

Chapitre 1. L'appréhension statistique

28

Exemple de problématique : on désire étudier les salaires des 8000 employés d'une entreprise.

Questions : qui va-t-on interroger ? que va-t-on mesurer ?

Définitions

- La **population statistique** (ou champ d'étude) est l'ensemble sur lequel porte l'étude statistique.
△la population n'est pas forcément un ensemble de personnes - ex : parc automobile français, ...
- Un **individu** est un élément de la population.
- Si la population est trop grande (i.e. trop d'individus) on fait généralement l'étude sur une partie de la population. Cette "partie" est appelée **échantillon**.
- La **variable statistique** est ce que l'on observe/mesure chez les individus de la population.

35

Définition

- Un tableau statistique donne pour chaque valeur (ou modalité) de la variable, l'effectif correspondant (c-a-d le nombre de fois où l'on observe la modalité).
- Il intègre en général également la fréquence de chaque modalité ainsi que l'effectif ou la fréquence cumulée (lorsqu'il s'agit d'une variable quantitative!).

- Soit p le nombre de modalités que l'on note en général x_i ($i = 1, \dots, p$)

- Effectif : n_i ; Effectif total : $n = \sum_{i=1}^p n_i$.

- Fréquence : $f_i = \frac{n_i}{n}$, $f_i \in [0, 1]$ que l'on exprime en %.

- Fréquence cumulée : $F_i = \sum_{j=1}^i f_j = \frac{1}{n} \sum_{j=1}^i n_j$ en %.

47

- Vocabulaire statistique
 - Définitions générales
 - Classification des variables statistiques
- Tableaux statistiques
 - Définition et formalisation
 - Exemples
- Représentation graphique
 - Introduction
 - Variable quantitative discrète : exemple du nbre de personnes par voiture
 - Variable quantitative continue : exemple du revenu de ménages

29

La nature des variables statistiques dépend de la nature de ses **modalités**, c-a-d des valeurs qu'elle peut prendre.

- Variables qualitatives** : grandeur non "mesurable", à laquelle on ne peut pas donner une valeur numérique.
 - variables **catégorielles** : aucune relation d'ordre ne peut être établie sur l'ensemble des modalités.
[ex : sexe, couleur des yeux, origine du baccalauréat]
 - variables **ordinales** : une relation d'ordres existe.
[ex : niveau de satisfaction (mauvais, bon, très bon, ...)]
- Variables quantitatives** : la grandeur observée est "mesurable".
 - Variables **quantitatives discrètes** : la variable prend ses valeurs dans un ensemble dénombrable.
[ex : nbre d'enfants par employé, ...]
 - Variables **quantitatives continues** : la variable prend ses valeurs dans un ensemble "continu" (intervalle, réels, ...) [ex : salaire des employés, taille, ...]

42

On désire étudier la répartition de 22 millions des actifs français selon leur catégorie socio-professionnelle.

Déterminez la population, l'individu, la variable statistique et sa nature.

Réponse : population = ens. des actifs, individu = 1 actif
variable = CSP, variable qualitative catégorielle.

Complétez le tableau statistique :

Modalité x_i	Effectif n_i ($\times 10^3$)	Fréquence f_i
Ouvriers		
Employés	6600	
Cadres	2640	
Autres	7260	
Total	22000	

49

On désire étudier la répartition de 22 millions des actifs français selon leur catégorie socio-professionnelle.

Déterminez la population, l'individu, la variable statistique et sa nature.

Réponse : population = ens. des actifs, individu = 1 actif
variable = CSP, variable qualitative catégorielle.

Complétez le tableau statistique :

Modalité x_i	Effectif n_i ($\times 10^3$)	Fréquence f_i
Ouvriers	5500	25%
Employés	6600	30%
Cadres	2640	12%
Autres	7260	33%
Total	22000	100%

51

On désire étudier le nbre de personnes présentes dans un véhicule à partir d'un échantillon de **400** voitures.

Déterminez population, individu, variable statistique.
Réponse : population = 400 voitures, individu = 1 voiture
variable = nbre de pers/voit., variable quantitative discrète.

Complétez le tableau statistique :

Modalité x_i	Effectif n_i	Fréq. f_i	Fréq. cumulée F_i
1	40	10%	10%
2	100	25%	35%
3	160	40%	75%
4	100	25%	100%
Total	400	100%	×

55

On désire étudier le revenu de l'ensemble des ménages français (**20** millions)

Déterminez population, individu, variable statistique.
Réponse : population = ensemble des ménages, individu = 1 ménage, variable = revenu, variable quantitative continue.

⚠ lorsque la variable est continue, il **faut** regrouper les observations en **classess** ; on parle de données **regroupées**.

Complétez le tableau statistique :

Modalité x_i (en euros)	Effectif n_i ($\times 10^6$)	Fréq. f_i	Fréq. cumulée F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×

60

On désire étudier le nbre de personnes présentes dans un véhicule à partir d'un échantillon de **400** voitures.

Déterminez population, individu, variable statistique.
Réponse : population = 400 voitures, individu = 1 voiture
variable = nbre de pers/voit., variable quantitative discrète.

Complétez le tableau statistique :

Modalité x_i	Effectif n_i	Fréq. f_i	Fréq. cumulée F_i
1		10%	
2		25%	
3		40%	
4		25%	
Total		100%	×

53

On désire étudier le revenu de l'ensemble des ménages français (**20** millions)

Déterminez population, individu, variable statistique.
Réponse : population = ensemble des ménages, individu = 1 ménage, variable = revenu, variable quantitative continue.

⚠ lorsque la variable est continue, il **faut** regrouper les observations en **classess** ; on parle de données **regroupées**.

Complétez le tableau statistique :

Modalité x_i (en euros)	Effectif n_i ($\times 10^6$)	Fréq. f_i	Fréq. cumulée F_i
[0, 1600[			45%
[1600, 2400[			80%
[2400, 3200[			100%
Total		100%	×

58

Modalité x_i (en euros)	Effectif n_i ($\times 10^6$)	Fréq. f_i	Fréq. cumulée F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×

Quelques interprétations :

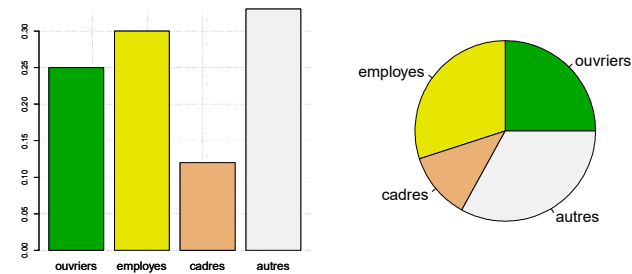
- 9 millions d'actifs ont un revenu **compris entre 0 et 1600 euros**
- 35% des actifs ont un revenu **compris entre 1600 et 2400 euros** .
- 80% des actifs ont un revenu **inférieur à 2400 euros** .

61

- Permet d’avoir une vision globale de la série statistique
- Permet de comparer deux séries statistiques en superposant les graphiques.
- Δ Graphique reste assez imprécis ; c’est un résumé qu’il faut absolument associer au tableau.
- Graphiques différents selon le **type de variable** étudiée.

65

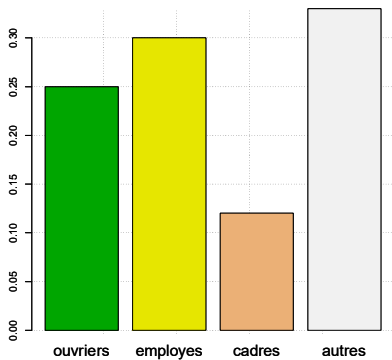
- Construction graphique : on représente chaque modalité de la variable par une surface proportionnelle à l’effectif.
- Pour ce faire, les **deux principales méthodes** sont
- ① Graphique en **tuyaux d’orgues**.
 - ② Graphique en secteurs ou **camembert**.



67

x_i	$n_i (\times 10^3)$	f_i
Ouvriers	5500	25%
Employés	6600	30%
Cadres	2640	12%
Autres	7260	33%
Total	22000	100%

On représente les **modalités** par des rectangles de base constante. Les hauteurs de ces rectangles correspondent aux **fréquences** (ou effectifs).



70

Les **modalités** sont représentées par des secteurs de disque dont l’**angle** est proportionnel aux **fréquences** ou **effectifs**.

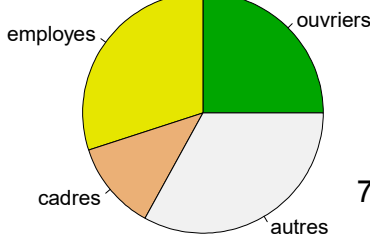
x_i	f_i	angle a_i
Ouvriers	25%	
Employés	30%	
Cadres	12%	
Autres	33%	
Total	100%	360°

- Méthodologie : faire une règle de 3 en sachant que $100\% \Leftrightarrow 360^\circ$
- Exemple : $25\% \Rightarrow 25/100 * 360 = 90^\circ$.

71

Les **modalités** sont représentées par des secteurs de disque dont l’**angle** est proportionnel aux **fréquences** ou **effectifs**.

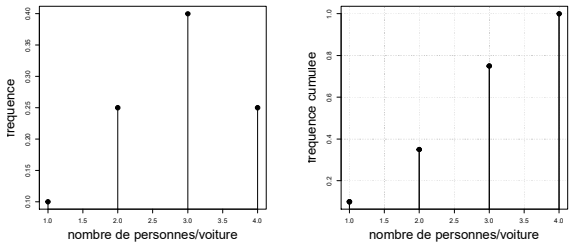
x_i	f_i	angle a_i
Ouvriers	25%	90°
Employés	30%	108°
Cadres	12%	43.2°
Autres	33%	118.8°
Total	100%	360°



73

x_i	n_i	f_i	F_i
1	40	10%	10%
2	100	25%	35%
3	160	40%	75%
4	100	25%	100%
Total	400	100%	×

- Deux types de graphique :
- ① diagramme en **bâtons fréquentiel** (ou d’effectifs) [**à gauche**]
 - ② diagramme des **fréquences cumulées** [**à droite**]



74

Revenu x_i (en euros)	n_i ($\times 10^6$)	Fréq. f_i	F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×

Pour représenter graphiquement une variable continue regroupée en classes, on dispose de deux outils

- ➊ **histogramme fréquentiel**
- ➋ **courbe des fréquences cumulées**

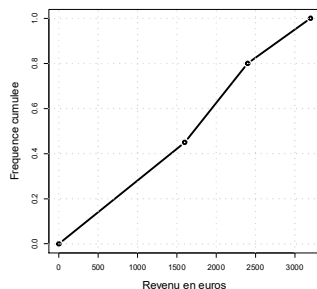
75

Revenu x_i (en euros)	n_i ($\times 10^6$)	Fréq. f_i	a_i (1 u.a.=800€)	d_i (par tranche de 800€)
[0, 1600[	9	45%		
[1600, 2400[	7	35%		
[2400, 3200[	4	20%		
Total	20	100%	×	×

81

⇒ graphique des fréquences cumulées (F_i) en fonction du revenu (x_i).

