recensement de la population 2008

- Les 1ères et dernières classes posent problème:
 - pour la première classe (- de 200), on prend 100 comme centre de classe
 - pour la dernière classe, arbitrairement 10000 à 50000
- Nombre d'habitants moyen par commune à partir des centres de classe: $\bar{x} = \frac{138286086}{36564} = 3782,2 \rightarrow$ Moyennes arithmétique
La taille moyenne des communes françaises est de 3782,2
- Moyenne par les moyennes réelles de chaque classe :
 $\bar{x} = \frac{62134963}{36564} = 1699,4 \rightarrow$ Moyenne réelles
- Le regroupement des données sous forme de tableau fait perdre de l'information, certaines fois elle est négligeable, d'autres fois elle est importante.

Propriétés de la moyenne arithmétique :

Propriété 1:

Soit $y = ax+b$ avec a et b deux paramètres non nuls, alors : $\bar{y} = a\bar{x} + b$

Exemple: $y_i = 2x_i + 1$:

Matière	n_i	x_i	$n_i \cdot x_i$	$y_i = 2x_i + 1$	$n_i \cdot y_i$
Mathématique	2	5	10	11	22
Statistique	1	4	4	9	9
Economie	5	6	30	13	65
Total	8		44		96

$$\bar{x} = \frac{44}{8} = 5,5$$

$$\bar{y} = \frac{96}{8} = 12 = 2 \times 5,5 + 1$$

Propriété 2:

Soit $n_1+n_2=n$ et \bar{x}_1, \bar{x}_2 les moyennes arithmétiques de chaque sous-échantillon, alors:

$$\bar{x} = \frac{n_1}{n} \cdot \bar{x}_1 + \frac{n_2}{n} \cdot \bar{x}_2$$

Exemple: 300 copies

- 120 copies avec une moyenne de 9/20
- 180 copies avec une moyenne de 11,5/20
- $\bar{x} = \frac{120}{300} \times 9 + \frac{180}{300} \times 11,5 = 10,5$

Propriété 3:

Somme des écarts à la moyenne est nulle soit :

$$\sum_{i=1}^k n_i(x_i - \bar{x}) = 0$$

En effet:

$$\begin{aligned} \sum_{i=1}^k n_i(x_i - \bar{x}) &= \sum_{i=1}^k n_i \cdot x_i - \sum_{i=1}^k n_i \cdot \bar{x} \\ &= n \cdot \bar{x} - \bar{x} \sum_{i=1}^k n_i \\ &= n \cdot \bar{x} - n \cdot \bar{x} \\ &= 0 \end{aligned}$$

→ La moyenne arithmétique est un indicateur qui est facilement calculable et interprétable. La moyenne arithmétique est la plus fréquente mais elle ne s'applique pas à toutes les variables

4. La moyenne géométrique

La moyenne géométrique est très utilisée dans l'analyse de l'évolution d'une variable dans le temps.

Moyenne géométrique simple d'une variable x , notée G On doit distinguer les données groupées des données non groupées, lorsque l'on a des données non groupées on parle de moyenne géométrique simple et pour des données groupées on parle de moyenne géométrique pondérée.

Pour les données non groupées:

- Moyenne géométrique simple d'une variable x , notée $G = \text{racine } n^{\text{ième}} \text{ du produit des } n \text{ observations} : G = \sqrt[n]{x_1 \cdot x_2 \dots x_n} = (x_1 \cdot x_2 \dots x_n)^{1/n}$

Exemple: $x_1=8, x_2=12 : G = \sqrt[2]{12 \times 8} = 9,8$

Pour les données groupées :

➤ Moyenne géométrique pondérée du caractère x

$$G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k}} = (x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k})^{1/n}$$

$$G = (x_1^{f_1} \cdot x_2^{f_2} \dots x_k^{f_k}) \text{ avec } n = n_1 + n_2 + \dots + n_k$$

Exemple:

x_i	n_i
8	2
12	3
Total	5

$$G = \sqrt[5]{8^2 \times 12^3} = 10,20$$

5. La moyenne harmonique

La moyenne harmonique se calcule dans le cadre de rapport, surtout vitesse ou distance moyenne.

Moyenne non pondérée:

- Un véhicule parcourt un trajet allé à 60km/h et le trajet retour à 30km/h

$$\text{Vitesse moyenne} = \frac{\text{Distance parcourue}}{\text{temps de parcours}}$$

➔ Distance parcourue: $D \times 2$

➔ Temps de parcours (en heures):

$$\frac{D}{60} + \frac{D}{30}$$

➔ Au total, la vitesse moyenne est :

$$\text{Vitesse moyenne} = \frac{D \times 2}{\frac{D}{60} + \frac{D}{30}}$$

Soit encore:

$$\text{Vitesse moyenne} = \frac{2}{\frac{1}{60} + \frac{1}{30}} = \frac{2}{3/60} = 120/3 = \frac{40km}{h}$$

➔ Il s'agit ici d'une moyenne harmonique non pondérée dont la formule s'écrit :

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

Avec n, le nombre d'observations (ici le nombre de trajets de même distance) et x_i , la vitesse sur chaque trajet (ici $x_1=60$, $x_2=30$)

Moyenne pondérée:

$$H = \frac{n}{\sum_{i=1}^n \frac{n_i}{x_i}}$$

Exemple: 300 kilomètres, 100 à 50km/h et 200 à 80km/h:

$$H = \frac{300}{\frac{100}{50} + \frac{200}{80}} = 66,67 \text{ km/h}$$

6. La moyenne quadratique

On utilise les moyennes quadratiques pour calculer des moyennes d'écart ou des moyennes d'amplitudes.

Exemple: action perd 3€ un jour (100 à 97€), regagne 5€ le lendemain puis reperd 2€ le jour suivant

$$x_1 = -3, x_2 = 5, \text{ et } x_3 = -2$$

Si on a des variations négatives, et des variations positives, on n'utilise pas de moyenne arithmétique.

Quelle a été la variation moyenne journalière au cours de ces trois jours ?

