Techniques d'analyse Statistique

(cours pour débutant)

Youssef LAHARACH

- - SOMMAIRE - -

1. Généralités

Éléments de vocabulaire:

• La plupart du temps les données se présentent sous la forme suivante:

- n unités ou observations
- p variables numériques.

• Lorsque n et p sont grands on cherche à synthétiser cette masse d'informations sous une forme exploitable et compréhensible.

• Une première étape consiste à décrire séparément les résultats obtenus pour chaque variable: c'est *la description unidimensionnelle*. On considérera donc ici qu'on ne s'intéresse qu'à une variable X, (appelée encore caractère), dont on possède n valeurs x_1, x_2,…x_n

1. Généralités

Population:

la population désigne l'ensemble des individus étudiés.

☼ Exemples:

- la population résidante au Maroc le 1 janvier 2005 à 0 heure.
- le parc automobile marocain au 31 décembre 2004.

Les unités statistiques ou individus:

L'individu est l'unité de base sur laquelle sont réalisées un certain nombre de mesures: on emploie aussi le terme d'observation.

Il arrivera donc que l'on désigne par individu des objets ou des événements:

- une consultation de médecin
- un accident de la route.

1. Généralités

Les variables:

Une variable est *une application* qui associe à chaque individu une valeur unique parmi un ensemble de valeurs possibles, appelées *modalités* de la variable.

Les modalités d'un même caractère doivent être à la fois incompatibles et exhaustives.

Chaque individu de la population présente une et une seulement des modalités du caractère.

Il existe différents types de variables, définis selon la nature de leurs modalités:

Variables numériques:

Les variables numériques sont des variables dont les modalités sont des nombres réels, *mesurables*, sur lesquels il est possible d'effectuer des opérations algébriques.

1. Généralités

- Une variable numérique est *Discrète* si ses modalités sont des valeurs isolées.

☼ Exemple: nombre d'enfants d'une famille
nombre de pièces d'un logement.

- Une variable est dite *continue* si ses modalités sont des valeurs quelconques d'un intervalle.

☼ Exemple: Poids, température, prix, salaire.

- Pour étudier une variable statistique continue, on définit des classes (ou tranches) de valeurs possibles, qui sont les modalités du caractère. Ces derniers peuvent avoir une amplitude constante ou variable.

1. Généralités

Variables qualitatives:

- Les variables qualitatives sont des variables dont les modalités sont non numériques, non mesurables (ou numériques, mais ne sont que des codes).
- Une variable qualitative est dite:

Ordinale: s'il existe un ordre sur ses modalités.

☼ Exemple: niveau de satisfaction, niveau hiérarchique.

Nominale: dans le case contraire.

☼ Exemple: sexe, profession.

2. Les distributions statistiques à un caractère

☞ Les tableaux statistiques:

Leur présentation diffère légèrement selon la nature des variables

1. Variables discrètes:

Soit X une variable observée sur une population de n individus, pouvant prendre les modalités $X_1,\ldots,X_p$.

On désigne par n_i le nombre d'individus prenant la modalité Xi appelé *effectif* de la modalité.

On appelle *fréquence* f_i de la modalité la proportion n_i/n.

On a les relations:

$$\sum_{i=1}^{p} n_i = n \quad \text{et} \quad \sum_{i=1}^{p} f_i = 1$$

2. Les distributions statistiques à un caractère

- On peut représenter la distribution de X sur la population par un tableau de la forme suivante:

X	Effectifs n_i	Fréquence f_i
X_1	n_1	f_1
…	…	…
X_i	n_i	f_i
…	…	…
X_p	n_p	f_p
Total	n	1

2. Les distributions statistiques à un caractère

2. Variables continues:

On regroupe les valeurs en k *classes* d'extrémités e_0,…e_k et l'on note pour chaque classe $[e_{i-1},e_i[$, l'effectif n_i et la fréquence f_i.

Les *effectifs cumulés* et *fréquences cumulées* F_i: effectif ou proportion d'individus pour lesquels $X<=e_i$ (ou $X<e_i$).

Le *centre de classe* C_i vaut $ci=(e_{i-1}+e_i)/2$.

L'amplitude de la classe C_i vaut $a_i=e_i-e_{i-1}$.

2. Les distributions statistiques à un caractère

X	Centre de classe c_i	Effectif n_i	Effectif cumulé	Fréquence f_i	Fréquence cumulée F_i
$[e_0,e_1[$	c_1	n_1	n_1	f_1	f_1
…	…	…	…	…	…
$[e_{i-1},e_i[$	c_i	n_i	$n_1+...n_i$	f_i	$f_1+…+f_i$
…	…	…	…	…	…
$[e_{p-1},e_p[$	c_p	n_p	n	f_p	1

2. Les distributions statistiques à un caractère

☞ Représentations graphiques:

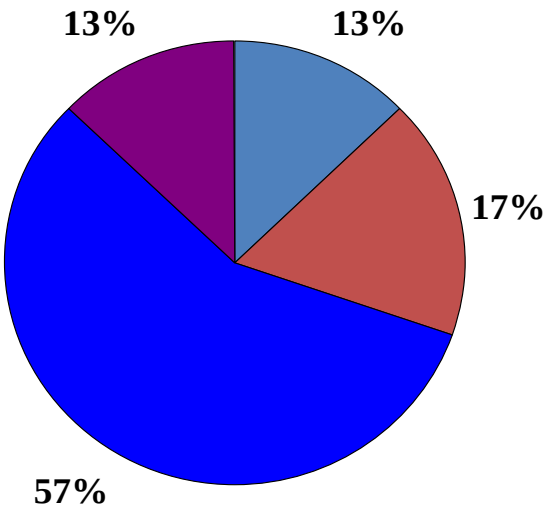
1. caractères qualitatifs:

Le principe de la représentation des caractères qualitatifs est la proportionnalités des aires aux effectifs.

Le *diagramme circulaire* («camembert »)

Chaque modalité est représentée par un secteur angulaire faisant un angle au centre.

Les secteurs circulaires ont un angle au centre proportionnel à l'effectif correspondant, et par conséquent une aire proportionnel à l'effectif.



$$\alpha_i = f_i * 360° = \frac{n_i}{n} * 360°$$

2. Les distributions statistiques à un caractère

2. Caractères quantitatives

Si le caractère est quantitatif, on utilise deux sortes de représentations:

- Le diagramme différentiel (diagramme en bâtons, histogramme)
- Le diagramme intégral (courbe cumulative).

a. Variable statistique discrète:

Le diagramme en bâtons: A chaque modalité, on fait correspondre un segment dont la longueur est proportionnelle à n_i ou f_i.

La courbe cumulative: Pour toute valeur x, on définit une fonction *cumulative* ou de *répartition.*

$$F(x) = \sum_{j=1}^{i} f_j \text{ pour } x_i \leq x \prec x_{i+1}$$

2. Les distributions statistiques à un caractère

b. Variable statistique continue:

- l'histogramme:

La surface des rectangles est proportionnelle aux effectifs ou aux fréquences (et non pas la hauteur des rectangles comme dans le digramme en tuyaux d'orgue).

L'histogramme est la courbe représentative des fréquences:

Si à la classe i d'amplitude a_i définie par $e_{i-1} <= x < e_i$ où $e_i = e_{i-1} + a_i$ correspond l'effectif n_i, c'est-à-dire la fréquence f_i, la fréquence moyenne par unité d'amplitude est f_i/a_i.

L'histogramme est constitué de tuyaux d' orgue juxtaposés dont les bases sont les clases ci et dont les hauteurs sont les fréquences moyennes par unité d'amplitude.

2. Les distributions statistiques à un caractère

☞ Exemple:

Répartition d'entreprises selon le nombre de leurs salariés:

Classes Ci	ai	ni	ni/ai	fi	Fi
0 à 10	10	350	35	0,35	0,35
10 à 20	10	240	24	0,24	0,59
20 à 50	30	150	5	0,15	0,74
50 à 100	50	90	1,8	0,09	0,83
100 à 200	100	70	0,7	0,07	0,9
200 à 500	300	60	0,2	0,06	0,96
500 à 1000	500	40	0,08	0,04	1
		1000			

2. Les distributions statistiques à un caractère

☞ La courbe cumulative :

- La fonction cumulative F(x) est la proportion d'individus de la population dont le caractère est inférieur à x.