Revenu x_i (en euros)	n_i ($\times 10^6$)	Fréq. f_i	F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×



88

Construction de l'histogramme

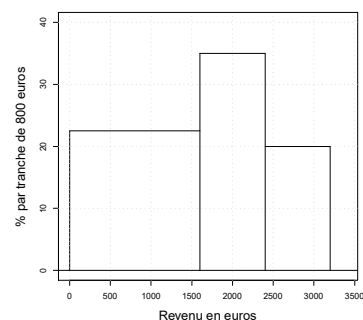
Définition

L'histogramme est un ensemble de **rectangles** juxtaposés tel que la **surface** de chacun des rectangles soit proportionnelle à la **fréquence** de chacune des classes. La longueur de la base de ces rectangles est **proportionnelle à l'amplitude** de la classe, notée a_i

- Exemple (revenu) : amplitude des classes= 1600 €, 800 €, 800€⇒ on peut tout renormaliser par une même **unité d'amplitude** (u.a.) valant ici 800 €. Ici, les facteurs d'unité d'amplitude valent 2 u.a., 1 u.a. et 1 u.a.
- Dans un histogramme la hauteur des rect. est aussi appelée **densité** et est noté d_i .
- Formule : $a_i \times d_i = f_i \Leftrightarrow d_i = \frac{f_i}{a_i}$
- La surface totale des rectangles vaut **1** (car $\sum_{i=1}^p f_i = 1$)

80

Revenu x_i (en euros)	n_i ($\times 10^6$)	Fréq. f_i	a_i (1 u.a.=800€)	d_i (par tranche de 800€)
[0, 1600[	9	45%	2	22.5%
[1600, 2400[	7	35%	1	35%
[2400, 3200[	4	20%	1	20%
Total	20	100%	×	×



85

Chapitre 2. Caractéristiques des distributions à une variable quantitative

89

Objectif général de ce chapitre

- Objectif : calculer des caractéristiques permettant de résumer les tableaux et graphiques.
- Trois catégories de caractéristiques :
 - 1 Tendance centrale
 - 2 Dispersion
 - 3 Concentration

90

- 1 Caractéristiques de tendance centrale
 - Mode
 - Médiane
 - Quantiles d'ordre quelconque
 - Moyenne
 - Synthèse : quelles caractéristiques pour résumer une série ?
 - Complément : méthode du "shift and share"
- 2 Caractéristiques de dispersion
 - Etendue (intervalle de variation)
 - Ecarts interquantiles
 - Ecart absolu
 - Ecart-type et variance
 - Comparaison de séries statistiques et synthèse
- 3 Caractéristiques de concentration
 - Courbe de Lorentz
 - Indice de Gini
 - Médiale

91

Mode d'une variable statistique

Définition

Le **mode** (ou **classe modale**) est la valeur (ou la classe) pour laquelle les individus sont le plus représentés.

Calcul du mode :

- variable discrète : modalité présentant le plus grand effectif (ou plus grande fréquence).
- variable continue : on cherche d'abord la classe ayant la plus grande densité : c'est la **classe modale**. Le mode peut ensuite être défini (par exemple comme le centre de cette classe).

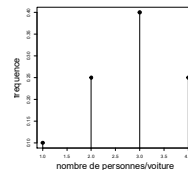
Remarques :

- pour une var. continue, en général on ne donne que la classe modale.
- Une série peut avoir plusieurs modes (en présence de maxima locaux de fréquence ou densité selon le type de variable) ; on parle de série **plurimodale**.

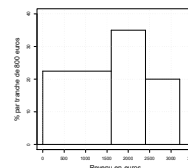
94

Application numérique sur deux exemples

Exemple Nbre pers./voiture	
x_i	f_i
1	10%
2	25%
3	40%
4	25%
Total	100%



Revenu des ménages français		
x_i (en euros)	f_i	d_i (/tr. de 800€)
[0, 1600[	45%	22.5%
[1600, 2400[	35%	35%
[2400, 3200[	20%	20%
Total	100%	×



95

Médiane - définition

Définition

La médiane est la valeur de la série (i.e. une modalité) qui partage la série en deux sous-ensembles de même effectif (ou de même fréquence).

△ Il faut distinguer deux cas :

- 1 les données sont observées de manière brute.
[le plus souvent une variable discrète]
- 2 les données sont regroupées en classes.
[le plus souvent une variable continue]

96

Médiane (2) - données brutes

Deux cas possibles en fonction du caractère pair ou impair de la taille de l'échantillon n :

- 1 n est impair : la médiane de la série de $n = 5$ âges : 17, 9, 19, 25, 21 est $Me = 19$ (ans).
- 2 n est pair : la médiane de la série de $n = 4$ âges : 17, 9, 19, 25 est entre 17 et 19 $\Rightarrow Me = (17 + 19)/2 = 18$ (ans)

Formule générale : Soient $x_1, \dots, x_n$ les valeurs de la série et soient $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ les versions ordonnées, i.e.

$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ alors

$$Me = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ est impair,} \\ \frac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{si } n \text{ est pair.} \end{cases}$$

102

Médiane - données brutes (2)

Quelle est la médiane de la série statistique suivante ?

Exemple nb personnes/voiture			
x_i	n_i	f_i	F_i
1	40	10%	10%
2	100	25%	35%
3	160	40%	75%
4	100	25%	100%
Total	400	100%	×

- $n = 400$ est pair $\Rightarrow$ il faut donc repérer la **200** -ème et **201** -ème observation dans la liste des observations ordonnées.

• $x_{(200)} = 3$, $x_{(201)} = 3 \Rightarrow Me = \frac{3+3}{2} = 3$ (pers./voiture)

105

Médiane (3) - données regroupées

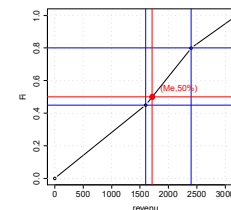
Exemple du revenu ménages			
x_i (en €)	n_i ($\times 10^6$)	f_i	F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×

Dans le cas où les données sont regroupées en classes, il faut suivre deux étapes :

- 1 repérer la **classe médiane**, i.e. la classe contenant la médiane.
Ici, 45% des ménage ont un revenu $< 1600€$ et 80% des ménages ont un revenu $< 2400€ \Rightarrow Me \in [1600, 2400[$
- 2 estimer la médiane par **interpolation linéaire**.

108

Médiane (5) - interpolation linéaire



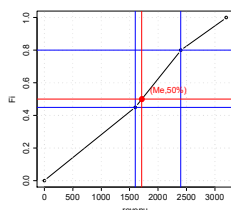
Graphiquement : la médiane correspond à l'abscisse du point d'intersection entre la courbe des (x_i, F_i) et la droite horizontale d'équation $y = 50\%$.

$\Rightarrow$ Formule générale : soit $]x_i, x_{i+1}[$ la classe médiane et soient F_i et F_{i+1} les fréquences cumulées évaluées en x_i et x_{i+1} , alors

$$Me = x_i + \frac{50\% - F_i}{F_{i+1} - F_i} \times (x_{i+1} - x_i)$$

110

Médiane (5) - interpolation linéaire



Graphiquement : la médiane correspond à l'abscisse du point d'intersection entre la courbe des (x_i, F_i) et la droite horizontale d'équation $y = 50\%$.

Application numérique :

- $x_i = 1600$, $x_{i+1} = 2400$, $F_i = 45\%$, $F_{i+1} = 80\%$.
- $Me = 1600 + \frac{50\% - 45\%}{80\% - 45\%} \times (2400 - 1600) \simeq 1714.28$ €.

113

Quantile

Définition

Un quantile d'ordre α (pour $\alpha \in (0, 1)$) notée en toute généralité Q_α est la valeur qui partage la série en deux sous-ensembles ; une proportion α se situe en dessous de Q_α et une proportion $1 - \alpha$ au-dessus strictement de Q_α .

Remarques :

- $Me = Q_{50\%}$.
- Quartiles (notés Q_1, Q_2, Q_3) : quantiles qui séparent la série en 4 sous-ensembles de même effectif/fréquence. Plus précisément

$$Q_1 = Q_{25\%}, Q_2 = Me, Q_3 = Q_{75\%}.$$

- Déciles (notés $D_1, D_2, \dots, D_9$) : quantiles qui séparent la série en 10 sous-ensembles de même fréquence. Plus précisément

$$D_1 = Q_{10\%}, D_2 = Q_{20\%}, \dots, D_9 = Q_{90\%}.$$

114

- Les quantiles se calculent de manière similaire à la médiane.
- Ainsi pour des données regroupées on a : si $Q_\alpha \in]x_i, x_{i+1}[$

$$Q_\alpha = x_i + \frac{\alpha - F_i}{F_{i+1} - F_i} \times (x_{i+1} - x_i)$$

Calculez le premier quartile de la série suivante

Exemple du revenu ménages			
x_i (en €)	n_i ($\times 10^6$)	f_i	F_i
[0, 1600[	9	45%	45%
[1600, 2400[	7	35%	80%
[2400, 3200[	4	20%	100%
Total	20	100%	×

- $Q_1 \in]0, 1600[$
- $Q_1 = 0 + \frac{25\% - 0}{45\% - 0} (1600 - 0) \simeq 888.89\text{€}.$

117

Il y a plusieurs types de moyenne dépendant essentiellement du problème considéré

- Moyenne arithmétique [la plus connue et la plus standard]
- Moyenne géométrique [utilisée par exemple pour calculer des taux moyens]
- Moyenne harmonique [utilisée pour calculer des moyennes de ratios]
- Moyenne quadratique [moyenne de carrés, notion moins utilisée]

118

Définition

Soit x_i ($i = 1, \dots, p$) les modalités d'une série brute, d'effectifs n_i ($i = 1, \dots, p$) et fréquence f_i , la moyenne arithmétique pondérée notée $\bar{x}$ est donnée par

$$\bar{x} = \frac{1}{n} \sum_{i=1}^p n_i x_i = \sum_{i=1}^p f_i x_i \text{ car } f_i = \frac{n_i}{n}.$$

- Si les données sont regroupées en classes, les x_i ne sont en général pas observées. Ces valeurs sont alors remplacées par les **centres de classes**, notés c_i pour $i = 1, \dots, p$.
- lorsque le nombre de modalités (ou nombre de classes) est grand, il devient intéressant d'utiliser la calculatrice (rentrer les données sous forme d'un tableau, configurer de manière appropriée et demander des résultats univariés).

119

Calculez la moyenne de la série

Exemple nb personnes/voiture			
x_i	n_i	f_i	F_i
1	40	10%	10%
2	100	25%	35%
3	160	40%	75%
4	100	25%	100%
Total	400	100%	×

Application :

$$\bar{x} = \frac{40 * 1 + 100 * 2 + 160 * 3 + 100 * 4}{400} = 2.8 \text{ pers./voiture.}$$

(Remarque : $10\% * 1 + 25\% * 2 + 40\% * 3 + 25\% * 4 = 2.8$)

121

Calculez la moyenne de la série

Exemple du revenu ménages				
x_i (en €)	c_i	n_i ($\times 10^6$)	f_i	F_i
[0, 1600[	800	9	45%	45%
[1600, 2400[	2000	7	35%	80%
[2400, 3200[	2800	4	20%	100%
Total	×	20	100%	×

Application :

$$\bar{x} = \frac{9 * 800 + 7 * 2000 + 4 * 2800}{20} = 1620 \text{ €}.$$

123

- La somme des écarts (pondérés) à la moyenne est nulle, c-a-d

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

Preuve :

$$\sum_{i=1}^p n_i (x_i - \bar{x}) = \sum_{i=1}^p n_i x_i - \left(\sum_{i=1}^p n_i \right) \bar{x} = n\bar{x} - n\bar{x} = 0.$$

124

- 1 La somme des écarts (pondérés) à la moyenne est nulle, c-a-d

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

- 2 Considérons une population P d'effectif total n composée de k sous-populations $P_1, \dots, P_k$ d'effectifs $n_1, \dots, n_k$ (donc $n = n_1 + \dots + n_k$). Notons $\bar{x}_1, \dots, \bar{x}_k$ les moyennes arithmétiques des sous-populations $P_1, \dots, P_k$ alors

$$\bar{x} = \frac{n_1 \bar{x}_1 + \dots + n_k \bar{x}_k}{n}$$

"la moyenne globale est égale à la moyenne pondérée des moyennes"

125

- Une action en bourse a évolué à la hausse de 10% l'année 1, puis a diminué de 5% l'année 2 et de 5% l'année 3.
- Question : Quel est le taux moyen (noté t_{moy}) d'évolution de cette action sur les trois années ?
 $\Delta \quad t_{moy} \neq 0!!!$
- La **moyenne géométrique** est le taux qui, appliqué durant les trois années donnera le même capital final selon l'évolution décrite précédemment.