Pour les données non pondérées:

Racine carrée de la moyenne arithmétique des carrés de la série

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

Exemple: $Q = \sqrt{\frac{1}{3} [(-3)^2 + 5^2 + (-2)^2]} = \sqrt{38/3} = 3,56$

Pour les données pondérées:

$$Q = \sqrt{\frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2}$$

Ou en fréquences: $Q = \sqrt{\sum_{i=1}^k f_i \cdot x_i^2}$

Exemple: Calcul des quatre moyennes

x_i	n_i
1	20
2	30
3	15
4	10
5	5
6	2
	82

$$\rightarrow \bar{x} = \frac{1}{82} [(20 \times 1) + (30 \times 2) + (15 \times 3) + (10 \times 4) + (5 \times 5) + (2 \times 6)] = 2,46$$

$$\rightarrow G = \sqrt[82]{1^{20} * 2^{30} * 3^{15} * 4^{10} * 5^5 * 6^2} = 2,15$$

$$\rightarrow H = \frac{82}{\frac{20}{1} + \frac{30}{2} + \frac{15}{3} + \frac{10}{4} + \frac{5}{5} + \frac{2}{6}} = 1,87$$

$$Q = \sqrt{\frac{1}{82} [(20 \times 1^2) + (30 \times 2^2) + (15 \times 3^2) + (10 \times 4^2) + (5 \times 5^2) + (2 \times 6^2)]}$$
$$= \sqrt{7,71} = 2,78$$

$$\underline{H < G < \bar{x} < Q}$$

→ On a calculé les quatre moyennes qui sont des indicateurs de tendance centrale, ces indicateurs de tendance centrale nous donnent une indication du milieu de la série mais ne nous disent rien sur la dispersion des valeurs autour de la tendance centrale, il n'indique pas entre autre si les variables prises par la variable sont proches ou éloignés de la tendance centrale.

Pour avoir cette information, nous allons calculer des indicateurs de dispersion.

II) Caractéristique de dispersion

Les indicateurs de tendance centrale comme la moyenne donnent une indication du milieu de la série mais ne disent rien sur la dispersion des valeurs autour de la tendance centrale.

Exemple: - $x_1 = 2000$ et $x_2 = 3000$ le revenu de deux personnes avec $\bar{x} = 2500$

→ Moyennes identiques pourtant la deuxième série est plus dispersée (ces valeurs sont davantage éloignées de la moyenne)

1. Les intervalles de variation

A. L'étendue

Différences entre la plus grande et la plus petite des valeurs observées de la variable
Exemple: notes obtenues par 5 étudiants à un examen:

2 8 12 14 18

→ Etendue de la série est égale à $18 - 2 = 16$

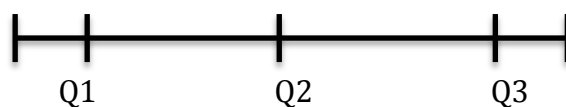
L'étendue est simple à calculer mais elle ne fournit qu'une information limitée sur le calcul de la variante.

B. Les intervalles de variation

Les quantiles sont les valeurs de la variable qui partagent la distribution classée par ordre croissant en plusieurs sous-ensembles de même effectif

Les quantiles les plus souvent utilisés sont les quartiles, les déciles et les centiles.

Les quartiles sont les valeurs de la variable qui partagent la série en quatre sous-ensembles de même effectif:



→ $Q1 = 25\%$ en dessous et 75% au-dessus

→ $Q3 = 75\%$ en dessous et 25% au-dessus

→ $Q2 =$ la médiane

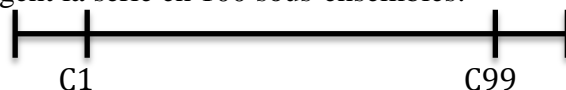
Les déciles partagent les observations classées en ordre croissant en 10 sous-ensembles



→ Ils sont au nombre

→ $D5$ correspond à la médiane

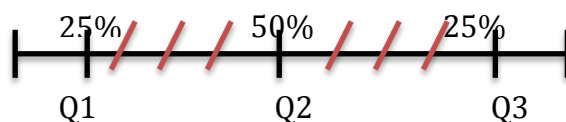
Les centiles partagent la série en 100 sous-ensembles:



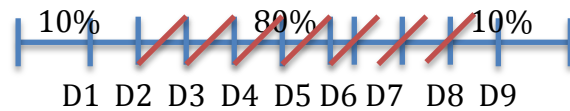
→ Ils sont au nombre de 99

→ $C50$ correspond à la médiane

Intervalles interquartiles $[Q1, Q3 [$:



Intervalles interdéciles [D1, D9 [



Intervalles intercentiles [C1, C99 [



Intervalles interquartiles : Exemple → Répartition des communes

$$Q_1 = 0 + \frac{0,25 - 0}{0,264 - 0} \times (200 - 0) = 189,39$$



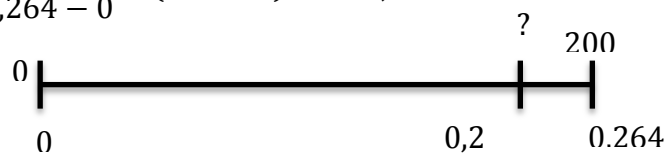
$$Q_3 = 1000 + \frac{0,75 - 0,743}{0,865 - 0,743} \times (1999 - 1000) = 1057,79$$



→ Intervalles interquartile [Q1 ; Q3 [= [189,1057[

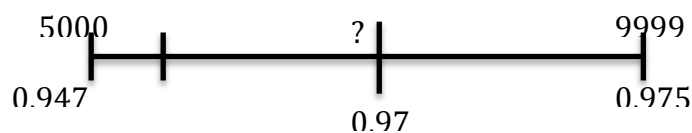
→ 75% des communes ont moins de 1057 habitants