- Cette fonction est connue seulement pour les valeurs de x qui sont des extrémités de classe:

  $x = e_0, \ldots, e_k$.

$$F(e_i) = \sum_{j=1}^{i} f_j$$

- La courbe cumulative est la courbe passant par les points de coordonnées e_i, $F(e_i)$. La fonction cumulative F(x) est monotone non décroissante.

3. Description numérique d'une variable statistique

A. Caractéristiques de tendance centrale:

Il s'agit de définir une valeur c autour de la quelle se répartissent les observations. Les plus utilisées sont *le mode, la médiane et la moyenne.*

☞ *La médiane:*

La médiane est la valeur de la variable statistique qui partage en deux effectifs égaux les individus de la population supposés rangés par valeur croissante du caractère.

La médiane Me est la valeur de la variable statistique telle que l'ordonnée de la courbe cumulative soit égale à 0.5: F(Me)= 0.5.

3. Description numérique d'une variable statistique

☞ Variables discrètes:

En général, l'équation F(M)=0.5 n'a pas de solution puisque la fonction F(x) varie par sauts.

Soit n la taille de l'échantillon et x la variable observée triée par ordre croissant:

si n est impair: $Me = x_{(n+1)/2}$;

si n est pair: $Me = (x_{n/2}+x_{n/2+1})/2$

☼ Exemple 1 (nombre de titres détenus pour un échantillon investisseurs):

Valeurs: 2, 15, 6, 18, 20, 1, 13,2,18.

Valeurs triées: 1, 2,2, 6, 13, 15, 18,18, 20.

n = 9; $Me= x_{(5)}= 13$

3. Description numérique d'une variable statistique

☼ Exemple 2:

X= nbre de titres	ni	fi	Fi	
0	15	0.12	0.12	
1	25	0.19	0.31	
2	56	0.43	0.74	
3	12	0.09	0.83	
4	12	0.09	0.93	
5	10	0.07	1	
Total	***130***			

3. Description numérique d'une variable statistique

☞Variables statistiques continues:

L'équation F(M)=1/2 a une racine unique qu'en général on ne peut situer qu'entre deux extrémités de classes. La classe i est la classe médiane si:

$n_1 + \ldots n_{i-1} < n/2 < n_1 + \ldots + n_i$

On détermine la classe médiane à partir des effectifs cumulés et par interpolation linéaire on déduit une évaluation de la médiane.:

$$Me = e(i-1) + ai * (0.5 - F(i-1))/fi$$

3. Description numérique d'une variable statistique

☼ Exemple:

Répartition d'entreprises selon le nombre de leurs salariés:

Classes Ci	ai	ni		fi	Fi
0 à 10	10	350		0.35	0.35
10 à 20	10	240		0.24	0.59
20 à 50	30	150		0.15	0.74
50 à 100	50	90		0.09	0.83
100 à 200	100	70		0.07	0.90
200 à 500	300	60		0.06	0.96
500 à 1000	500	40		0.04	1.00
		1000			

Médiane = 10+ 10*(**0,5**-0,35)/0,24 = **16,25**

Propriétés:

- Insensible aux variations des valeurs extrêmes,
- Calcul rapide,
- Interprétation aisée,
- La médiane minimise la quantité: $\sum_{i=1}^{n} |x_i - Me|$

3. Description numérique d'une variable statistique

☞ La moyenne:

La moyenne arithmétique d'une variable statistique est la somme pondérée des valeurs par les fréquences, c'est-à-dire encore la somme des observations divisée par l'effectif n de la population:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{p} n_j x_j = \sum_{j=1}^{p} f_j x_j = \frac{\sum_{j=1}^{p} n_j x_j}{\sum_{j=1}^{p} n_j}$$

Dans le cas de variables continues, si on ne connaît pas les valeurs x_i mais seulement leur nombre dans chaque intervalle de classe (e_{i-1},e_i), on prend dans ce cas le centre de classe comme valeur de x_i.

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{p} n_j c_j$$

3. Description numérique d'une variable statistique

☞ Propriétés de la moyenne:

- la moyenne des différences à la moyenne est nulle: $\sum_{j=1}^{p} f_j \left(x_j - \bar{x}\right) = 0$
- la moyenne minimise la quantité: $\sum_{j=1}^{p} f_j \left(x_j - a\right)^2$
- Le minimum s'appelle la variance: $Var = \sum_{j=1}^{p} f_j \left(x_j - \bar{x}\right)^2$
- On a également: $\sum_{j=1}^{p} f_j \left(x_j - a\right)^2 = \sum_{j=1}^{p} f_j \left(x_j - \bar{x}\right)^2 + \left(\bar{x} - a\right)^2$
- La moyenne est un opérateur linéaire: $\overline{\alpha X} = \alpha . \overline{X}$ et $\overline{X + Y} = \overline{X} + \overline{Y}$

3. Description numérique d'une variable statistique

☼Exemple:

Répartition d'entreprises selon le nombre de leurs salariés:

Classes Ci	ci	ni	ci*ni	fi	ci*fi
0 à 10	5	350	1750	0,35	1,75
10 à 20	15	240	3600	0,24	3,6
20 à 50	35	150	5250	0,15	5,25
50 à 100	75	90	6750	0,09	6,75
100 à 200	150	70	10500	0,07	10,5
200 à 500	350	60	21000	0,06	21
500 à 1000	750	40	30000	0,04	30
Total		**1000**	**78850**		**78,85**

Moyenne = 78850/1000= 78,85

3. Description numérique d'une variable statistique

☞ Moyenne géométrique:

On appelle moyenne géométrique G d'une variable statistique discrète X a valeurs xi et a effectif ni la quantité:

$$G = \sqrt[n]{x_1^{n_1} . x_2^{n_2} \ldots x_p^{n_p}}$$

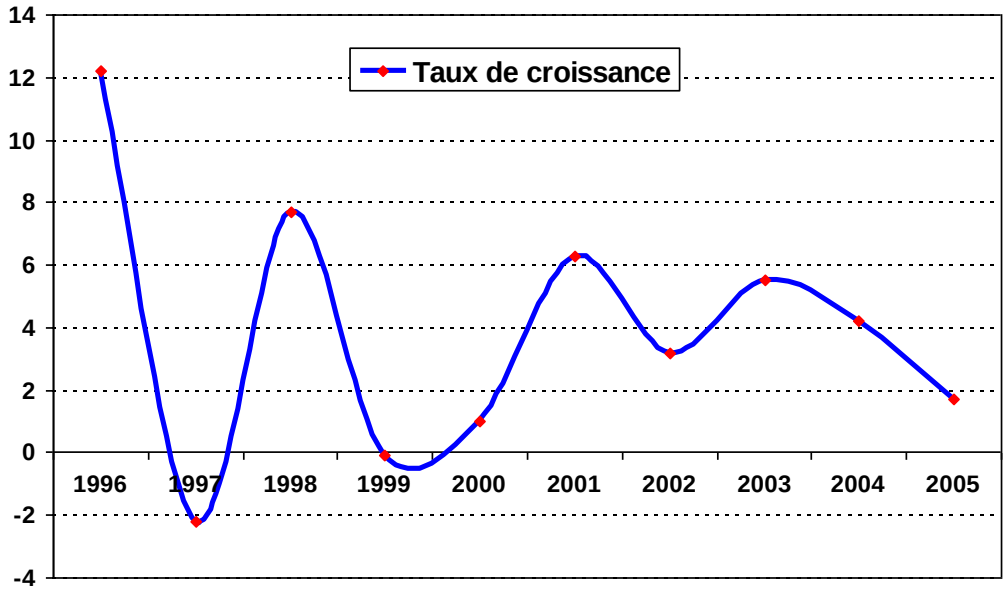
Où n est la taille de l'échantillon. $n = \sum_{i=1}^{p} n_i$

Ou encore: $\log(G) = \frac{1}{n} \sum_{i=1}^{p} n_i \log(x_i)$

3. Description numérique d'une variable statistique

☼ Exemple: calcul du taux de croissance annuel moyen

	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
PIB réel (1980= 100)	12.2	-2.2	7.7	-0.1	1.0	6.3	3.2	5.5	4.2	1.7



3. Description numérique d'une variable statistique

☼ Exemple: calcul du taux de croissance annuel moyen

Par définition le taux de croissance moyen entre 1996 et 2005 est donné par:

$$(1+ T_{05/96})^{9}=(1+T_{97/96})*(1+T_{98/97})*(1+T_{99/98})*(1+T_{00/99})*(1+T_{01/00})*$$
$$(1+T_{02/01})*(1+T_{03/02})*(1+T_{04/03})*(1+T_{05/04})$$

$T_{05/96} = 4.314\%$

☹ Erreur: ne calculez pas la moyenne arithmétique des taux de croissance d'une année à l'autre.