130

Définition

Soit la série statistique $x_1, \dots, x_p$ d'effectif $n_1, \dots, n_p$ alors la moyenne géométrique notée en général $\bar{x}_G$ est définie par

$$\bar{x}_G = \left(x_1^{n_1} \times x_2^{n_2} \times \dots \times x_p^{n_p} \right)^{1/n}$$

où $n = n_1 + \dots + n_p$.

132

Ex : salaire de $n_H = 200$ hommes et $n_F = 100$ femmes d'une entreprise.

Calculez la moyenne de la série Ensemble de deux façons différentes :

x_i (en €)	c_i	$n_{i,H}$	$n_{i,F}$	$n_{i,E}$
[0, 1500[	750	70	60	130
[1500, 3000[	2250	130	40	170
Total	x	200	100	300

Méthode 1 (méthode directe) :

$$\bar{x}_E = \frac{1}{300} (750 \times 130 + 2250 \times 170) = 1600€.$$

Méthode 2 (en utilisant la propriété précédente) :

$$\bar{x}_H = \frac{1}{200} (750 \times 70 + 2250 \times 130) = 1725€.$$

$$\bar{x}_F = \frac{1}{100} (750 \times 60 + 2250 \times 40) = 1350€.$$

$$\bar{x}_E = \frac{1}{300} (200 \times \bar{x}_H + 100 \times \bar{x}_F) = \frac{1}{300} (200 \times 1725 + 100 \times 1350) = 1600€$$

128

Soit C_0 le capital initial et soient C_1, C_2, C_3 les capitaux après 1, 2 ou 3 années. On a

- selon l'énoncé $C_1 = (1 + 10\%)C_0$, $C_2 = (1 - 5\%)C_1$ et $C_3 = (1 - 5\%)C_2$, c-a-d

$$C_3 = (1 + 10\%)(1 - 5\%)(1 - 5\%)C_0.$$

- selon la définition du taux moyen : $C_1 = (1 + t_{moy})C_0$, $C_2 = (1 + t_{moy})C_1$ et $C_3 = (1 + t_{moy})C_2$, c-a-d

$$C_3 = (1 + t_{moy})^3 C_0.$$

Par identification des deux identités, il vient que pour tout capital initial C_0

$$(1 + 10\%)(1 - 5\%)(1 - 5\%) = (1 + t_{moy})^3$$

$\Leftrightarrow$

$$t_{moy} = ((1 + 10\%)(1 - 5\%)(1 - 5\%))^{1/3} - 1.$$

131

- Elle permet de calculer des moyennes de ratios.
- Exemple : Un coureur monte une côte de 1km à la vitesse de 10km/h et descend cette même côte à la vitesse de 30km/h.
- Question : Quelle est la vitesse moyenne du coureur ?

$$v_{moy} \neq 20 \text{ km/h} !!$$

car il a passé plus de temps à 10km/h qu'à 30km/h.

- On cherche v_{moy} telle que la somme des temps passés à la montée et la descente soit égal au temps passé à la vitesse

v_{moy} :

$$t_{montée} = \frac{1}{10}, \quad t_{desc.} = \frac{1}{30}, \quad t_{v_{moy}} = \frac{2}{v_{moy}}$$

$$\Leftrightarrow \frac{2}{v_{moy}} = \frac{1}{10} + \frac{1}{30} \Leftrightarrow v_{moy} = \frac{2}{\frac{1}{10} + \frac{1}{30}} = 15 \text{ km/h.}$$

135

Définition

Soit la série statistique $x_1, \dots, x_p$ d'effectif $n_1, \dots, n_p$ alors la moyenne harmonique notée en général $\bar{x}_H$ est définie par

$$\bar{x}_H = \frac{n}{\frac{n_1}{x_1} + \dots + \frac{n_p}{x_p}}$$

où $n = n_1 + \dots + n_p$.

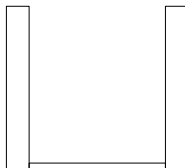
136

- Mode(s), médiane, moyenne(s) : quel(s) indicateur(s) utiliser pour résumer une série et en donner des tendances centrales ?
- Cela dépend de la "forme" générale de la série statistique étudiée selon qu'elle soit :
 - plurimodale,
 - symétrique,
 - asymétrique.

137

... quel est l'indicateur pertinent ?

Salaires x_i en €	c_i	n_i	a_i (1 u.a. 4000€)
[0, 4000[	2000	45	1
[4000, 8000[	16000	10	6
[8000, 32000[	30000	45	1



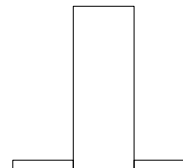
série pluri-modale

- $\bar{x} = 16000\text{€}$, $Me = 16000\text{€}$.
 - 2 classes modales : [0, 4000[, [8000, 32000[.
- ⇒
- Moyenne et médiane non représentatives de la série.
 - Modes informatifs.

138

... quel est l'indicateur pertinent ?

Salaires x_i en €	c_i	n_i	a_i (1 u.a. 1000€)
[0, 1000[	500	5	1
[1000, 2000[	1500	90	1
[2000, 3000[	2500	5	1



série symétrique

- $\bar{x} = 1500\text{€}$, $Me = 1500\text{€}$.
 - classes modales : [1000, 2000[.
- ⇒
- les trois indicateurs peuvent être utilisés.
 - on préférera la moyenne qui possède des propriétés intéressantes (calcul algébrique)

139

... quel est l'indicateur pertinent ?

Salaires x_i en €	c_i	n_i	a_i (1 u.a. 2000€)
[0, 2000[	1000	90	1
[2000, 38000[	18000	10	18



série asymétrique

- $\bar{x} = 2900\text{€}$, $Me = 1100\text{€}$.
- ⇒
- La moyenne n'est pas représentative car trop influencée par les gros salaires.
 - la médiane est plus adaptée.

140

- méthode utilisée pour comparer plusieurs moyennes pondérées lorsque les coefficients de pondération sont très $\neq$, par exemple lorsqu'ils évoluent au cours du temps.
- permet de lisser l'effet structure.

Exemples : salaires de 2 CSP en 2010 et 2011.

CSP	Année 2010 f_i	Année 2010 $\bar{x}_i$ (€)	Année 2011 f_i	Année 2011 $\bar{x}_i$ (€)
Cadres	10%	2000	50%	1300
Employés	90%	1000	50%	900

- $\bar{x}_{2010} = 1100\text{€}$, $\bar{x}_{2011} = 1100\text{€}$.
- peut-on conclure qu'il n'y a pas d'évolution de salaires de 2010 à 2011 ?

142

Complement I : Comparaison de séries (1)

- soit x la série statistique de 4 produits en Francs : 100F, 200F, 300F et 400F.
- soit y la série statistique des 4 produits en € : 15€, 30€, 45€, 60€.
- Intuitivement, ces deux séries sont dispersées de la même manière. Or,

$$\sigma_x = 111.8F \quad \text{et} \quad \sigma_y = 16.8\epsilon.$$

- **Conclusion** : pour comparer les deux séries qui ne sont pas dans la même unité, il faut transformer les caractéristiques de dispersion.
- **Coefficient de variation** : $\frac{\sigma_x}{\bar{x}}$ = c'est le % de variation par rapport à la moyenne, sans unité.

$$\frac{\sigma_x}{\bar{x}} \simeq \frac{111.8}{250} \simeq 0.45 \quad \text{et} \quad \frac{\sigma_y}{\bar{y}} \simeq \frac{16.8}{37.5} \simeq 0.45. \quad 163$$

Complement I : comparaison de séries (2)

D'autres indicateurs de comparaison de séries statistiques :

- Coefficient de dispersion :

$$\frac{Q_3 - Q_1}{Me} \quad \text{ou} \quad \frac{D_9 - D_1}{Me}.$$

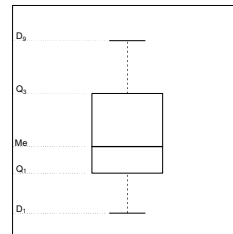
- Rapport interquartile ou rapport interdécile :

$$\frac{Q_3}{Q_1} \quad \text{ou} \quad \frac{D_9}{D_1}$$

Caractéristiques de tendance centrale	Caractéristiques de dispersion	Caractéristiques de concentration
oooooooooooooooooooooooooooo	oooooooooooo●oo	oooooooo
Comparaison de séries statistiques et synthèse		

Complement II : la boîte à moustaches (1)

- aussi appelée box plot ou diagramme de Tukey.
- moyen rapide de visualiser des caractéristiques centrale et de dispersion d'une **série quantitative**.
- principalement utilisée pour comparer un **même caractère** pour plusieurs populations.
- basée sur le calcul de D_1, Q_1, Me, Q_3 et D_9 .



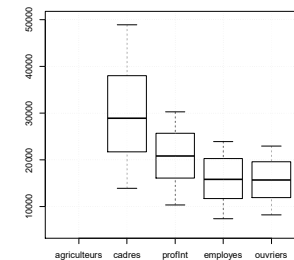
Caractéristiques de tendance centrale oooooooooooooooooooooooooooo	Caractéristiques de dispersion oooooooooooo●●●●	Caractéristiques de concentration oooooooo
Comparaison de séries statistiques et synthèse		

Complement II : la boîte à moustaches (2)

Etude sur le niveau de vie des ménages en euros par CSP (personne de référence) en 2010. Application : complétez le graphique suivant avec les revenus des agriculteurs ...

sachant que pour les
agriculteurs

- $D_1 = 6040$
- $Q_1 = 11135$
- $Me = 18010$
- $Q_3 = 27140$
- $D_9 = 39010$



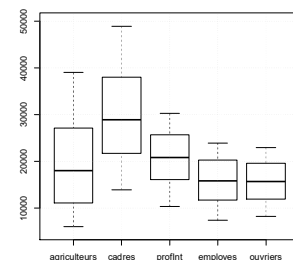
Caractéristiques de tendance centrale oooooooooooooooooooooooooooo	Caractéristiques de dispersion ooooooooooooo●oooo	Caractéristiques de concentration oooooooo
Comparaison de séries statistiques et synthèse		

Complement II : la boîte à moustaches (2)

Etude sur le niveau de vie des ménages en euros par CSP (personne de référence) en 2010. Application : complétez le graphique suivant avec les revenus des agriculteurs ...

sachant que pour les
agriculteurs

- $D_1 = 6040$
- $Q_1 = 11135$
- $Me = 18010$
- $Q_3 = 27140$
- $D_9 = 39010$



Caractéristiques de tendance centrale	Caractéristiques de dispersion	Caractéristiques de concentration
oooooooooooooooooooooooooooo	oooooooooooooooooooo	oooooooo

Introduction

- Elles sont utilisées pour mesurer (essentiellement) la répartition de la masse salariale. La répartition de la masse salariale se situe entre les deux cas extrêmes suivants
 - Répartition des salaires parfaitement équitables : un certain pourcentage de salariés reçoit le même pourcentage de la masse salariale. *On dit que la concentration est nulle.*
 - Un seul salarié reçoit toute la masse salariale (et les autres rien). *On dit que la concentration est maximale.*
- Trois indicateurs pour quantifier la concentration
 - ① courbe de Lorentz
 - ② Indice de Gini
 - ③ Médiane.