$$D_2 = 0 + \frac{0,2 - 0}{0,264 - 0} \times (200 - 0) = 151,51$$



→ 20% des communes ont moins de 151 habitants

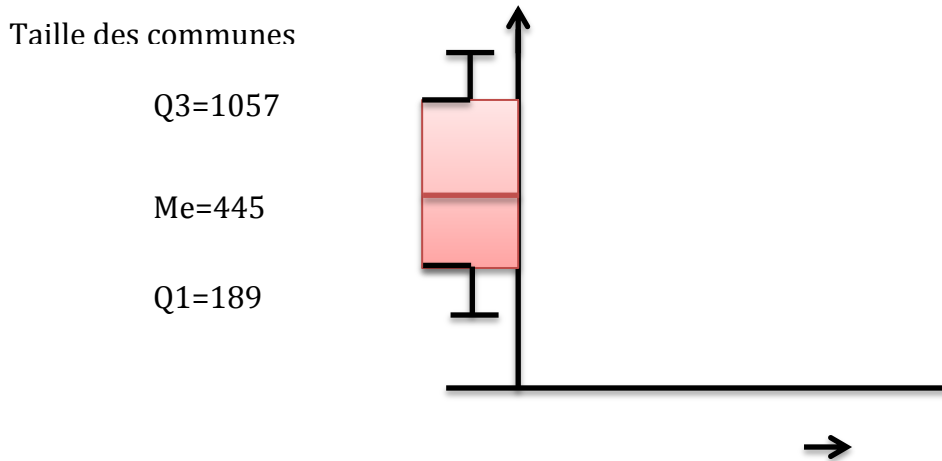
$$C_{97} = 5000 + \frac{0,97 - 0,947}{0,975 - 0,947} \times (9999 - 5000) = 9106,32$$



La boîte à moustache (TURKEY 1977):

- Graphique très simple qui résume la série à partir de ses valeurs extrêmes, de ses quartiles et de sa médiane.
- Echelle verticale, pour les observations de la variable (aucune échelle ne figure sur l'axe horizontal)
- La « boîte » est la partie du graphique comprise entre le premier et le troisième quartile
- La médiane, nécessairement située à l'intérieur de la boîte, est représentée par un trait horizontal
- Dans les parties basse et haute du graphique figurent les « moustaches », joignant le minimum au premier quartile (en bas) et le troisième quartile au maximum (en haut)

Exemple: Q1=189, Q2= 445=Me, Q3=1057



2. La variance et l'écart-type

La variance et l'écart-type sont très proches puisque l'écart-type est la racine de la variance.

C'est la moyenne arithmétique des carrés des écarts des valeurs de la variable à sa moyenne arithmétique : $V(x)$

Données non groupées:

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

Exemple: Prix d'un sac de 5 kg de pommes

I	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	6	-3	9
2	10	1	1
3	8	-1	1
4	12	3	9
	36	0	20

$$\bar{x} = \frac{36}{4} = 9\text{€}$$

$$V(x) = \frac{1}{4} \sum_{i=1}^4 (x_i - 9)^2 = \frac{20}{4} = 5\text{€}$$

$$\sigma = \sqrt{V(x)} = \sqrt{5} = 2,24\text{€}$$

→ La valeur de la variance ne s'interprète pas directement, on peut juste comparer les variances de plusieurs séries ayant les mêmes unités.

La série ayant la variance la plus élevée sera celle qui a la plus forte dispersion

→ Pour l'interprétation de l'écart-type, en moyenne l'écart des prix observés par rapport au prix moyen est de 2,24€

Données groupées:

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

Ou

$$V(x) = \sum_{i=1}^n f_i \cdot (x_i - \bar{x})^2$$

Exemple: Notes obtenues par 60 étudiants

I	x_i	n_i	$n_i \cdot x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i(x_i - \bar{x})^2$
1	4	6	24	-8	64	384
2	8	12	96	-4	16	192
3	12	26	312	0	0	0
4	18	16	288	6	36	576
	36	60	720			1152

$$\bar{x} = \frac{720}{60} = 12$$

$$V(x) = \frac{1152}{60} = 19,2$$

$$\sigma = \sqrt{V(x)} = \sqrt{19,2} = 4,38$$

A. La variance

Propriété 1:

a désigne un paramètre non-nul :

$$\begin{aligned}V(a \cdot x) &= a^2 \cdot V(x) \\ \sigma_y &= a \cdot \sigma_x\end{aligned}$$

Exemple: augmentation des notes de 10% :

$$\begin{aligned}V(y) &= 1,1^2 \cdot V(x) = 1,21 \cdot V(x) \text{ soit } V(y) = 23,23 \\ \sigma_y &= 1,1 \cdot \sigma_x \text{ soit } \sigma_y = 4,82\end{aligned}$$

→ La conséquence de cette propriété est que la variance dépend fortement de l'unité de mesure. Si on change l'unité de ma variable, la variance est alors fortement modifiée

Propriété 2:

Si a et b désignent des paramètres non nuls :

$$V(x + b) = V(x)$$

Exemple: Augmentation des notes de 1 point: variance et écart-type restent inchangés

De façon plus générale, en généralisant les deux propriétés suivantes :

$$V(a \cdot x + b) = a^2 \cdot V(x)$$

Propriété 3:

Formule développée de la variance :

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

Ou

$$V(x) = \sum_{i=1}^n f_i \cdot x_i^2 - \bar{x}^2$$

Démonstration :

$$\begin{aligned} V(x) &= \frac{1}{n} \sum_{i=1}^n n_i \cdot (x_i - \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2 + \frac{1}{n} \sum_{i=1}^n n_i \cdot \bar{x}^2 - 2 \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i \cdot \bar{x} \\ &= \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2 + \bar{x}^2 + \bar{x}^2 \frac{1}{n} \sum_{i=1}^n n_i - 2\bar{x} \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i \\ &= \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2 + \bar{x}^2 - 2\bar{x} \cdot \bar{x} \\ &= \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2 + \bar{x}^2 - 2\bar{x}^2 \\ V(x) &= \frac{1}{n} \sum_{i=1}^n n_i \cdot x_i^2 - \bar{x}^2 \end{aligned}$$

Propriété 4:

- \bar{x} tirée d'une série de n observations.
- La variance de la population totale est égale à la moyenne des variances des différentes sous-populations + la variance des moyennes des différentes sous population.

$$n_1 + n_2 = n ; \bar{x}_1 \text{ et } \bar{x}_2, V(x_1) \text{ et } V(x_2)$$

$$V(x) = \frac{1}{n} [n_1 V(x_1) + n_2 V(x_2)] + \frac{1}{n} [n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2]$$

$$V(x) = V(\bar{x}_1) + V(\bar{x}_2)$$

Coefficient de variation:

- Un des problèmes qui se posent entre la variance et l'écart-type vient du fait qu'ils dépendent des unités de mesure, ce qui empêche les comparaisons de variable exprimées dans différentes unités

- Coefficient de variation:

$$cv(x) = \frac{\sigma}{\bar{x}}$$

Exemple: Revenu moyen de 2500€ avec un écart-type de 500€:

$$cv(x) = \frac{500}{2500} = 0,2$$

Revenu moyen de 2000\$ avec un écart-type de 500\$:

$$cv(x) = \frac{500}{2000} = 0,25$$

→ Les revenus aux Etats-Unis sont plus dispersés qu'en France.