Moyenne arithmétique = 3.950 %

3.Description numérique d'une variable statistique

B. Caractéristiques de dispersion

☞ Différences et écarts:

- Considérons une caractéristique de tendance centrale C et une valeur possible x_i. Les quantités: x_i-C et $|x_i$-C| sont respectivement la différence à la tendance et l'écart à la tendance centrale.

- Par construction, la tendance centrale de la série des différences x_i-C est nulle.

- La série des écarts $|x_i$-C| définit en revanche une variable statistique positive.

- Suivant que les écarts sont pris par rapport à la médiane ou par rapport à la moyenne, on aboutit à plusieurs indices de dispersion.

3.Description numérique d'une variable statistique

☞ l'écart quadratique ou écart type:

- L'écart type est définie par la formule: $$\sigma = \sqrt{\sum_{i=1}^{p} f_i \left(x_i - \bar{x}\right)^2}$$

- L'écart type d'une distribution est la racine carrée de la variance; à la différence de la variance, il s'exprime dans la même unité que la variable sur laquelle elle est calculée.

$$v(X) = \sum_{i=1}^{p} f_i \left(x_i - \bar{x}\right)^2$$

- La variance est donc égale à la moyenne arithmétique des carrés des écarts à la moyenne.

On montre la formule suivante: $$v(X) = \frac{1}{n}\left(\sum_{j=1}^{p} x_j^2\right) - \bar{x}^2$$

- Dans le cas de variables continues, il convient ici encore d'affecter les individus à son centre de classes, d'où la formule:

$$v(X) = \sum_{i=1}^{p} f_i \left(c_i - \bar{c}\right)^2$$

3.Description numérique d'une variable statistique

☞ Le coefficient de variation:

- La moyenne comme l'écart type, s'expriment dans la même unité que X. On définit le coefficient de variation, en général pour des variables positives seulement, comme le rapport de l'écart type à la moyenne:

$$CV = \frac{\sigma_x}{\bar{x}}$$

- Cette quantité est sans dimension, indépendante des unités choisis. Elle permet de comparer la variabilité relative de distributions statistiques dont les ordres de grandeur sont différents (ex: comparaison par pays ou par secteur d'activité,..).

3.Description numérique d'une variable statistique

☞ Autres mesures de dispersion

♦ les quantiles (ou fractiles):

- la quantile d'ordre a noté x_a est la racine de l'équation : $F(x_a)=a$. c'est-à-dire $x_a=F^{-1}(a)$, avec F^{-1} est la fonction inverse de F.

- C'est un nombre qui partage l'échantillon des valeurs de la variable X, rangées dans l'ordre croissant, en deux sous échantillon, l'un comprenant une fraction *(a)* de l'échantillon et l'autres une fraction *(1-a).*

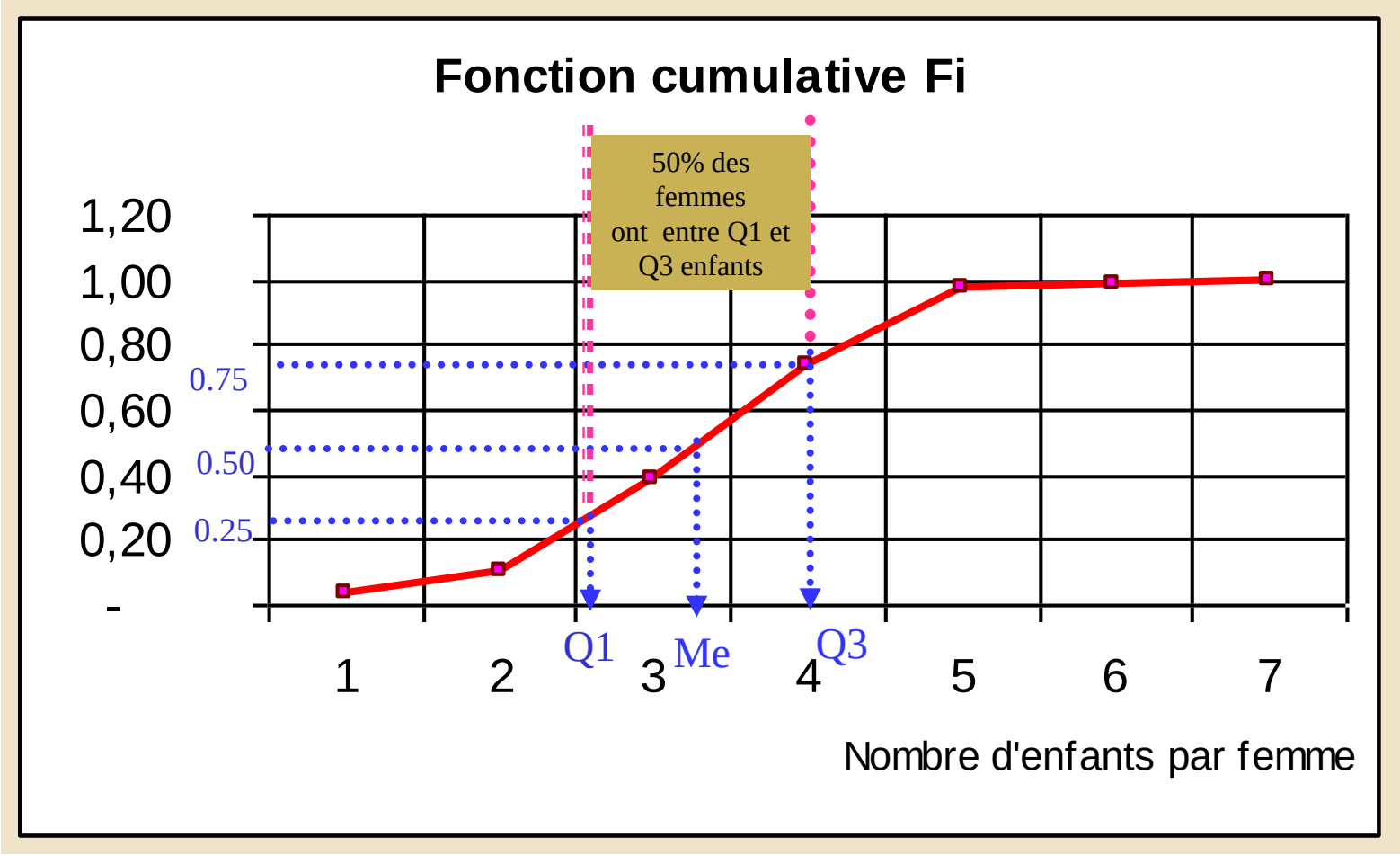
- Les quantiles utilisés le plus souvent sont les suivants:

Médiane: a=1/2, Quartiles: a=1/4, ¾.

- La détermination des quantiles se fait de manière graphique ou calculatoire. On utilise une interpolation linéaire pour les variables continues regroupées en classes.
- L'intervalle interquartile Q3-Q1 contient 50% de la population; laissant à gauche 25% et à droite 25%.