On étudie les salaires de 50 employés d'une entreprise.

x_i (en €)	c_i	n_i	f_i	F_i			
[600, 1200[	900	15	30%	30 %			
[1200, 1800[	1500	25	50%	80%			
[1800, 2100[	1950	10	20%	100%			
Total	×	50	100%	×			

1

2

172

On étudie les salaires de 50 employés d'une entreprise.

x_i (en €)	c_i	n_i	f_i	F_i	$n_i c_i$	g_i	G_i
[600, 1200[	900	15	30%	30 %	13500	19.1%	19.1%
[1200, 1800[	1500	25	50%	80%	37500	53.2%	72.3%
[1800, 2100[	1950	10	20%	100%	19500	27.7%	100%
Total	×	50	100%	×	70500	100%	×

1 on calcule la masse salariale = $n_i \times c_i$.

2 on calcule le % de la masse salariale g_i , ainsi que les fréquences cumulées G_i .

174

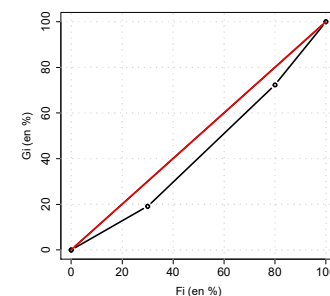
On étudie les salaires de 50 employés d'une entreprise.

x_i (en €)	c_i	n_i	f_i	F_i	$n_i c_i$	g_i	G_i
[600, 1200[	900	15	30%	30 %	13500	19.1%	19.1%
[1200, 1800[	1500	25	50%	80%	37500	53.2%	72.3%
[1800, 2100[	1950	10	20%	100%	19500	27.7%	100%
Total	×	50	100%	×	70500	100%	×

Définition

La courbe de Lorentz est obtenue en faisant correspondre à la fréquence cumulée F_i à la fréquence cumulée G_i de la masse salariale.

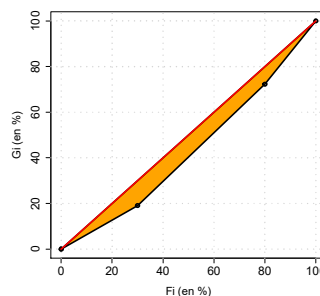
175



1 droite rouge = répartition parfaitement équitable.

2 Plus la courbe de Lorentz est éloignée de la droite rouge et plus la concentration est forte (répartition de moins en moins équitable).

176



Soit S la surface orange.

$$I_{Gini} = \frac{S}{\text{Surf. Demi-carré}} = 2S \in [0, 1]$$

1 Plus I_{Gini} est proche de 0, plus la concentration est faible (proche de équirépartition).

2 Dans notre cas, $I_{Gini} \approx 14\%$ (on ne cherchera pas à calculer l'indice)

177

x_i (en €)	c_i	n_i	f_i	F_i	$n_i c_i$	g_i	G_i
[600, 1200[	900	15	30%	30 %	13500	19.1%	19.1%
[1200, 1800[	1500	25	50%	80%	37500	53.2%	72.3%
[1800, 2100[	1950	10	20%	100%	19500	27.7%	100%
Total	×	50	100%	×	70500	100%	×

1 La médiane est la médiane de la série masse associée. Dans notre exemple

178

x_i (en €)	c_i	n_i	f_i	F_i	$n_i c_i$	g_i	G_i
[600, 1200[	900	15	30%	30 %	13500	19.1%	19.1%
[1200, 1800[	1500	25	50%	80%	37500	53.2%	72.3%
[1800, 2100[	1950	10	20%	100%	19500	27.7%	100%
Total	×	50	100%	×	70500	100%	×

- La médiale est la médiane de la série masse associée. Dans notre exemple

$$\text{Médiale} = 1200 + \frac{50\% - 19.1\%}{72.3\% - 19.1\%} \times (1800 - 1200) \simeq 1548\text{€}.$$

Les salariés recevant moins de 1548 €représentent 50% de la masse salariale.

- Mesure de concentration :

$$\Delta = \frac{\text{Médiale} - Me}{Etendue} \geq 0.$$

Δ petit = faible concentration, Δ grand= grande concentration180

Ici, on peut vérifier que $\Delta \simeq (1548 - 1440)/(2100 - 600) \simeq 7.2\%$.

Chapitre 3. Les distributions à deux variables

1 Autour des tableaux de contingence

- Définition
- Distributions conditionnelles
- Relations entre les différentes fréquences
- Moyennes et Variances conditionnelles

② Etude de la liaison entre deux variables

- Mesure de la dépendance entre deux variables
 - Définition de l'indépendance totale
 - Définition de la dépendance totale
 - χ^2 et coefficient de Cramer
- Mesure de la liaison fonctionnelle
 - Courbes de régression
 - Rapport de corrélation
 - Régression linéaire

Tableau de contingence

- = tableau statistique permettant de présenter **deux** séries statistiques **simultanément** et de **manière croisée**.
- exemple : dans une entreprise de 200 salariés, on étudie les variables X =âge et Y =salaires.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14	6	20
[22, 24[(i = 2)	28	46	74
[24, 26[(i = 3)	20	86	106
Total	62	138	200

- X et Y sont des variables continues (regroupées en classes)
- On note I le nombre de modalités de X (ici $I = 3$) et J le nombre de modalités de Y (ici $J = 2$). 18

Tableau de contingence (2)

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14	6	20
[22, 24[(i = 2)	28	46	74
[24, 26[(i = 3)	20	86	106
Total	62	138	200

- i désigne l'indice d'une **ligne** et j désigne l'indice d'une **colonne**.

- n_{ij} désigne l'effectif partiel.

Exemple : $n_{12} = 6$ salariés sont âgés entre 20 et 22 ans **et** ont un salaire compris entre 1000 et 1200 €.

- on note $n_{i\bullet}$ l'effectif marginal de X (eff. total en lignes)
et $n_{\bullet j}$ l'effectif marginal de Y (effectif total en colonnes).

Exemple : $n_{2\bullet} = 74$ salariés sont âgés entre 22 et 24 ans ;
 $n_{\bullet 1} = 62$ salariés ont un salaire entre 800 et 1000€.

- $n = n_{\bullet\bullet}$ correspond à l'effectif total.

Tableau de contingence (3)

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14	6	20
[22, 24[(i = 2)	28	46	74
[24, 26[(i = 3)	20	86	106
Total	62	138	200

Formules : Pour $i = 1, \dots, I$ et pour $j = 1, \dots, J$

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

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

$$n = n_{\bullet\bullet} = \sum_{i=1}^I n_{i\bullet} = \sum_{j=1}^J n_{\bullet j} = \sum_{i=1}^I \sum_{j=1}^J n_{ij}.$$

Fréquences partielles et marginales

⚠ Les fréquences sont notées entre parenthèses.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (7 %)	6 (3 %)	20 (10 %)
[22, 24[(i = 2)	28 (14 %)	46 (23 %)	74 (37 %)
[24, 26[(i = 3)	20 (10 %)	86 (43 %)	106 (53 %)
Total	62 (31 %)	138 (69 %)	200 (100%)

- f_{ij} désigne la fréquence **partielle**.

Exemple : $f_{12} = 3\%$ des salariés sont âgés entre 20 et 22 ans et ont un salaire compris entre 1000 et 1200 €.

- on note $f_{i\bullet}$ la fréquence marginale de X (fréq. totale en lignes) et $f_{\bullet j}$ la fréquence marginale de Y (fréq. totale en colonnes).

Exemple : $f_{2\bullet} = 37\%$ des salariés sont âgés entre 22 et 24 ans; $f_{\bullet 1} = 31\%$ des individus ont un salaire entre 800 et 1000€.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (7%)	6 (3%)	20 (10%)
[22, 24[(i = 2)	28 (14%)	46 (23%)	74 (37%)
[24, 26[(i = 3)	20 (10%)	86 (43%)	106 (53%)
Total	62 (31%)	138 (69%)	200 (100%)

Formules : Pour $i = 1, \dots, I$ et pour $j = 1, \dots, J$

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

$$f_{i\bullet} = \frac{n_{i\bullet}}{n} = \sum_{j=1}^J \frac{n_{ij}}{n} = \sum_{j=1}^J f_{ij}$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{n} = \sum_{i=1}^I \frac{n_{ij}}{n} = \sum_{i=1}^I f_{ij}$$

(ex : 3% = $\frac{6}{200}$)

(ex : 37% = $\frac{74}{200} = 14\% + 23\%$)

(ex : 31% = $\frac{62}{200} = 7\% + 14\% + 10\%$)

187

- Une distribution conditionnelle est une distribution statistique obtenue en **restreignant** la population à un **événement particulier** (une classe par exemple).
- $J = 2 \Rightarrow$ il y a **deux distributions** conditionnelles de X par rapport à Y .
 - la distribution de X sachant $Y \in [800, 1000[$.
 - la distribution de X sachant $Y \in [1000, 1200[$.
- $I = 3 \Rightarrow$ il y a **trois** distributions conditionnelles de Y par rapport à X
 - la distribution de Y sachant $X \in [20, 22[$.
 - la distribution de Y sachant $X \in [22, 24[$.
 - la distribution de Y sachant $X \in [24, 26[$.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (22.6%)	6 (4.3%)	20
[22, 24[(i = 2)	28 (45.2%)	46 (33.3%)	74
[24, 26[(i = 3)	20 (32.2%)	86 (62.4%)	106
Total	62 (100%)	138 (100%)	200

Formules : Pour $i = 1, \dots, I$ et pour $j = 1, \dots, J$

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$$

(ex : 22.6% = $\frac{14}{62}$)

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (7%)	6 (3%)	20 (10%)
[22, 24[(i = 2)	28 (14%)	46 (23%)	74 (37%)
[24, 26[(i = 3)	20 (10%)	86 (43%)	106 (53%)
Total	62 (31%)	138 (69%)	200 (100%)

- La distribution marginale de X est représentée par la colonne “total” (fréquences bleues).
- La distribution marginale de Y est représentée par la ligne “total” (fréquences vertes).
- Ce sont bien des **distributions** car lorsque l’on somme les $f_{i\bullet}$ ou les $f_{\bullet j}$, on obtient **100%**.
- $\Rightarrow$ puisqu’on a une distribution, on peut calculer tous les indicateurs du chapitre précédent.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (22.6 %)	6 (4.3 %)	20
[22, 24[(i = 2)	28 (45.2 %)	46 (33.3 %)	74
[24, 26[(i = 3)	20 (32.2 %)	86 (62.4 %)	106
Total	62 (100%)	138 (100%)	200

- On calcule les fréquences des âges en se restreignant à la sous-population des individus ayant un salaire entre **800 et 1000 €**, puis à la sous-population des individus ayant un salaire entre **1000 et 1200 €**.
- Les fréquences conditionnelles sont en général notées $f_{i|j}$
- Interprétation :
 - 22.6%** des employés ayant un salaire entre 800 et 1000 €sont âgés entre 20 et 22 ans.
 - Parmi les employés ayant un salaire entre 1000 et 1200 €, **62.4%** d’entre eux sont âgés entre 24 et 26 ans.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (70 %)	6 (30 %)	20 100%
[22, 24[(i = 2)	28 (37.8 %)	46 (62.2 %)	74 100%
[24, 26[(i = 3)	20 (18.9 %)	86 (81.1 %)	106 100%
Total	62	138	200

- Ces fréquences conditionnelles sont en général notées $f_{j|i}$
- Interprétation :
 - 70%** des employés âgés entre 20 et 22 ans ont un salaire compris entre 800 et 1000 €.
 - Parmi les employés âgés entre 22 et 24 ans, **62.2%** d’entre eux ont un salaire compris entre 1000 et 1200 €.