Chapitre 3: Indices et taux de croissance

Introduction :

Indices: nombre sans unités de mesure qui permet d'apprécier l'évolution d'une variable dans le temps et facilite les comparaisons

Exemple: PIB base 100 en 2010 = 116,7

- Indice élémentaire: nombre sans dimension qui retrace l'évolution d'une seule variable

Exemple: PIB, le chômage en France

- Indice synthétique: indice qui retrace l'évolution d'une variable calculée à partir de plusieurs autres variables statistiques

Exemple: l'indice des prix à la consommation

I) Indice élémentaire

Définitions :

- x une variable avec des valeurs dans le temps
 - x_t la valeur prise à la date t
- ➔ Indice élémentaire de la variable x à la date t base 100, noté $I_{t/0}$ (ou $I(x)_{t/0}$) :

$$I_{t/0} = \frac{x_t}{x_0}$$

Exemple: Nombre de créations d'entreprises par an en France entre 2003 et 2010

Année	2003	2004	2005	2006
Créations (en milliers)	239,7	268,4	270,3	285,7
Créations (en milliers)	321,5	331,4	580,1	622

➔ L'indice du nombre de créations en 2010 base 2003 vaut:

$$I_{2010/2003} = \frac{x_{2010}}{x_{2003}} \times 100 = \frac{622}{239,7} \times 100 = 259,49$$

➔ Entre 2003 et 2010, le nombre de créations d'entreprise a augmenté de 259,49

Année	2003	2004	2005	2006
Créations en milliers	100	112,0	112,8	119,1
Créations en milliers	134,1	138,2	242,0	259,5

A) Propriété:

Circularité: Cette propriété nous permet de comparer l'évolution d'une variable entre deux dates t et t' à partir de ces indices: $I_{t/0} = I_{t/t'} \times I_{t'/0} \times \frac{1}{100}$

Démonstration :

$$\begin{aligned} I_{t/0} &= \frac{x_t}{x_0} \cdot 100 \\ &= \frac{x_{t'}}{x_0} \cdot \frac{x_t}{x_{t'}} \cdot 100 \\ &= I_{t'/0} \times I_{t/t'} \times \frac{1}{100} \end{aligned}$$

Exemple: Indice du nombre de création en 2010 base 100 en 2004 et 2003

$$I_{2010/2004} = \frac{x_{2010}}{x_{2004}} \times 100 = \frac{622}{268.4} \times 100 = 231,7$$

Et

$$I_{2010/2003} = I_{2010/2004} \times I_{2004/2003} \times \frac{1}{100} = 259,5$$

Autre écriture:

$$I_{t/t'} = \frac{I_{t/0}}{I_{t'/0}} \cdot 100$$

Par exemple:

$$I_{2010/2006} = \frac{I_{2010/2003}}{I_{2006/2003}} \times 100 = \frac{259,5}{119,1} = 217,9$$

Et

$$I_{2010/2006} = \frac{x_{2010}}{x_{2006}} \times 100 = \frac{622}{285,7} \times 100 = 217,7$$

B) Propriété: Enchaînement des indices

$$I_{t/0} = \left[\frac{I_{t/(t-1)}}{100} \times \frac{I_{(t-1)/(t-2)}}{100} \times \dots \times \frac{I_{3/2}}{100} \times \frac{I_{2/1}}{100} \times \frac{I_{1/0}}{100} \right] \times 100$$

Exemple:

$$I_{2006/2003} = \left[\frac{I_{2006/2005}}{100} \times \frac{I_{2005/2004}}{100} \times \frac{I_{2004/2003}}{100} \right] \times 100$$

$$I_{2006/2003} = \left[\frac{105,7}{100} \times \frac{100,7}{100} \times \frac{112}{100} \right] \times 100 = 119,2$$

• Produit et division:

➤ **Définition produit:** On considère une variable x qui est le produit de deux autres variables: y et z soit $x = y \cdot z$

$$I_{t/0}(x) = \frac{I_{t/0}(y) \times I_{t/0}(z)}{100}$$

Exemple: Chiffres d'affaires d'une entreprise $CA = p \times q$

$$I_{2008/2006}(CA) = \frac{I_{2008/2006}(p) \times I_{2008/2006}(q)}{100}$$

$$I_{2008/2006}(CA) = \frac{105 \times 110}{100} = 115,5$$

➔ Mon CA entre 2006 et 2008 a augmenté de 15,5 %

➤ **Définition division:** On considère une variable x qui est le rapport de deux autres variables: y et z soit $x = y / z$

$$I_{t/0}(x) = \frac{I_{t/0}(y)}{I_{t/0}(z)} \times 100$$

C) Taux de croissance

- Le taux de croissance entre la date t et la date 0 d'une variable x s'écrit:

$$g = \frac{x_t - x_0}{x_0}$$

Exemple: Taux de croissance du nombre de créations entre 2003 et 2007:

$g = \frac{321,5 - 239,7}{239,7} = 0,341$ Entre 2003 et 2007, le nombre de création d'entreprise est augmenté de 34,1 %.

- On a aussi: $g = \frac{x_t - x_0}{x_0} = \frac{x_t}{x_0} - 1 = \frac{I_{t/0}}{100} - 1$

Remarque 1: Une augmentation d'une variable d'un certain pourcentage n'est pas compensée par une baisse du même pourcentage car les bases ne sont plus les mêmes.

Exemple: Revenu de 2000€ augmente la première année de 10% puis diminue l'année suivante de 10% :

- La première année : $2000 \times 1,1 = 2200 \text{ €}$
- L'année suivante : $2200 - 0,1 \times 2200 = 2200 \times 0,9 = 1980 \text{ €}$

Remarque 2: On appelle **multiplicateur** associé à un taux de croissance g , la valeur $1+g$

Exemple: $x_0 = 100$ et $g = 0,2$, alors $x_t = 100 \times 1,2 = 120$

Le diviseur $\frac{1}{1+g}$

Exemple : $x_0 = 200$, alors $x_t = 200 - 0,75 \times 200 = 50$

- Taux de croissance global: taux de croissance d'une variable sur plusieurs périodes, calculé à partir des différents taux de croissance sur chaque période.