3.Description numérique d'une variable statistique

☼ Exemple:



3.Description numérique d'une variable statistique

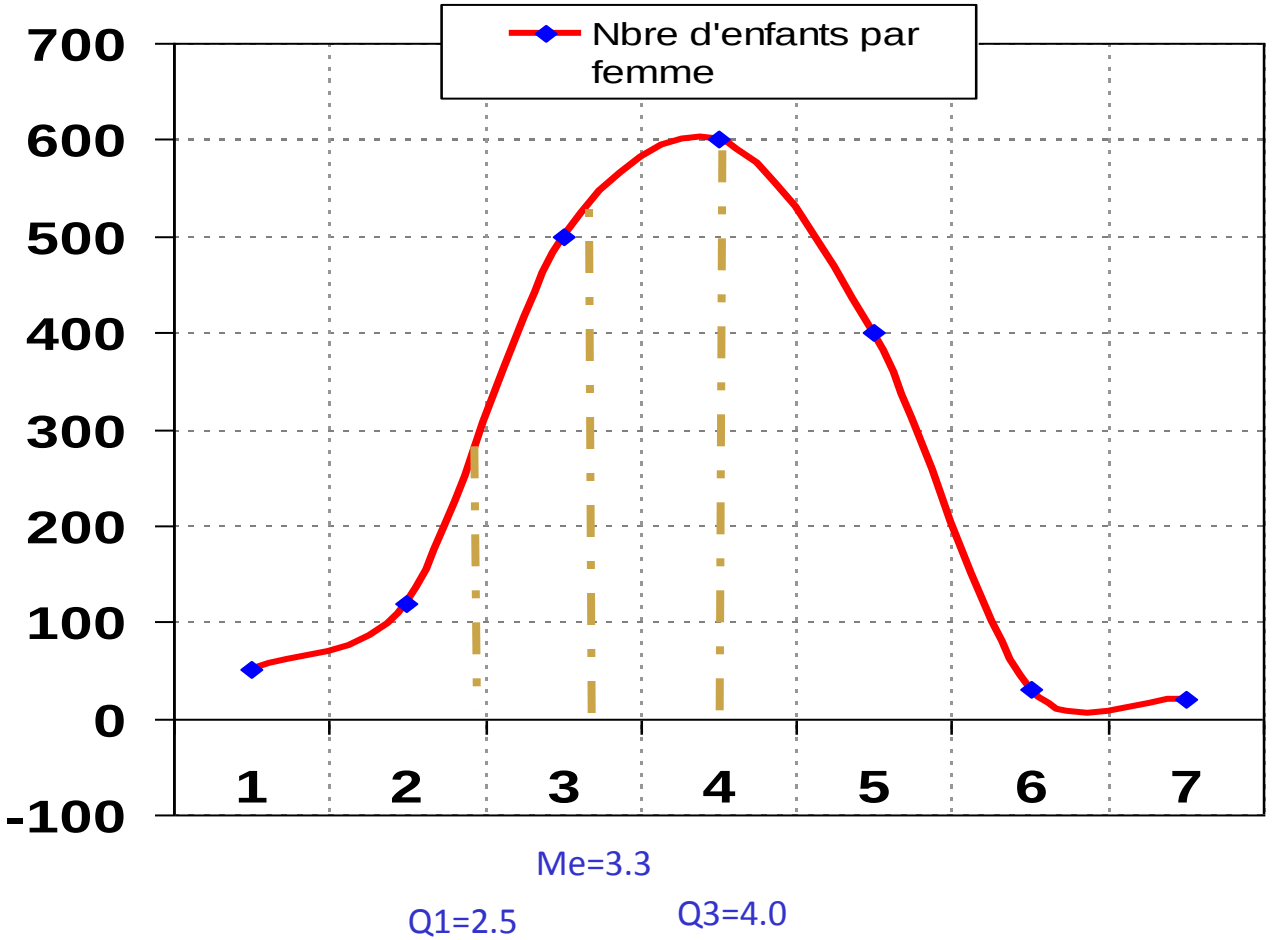
☞ Exemple:

Calcul par interpolation linéaire des quantiles d'ordre 0.25 (Q3), la médiane (quantile d'ordre 0.50) et d'ordre 0.75.

Nombre d'enfants par femme			
Xi	**ni**	**fi**	**Fi**
1	**50**	**0,03**	**0,03**
2	**120**	**0,07**	**0,10**
3	**500**	**0,29**	**0,39**
4	**600**	**0,35**	**0,74**
5	**400**	**0,23**	**0,97**
6	**30**	**0,02**	**0,99**
7	**20**	**0,01**	**1,00**
Total	**1720**	**1,00**	

Q1 = 2 + (3-2)*(0.25-0.10) / (0.39-0.10) = 2.52
Q3 = 3 + (4-3)*(0.5-0.39) / (0.74-0.39) = 3.31
Q3 = 4 + (5-4)*(0.75-0.74) / (0.97-0.74) = 4.04.

3.Description numérique d'une variable statistique



3.Description numérique d'une variable statistique

☞ Les moments

- On appelle moment d'ordre r par rapport à la valeur a la quantité:

$$m_r = \sum_{i=1}^{p} f_i (x_i - a)^r$$

- Suivant les valeurs de a, on définit plusieurs sortes de moments:
  - Moments non centrés: $m_r = \sum_{i=1}^{p} f_i x_i^{\,r}$
  - Moments centrés: $\mu_r = \sum_{i=1}^{p} f_i (x_i - \bar{x})^r$
  - Exemples:

$$m_o = 1 \qquad \mu_0 = 1$$
$$m_1 = \bar{x} \qquad \mu_1 = 0$$
$$m_2 = \sigma^2 + \bar{x}^2 \qquad \mu_2 = \sigma^2$$

3.Description numérique d'une variable statistique

☞ Caractéristiques de forme

Outre les mesures de tendance centrale et de dispersion, on peut chercher à caractériser la forme d'une distribution au moyen d'un indice résumé.

1. Coefficient d'asymétrie:

- Si une distribution est symétrique , ses divers moments centrés d'ordre impair sont nuls. Fisher a proposé le coefficient d'asymétrie:

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

- Ce coefficient est sans dimension, invariant par changement d'échelle et d'origine et nul pour les distributions symétriques.

3.Description numérique d'une variable statistique

☞ Coefficient d'aplatissement:

- Le coefficient d'aplatissement de Fisher est: $$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

- Ce coefficient est sans dimension, invariant par changement d'échelle et d'origine. La constante 3 est choisie de telle façon que le coefficient soit nul lorsque la distribution est normale.

- Le coefficient est positif si la distribution est moins aplatie que la distribution normale et négatif dans le cas contraire.

3.Description numérique d'une variable statistique

☞ Exemple:

Xi	ni	fi	fi*xi	xi-moy	fi*(xi-moy)^2	fi*(xi-moy)^3	fi*(xi-oy)^4
1	50	0,03	0,03	- 2,78	0,23	- 0,63	1,75
2	120	0,07	0,14	- 1,78	0,22	- 0,40	0,71
3	500	0,29	0,87	- 0,78	0,18	- 0,14	0,11
4	600	0,35	1,40	0,22	0,02	0,00	0,00
5	400	0,23	1,16	1,22	0,34	0,42	0,51
6	30	0,02	0,10	2,22	0,09	0,19	0,42
7	20	0,01	0,08	3,22	0,12	0,39	1,24
Total	1720	1,00	3,78		1,19	- 0,17	4,74

écart type = 1.09
Coefficient d'asymétrie: -0.13
Coefficient d'aplatissement : 0.33

3.Description numérique d'une variable statistique

☞ Caractéristiques de concentration

- On s'intéresse, par exemple, à la masse des salaires correspondant à chaque classe (ne se rapporte plus à n_i mais à $n_i x_i$):

- S_i est le total des salaires gagnés par les ni ouvriers dont le salaire est compris entre e_{i-1} et e_i.