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (70%)	6 (30%)	20 100%
[22, 24[(i = 2)	28 (37.8%)	46 (62.2%)	74 100%
[24, 26[(i = 3)	20 (18.9%)	86 (81.1%)	106 100%
Total	62	138	200

Formules : Pour $i = 1, \dots, I$ et pour $j = 1, \dots, J$

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}} \quad \left(\text{ex : } 30\% = \frac{6}{20} \right)$$

193

Autour des tableaux de contingence
 Moyennes et Variances conditionnelles

Étude de la liaison entre deux variables

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14	6	20
[22, 24[(i = 2)	28	46	74
[24, 26[(i = 3)	20	86	106
Total	62	138	200

Concentrons-nous sur la variable X : on notera $\bar{x}_1$ (ou $\bar{x}_{|Y \in [800, 1000[}$) et $\bar{x}_2$ (ou $\bar{x}_{|Y \in [1000, 1200[}$) les deux moy. cond. de X sachant Y :

- La moyenne de X = la moyenne des moyennes conditionnelles

$$\bar{x} = \frac{1}{n} \sum_{j=1}^J n_{\bullet j} \bar{x}_j.$$

- Vérification :

- En utilisant la distribution marginale : $\bar{x} \approx 23.86$ ans.
- En utilisant les fréq. conditionnelles, $\bar{x}_1 \approx 23.19$ ans et $\bar{x}_2 \approx 24.16$ ans.
- En combinant $\frac{62 \times 23.19 + 138 \times 24.16}{200} \approx 23.86$ ans.

195

Il y a deux extrêmes du niveau de liaison entre deux variables (quelles que soient la ou les natures des variables) :

- l'indépendance totale (ou liaison nulle).
- la dépendance totale (ou liaison fonctionnelle).

Le but de cette section est de mesurer la dépendance, et de quantifier en particulier le niveau de proximité par rapport aux deux cas précédents.

197

Rappelons que

$$f_{ij} = \frac{n_{ij}}{n}, \quad f_{\bullet j} = \frac{n_{\bullet j}}{n}, \quad f_{i|j} = \frac{n_{ij}}{n_{\bullet j}}$$

⇒

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

De la même façon on peut obtenir

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

194

- Notons $Var_j(X)$ les variances conditionnelles de X sachant Y . Rappelons la formule de décomposition de la variance (qui peut s'exprimer en fonction des variances conditionnelles) :

$$Var(X) = \underbrace{\frac{1}{n} \sum_{j=1}^J n_{\bullet j} Var_j(X)}_{\text{variance intra}} + \underbrace{\frac{1}{n} \sum_{j=1}^J n_{\bullet j} (\bar{x}_j - \bar{x})^2}_{\text{variance inter}}$$

- La vérification sur l'exemple considéré est laissée en exercice.
- Des résultats tout à fait similaires sont bien évidemment valables pour la variable Y (notez que ceci est possible car Y est quantitative).

196

Définition

- La variable Y est totalement indépendante de la variable X si les variations de X n'entraînent pas de variations de Y .
- La variable X est totalement indépendante de la variable Y si les variations de Y n'entraînent pas de variations de X .

Théorème

- Y est totalement indépendante de X si et seulement si

$$f_{j|i} = f_{\bullet j}$$

(c-a-d les fréquences conditionnelles ne dépendent pas des lignes du tableau de contingence et sont égales aux fréquences marginales).

- X est totalement indépendante de Y si et seulement si

$$f_{i|j} = f_{i\bullet}$$

- L'indépendance est réciproque.

198

Théorème

Les variables X et Y sont indépendantes si et seulement si

$$f_{ij} = f_{i\bullet} \times f_{\bullet j} \iff n_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Corollaire

Un tableau de contingence est associé à deux variables X et Y indépendantes si et seulement si les **lignes (resp. colonnes)** sont **proportionnelles** entre elles.

Exemple : tableau associé à deux var. indépendantes

$X \setminus Y$	y_1	y_2	y_3	Total
x_1	2	4	12	18
x_2	4	8	24	36
Total	6	12	36	54

On peut par exemple vérifier que

$$\frac{n_{2\bullet} \times n_{\bullet 3}}{n} = \frac{36 \times 36}{54} = 24 = n_{23}$$

Exemple 1 :

$X \setminus Y$	y_1	y_2
x_1	2	0
x_2	1	0
x_3	0	1

$\Rightarrow Y$ est **totale-ment dépendante** de X et la réciproque est **fausse**.

Exemple 2 :

$X \setminus Y$	y_1	y_2	y_3
x_1	2	0	0
x_2	0	1	4

$\Rightarrow X$ est **totale-ment dépendante** de Y et la réciproque est **fausse**.

Exemple 3 :

$X \setminus Y$	y_1	y_2
x_1	2	0
x_2	0	1

$\Rightarrow X$ est **totale-ment dépendante** de Y et la réciproque est **vraie**.

Théorème

$$0 \leq \chi^2 \leq \chi^2_{\max} = n \times \min(I-1, J-1).$$

La quantité $\chi^2_{\max}$ est la valeur du χ^2 si la dépendance entre X et Y était totale et réciproque.

Définition

Le coefficient de Cramer $C \in [0, 1]$ est défini par

$$C = \sqrt{\frac{\chi^2}{\chi^2_{\max}}}$$

- Si C est proche de **0** alors les variables X et Y sont presque **indépendantes**.
- Si C est proche de **1**, alors les variables X et Y sont fortement **liées** (pas nécessairement liées fonctionnellement).
- Le C de Cramer peut être calculé pour n'importe quel type de variables et Y .

Définition

- Y est **totale-ment dépendante** de X (ou fonctionnellement liée à X) si à chaque valeur x_i de X correspond une unique valeur y_j de Y , autrement dit si chaque ligne du tableau de contingence ne contient qu'un seul effectif n_{ij} non nul.
- X est **totale-ment dépendante** de Y (ou fonctionnellement liée à Y) si à chaque valeur y_j de Y correspond une unique valeur x_i de X , autrement dit si chaque colonne du tableau de contingence ne contient qu'un seul effectif n_{ij} non nul.
- La dépendance totale n'est pas une **notion réciproque**.

Définition

Le χ^2 est un nombre mesurant l'écart entre la situation observée et la situation si les variables avaient été théoriquement **indépendantes**.

Méthodologie :

- construction du tableau de contingence sous hypothèse d'indépendance, c-a-d calcul des

$$n'_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

- on calcule ensuite

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - n'_{ij})^2}{n'_{ij}}$$

$X = \text{Age} \setminus Y = \text{Salaire}$	[800, 1000[($j = 1$)	[1000, 1200[($j = 2$)	Total
[20, 22[($i = 1$)	14 (6.2)	6 (13.80)	20
[22, 24[($i = 2$)	28 (22.94)	46 (51.06)	74
[24, 26[($i = 3$)	20 (32.86)	86 (73.14)	106
Total	62	138	200

- calcul des effectifs théoriques n'_{ij} .
Exemple : $n'_{32} = \frac{n_{3\bullet} \times n_{\bullet 2}}{n} = \frac{138 \times 106}{200} \approx 73.14$.
- Calcul du χ^2
$$\chi^2 = \frac{(14 - 6.2)^2}{6.2} + \frac{(6 - 13.8)^2}{13.8} + \dots + \frac{(86 - 73.14)^2}{73.14} \approx 23.13$$
- $\chi^2_{\max} = 200 \times \min(3-1, 2-1) = 200 \times 1 = 200$.
- $C = \sqrt{\frac{23.13}{200}} \approx 34\%$ (dépendance modérée).

Question

Quels sont les couples (x_i, y_j) qui contribuent le plus au χ^2 ?

Réponse : il suffit de calculer pour chaque case le rapport

$$\frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e}$$

$$\chi^2$$

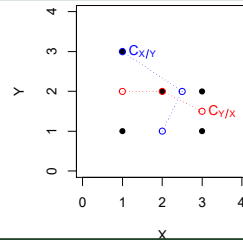
X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14 (42.4%)	6 (19.1%)	20
[22, 24[(i = 2)	28 (4.8%)	46 (2.2%)	74
[24, 26[(i = 3)	20 (21.8%)	86 (9.8%)	106
Total	62	138	200

● Exemple 1ère case : $((6.2 - 14)^2 / 6.2) / 23.13 \approx 42.4\%$.

● La case des individus **les plus jeunes et touchant le plus bas salaires** s'écarte le plus de l'hypothèse d'indépendance.

209

X \ Y	1	2	3	Total
1	1	0	1	2
2	0	1	0	1
3	1	1	0	2
Total	2	2	1	5



Définition

- la **courbe de régression de Y en X** est obtenue en faisant correspondre à chaque valeur de x_i de X la moy. conditionnelle de Y sachant $X = x_i$. Cette courbe est notée $C_{Y/X}$.
- la **courbe de régression de X en Y** est obtenue en faisant correspondre à chaque valeur de y_j de Y la moy. conditionnelle de X sachant $Y = y_j$. Cette courbe est notée $C_{X/Y}$.

213

Concept basé sur la formule de décomposition de la variance

$$\text{Var. Totale} = \text{Var(moy. cond.)} + \text{Moy(Var. Cond.)}$$

Définition

- Le rapport de corrélation de Y en X est défini par

$$\eta_{Y/X}^2 = \frac{\text{Var(moy. cond. de Y|X)}}{\text{Var(Y)}} = \frac{\frac{1}{n} \sum_i n_{i\bullet} (\bar{Y}_i - \bar{Y})^2}{\text{Var(Y)}}$$

- Le rapport de corrélation de X en Y est défini par

$$\eta_{X/Y}^2 = \frac{\text{Var(moy. cond. de X|Y)}}{\text{Var(X)}} = \frac{\frac{1}{n} \sum_j n_{\bullet j} (\bar{X}_j - \bar{X})^2}{\text{Var(X)}}$$

● $0 \leq \eta_{Y/X}^2 \leq 1$ et $0 \leq \eta_{X/Y}^2 \leq 1$

● Plus η^2 est **élevé** (resp. **faible**) et plus la liaison fonctionnelle est **forte** (resp. **faible**)

215

- pour savoir si X et Y sont liées fonctionnellement, on trace le nuage de points (x_i, y_i) .
- $\Rightarrow$ section valable uniquement pour X et Y **quantitatives**.
- $\Rightarrow$ il faut disposer des données brutes, autrement dit chaque couple (x_i, y_i) est observée une et une seule fois. Autrement dit, la table de contingence correspondante ne contient que des **0 ou 1**.
- On trace alors le nuage de points (x_i, y_j) et on essaie d'estimer la fonction de lien éventuelle.

210

Théorème

- Si X et Y sont deux variables indépendantes alors $C_{Y/X}$ est parallèle à l'axe des abscisses et la courbe $C_{X/Y}$ est parallèle à l'axe des ordonnées (**△**réciroque fausse).
- Si aucun point ne s'écarte de $C_{Y/X}$, Y totalement dépendante de X (**Y = f(X)**).
- Si aucun point ne s'écarte de $C_{X/Y}$, X totalement dépendante de Y (**X = f(Y)**).

214

X=Age \ Y=Salaire	[800, 1000[(j = 1)	[1000, 1200[(j = 2)	Total
[20, 22[(i = 1)	14	6	20
[22, 24[(i = 2)	28	46	74
[24, 26[(i = 3)	20	86	106
Total	62	138	200

Démarche pour calculer le rapport de corrélation de X en Y :

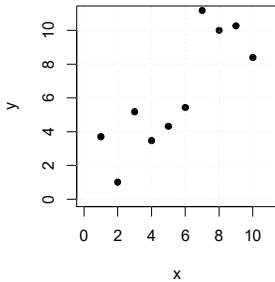
- calcul des moyenne et variance marginale de X : $\bar{x} \approx 23.86$ (ans) et $\text{Var}(X) \approx 1.78$ (ans²).
 - calcul des moyennes conditionnelles de X sachant $Y \in [800, 1000[$ et de X sachant $Y \in [1000, 1200[$: $\bar{x}_1 \approx 23.19$ (ans) et $\bar{x}_2 \approx 24.16$ (ans).
 - calcul de la variance interpopulation (var. moy. cond.)
- $$\text{Var. Inter} = \frac{62 \times (23.19 - 23.86)^2 + 138 \times (24.16 - 23.86)^2}{200} \approx 0.20 \text{ (ans}^2\text{)}.$$

● $\eta_{X/Y}^2 \approx \frac{0.2}{1.78} \approx 11.2\%$
(11.2 % de la variance de X est expliquée par la variable Y).