Exemple:

- x_0 augmente au taux g_1 la 1^{ère} année, au taux g_2 la 2^{ème} année et au taux g_3 la 3^{ème} année.
 - Fin 1^{ère}: $x_1 = (1 + g_1) x_0$
 - Fin 2^{ème}: $x_2 = (1 + g_2) x_1 = (1 + g_1) \cdot (1 + g_2) x_0$
 - Fin 3^{ème}: $x_3 = (1 + g_3) x_2 = (1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3) x_0$
- Le taux de croissance global: $g = (1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3) - 1$
- $x_0 = 100$, $g_1 = 5 \%$, $g_2 = 3 \%$ et $g_3 = 8 \%$
 $x_3 = 100 \times 1,05 \times 1,03 \times 1,08$
 $= 100 \times 1,16802$
 $= 116,802 \%$
 $g = 0,16802 = 16,802 \%$

- Taux de croissance périodique moyen: est tel que tout se passe comme si la variable avait évolué en moyenne à ce taux à chaque période.
 - Variable augmente sur 3 périodes aux taux respectifs g_1 , g_2 et g_3 , au taux global g , donc le taux de croissance périodique moyen \bar{g}
 $x_0 = (1 + \bar{g})^3 = (1 + g) x_0$
 $\Leftrightarrow \bar{g} = \sqrt[3]{(1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3)} - 1$
 $\Leftrightarrow \bar{g} = [(1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3)]^{1/3} - 1$
 - Le multiplicateur $(1 + \bar{g})$ est la moyenne géométrique des multiplicateurs

- Taux moyen / Taux global: $(1 + g) = (1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3)$
Donc $\bar{g} = \sqrt[3]{(1 + g)} - 1 = (1 + g)^{1/3} - 1$
Ou de façon générale: $\bar{g} = (1 + g)^{1/t} - 1$

Exemple: nombre de créations d'entreprises par an en France entre 2003 et 2007

- Le taux de croissance global entre 2003 et 2007 :

$$g = \frac{321,5 - 239,7}{239,7} = 0,341$$

- Le taux de croissance annuel moyen :

$$\bar{g} = (1 + 0,341)^{1/4} - 1 = 0,0761$$

→ En moyenne le nombre de création d'entreprise a augmenté de 7,61 % par an.

- Taux moyen / moyenne géométrique: on a vu dans le chapitre 2 que la moyenne géométrique était utilisée pour calculer des taux de croissance moyen, on avait distingué le cas des données groupées et non groupées.

Rappel: $G = \sqrt[n]{x_1 \cdot x_2 \dots x_n} = (x_1 \cdot x_2 \dots x_n)^{1/n}$

➤ Données non groupées:

Exemple: action augmente de 6 % la première année, de 4 % la deuxième année et de 8 % la troisième

- Taux de croissance global: $g = (1 + g_1) \cdot (1 + g_2) \cdot (1 + g_3) - 1$

- Taux de croissance annuel moyen:

$$(1 + \bar{g})^3 = 1,06 \times 1,04 \times 1,08$$

$$(1 + \bar{g})^3 = \sqrt[3]{1,06 \cdot 1,04 \cdot 1,08}$$

$$\bar{g} = \sqrt[3]{1,06 \cdot 1,04 \cdot 1,08} - 1 = 0,0599$$

→ En moyenne, l'action augmente de 5,99 % par an

→ $\bar{g} = G - 1$ Avec G la moyenne géométrique simple des multiplicateurs.

➤ Données groupées:

Rappel: $G = \sqrt[n]{x_1^n \cdot x_2^n \dots x_k^n} = (x_1^{n_1} \cdot x_2^{n_2} \dots x_k^{n_k})^{1/n}$

Exemple: bénéfices augmentent de 5 % par an les deux premières années, de 9 % par an les cinq suivantes et de 12 % par an les trois suivantes.

$$(1 + \bar{g})^{10} = 1,05^2 \times 1,09^5 \times 1,12^3$$

$$1 + \bar{g} = (1,05^2 \times 1,09^5 \times 1,12^3)^{1/10}$$

$$\bar{g} = (1,05^2 \times 1,09^5 \times 1,12^3)^{1/10} - 1 = 2,383^{1/10} - 1 = 0,0907$$

→ En moyenne, le bénéfice de ma société ont eu un taux de croissance par année est 9,07 %

→ $\bar{g} = G - 1$ Avec G la moyenne géométrique pondérée des multiplicateurs des taux de croissance annuels.

II) Indice synthétique

- Il vise à représenter l'évolution d'une grandeur complexe déterminée à partir de plusieurs variables.
- Les indices synthétiques sont des moyennes d'indices élémentaires.

Exemple:

- **CAC 40:** vise à retracer l'évolution de la valeur d'un portefeuille composé des 40 plus grandes sociétés françaises (par leur capitalisation boursière).
Si on calcule pour chaque entreprise l'évolution du cours de leur action, on calculera un indice élémentaire.
- Indice des prix de l'ensemble des biens consommés par un ménage (indice des prix à la consommation).

→ Si on s'intéresse à l'évolution du prix d'un seul bien, on a un indice élémentaire. Au contraire, si on s'intéresse à l'évolution du prix des biens consommés par un ménage, on aura alors un indice synthétique.

- On va calculer 4 indices différents:
 - L'indice de valeur
 - L'indice de Laspeyres
 - L'indice de Paasche
 - L'indice de Fisher

Exemple:

p_0^i le prix du bien i à la date 0 / P_t^i le prix du bien i à la date t

q_0^i la quantité de bien i consommée à la date 0 / q_t^i la quantité consommée à la date t

	0		T	
i	p	Q	P	q
1	2	14	3	16
2	8	12	6	15
3	4	20	5	22

→ On va calculer les 4 indices à partir de cet exemple :

- **L'indice de valeur**
 - L'indice de valeur mesure l'évolution de la valeur de la consommation entre les dates 0 et t
 - Il s'écrit s'il y'a n biens :

$$I_{t/0}(pq) = \frac{\sum_{i=1}^n P_t^i \cdot q_t^i}{\sum_{i=1}^n P_0^i \cdot q_0^i} \times 100$$

→ Le numérateur représente la valeur de la consommation à la date t et le dénominateur la valeur de la consommation à la date 0.