On détermine:

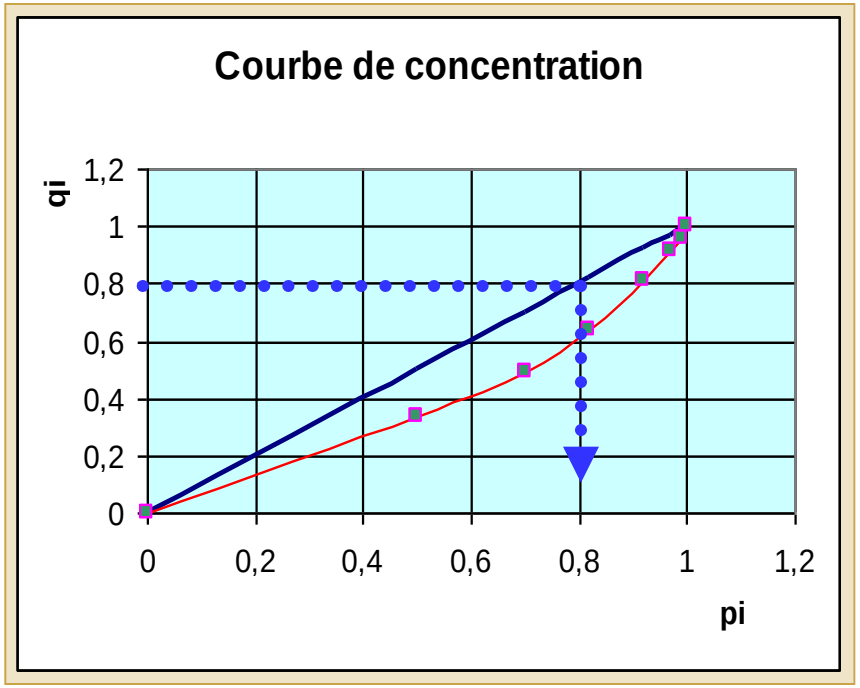
- p_i: proportion de personnes dont le salaire est inférieure à e_i,

- q_i: proportion de la masse totale des salaires gagnée par les personnes dont le salaire est inférieure à e_i .

3.Description numérique d'une variable statistique

☞ Courbe de concentration:

♦ La courbe de concentration de la distribution des salaires est la courbe représentative de q_i en fonction de p_i. Comme p_i et q_i ne sont connus que pour les extrémités de classes e_i, on ne dispose que des points correspondants pour tracer la courbe de concentration.



60% de la masse salariale est partagée entre 80% des salariés.

3.Description numérique d'une variable statistique

☞ L'indice de concentration

- L'indice de concentration est le double de l'aire comprise entre la courbe de concentration et la première bissectrice. C'est un nombre sans dimension compris entre 0 et 1.
- Il vaut 0 dans le cas d'une répartition uniforme: la concentration nulle correspond à la distribution égalitaire.
- Il tend vers 1 dans le cas d'une distribution très concentrée (peu d'individus accaparent une grande proportion de la masse totale).

3.Description numérique d'une variable statistique

☞ La médiale

♦ La médiale de la distribution des salaires est le salaire qui fractionne la masse salariale en deux portions égales.

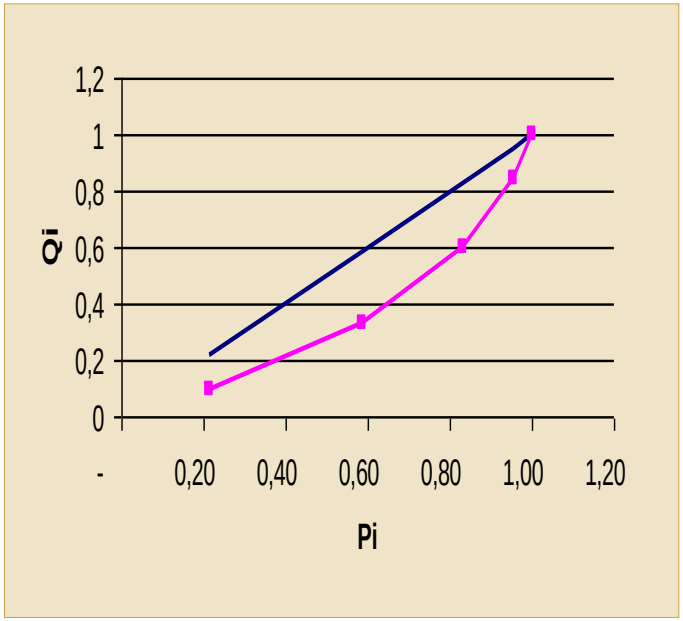
Q(x)=1/2

Comme la médiane, la médiale est évalué par interpolation linéaire à partir des extrémités de la classe médiale:

$$Ml = e_{i-1} + a_i \frac{0.5 - Q_{i-1}}{q_i}$$

☼ Exemple: Concentration des salaires dans une entreprise.

3.Description numérique d'une variable statistique



ei		ni	pi	ci	ci*ni	qi	Qi
1 700	2 500	70	0,22	2 100	147 000	0,10	0,10
2 500	3 500	120	0,37	3 000	360 000	0,23	0,33
3 500	7 000	80	0,25	5 250	420 000	0,27	0,60
7 000	12 000	40	0,12	9 500	380 000	0,25	0,84
12 000	20 000	15	0,05	16 000	240 000	0,16	1,00
Total		325	1,00		1 547 000	1,00	

Classe médiale: [3500 – 7000[.

Médiale: Ml= 3500+3500* (0.5- 0.33)/0.27 = 5703.7

4. Les indices statistiques

☞ Définition

- Considérons l'évolution temporelle d'une grandeur G. Soit: G0,…Gi,…

  On appelle indice élémentaire de la grandeur G à la date t par rapport à la date 0, le rapport:

$$I_{t/0}(G) = \frac{G_t}{G_0}$$

- La date 0, utilisée comme date de comparaison, s'appelle date de référence. La date t qui lui est comparée est la date courante.

- On exprime habituellement un indice élémentaire en pourcentage, la valeur 100 correspondant à la date de référence.

☞ Propriétés d'un indice élémentaire

A. Circularité:

- La circularité est une propriété fondamentale qui permet de comparer non seulement les dates 0 et t, 0 et j d'autre part, mais aussi t et j:

$$I_{t/j}(G) = \frac{I_{t/0}(G)}{I_{j/0}(G)} \qquad I_{t/0}(G) = I_{t/j}(G) * I_{j/0}(G)$$

- La comparaison s'effectue ainsi, indépendamment du choix de la date de référence 0, entre les indices comme entre les valeurs:

- La circularité entraîne deux autres propriétés:

- *La réversibilité:*

$$I_{0/t}(G) = \frac{1}{I_{t/0}(G)}$$

-*L'enchaînement:*

$$I_{t/0}(G) = I_{t/t-1}(G) * I_{t-1/t-2}(G) \ldots I_{1/0}(G)$$

On obtient ainsi l'indice à la date t par rapport à la date 0 en faisant le produit des indices intermédiaires d'une date par rapport à la date précédente.

B. Addition:

- G_t et G_0 sont des sommes pondérées à coefficients constants ou variables:

$$G_t = \sum_i a^i G_t^i$$

$$G_0 = \sum_i a^i G_0^i$$

$$I_{t/0}(G) = \frac{\sum_i a^i G_t^i}{\sum_i a^i G_0^i} = \frac{\sum_i a^i G_o^i \frac{G_t^i}{G_o^i}}{\sum_i a^i G_0^i} = \frac{\sum_i a^i G_o^i I_{t/0}(G^i)}{\sum_i a^i G_0^i}$$

- Alors l'indice élémentaire d'une somme pondérée est la moyenne arithmétique pondérée des indices élémentaires. En effet:

  Le coefficient de pondération de l'indice partiel $I_{t/0}(G^i)$ est égal à : $\omega^i = \frac{a^i G_0^i}{\sum_i a^i G_0^i}$

- L'indice de la moyenne G est la moyenne arithmétique pondérée des indices partiels mais avec des coefficients de pondération différents:

$$\frac{a^i G_0^i}{\sum_i a^i G_0^i} \neq a^i$$

4. Les indices statistiques

c. Multiplication:

- L'indice élémentaire d'un produit est égal au produit des indices élémentaires:

$$I_{t/0}(AB) = I_{t/0}(A)I_{t/0}(B)$$

D. Division:

- L'indice élémentaire d'un quotient est égal au quotient des indices élémentaires:

$$I_{t/0}(A/B) = \frac{I_{t/0}(A)}{I_{t/0}(B)}$$

4. Les indices statistiques

☞ Indice synthétique

1. Définition: considérons une grandeur G complexe, c'est-à-dire constitué d'éléments G^1,…G^i,..Par exemple, G est le niveau général des prix : les constituants G^i sont les prix des différents articles au stade final de leur commercialisation.