219

Régression linéaire

- Si le nuage de points observé est "presque" linéaire, il y a de fortes chances que la liaison entre X et Y soit linéaire (et que celle de Y à X soit linéaire).
- Exemple : imaginons observer le nuage suivant :



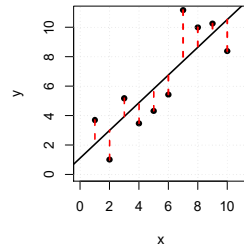
⇒ On peut suspecter une **liaison linéaire**. Pour mesurer ceci on utilise le coefficient de **corrélation linéaire**.

220

Méthode des moindres carrés

Si le coefficient r est jugé acceptable, on peut tenter d'estimer la droite de régression (de Y en X) en utilisant la **méthode des moindres carrés**

on se donne une droite d'équation $y = ax + b$, la MMC consiste à minimiser la somme des écarts rouges au carré.



Autrement dit, on va chercher le minimum en a et b de la fonction

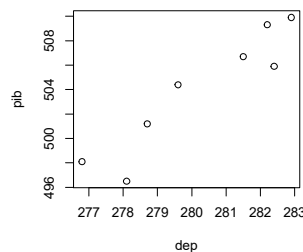
$$f(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2$$

222

Exemple d'application

Le tableau suivant présente les dépenses (**dep**) des ménages et PIB (**pib**) en milliards d'euros pour les 4 trimestres de 2011 et 2012. Peut-on expliquer l'évolution du PIB en fonction des dépenses ?

dep	278.1	276.8	278.7	279.6	282.4	281.5	282.2	282.9
pib	496.5	498.1	501.2	504.4	505.9	506.7	509.3	509.9



⇒ L'ajustement linéaire semble **adéquat et pertinent**.

224

Coefficient de corrélation linéaire

Soit (x_i, y_i) pour $i = 1, \dots, n$ un nuage de points. Ce coefficient est défini par

$$r = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

où

$$\text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \bar{xy} - \bar{x} \times \bar{y}.$$

- Si r est proche de **1 ou -1**, X et Y sont (certainement) **liées linéairement**.
- Si $r > 0$ la pente de la droite est > 0 . Si $r < 0$, la pente de la droite est < 0 .
- Si r est proche de **0**, l'ajustement linéaire n'est pas **acceptable** (ce qui ne signifie pas que X et Y ne puissent pas être liées par une fonction).
- r^2 est appelé coefficient de **détermination** ($0 \leq r^2 \leq 1$).

221

Solutions au problème

La droite de régression ...

- ... de Y en X a pour équation $y = \hat{a}x + \hat{b}$ avec

$$\hat{a} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \quad \text{et} \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

- ... de X en Y a pour équation $x = \hat{a}'y + \hat{b}'$ avec

$$\hat{a}' = \frac{\text{Cov}(X, Y)}{\text{Var}(Y)} \quad \text{et} \quad \hat{b}' = \bar{x} - \hat{a}'\bar{y}.$$

- les deux droites de régression passent par le point $(\bar{x}, \bar{y})$.
- On peut remarquer que

$$\hat{a} \times \hat{a}' = \frac{\text{Cov}(X, Y)^2}{\text{Var}(X) \text{Var}(Y)} = r^2.$$

223

Exemple d'application (2)

dep	278.1	276.8	278.7	279.6	282.4	281.5	282.2	282.9
pib	496.5	498.1	501.2	504.4	505.9	506.7	509.3	509.9

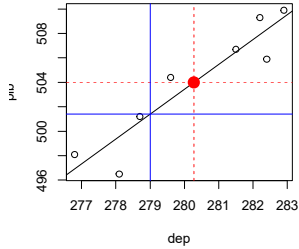
Démarche

- Calculez $\overline{dep}$, $\overline{pib}$, $\text{Var}(dep)$ et $\text{Var}(pib)$
 $\overline{dep} \approx 280.28$ (Me), $\overline{pib} \approx 504$ (Me), $\text{Var}(dep) \approx 4.54$ (Me)², $\text{Var}(pib) \approx 21.61$ (Me)²
- Calcul intermédiaire
 $\overline{dep \times pib} = \frac{1}{8}(278 \times 496 + \dots + 283 \times 510) = 141267.8$ (Me)².
- Calcul de la covariance
 $\text{Cov}(dep, pib) = \overline{dep \times pib} - \overline{dep} \times \overline{pib} \approx 9.25$ (Me)².
- Calcul du coefficient de corrélation linéaire
 $R = \frac{9.25}{\sqrt{21.61 \times 4.54}} \approx 93.38\%$ (ajustement linéaire très pertinent)
- Puisque l'ajustement linéaire est très bon, calculons la droite de régression

$$\hat{a} = \frac{9.25}{4.54} \approx 2.04 \quad \text{et} \quad \hat{b} = 504 - 2.04 \times 280.28 \approx -67.77$$
 (Me).

232

dep	278.1	276.8	278.7	279.6	282.4	281.5	282.2	282.9
pib	496.5	498.1	501.2	504.4	505.9	506.7	509.3	509.9



- La droite de régression

$$pib = 2.04 \times dep - 67.77.$$

passe par le point $(\overline{dep}, \overline{pib})$.
- Quelle estimation du PIB proposer pour une $dep = 279$ (M€) ? $\Rightarrow$

$$\widehat{pib} = 2.04 \times 279 - 67.77$$

$$= 501.39 \text{ (Me)}.$$

Chapitre 4. Les séries chronologiques

236

- 1 Présentation et analyse théorique des séries chronologiques
 - Généralités
 - Décomposition d'une série chronologique
 - Modélisation d'une série chronologique
- 2 Estimation de la tendance et des variations saisonnières
 - Estimation de la tendance
 - Estimaion des variations saisonnières
- 3 Série corrigée des variations saisonnières et prédiction
 - Série C.V.S.
 - Prédiction

237

Généralités

Définition

Une série chronologique (ou série temporelle) est une suite d'observations numériques **ordonnées dans le temps**.

Exemples :

- Consommation d'électricité mensuelle.
- Chiffre d'affaires trimestriel d'une entreprise.

Notations :

- la variable étudiée est notée y .
- le temps est repéré par la lettre t , $t = 1, 2, \dots$
- $y(t)$ est la valeur de la série étudiée à l'instant t .

238

Un premier exemple : tableaux

On s'intéresse ici au Chiffre d'Affaires trimestriel de l'entreprise Biomérieux entre 2010 et 2012. Deux façons **équivalentes** de présenter ce jeu de données :

	2010	2011	2012
I	306	327	362
II	344	345	387
III	333	347	382
IV	373	406	437

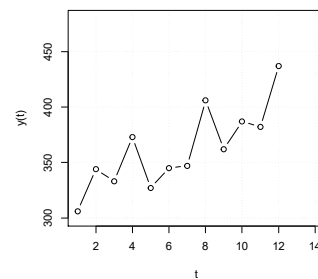
t	$y(t)$
1	306
2	344
3	333
4	373
5	327
6	345
7	347
8	406
9	362
10	387
11	382
12	437

239

Un premier exemple : représentation graphique

t	$y(t)$
1	306
2	344
3	333
4	373
5	327
6	345
7	347
8	406
9	362
10	387
11	382
12	437

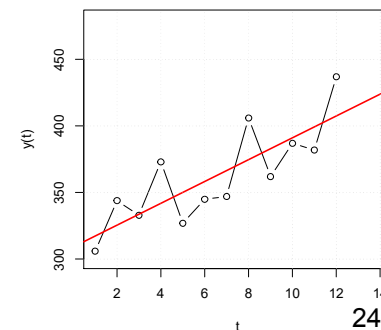
- En abscisse : **le temps t** .
- En ordonnée : **la variable $y(t)$** .



240

Différentes composantes : I. la tendance

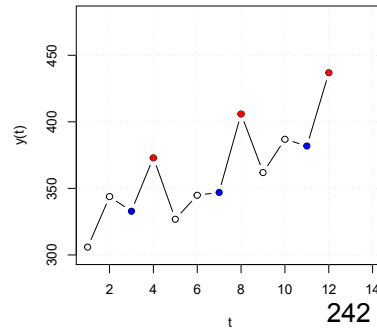
La **tendance** ou **trend** : traduit généralement l'évolution générale du phénomène observé (croissance, décroissance, stagnation, ...). C'est une courbe (ici une droite) que l'on note $T(t)$ (fonction de t).



241

Les **variations saisonnières** : ce sont des **pics** ou des **creux** qui se répètent de cycle en cycle. Ici le cycle est de un an, c-a-d 4 trimestres. La fonction de saisonnalité se note $S(t)$:

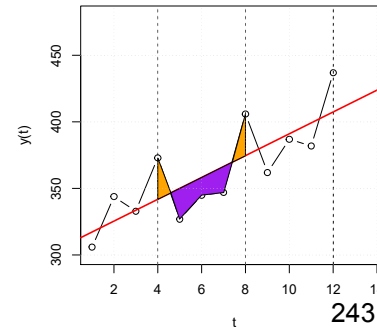
- 1 $S(t)$ doit être une fonction **périodique**. Autrement dit, si le cycle est de un an et si les observations sont
 - mensuelles : $S(t + 12) = S(t)$.
 - trimestrielles : $S(t + 4) = S(t)$.



242

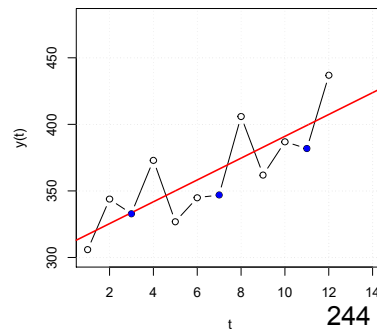
Les **variations saisonnières** : ce sont des **pics** ou des **creux** qui se répètent de cycle en cycle. Ici le cycle est de un an, c-a-d 4 trimestres. La fonction de saisonnalité se note $S(t)$:

- 1 $S(t)$ doit être une fonction **périodique**.
- 2 Les variations saisonnières se **compensent** sur un cycle. C'est le **principe de conservation des aires** : surface violette = surface orange pour tous les cycles.



243

Ce sont des pics ou des creux qui "cassent" la monotonie des cycles. On les note $\varepsilon(t)$. Dans la suite, on ne tiendra pas compte de ces variations accidentelles.



244

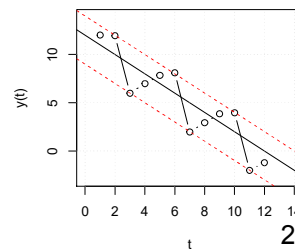
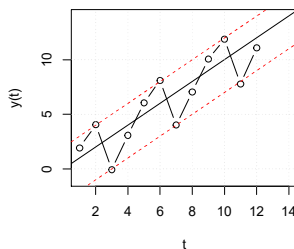
- 1 **Modéliser** une série temporelle consiste à supposer qu'elle est régie par une certaine fonction.
- 2 En particulier, on supposera que $y(t) = f(T(t), S(t))$, c-a-d fonction d'une certaine tendance et de variations saisonnières.
- 3 **Restriction du cours** : ne seront considérés ici que
 - 1 les modèles **additifs**.
 - 2 les modèles **multiplicatifs**.