Exemple:

$$I_{t/0}(pq) = \frac{\sum_{i=1}^3 P_t^i \cdot q_t^i}{\sum_{i=1}^3 P_0^i \cdot q_0^i} \times 100$$

$$I_{t/0}(pq) = \frac{3.16 + 6.15 + 5.22}{2.14 + 8.12 + 4.20} \times 100 = \frac{248}{204} \times 100 = 121,6$$

→ Entre la période initiale et la période finale, les dépenses de consommation ont augmenté de 21,6 %.

- Un indice de valeur présente un intérêt limité car il résulte des variations combinées des prix et des quantités.
- Préférable de mesurer distinctement la variation des prix et la variation des quantités
- Les prix ne peuvent être additionnés ni les quantités
- Calcul indice représentatif de l'évolution du pouvoir d'achat: Résultat différent si l'on considère le poids à la date de référence et le poids à la date finale ? On va pondérer l'évolution des prix en fonction du poids de chaque produit dans la consommation des ménages
- Indices de Paasche (prend la période finale comme pondération), Laspeyres (prend en compte la période initiale pour la pondération), et Fischer (fait une moyenne des deux autres indices).
Pour chaque indice, on va calculer un indice des prix mais aussi un indice des quantités.

→ On calcule trois autres indices car le choix de la période référence pour pondérer nos prix pose problème.

• L'indice de Laspeyres

→ Indice des prix

- Lorsque l'on calcule un indice des prix, on suppose que les quantités sont fixées à la période initiale
- Ainsi, quantités constantes et égales à leur valeur initiales (q_0^i)
- Valeurs de la consommation:

- ❖ A la date de base: $\sum_{i=1}^n P_0^i \cdot q_0^i$

- ❖ A la date t: $\sum_{i=1}^n P_t^i \cdot q_0^i$

- L'indice: $L_{t/0}^p = \frac{\sum_{i=1}^n P_t^i \cdot q_0^i}{\sum_{i=1}^n P_0^i \cdot q_0^i} \times 100$

Exemple: $L_{t/0}^p = \frac{3.14 + 6.12 + 5.20}{2.14 + 8.12 + 4.20} \times 100 = \frac{214}{204} \times 100 = 104,9$

→ Les prix ont évolué, entre ma date initiale et ma date finale, selon Laspeyres de 4,9 %.

→ Lien entre les indices de Laspeyres et les indices élémentaires

Un indice élémentaire du prix d'un bien i: $I_{t/0} = \frac{P_t}{P_0} \times 100$

L'indice de laspeyres est une moyenne pondérée des indices élémentaires de prix:

$$L_{t/0}^p = \frac{\sum_{i=1}^n P_0^i \cdot q_0^i \cdot \frac{P_t^i}{P_0^i}}{\sum_{i=1}^n P_0^i \cdot q_0^i} \times 100 \Rightarrow L_{t/0}^p = \sum_i^n \left(\frac{P_0^i \cdot q_0^i}{\sum_{i=1}^n P_0^i \cdot q_0^i} \right) \times \frac{P_t^i}{P_0^i} \times 100$$

$$L_{t/0}^p = \sum_i^n c_0^i \cdot I_{t/0}^i$$

$$\text{Avec } c_0^i = \frac{P_0^i \cdot q_0^i}{\sum_{i=1}^n P_0^i \cdot q_0^i}$$

→ C représente un coefficient budgétaire, c'est à dire la part de la dépense en bien i dans la dépense totale.

$$\rightarrow c_0^1 = \frac{28}{204} = 0,137 ; c_0^2 = \frac{28}{204} = 0,471 ; c_0^3 = \frac{28}{204} = 0,392$$

$$\rightarrow I_{t/0}^1 = \left(\frac{3}{2}\right) \times 100 = 150 ; I_{t/0}^2 = 75 ; I_{t/0}^3 = 125$$

$$\rightarrow \text{On peut alors vérifier que: } L_{t/0}^p = 0,137 \times 150 + 0,392 \times 125 + 0,471 \times 75 = 104,9$$

→ Indice des quantités

- Prix sont constants et prix à la date 0 qui sont fixes

- L'indice s'écrit: $L_{t/0}^q = \frac{\sum_{i=1}^n P_0^i \cdot q_t^i}{\sum_{i=1}^n P_0^i \cdot q_0^i} \times 100$

$$\text{Exemple: } L_{t/0}^q = \frac{2.16 + 8.15 + 4.22}{2.14 + 8.12 + 4.20} \times 100 = \frac{240}{204} \times 100 = 117,7$$

→ Les quantités entre la date initiale et la date finale ont augmentés de 17,7%, toujours selon Laspeyres.

• L'indice de Paasche

→ Indice des prix

- Les quantités sont fixes à la période t

- Il s'écrit $p_{t/0}^p = \frac{\sum_{i=1}^n P_t^i \cdot q_t^i}{\sum_{i=1}^n P_0^i \cdot q_t^i} \times 100$

$$\text{Exemple: } p_{t/0}^p = \frac{3.16 + 6.15 + 5.22}{3.14 + 6.12 + 5.20} \times 100 = \frac{248}{214} \times 100 = 115,9$$

→ Selon Paasche, les quantités consommées ont augmenté de 15,9 %

→ Pour un indice des prix fixes, on fait varier les quantités et vice versa

- **L'indice de Fischer**

Il n'y aucune raison objective de choisir entre l'indice de Laspeyres et l'indice de Paasche.

L'indice de Laspeyres a tendance à surestimer les résultats, alors que l'indice de Paasche les sous- estime.

L'indice de Fischer va donc faire une moyenne des deux indices, plus exactement il va faire une moyenne géométrique des indices.

→ **Indice des prix**

-
$$F_{t/0}^p = \sqrt{L_{t/0}^p \cdot P_{t/0}^p}$$

Exemple: $F_{t/0}^p = \sqrt{104,9 \cdot 103,3} = 104,1$

→ Les prix pendant ma période ont augmenté de 4,1 % selon Fischer.