- Les indices élémentaires des constituants G^i sont définis par:

$$I_{t/0}(G^i) = \frac{G_t^i}{G_0^i}$$

- Le problème est de synthétiser en un unique indice les indices élémentaires des constituants de G. L'indice synthétique I(G) devra si possible posséder des propriétés analogues à celles des indices élémentaires.

4. Les indices statistiques

☞ Les indices synthétiques utilisés en pratique:

- Soit, à la date 0, ω_0^i l'importance relative du constituant i dans la grandeur complexe G et ω_1^i la quantité analogue à la date 1:

- L'indice de Laspeyres est la moyenne arithmétique pondérée des indices élémentaires par les coefficients ω_0^i de la date de référence:

$$L_{1/0}(G) = \sum_i \omega_0^i I_{1/0}(G^i) = \sum_i \omega_0^i \frac{G_1^i}{G_0^i} \qquad \sum_i \omega_0^i = \sum_i \omega_1^i = 1$$

- L'indice de Paasche est la moyenne harmonique pondérée des indices élémentaires par les coefficients de la date courante:

$$\frac{1}{P_{1/0}(G)} = \sum_i \frac{\omega_1^i}{I_{1/0}(G^i)} = \sum_i \omega_1^i \frac{G_0^i}{G_1^i}$$

- L'indice de Fisher est la moyenne géométrique simple des indices de Laspeyres et de Paasche:

$$F_{1/0}(G) = \sqrt{L_{1/0}(G) * P_{1/0}(G)}$$

4. Les indices statistiques

☞ Les indices de prix, de quantité et de valeur

- Considérons l'évolution des dépenses d'une famille donnée entre les dates 0 et 1. Admettons pour simplifier que les articles consommés à l'une des dates sont encore sur le marché et sous la même forme à l'autre date.

- Soit p^i le prix de l'article i et q^i la quantité de cet article achetée par la famille:

$$p_0^i, q_0^i \quad \text{à la date } 0$$
$$p_1^i, q_1^i \quad \text{à la date } 1$$

- Les dépenses consacrées à l'article i sont respectivement:

$$D_0^i = p_0^i * q_0^i \quad \text{à la date } 0$$
$$D_1^i = p_1^i * q_1^i \quad \text{à la date } 1$$

- Et les dépenses totales:

$$D_0 = \sum_i p_0^i * q_0^i \quad \text{à la date } 0$$
$$D_1 = \sum_i p_1^i * q_1^i \quad \text{à la date } 1$$

4. Les indices statistiques

- On appelle coefficient budgétaire de l'article i la part de la dépense totale consacrée à cet article:

$$\omega_0^i = \frac{p_0^i q_0^i}{\sum_i p_0^i q_0^i} \text{ à la date } 0$$

$$\omega_1^i = \frac{p_1^i q_1^i}{\sum_i p_1^i q_1^i} \text{ à la date } 1$$

- Les coefficients budgétaires, de somme égale à 1, mesurent l'importance relative des différents articles dans le budget familial.
- Les indices élémentaires des grandeurs considérées sont par définition:

$$I_{1/0}(p^i) = \frac{p_1^i}{p_0^i} : \text{indice de prix de l'article i}$$

$$I_{1/0}(q^i) = \frac{q_1^i}{q_0^i} : \text{indice de quantité de l'article i}$$

$$I_{1/0}(D^i) = \frac{p_1^i * q_1^i}{p_0^i * q_0^i} : \text{indice de dépense de l'article i}$$

- Ces trois indices élémentaires sont liés par la relation: $I_{1/0}(D^i) = I_{1/0}(p^i) I_{1/0}(q^i)$

4. Les indices statistiques

Indice de	Prix	Quantité
Laspeyres	$L_{1/0}(p) = \frac{\sum_i p_1^i * q_0^i}{\sum_i p_0^i * q_0^i}$	$L_{1/0}(q) = \frac{\sum_i q_1^i * p_0^i}{\sum_i q_0^i * p_0^i}$
Paasche	$P_{1/0}(p) = \frac{\sum_i p_1^i * q_1^i}{\sum_i p_0^i * q_1^i}$	$P_{1/0}(q) = \frac{\sum_i q_1^i * p_1^i}{\sum_i q_0^i * p_1^i}$

Les indices de Laspeyres et de Paasche se présentent ainsi comme des rapports de dépenses totales où le facteur (prix ou quantité) autre que celui considéré est constant:

- Pour les indices de prix: dépenses totales à quantités constantes et système de prix variable,
- Pour l'indice de quantité: dépenses totales à prix constants et quantités variables.
- L'indice de Laspeyres utilise les constantes de la date de référence tandis que l'indice de Paasche utilise celles de la date courante.

4. Les indices statistiques

☼ Exemple :

	date 0		**date 1**	
	p0	**q0**	**p1**	**q1**
Produit 1	12	9	13	10
Produit 2	40	10	39	12
Produit 3	2	25	5	24
Produit 4	15	36	20	34
	p1*q0	**p0*q0**	**p1*q1**	**p0*q1**
Produit 1	117	108	130	120
produit 2	390	400	468	480
Produit 3	125	50	120	48
produit 4	720	540	680	510
Total	1352	1098	1398	1158
	Laspeyres	**123,1**	**Paasche**	**120,7**

5. Mesures de liaison entre variables

- On s'intéresse aux liaisons entre variables: c'est ce qu'on appelle l'étude des corrélations. Les méthodes et les indices de dépendance varient selon la nature des variables étudiées. Elles peuvent être classées en 4 catégories:
- La statistique du chi-deux et ses dérivées, utilisable quel que soit le type des variables, à condition qu'elles aient chacune un nombre de modalités distinctes pas trop élevée, le plus souvent utilisée pour mesurer la liaison entre deux variables qualitatives,

- Le coefficient de corrélation (ou de Pearson) calculable pour des variables numériques,

- Les coefficients de corrélation des rangs de Spearman et de Kendall et des statistiques de même type adaptées aux variables ordinales.

- Divers indicateurs utilisables pour toutes variables.

5. Mesures de liaison entre variables

A. Liaison entre deux variables numériques

- Supposons que l'on observe pour n individus deux variables X et Y (ex. la consommation et le revenu). On a donc n couples (x_i, y_i).

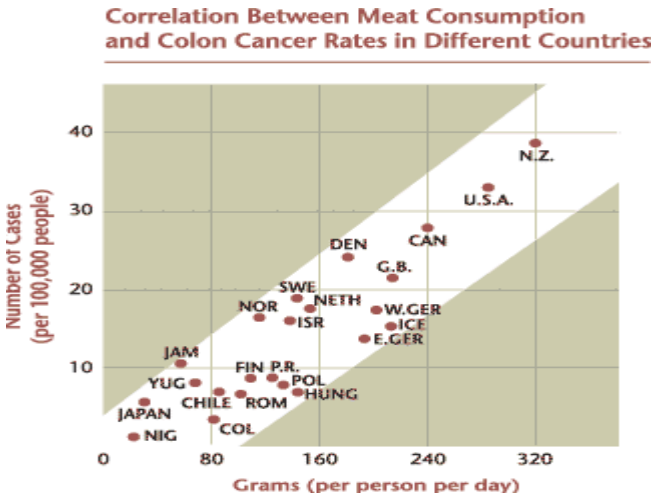
1. Etude graphique de la corrélation:

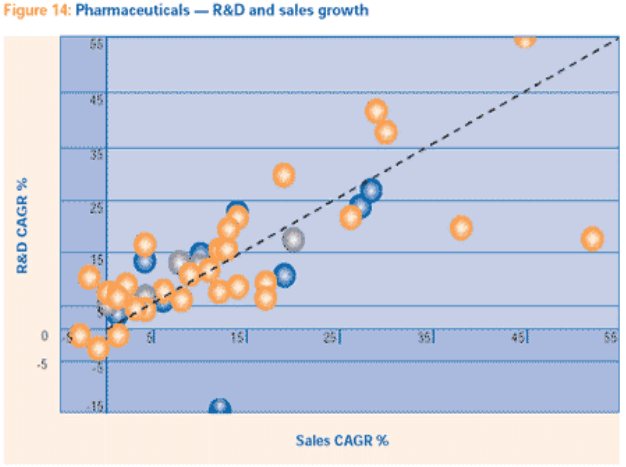
- Afin d'examiner s'il existe une liaison entre X et Y on représente chaque observation i comme un point de coordonnées (x_i, y_i).la forme du nuage de points ainsi tracé est fondamentale pour la suite.