247

- 1 Il est défini par

$$y(t) = T(t) + S(t)$$

- 2 Il est utilisé lorsque les **fluctuations/variations** autour de la tendance sont **constantes**.

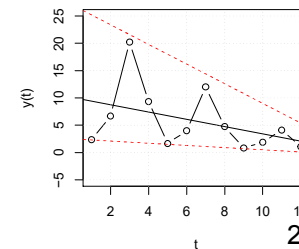
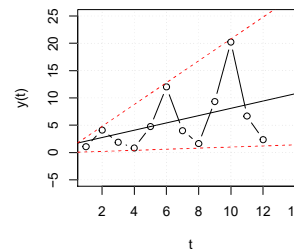


248

- 1 Il est défini par

$$y(t) = T(t) \times S(t)$$

- 2 Il est utilisé lorsque les **fluctuations/variations** autour de la tendance sont **croissantes ou décroissantes**.



249

L'estimation des variations saisonnières ...

- se fait **après** avoir estimé la tendance.
- est **dépendante d'un modèle envisagé** : additif ou multiplicatif (**dans ce cours**).

On suppose ci-dessous que $p = 4$ (4 observations par cycle)

<u>Modèle additif</u> : $y(t) = T(t) + S(t)$	<u>Modèle multiplicatif</u> : $y(t) = T(t) \times S(t)$
Calculer la fonction $S(t) = y(t) - T(t)$.	Calculer la fonction $S(t) = y(t)/T(t)$.
Périodiser la fonction S : définir $p = 4$ coefficients saisonniers par exemple : $S_1 = (S(1) + S(5) + S(9))/3$	
Principe de conservations des aires : calculer $\bar{S} = (S_1 + \dots + S_4)/4$.	
et définir les coefficients saisonniers modifiés S'_j pour $j = 1, \dots, 4$	
$S'_j = S_j - \bar{S}$	$S'_j = S_j / \bar{S}$

...en supposant un **modèle additif** car les fluctuations de la série autour de la tendance semblent **constantes**.

t	$y(t)$	$M_4(t)$	$T(t)$	$S(t)$	S_j	S'_j
1	306	$\times$	322.2	-16.2		
2	344	$\times$	329.1	14.9		
3	333	341.6	336.1	-3.1		
4	373	344.4	343	30		
5	327	346.2	350	-23		
6	345	352.1	356.9	-11.9		
7	347	360.6	363.9	-16.9		
8	406	370.2	370.8	35.2		
9	362	379.9	377.8	-15.8		
10	387	388.1	384.7	2.3		
11	382	$\times$	391.7	-9.7		
12	437	$\times$	398.6	38.4		

- ➊ Etape 1 :
- $$S(t) = y(t) - T(t) .$$

...en supposant un **modèle additif** car les fluctuations de la série autour de la tendance semblent **constantes**.

t	$y(t)$	$M_4(t)$	$T(t)$	$S(t)$	S_j	S
1	306	$\times$	322.2	-16.2	-18.3	
2	344	$\times$	329.1	14.9	1.8	
3	333	341.6	336.1	-3.1	-9.9	
4	373	344.4	343	30	34.5	
5	327	346.2	350	-23	-18.3	
6	345	352.1	356.9	-11.9	1.8	
7	347	360.6	363.9	-16.9	-9.9	
8	406	370.2	370.8	35.2	34.5	
9	362	379.9	377.8	-15.8	-18.3	
10	387	388.1	384.7	2.3	1.8	
11	382	$\times$	391.7	-9.7	-9.9	
12	437	$\times$	398.6	38.4	34.5	

- ① Etape 1 : $S(t) = y(t) - T(t)$.
 ② Etape 2 : calcul des S_j .
 ex :
 $S_1 = (-16.2 - 23 - 15.8)/3$

...en supposant un **modèle additif** car les fluctuations de la série autour de la tendance semblent **constantes**.

t	$y(t)$	$M_4(t)$	$T(t)$	$S(t)$	S_j	S'_j
1	306	$\times$	322.2	-16.2	-18.3	-20.3
2	344	$\times$	329.1	14.9	1.8	-0.2
3	333	341.6	336.1	-3.1	-9.9	-11.9
4	373	344.4	343	30	34.5	32.5
5	327	346.2	350	-23	-18.3	-20.3
6	345	352.1	356.9	-11.9	1.8	-0.2
7	347	360.6	363.9	-16.9	-9.9	-11.9
8	406	370.2	370.8	35.2	34.5	32.5
9	362	379.9	377.8	-15.8	-18.3	-20.3
10	387	388.1	384.7	2.3	1.8	-0.2
11	382	$\times$	391.7	-9.7	-9.9	-11.9
12	437	$\times$	398.6	38.4	34.5	32.5

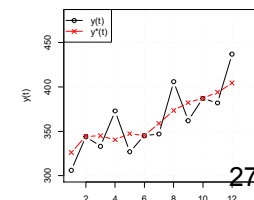
- ➊ Etape 1 :
 $S(t) = y(t) - T(t).$
- ➋ Etape 2 : calcul des S_j .
- ➌ Etape 3 : calcul de $\overline{S}$ et de S'_j
$$\overline{S} = (S_1 + \dots + S_4)/4 \approx 2$$
$$S'_j = S_j - \overline{S}$$

Définition

- La série corrigée des variations saisonnières est une série temporelle permettant de pouvoir comparer les valeurs numériques de la série initiale indépendamment des fluctuations saisonnières. Cette série est notée $y^*(t)$.
- La définition dépend du modèle choisi :
 - modèle additif : $y^*(t) = y(t) - S'_t$.
 - modèle multiplicatif : $y^*(t) = y(t)/S'_t$.

Exemple de calcul sur l'exemple : $y^*(1) = 306 - (-20.3) = 326.3$

t	$y(t)$	$T(t)$	S'_j	$y^*(t)$
1	306	322.2	-20.3	326.3
2	344	329.1	-0.2	344.2
3	333	336.1	-11.9	344.9
4	373	343	32.5	340.5
5	327	350	-20.3	347.3
6	345	356.9	-0.2	345.2
7	347	363.9	-11.9	358.9
8	406	370.8	32.5	373.5
9	362	377.8	-20.3	382.3
10	387	384.7	-0.2	387.2
11	382	391.7	-11.9	393.9
12	437	398.6	32.5	404.5



276

Chapitre 5. Indices élémentaires et synthétiques

278

Indices élémentaires

Indices synthétiques

Définitions générales

Objectif et Notations

Objectif

Comparer des **grandeurs numériques** qui évoluent au cours du **temps et/ou de l'espace**.

Notations :

- V_t est la valeur de la grandeur étudiée au temps t (t est la **date courante**).
- V_0 est la valeur de la grandeur étudiée au temps 0 (0 est la **date de référence**).

280

Indices élémentaires

Indices synthétiques

Définitions générales

Comment interpréter la valeur d'un indice ?

Interprétation

Soit $I_{t/0}$ l'indice élémentaire d'une valeur V à la date t base 100 à la date 0 et soit $\Delta = I_{t/0} - 100$

- $\Delta > 0 \Leftrightarrow$ la valeur V a **augmenté** de $\Delta\%$ de la date 0 à la date t .
- $\Delta < 0 \Leftrightarrow$ la valeur V a **diminué** de $\Delta\%$ de la date 0 à la date t .

Retour sur exemple :

$I_{12/10}^P = 120 \Leftrightarrow$ le prix a **augmenté de 20%** de 2010 à 2012.

286

- Indices élémentaires
 - Définitions générales
 - Propriétés

- Indices synthétiques
 - Définitions et objectifs
 - Indices de Laspeyres, Paasche et Fisher

279

Indices élémentaires

Indices synthétiques

Définitions générales

Définition

L'indice élémentaire d'une grandeur V à la date t base 100 à la date 0 est

$$I_{t/0} = I_{t/0}^V = \frac{V_t}{V_0} \times 100.$$

Exemple : Quel est l'indice du prix base 100 en 2010 d'un produit valant 250 € en 2010 et 300 € en 2012 ?

Réponse :

$$I_{12/10}^P = \frac{P_{12}}{P_{10}} \times 100 = \frac{300}{250} \times 100 = 120.$$

282

Indices élémentaires

Indices synthétiques

Propriétés

Circularité ou transférabilité

Proposition

Si une grandeur numériques V prend les valeurs V_0 , V_t et $V_{t'}$ aux instants 0, t , t' alors

$$I_{t/0} = I_{t/t'} \times I_{t'/0} \times \frac{1}{100}$$

Remarque :

- on peut connaître l'évolution de 0 à t connaissant les évolutions de 0 à t' et de t' à t . Pas besoin d'observer V_0 , V_t et $V_{t'}$.
- cette propriété peut être généralisée ...

Preuve : pour démontrer ce résultat, on part du terme de droite

$$\begin{aligned} I_{t/t'} \times I_{t'/0} \times \frac{1}{100} &= \left(\frac{V_t}{V_{t'}} \times 100 \right) \left(\frac{V_{t'}}{V_0} \times 100 \right) \times \frac{1}{100} \\ &= \frac{V_t}{V_0} \times 100 = I_{t/0}. \end{aligned}$$

288

Exemple

Le chiffre d'affaires d'une entreprise a augmenté de 30% de 2010 à 2011 et diminué de 25% de 2011 à 2012. Le CA a-t-il diminué ou augmenté de 2010 à 2012 ?

⚠ La réponse n'est pas 30%-25% = augmentation de 5% ! Les pourcentages ne s'additionnent pas.

Formalisation énoncé : on a $I_{11/10} = 130$, $I_{12/11} = 75$ et on cherche $I_{12/10}$. Par la propriété de circularité

$$\begin{aligned} I_{12/10} &= I_{12/11} \times I_{11/10} \times \frac{1}{100} \\ &= \frac{130 \times 75}{100} = 97.5. \end{aligned}$$

Autrement dit, le CA a **diminué de 2.5 %** de 2010 à 2012.

296

Exemple

Si un prix augmente de 20% de 2010 à 2012, que dire de son évolution de 2012 à 2010 ?

⚠ La réponse n'est pas diminution de 20%. Les pourcentages ne sont pas **réversibles**.

Formalisation : on a $I_{12/10} = 120$ et on cherche $I_{10/12}$. Par la propriété de réversibilité

$$I_{10/12} = \frac{100^2}{I_{12/10}} = \frac{100^2}{120} \approx 83.33.$$

Autrement dit, le prix a **diminué de 16.67%** de 2012 à 2010.

302

Exemple

Soit P et Q les prix et quantités d'un produit vendu par une entreprise. Si le prix de ce produit augmente de 60% de 2000 à 2010 et si les quantités vendues ont diminué de 50% de 2000 à 2010, quelle est l'évolution des recettes de 2000 à 2010 ?

Formalisation : on a

$$I_{10/00}^P = 160 \quad \text{et} \quad I_{10/00}^Q = 50$$

et on cherche $I_{10/00}^R$ où $R = P \times Q$. Par la propriété précédente on a

$$I_{10/00}^R = I_{10/00}^P \times I_{10/00}^Q \times \frac{1}{100} = \frac{160 \times 50}{100} = 80.$$

Autrement dit, les recettes ont **diminué de 20%** de 2000 à 2010.

305

Problème : connaissant l'évolution de 0 à t d'une certaine quantité, peut-on connaître l'évolution de t à 0 ?