- De manière générale: $p_{t/0}^p \leq F_{t/0}^p \leq L_{t/0}^p$

→ **L'indice des quantités**

- Il s'écrit:
$$F_{t/0}^q = \sqrt{L_{t/0}^q \cdot P_{t/0}^q}$$

Exemple: $F_{t/0}^q = \sqrt{117,7 \cdot 115,9} = 116,8$

→ Les quantités pendant ma période ont augmenté de 16,8 % selon Fischer.

- De manière générale: $p_{t/0}^q \leq F_{t/0}^q \leq L_{t/0}^q$

Chapitre 4: Les distributions statistiques à deux caractères

- **Introduction:**

- Une distribution à deux caractères est une série portant sur deux caractères qui peuvent prendre chacun plusieurs modalités.

Exemple:

- La distribution des revenus selon l'âge (quantitatif)
 - La distribution des revenus selon la catégorie socio-professionnelle (qualitatif)
- On représente ces distributions dans des tableaux à double entrée que l'on appelle tableau de contingence.
 - La plupart du temps, on s'intéresse à des distributions avec deux caractères quantitatifs, on peut trouver des distributions avec un caractère qualitatif et un caractère quantitatif mais ce n'est pas intéressant.

I) Les tableaux de contingence

Exemple: revenus selon l'âge dans une population de 200 personnes

x_i	y_j	[1000, 2000[[2000, 4000[[4000, 6000[
[20, 40[40	30	10	80
[40, 60[20	60	40	120
		60	90	50	200

A) **Effectifs**

- On considère une série statistique à deux caractères x et y
 - Le caractère x peut prendre p modalités $x_1, x_2, x_3, \dots, x_p$
 - Le caractère y peut prendre q modalités $y_1, y_2, y_3, \dots, y_p$
- Selon l'exemple: $p = 2$: [20, 40[et [40, 60[et $q = 3$: [1000, 2000[, [2000, 4000[, [4000, 6000[

- Tableau de contingence:

	y _j	y ₁	y ₂	...	y _j	...	y _q	n _i
x _i								
x ₁		n ₁₁	n ₁₂	...	n _{1j}	...	n _{1q}	n _{1.}
x ₂		n ₂₁	n ₂₂	...	n _{2j}	...	n _{2q}	n _{2.}
...								...
x _i		n _{ij}	...	n _{iq}	n _{i.}
...								...
x _p		n _{p1}	n _{p2}		...		n _{pq}	n _{p.}
n _j		n _{.1}	n _{.2}	...	n _{.j}	...	n _{.q}	n _{..}

Modalité de x

Effectif total

Modalité de y

Effectif croisé

Effectifs marginaux

- **Effectif croisé**

L'effectif n_{ij}, appelé effectif croisé (ou partiel), désigne l'effectif vérifiant à la fois la modalité i du caractère x et la modalité j du caractère y, c'est à dire l'effectif à l'intersection de la ligne i et la colonne j.

Exemple: n₁₂ = 30 signifie que 30 individus ont entre 20 et 40 ans (modalité 1) et gagnent entre 2000 et 4000 €.

- **Effectifs marginaux**

→ Effectif marginal de x

- L'effectif total de la modalité i du caractère x, appelée effectif marginal du caractère x, est noté n_{i.}

Exemple: n_{1.} = 80, ce qui signifie que parmi la population 80 individus ont entre 20 et 40 ans.

- Formule générale: $n_{i.} = \sum_{j=1}^q n_{ij}$

Exemple: n_{1.} = n₁₁ + n₁₂ + ... + n_{1q}

$$n_{1.} = n_{11} + n_{12} + n_{13} = 40 + 30 + 10 = 80$$

→ Effectif marginal de y

- L'effectif total de la modalité j du caractère y, appelé effectif marginal du caractère y, est noté n_{.j}

Exemple: n_{.2} = 90 signifie que 90 personnes gagnent entre 2000 et 4000€.

- Formule générale: $n_{.j} = \sum_{i=1}^p n_{ij}$

Exemple: n_{.j} = n_{1j} + n_{2j} + ... + n_{pj}

- **Effectif total**

- 1^{ère} façon de le calculer: somme de tous les effectifs croisés:

$$\mathbf{n}_{..} = \sum_{i=1}^p \sum_{j=1}^q \mathbf{n}_{ij}$$

- 2^{ème} façon de le calculer: somme des effectifs marginaux de x :

$$\mathbf{n}_{..} = \sum_{i=1}^p \mathbf{n}_i$$

Exemple: $n_{..} = n_{1.} + n_{2.} = 80 + 120 = 200$

- 3^{ème} façon de le calculer: somme des effectifs marginaux de y :

$$\mathbf{n}_{..} = \sum_{j=1}^q \mathbf{n}_{.j}$$

Exemple: $n_{..} = n_{.1} + n_{.2} + n_{.3} = 60 + 90 + 50 = 200$

B) **Les fréquences**

➤ Comme pour toute distribution statistique, on peut calculer à partir des effectifs des fréquences, ici on va calculer trois types de fréquences

- Des fréquences croisées
- Des fréquences marginales pour x et y
- Des fréquences conditionnelles

- **Fréquences croisées**

- Correspondent au rapport de l'effectif croisé (ou partiel) sur l'effectif total:

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

- La fréquence croisée va mesurer la proportion des individus qui vérifie à la fois la modalité i du caractère x et la modalité j du caractère y

Exemple:

$$f_{12} = \frac{n_{12}}{n_{..}} = \frac{30}{200} = 0,15$$

➔ 15% des individus de ma population ont entre 20 et 40 ans et gagnent entre 2000 et 4000 €.

- Somme des fréquences:

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

- **Fréquences marginales**

- La distribution marginale d'un caractère est la distribution des effectifs d'un seul caractère.

Exemple: la distribution des individus de la population selon l'âge, quel que soit leur revenu, est la distribution marginale du caractère x

X_i	$n_{i.}$	F_i
[20, 40[80	0,40
[40, 60[120	0,60
$n_{..}$	200	1

- A partir de cette distribution marginale, on peut calculer les fréquences marginales, on obtient les fréquences marginales en faisant le rapport entre effectif marginal de x et effectif total.

- Pour le caractère x : $f_i = \frac{n_i}{n_{..}}$

- $f_{1.} = \frac{n_{1.}}{n_{..}} = \frac{80}{200} = 0,4$
→ 40 % ont entre 20 et 40 ans.

- $f_{2.} = \frac{n_{2.}}{n_{..}} = \frac{120}{200} = 0,6$
→ 60 % ont entre 40 et 60 ans.