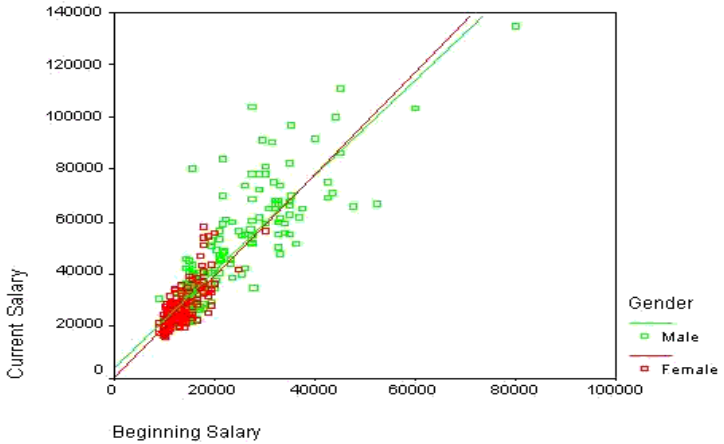
- Différents cas peuvent se présenter. Ainsi, on peut, entre autres, observer une absence de liaison, une absence de liaison en moyenne mais pas en dispersion, une corrélation linéaire positive, une corrélation non linéaire,..

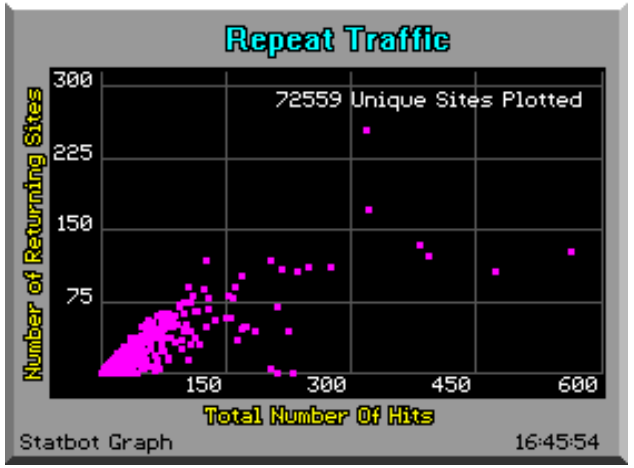
- Rappelons que la non corrélation n'implique pas nécessairement l'indépendance.

5. Mesures de liaison entre variables









5. Mesures de liaison entre variables

2. Le coefficient de corrélation linéaire

- Lorsque les variables X et Y sont numériques, on peut calculer le coefficient de corrélation linéaire qui mesure le caractère plus au moins linéaire de la dépendance entre les deux variables.
- Ce coefficient est défini par:

$$r_{XY} = \frac{s_{XY=}}{S_X S_Y} = \frac{\sum_{k=1}^{n}\left(x_k - \bar{x}\right)\left(y_k - \bar{y}\right)}{\sqrt{\sum_{k=1}^{n}\left(x_k - \bar{x}\right)^2} * \sqrt{\sum_{k=1}^{n}\left(y_k - \bar{y}\right)^2}}$$

Où s_X et s_Y sont les écarts-type de X et de Y.

s_{XY} est la covariance empirique entre X et Y.

- Le coefficient r est compris entre -1 et 1. Il est égal à 1 en valeur absolue si et seulement si il existe une relation linéaire exacte entre les variables X et Y (ex. $ax_i+by_i+c=0$ pour tout i).

- Le CCL mesure la forme plus au mojns linéaire du nuage de points que l'on obtient en représentant sur un plan les individus par des coordonnées (xi,yi).
- En particulier, il peut être nul alors qu'il existe une relation non linéaire entre les variables.

5. Mesures de liaison entre variables

3. Liaison entre une variable numérique et une variable qualitative

- Si X a k catégories on notera n_1, …n_k les effectifs observés et $\overline{y_1}$,…, $\overline{y_k}$ les moyennes de Y pour chaque catégorie et $\overline{y}$ la moyenne totale.
- On veut étudier la liaison entre une variable Y, numérique, et une variable X qualitative . La mesure de la liaison est le rapport de corrélation : $\eta^2_{Y/X}$

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{k} ni*\left(\overline{y_i} - \overline{y}\right)^2 + \frac{1}{n}\sum_{i=1}^{k} ni*s_i^2$$

$$\frac{1}{n}\sum_{i=1}^{k} ni*\left(\overline{y_i} - \overline{y}\right)^2 \text{ est appelée variance inter-catégories}$$

$$\frac{1}{n}\sum_{i=1}^{k} ni*s_i^2 \text{ est appelée variance intra-catégories}$$

s_i^2 sont les variances de Y à l'intérieur de chaque catégorie.

$\eta^2 = 0$ si $\overline{y_1} = .. = \overline{y_k}$ d'où absence de dépendance en moyenne.

$\eta^2 = 1$ si tous les individus d'une catégorie de X ont même valeur de Y.

$$\eta^2_{Y/X} = \frac{\sum_{i=1}^{k} f_i * \left(\overline{y_i} - \overline{y}\right)^2}{s_y^2}$$

5. Mesures de liaison entre variables

4. Liaison entre deux variables qualitatives

- La statistique du Chi-deux D^2 est une mesure des écart entre les effectifs observés n_{ij} et les effectifs théoriques: elle caractérise donc, d'une certaine façon, l'écart entre la situation observée et la situation d'indépendance. L'expression de cette statistique, est donnée par:

$$D^2 = n * \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - f_{i.} * f_{.j}\right)^2}{f_{i.} * f_{.j}}$$

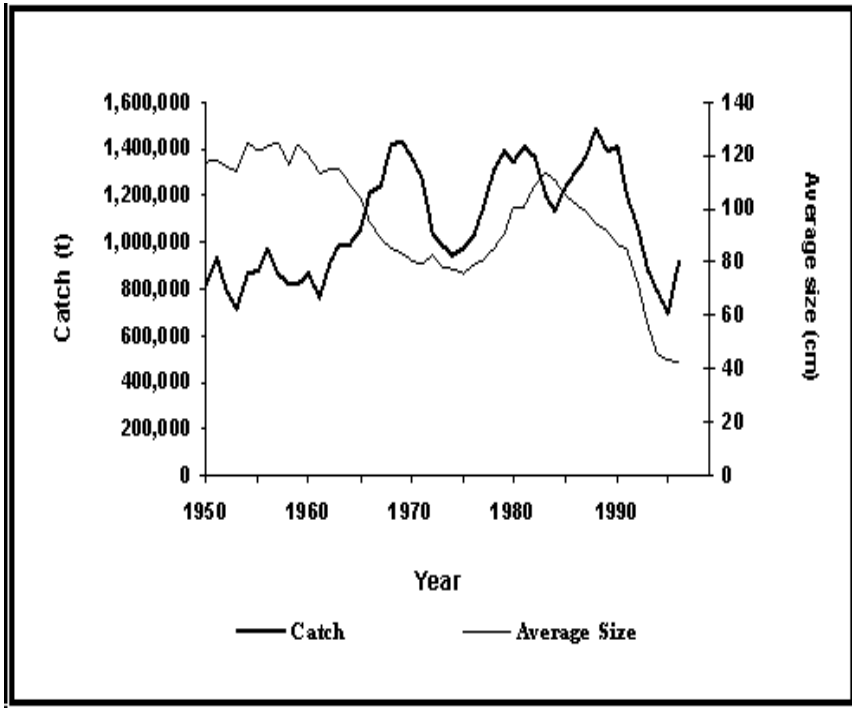
- Chaque élément de la somme dou ion au chi-deux de la case (Xi,Yj).