Proposition

Si une grandeur numérique V prend les valeurs V_0 et V_t aux instants 0 et t alors

$$I_{0/t} = \frac{100^2}{I_{t/0}}.$$

Preuve : à nouveau on part du terme de droite

$$\begin{aligned} \frac{100^2}{I_{t/0}} &= \frac{100^2}{\frac{V_t}{V_0} \times 100} \\ &= \frac{V_0}{V_t} \times 100 = I_{0/t}. \end{aligned}$$

298

Proposition

Si une grandeur a est égale à tout instant

- 1 au produit des grandeurs b et c , c-a-d $a = b \times c$ alors

$$I_{t/0}^a = (I_{t/0}^b \times I_{t/0}^c) \times \frac{1}{100}.$$

- 2 au rapport des grandeurs b et c , c-a-d $a = b/c$ alors

$$I_{t/0}^a = \frac{I_{t/0}^b}{I_{t/0}^c} \times 100.$$

Preuve : démontrons la 1ère propriété en partant du terme de droite

$$\begin{aligned} (I_{t/0}^b \times I_{t/0}^c) \times \frac{1}{100} &= \left(\frac{b_t}{b_0} \times 100 \right) \left(\frac{c_t}{c_0} \times 100 \right) \times \frac{1}{100} \\ &= \frac{b_t \times c_t}{b_0 \times c_0} \times 100 = \frac{a_t}{a_0} \times 100 = I_{t/0}^a. \end{aligned}$$

303

- Utilisation : étude de **grandeurs complexes** c-a-d de grandeurs composées de plusieurs grandeurs simples.
- Exemple : comprendre l'évolution

- du prix d'un plat : grandeur composée des prix et quantités de matières premières (farine, huile, ...).
- de la recette globale d'une entreprise : grandeur composée des prix et quantités de l'ensemble des produits vendus par cette entreprise.
- du nombre d'entrées au cinéma : grandeur dépendant du nombre de films et du nombre moyen d'entrées par film pour différents types de films (français, américain, ...).
- ...

306

- i : indice d'un produit parmi un panier de I produits.
- $P_i(t)$ (ou $P_{i,t}$) : Prix du produit i à l'instant t .
- $Q_i(t)$ (ou $Q_{i,t}$) : Quantités du produit i à l'instant t .

Définitions

- $P_i(t) \times Q_i(t) = V_i(t)$ est la **valeur globale du produit** i à l'instant t .
- $V(t) = \sum_i P_i(t)Q_i(t)$ est la **valeur globale de l'ensemble des produits**.
- L'indice de valeur globale est

$$I_{t/0}^V = \frac{V(t)}{V(0)} \times 100 = \frac{\sum_i P_i(t)Q_i(t)}{\sum_i P_i(0)Q_i(0)} \times 100$$

307

- Pour pouvoir déterminer l'influence des prix ou des quantités, il faut définir des scénari permettant d'isoler soit les prix soit les quantités.
- C'est l'idée des indices de Laspeyres et de Paasche :
 - l'indice de **Laspeyres** des **prix**, par exemple, regarde l'évolution des prix de **0** à **t** en supposant que les **quantités** sont fixées à la date **0**.
 - l'indice de **Paasche** des **quantités**, par exemple, regarde l'évolution des quantités de **0** à **t** en supposant que les **prix** sont fixées à la date **t**.

313

Définition

Les indices de Laspeyres des prix et des quantités sont respectivement définis par :

$$L_{t/0}^P = \frac{\sum_i P_i(t)Q_i(0)}{\sum_i P_i(0)Q_i(0)} \times 100 \quad \text{et} \quad L_{t/0}^Q = \frac{\sum_i P_i(0)Q_i(t)}{\sum_i P_i(0)Q_i(0)} \times 100.$$

Comment interpréter les valeurs $L_{12/10}^P \approx 136.48$ et $L_{12/10}^Q \approx 116.39$?

- A **quantités fixées en 2010**, les **prix** ont augmenté de 36.48% de 2010 à 2012.
- A **prix fixés en 2010**, les **quantités** ont augmenté de 16.39% de 2010 à 2012.

316

Etudions le panier de trois produits de la TAG vendus à une sous-population (non précisée ici) en 2010 et 2012.

Produit	2010		2012	
	Prix	Quantités	Prix	Quantités
Ticket simple	1.5	100	2	100
Abonnement hebdomadaire	35	40	50	50
Abonnement mensuel	130	35	175	40

Question : quelle est l'évolution des recettes de la TAG pour ce panier et la sous-population étudiée ?

Réponse :

$$I_{12/10}^V = \frac{2 \times 100 + 50 \times 50 + 175 \times 40}{1.5 \times 100 + 35 \times 40 + 130 \times 35} \times 100 \approx 159.02$$

Les recettes ont donc **augmenté de 59.02%** de 2010 à 2012.

Mais cette augmentation est-elle plus due à une augmentation des prix ou des quantités ?

311

Définition

Les indices de Laspeyres des prix et des quantités sont respectivement définis par :

$$L_{t/0}^P = \frac{\sum_i P_i(t)Q_i(0)}{\sum_i P_i(0)Q_i(0)} \times 100 \quad \text{et} \quad L_{t/0}^Q = \frac{\sum_i P_i(0)Q_i(t)}{\sum_i P_i(0)Q_i(0)} \times 100.$$

Application : indices de Laspeyres des prix et quantités sur l'exemple

Produit	2010		2012	
	Prix	Quantités	Prix	Quantités
Ticket simple	1.5	100	2	100
Abonnement hebdomadaire	35	40	50	50
Abonnement mensuel	130	35	175	40

$$L_{12/10}^P = \frac{2 \times 100 + 50 \times 40 + 175 \times 35}{1.5 \times 100 + 35 \times 40 + 130 \times 35} \times 100 \approx 136.48.$$

$$L_{12/10}^Q = \frac{1.5 \times 100 + 35 \times 50 + 130 \times 40}{1.5 \times 100 + 35 \times 40 + 130 \times 35} \times 100 \approx 116.39.$$

315

Définition

Les indices de Paasche des prix et des quantités sont respectivement définis par :

$$P_{t/0}^P = \frac{\sum_i P_i(t)Q_i(t)}{\sum_i P_i(0)Q_i(t)} \times 100 \quad \text{et} \quad P_{t/0}^Q = \frac{\sum_i P_i(t)Q_i(t)}{\sum_i P_i(t)Q_i(0)} \times 100.$$

Application : indices de Paasche des prix et quantités sur l'exemple

Produit	2010		2012	
	Prix	Quantités	Prix	Quantités
Ticket simple	1.5	100	2	100
Abonnement hebdomadaire	35	40	50	50
Abonnement mensuel	130	35	175	40

$$P_{12/10}^P = \frac{2 \times 100 + 50 \times 50 + 175 \times 40}{1.5 \times 100 + 35 \times 50 + 130 \times 40} \times 100 \approx 136.62.$$

$$P_{12/10}^Q = \frac{2 \times 100 + 50 \times 50 + 175 \times 40}{2 \times 100 + 50 \times 40 + 175 \times 35} \times 100 \approx 116.52.$$

318

Définition

Les indices de Paasche des prix et des quantités sont respectivement définis par :

$$\mathbf{P}_{t/0}^P = \frac{\sum_i P_i(t) Q_i(t)}{\sum_i P_i(0) Q_i(t)} \times 100 \quad \text{et} \quad \mathbf{P}_{t/0}^Q = \frac{\sum_i P_i(t) Q_i(t)}{\sum_i P_i(t) Q_i(0)} \times 100.$$

Comment interpréter les valeurs $P_{12/10}^P \approx 136.62$ et $P_{12/10}^Q \approx 116.52$?

- A **quantités fixées en 2012**, les **prix** ont augmenté de 36.62% de 2010 à 2012.
- A **prix fixés en 2012**, les **quantités** ont augmenté de 16.52% de 2010 à 2012.

319

$I_{12/10}^V$	$L_{12/10}^P$	$P_{12/10}^P$	$F_{12/10}^P$	$L_{12/10}^Q$	$P_{12/10}^Q$	$F_{12/10}^Q$
159.02	136.48	136.62	136.55	116.39	116.52	116.45

Interprétation :

- D'un point de vue global, l' **augmentation de 59.02%** des recettes globales de la TAG est **d'avantage due à l'augmentation des prix** (de l'ordre de 36.5%) qu'à l' **augmentation des quantités** (de l'ordre de 16.4%).

321

$I_{12/10}^V$	$L_{12/10}^P$	$P_{12/10}^P$	$F_{12/10}^P$	$L_{12/10}^Q$	$P_{12/10}^Q$	$F_{12/10}^Q$
159.02	136.48	136.62	136.55	116.39	116.52	116.45

Par exemple,

$$L_{12/10}^P \times P_{12/10}^Q \times \frac{1}{100} = \frac{136.48 \times 116.52}{100} \approx 159.03 \approx I_{12/10}^V$$

$$F_{12/10}^P \times F_{12/10}^Q \times \frac{1}{100} = \frac{136.55 \times 116.45}{100} \approx 159.01 \approx I_{12/10}^V$$

323

Résumé :

$I_{12/10}^V$	$L_{12/10}^P$	$P_{12/10}^P$	$L_{12/10}^Q$	$P_{12/10}^Q$
159.02	136.48	136.62	116.39	116.52

- Il est possible que les indices de Laspeyres et de Paasche soient très différents.
- Pour nuancer cette possible situation, Fisher a défini des indices **moyens** des indices précédents.

Définition

L'indice de Fisher des prix (resp. des quantités) est défini comme la **moyenne géométrique** des indices de Laspeyres et de Paasche des prix (resp. des quantités), c-a-d

$$F_{t/0}^P = \sqrt{L_{t/0}^P \times P_{t/0}^P} \quad \text{et} \quad F_{t/0}^Q = \sqrt{L_{t/0}^Q \times P_{t/0}^Q}$$

320

S'il n'y avait qu'un seul produit

$$I_{t/0}^V = I_{t/0}^P \times I_{t/0}^Q \times \frac{1}{100}$$

car $V = P \times Q$. Lorsque l'on a plusieurs produits, on a les relations suivantes

Proposition

$$\begin{aligned} I_{t/0}^V &= L_{t/0}^P \times P_{t/0}^Q \times \frac{1}{100} \\ &= P_{t/0}^P \times L_{t/0}^Q \times \frac{1}{100} \\ &= F_{t/0}^P \times F_{t/0}^Q \times \frac{1}{100} \end{aligned}$$

322

Rappel : lorsqu'il n'y a qu'un seul produit, on a par exemple pour les prix (valable aussi pour les quantités)

$$I_{0/t}^P = \frac{100^2}{I_{t/0}^P}.$$

En présence de plusieurs produits, ceci n'est plus tout à fait vrai. Cependant,

proposition

Notons $\bullet = P$ ou Q (prix ou quantités)

$$L_{0/t}^\bullet = \frac{100^2}{P_{t/0}^\bullet} \quad \text{et} \quad P_{0/t}^\bullet = \frac{100^2}{L_{t/0}^\bullet}.$$

Par contre les indices de Fisher sont réversibles. Ainsi

$$F_{0/t}^\bullet = \frac{100^2}{F_{t/0}^\bullet}.$$

324

$I_{12/10}^V$	$L_{12/10}^P$	$P_{12/10}^P$	$F_{12/10}^P$	$L_{12/10}^Q$	$P_{12/10}^Q$	$F_{12/10}^Q$
159.02	136.48	136.62	136.55	116.39	116.52	116.45

Par exemple,

- ❶ quel est l'indice de Laspeyres des prix en 2010 base 100 en 2012 ?

Réponse :

$$L_{10/12}^P = \frac{100^2}{P_{12/10}^P} = \frac{100^2}{136.62} \simeq 73.2$$

- ❷ quel est l'indice de Fisher des quantités en 2010 base 100 en 2012 ?

Réponse :

$$F_{10/12}^Q = \frac{100^2}{F_{12/10}^Q} = \frac{100^2}{116.45} \simeq 85.87.$$