- Pour le caractère y : $f_{.j} = \frac{n_{.j}}{n_{..}}$

- $f_{.3} = \frac{n_{.3}}{n_{..}} = \frac{50}{200} = 0,25$
→ 25 % des personnes de ma population ont un salaire entre 4000 et 6000€.

- La somme des fréquences:

$$\sum_{i=1}^p f_{i.} = 1 \text{ et } \sum_{j=1}^q f_{.j} = 1$$

On résume les résultats trouvés dans ce tableau:

	y_i	[1000, 2000[[2000, 4000[[4000, 6000[f_i
x_i					
[20, 40[0,20	0,15	0,05	0,40
[40, 60[0,10	0,30	0,20	0,60
$f_{.j}$		0,3	0,45	0,25	1

	y_j	y_1	y_2	...	y_j	...	y_q	f_i
x_i								
x_1		f_{11}	f_{12}	...	f_{1j}	...	f_{1q}	$f_{1.}$
x_2		f_{21}	f_{22}	...	f_{2j}	...	f_{2q}	$f_{2.}$
...								...
x_i		f_{ij}	...	f_{iq}	$f_{i.}$
...								...
x_p		f_{p1}	f_{p2}	f_{pq}	$f_{p.}$
n_j		$f_{.1}$	$f_{.2}$...	$f_{.j}$...	$f_{.q}$	1

• **Fréquences conditionnelles**

- Les distributions conditionnelles d'un caractère sont les distributions des effectifs selon les modalités de ce caractère, pour une valeur donnée de l'autre caractère

Exemple: la répartition par âge des individus gagnant entre 4000 et 6000€ est la distribution conditionnelle de x pour $y = y_3$

X_i	$n_{j/ (j=3)}$	$f_{j/ (j=3)}$
[20, 40[10	0,20
[40, 60[40	0,80
$n_{.3}$	50	1

→ On aura autant de distribution conditionnelle de x sachant y que de modalité q , et à l'inverse on aura autant de distribution conditionnelle de y sachant x que de modalité p

- Fréquences conditionnelles de x sachant y

- De façon générale, si $y = y_j$:

$$f_{i \over j} = \frac{n_{ij}}{n_j}$$

Exemple: sachant que mes individus gagnent entre 4000 et 6000 €, 20 % sont âgés de 20 à 40 ans.

$$f_{(i=1)/(j=3)} = \frac{n_{13}}{n_{.3}} = \frac{10}{50} = 0,20$$

Exemple: 80 % des individus qui gagnent entre 4000 € et 6000 €, ($j = 3$) ont entre 40 et 60 ans ($i = 2$): $f_{(i=2)/(j=3)} = \frac{n_{23}}{n_{.3}} = \frac{40}{50} = 0,80$

Exemple: 33,3 % des personnes gagnant entre 1000 € et 2000 €, ont entre 40 et 60 ans: $f_{(i=2)/(j=1)} = \frac{n_{21}}{n_{.1}} = \frac{20}{60} = 0,333$

Moyennes marginales : $1/n_{.} \cdot \text{sigma } (i=1 \text{ à } p) n_{i.} \cdot x_i$ (ou $f_{i.} \cdot x_i$)

Variance marginale : $V(x) = 1/n_{.} \cdot \text{sigma } (i=1 \text{ à } p) n_{i.} (x_i - \bar{x})^2$ (ou $n_{i.} (x_i - \bar{x})^2$)

Même formule pour ses formules en y

B. Lois conditionnelles

Il y a q moyennes conditionnelles de x selon y :

$$\bar{x}_{j} = 1/n_{.j} \cdot \text{sigma } n_{ij} \cdot x_i$$

C. Moyenne conditionnelle

Il y a p moyennes conditionnelles de y selon x :

$$\bar{y}_{i} = 1/n_{i.} \cdot \text{Sigma } n_{ij} \cdot y_j$$

D. La variance conditionnelle de x

$$V_j(x) = 1/n_{.j} \cdot \text{sigma } n_{ij} (x_i - \bar{x}_j)^2$$

E. La variance conditionnelle de y

$$V_i(y) = 1/n_{i.} \cdot \text{Sigma } n_{ij} (y_j - \bar{y}_i)^2$$

La covariance :

La covariance est un indicateur qui permet de mesurer la dépendance linéaire entre deux variables :

$$\text{Cov}(x,y) = 1/n_{.} \cdot \text{sigma } (i=1 \text{ à } p) \cdot \text{sigma } (j=1 \text{ à } q) n_{ij} (x_i - \bar{x}) (y_j - \bar{y})$$

Les courbes de régressions :

La dépendance est totale quand $C_{x/y}$ et $C_{y/x}$ sont confondus. La liaison est relative quand il y a une corrélation positive quand les variantes se produisent dans le même sens (variation positive), ou quand il y a une corrélation négative quand les variations se produisent en sens opposé (courbe décroissante).

Si les courbes de régression sont des droites on dit que la corrélation est linéaire sauf si ces droites sont parallèles aux axes (indépendance totale dans ce cas).

L'ajustement linéaire :

L'objectif de l'ajustement linéaire est destiné à partir d'un échantillon donné, à estimer les paramètres d'une droite qui relie deux variables x et y. La méthode utilisée est celle des moindres carrés. A et \hat{b} sont les estimations des paramètres a et b.

Par Julien Nardelli et diffuser par Cours-walid.fr

Le calcul des estimations de la variable y est $\hat{y} = \hat{a}x + \hat{b}$ qui est la différence entre la valeur observé de y et son estimation par les moindres carrés. Cette méthode est tel que la quantité : $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ est minimale.

Formule : $\hat{a} = \text{Cov}(x ; y) / V(x)$ et $\hat{b} = \bar{y} - \hat{a} \cdot \bar{x}$

$\bar{x} = \sum x_i / \text{variance}$ (tout comme \bar{y})

S'il existe une relation linéaire, la covariance est non-nulle. S'il n'en existe pas, la covariance est nulle.

La qualité de la régression est donné par deux indicateurs : le coefficient de corrélation et le coefficient de détermination :

$$R = \text{cov}(x,y) / \text{écart type}(x) \text{écart type}(y)$$

$$R^2 = \text{cov}(x,y)^2 / V(x) V(y)$$