- On peut montrer que la valeur maximale de D^2 est n*inf(p-1, q-1), qui dépend à la fois du nombre d'individus n et des dimensions p et q du tableau. Diverses statistiques, dérivées du D^2, ont été proposées pour comparer des tableaux de taille différents ou portant sur des populations d'effectifs différents pour obtenir une mesure comprise entre 0 (indépendance) et 1 (liaison fonctionnelle).

$$\Phi = \sqrt{\frac{D^2}{n}}$$

6. Les séries chronologiques (Introduction)

1. Définition*:* on appelle série chronologique ou série temporelle une série d'observations chiffrées ordonnées dans le temps.



♦ Exemple:

- Série journalière des températures relevées à zéro heure en un point donné,
- Série mensuelle des consommations nationales d'électricité,
- Série trimestrielle des livraisons d'essence automobile.

6. Les séries chronologiques

2. But de l'étude des séries chronologiques

- Prévoir:

  - La fonction première pour laquelle il est intéressant d'observer l'historique d'une variable vise à en découvrir certaines régularités afin de pouvoir extrapoler et d'établir une prévision.

  - Comprendre la dynamique qui relie une observation à celles qui l'ont précédée et de supposer que les mêmes causes produisent les mêmes effets.

- Relier les variables:

  - Il est important de savoir a priori si certaines relations sont économétriquement possibles et d'éviter les équations qui ne présentent aucun sens.

  - Exemple de la relation entre la demande et l'inflation. Peut-on faire l'hypothèse que l'inflation influence positivement la demande? Ce qui reviendrait à dire qu'en période de forte inflation, les citoyens souhaitent consommer davantage qu'en période où elle est faible.

6. Les séries chronologiques

2. But de l'étude des séries chronologiques

- Déterminer la causalité:

  - Une approche dynamique permet de s'intéresser aux relations de causalité. Pour qu'un mouvement en provoque un autre, il est nécessaire qu'il le précède.

  - L'utilisation de *retards* d'une variable, de ses valeurs aux périodes précédentes, dans les équations autorise la mesure des effets de causalité et permet également de connaître la *durée de transmission* entre une source et son effet.

- Distinguer entre court terme et long terme:

  - Certaines lois de comportement ne sont jamais vérifiées en pratique car elles ne s'appliquent que sur les équilibre de long terme. A plus courte échéance, des variations contrarient perpétuellement leur mise en œuvre.

  - Les ajustements transitoires s'opèrent continuellement afin de s'approcher de ces équilibres.

6. Les séries chronologiques

2. But de l'étude des séries chronologiques

- Etudier des anticipations des agents:

  - Comment prendre en compte les anticipations des agents? Dans une décision entre épargne et consommation (par exemple), ce ne sont pas seulement les revenus actuel et passé qui comptent, mais aussi l'idée qu'on se fait de l'avenir.

  - Il faut donc faire intervenir des valeurs avancées des variables, via leur anticipation en utilisant la manière dont celles-ci été formées dans le passé.

- Repérer les tendances et cycles
- Corriger des variations saisonnières (CVS)

- Détecter les chocs structurels:

  Un choc structurel est défini comme une modification permanente ou temporaire de la façon dont est générée une variable. Ils sont souvent non anticipables et difficiles à mesurer.

6. Les séries chronologiques

2. But de l'étude des séries chronologiques

- Contrôler les processus:

  Lorsq'une autorité fixe librement le niveau d'une variable ayant une forte influence sur le reste de l'économie, comme par exemple le taux d'intérêts directeur sur lequel la banque centrale a autorité, il lui faut à la fois:

  - ☞ quantifier l'ampleur de son impact,
  - ☞ et mesurer la durée de transmission de son effet dans l'économie.

En retour, cette autorité peut prendre en compte son propre comportement afin d'anticiper les évolutions d'une variable cible, comme *l'inflation*.

6. Les séries chronologiques

3. Les éléments constitutifs d'une série chronologique

- On distingue trois composantes principales:

- Le mouvement extra-saisonnier correspond à l'évolution fondamentale de la série. On décompose parfois ce mouvement en deux éléments: le trend ou la tendance à long terme, désigné par F_t, le cycle, mouvement oscillatoire d'amplitude et de périodicité variables, la périodicité étant supérieure à l'année.

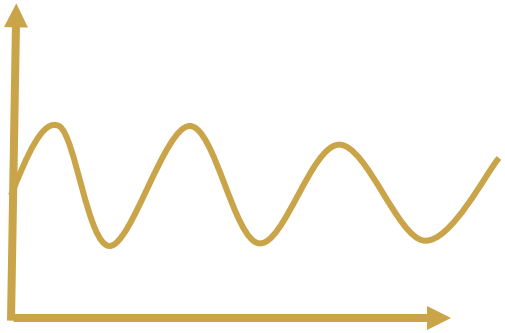
- Les variations saisonnières, désignées par S_t, sont des fluctuations périodiques qui se reproduisent de façon plus au moins permanente d'une année à l'autre.

- Les variations résiduelles ou accidentelles, notez Z_t, sont des fluctuations irrégulières et imprévisibles, supposées en général de faible amplitude, qui traduisent l'effet des facteurs perturbateurs non permanents.

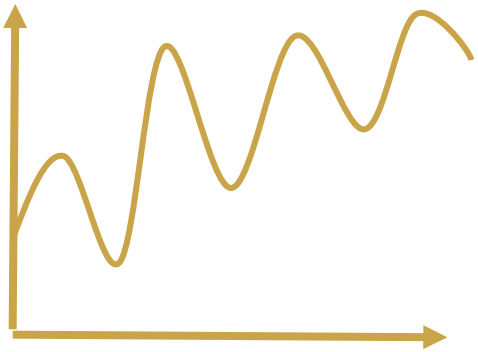
6. Les séries chronologiques

4. Concepts des séries temporelles:

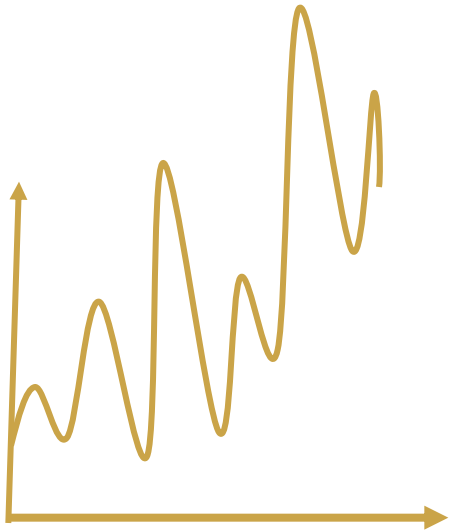
- Stationnarité : Une série est stationnaire si sa moyenne et sa variance existent et ne dépendent pas du temps. La stationnarité est une propriété de stabilité. La série oscille autour de sa moyenne avec une variance constante.
- Stationnarité en moyenne et en variance.



Stationnaire en moyenne



Stationnaire en variance
Non stationnaire en moyenne



Non stationnaire en moyenne
Non stationnaire en variance

6. Les séries chronologiques

☞ Exercice:

On se donne deux séries chronologique: la masse monétaire M3 et l'indice du coût de la vie à fréquence mensuelle:

- Représenter et étudier séparément l'évolution mensuelle des séries;
- Calculer les mesures de tendance centrale: moyenne, écart-type, coefficients de variation;
- Calculer les taux d'accroissement mensuels en glissement annuel;
- Représenter et étudier dans le même graphique l'évolution des accroissements mensuels en glissements annuels;
- Calculer les coefficients de corrélation temporelle entre l'ICV et la masse monétaire M3;
- Déduire à partir des coefficients de corrélation temporelle retardés le *délai de transmission* entre la masse monétaire et l'inflation.