Distribution statistique à un seul caractère

Terminologie de base

A-1 • Population (ou population statistique): *Ensemble (de même nature) concerné par une étude statistique. On parle aussi de champ de l'étude.*

Si l'on s'intéresse aux notes d'un groupe d'étudiants, ce groupe constitue la population.

A-2 • Individu (ou unité statistique): *On désigne ainsi tout élément de la population considérée. Dans l'exemple indiqué ci-dessus, un individu est tout étudiant du groupe.*

A-3 • Échantillon : *Dans une étude statistique, il est fréquent que l'on n'observe pas la population tout entière, les observations d'un phénomène considéré sont réalisées sur une partie restreinte de la population, appelée échantillon.*

On appelle donc échantillon le sous-ensemble de la population sur lequel sont effectivement réalisées les observations.

A-4 • Taille de l'échantillon : *C'est le cardinal de l'échantillon, autrement dit c'est le nombre d'individus qu'il contient (échantillon de taille 800, de taille 1000...).*

En général, on note n la taille de l'échantillon considéré.

A-5 • Enquête (statistique): *C'est l'opération consistant à observer (ou mesurer, ou questionner...) l'ensemble des individus d'un échantillon (ou, éventuellement, de la population complète.)*

A-6 • Recensement : *Enquête dans laquelle l'échantillon observé est en fait la population tout entière (on parle aussi d'enquête exhaustive).*

A-7 • Sondage : *C'est, au contraire, une enquête dans laquelle l'échantillon observé est un sous-ensemble strict de la population (on parle, dans ce cas, d'enquête non exhaustive).*

$\mathcal{A}-8$ • Variable (statistique) : *C'est une caractéristique (âge, salaire, sexe...), définie sur la population et observée sur l'échantillon.*

D'un point de vue mathématique, une variable est une application définie sur l'échantillon.

Si cette application est à valeurs dans $\mathbb{R}$ (ensemble des nombres réels), ou dans une partie de $\mathbb{R}$, elle est dite quantitative (âge, salaire, taille...); sinon elle est dite qualitative (sexe, catégorie socioprofessionnelle...).

On retiendra que les variables quantitatives sont celles prenant des valeurs numériques et que les variables qualitatives sont celles prenant des valeurs non numériques (en faisant bien attention au fait qu'un codage ne représente pas une valeur : même si on code 1 les hommes et 2 les femmes, la variable « sexe » demeure qualitative).

$\mathcal{A}-9$ • Série statistique : *Ensemble de mesures d'une ou plusieurs variables faites sur une population ou un échantillon d'individus.*

$\mathcal{A}-10$ • Données (statistiques) : *Le terme de données est tries utilisé en statistique.*

Il désigne l'ensemble des individus observés (ceux de l'échantillon), l'ensemble des variables considérées et les observations de ces variables sur ces individus.

Les données sont en général présentées sous forme de tableaux (individus en lignes et variables en colonnes) et stockées dans un fichier informatique.

Étude d'une variable qualitative

$\mathcal{B}-1$ • Variables nominales et variables ordinales :

Par définition, les observations d'une variable qualitative ne sont pas des valeurs numériques, mais des caractéristiques, appelées modalités. Lorsque ces modalités sont naturellement ordonnées (par exemple, la mention au bac dans une population d'étudiants), la variable est dite ordinale, elle est construite de manière analogue à celle d'une variable quantitative discrète.

Dans le cas contraire (par exemple, la profession dans une population de personnes actives) la variable est dite nominale.

B– 2 • Traitements statistiques :

Il est clair qu'on ne peut pas envisager de calculer des caractéristiques numériques avec une variable qualitative (qu'elle soit nominale ou ordinale). Dans l'étude statistique d'une telle variable, on se contentera donc de faire des tableaux statistiques et des représentations graphiques. Encore faut-il noter que les notions d'éjectifs cumulés et de fréquences cumulées n'ont de sens que pour des variables ordinales (elles ne sont pas définies pour les variables nominales).

a • Construction : *Le tableau ci-dessous donne la répartition de la population active*

☞ On recense les k différentes modalités $M_1, M_2, \dots, M_k$ prises par la variable.

☞ Pour chaque modalité, on compte le nombre d'individus pour lesquels la variable prend cette modalité. On appelle ce nombre effectif de la modalité et on note n_i l'effectif de la i-ème modalité M_i.

☞ On regroupe dans un tableau les différentes modalités et leurs effectifs respectifs.

☞ La somme des effectifs des différentes modalités doit être égale à l'effectif total :

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + \cdots + n_k = n$$

☞ On note f_i la proportion (ou fréquence) de la i-ème modalité : $f_i = n_i / n$

On note aussi : $P(X = M_i) = f_i = n_i / n$

☞ La somme des proportions est égale à 1 ou 100% : $\sum_{i=1}^{k} f_i = f_1 + f_2 + \cdots + f_k = 1$

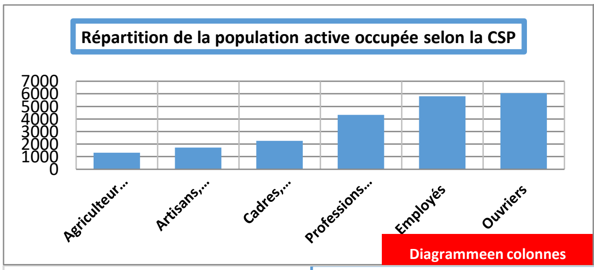
Modalités: M_i	Effectifs: n_i	Fréquences: f_i
M_1	n_1	f_1
M_2	n_2	f_2
$\vdots$	$\vdots$	$\vdots$
M_k	n_k	f_k
$\sum$	n	1

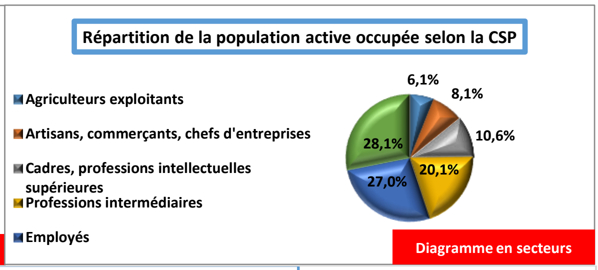
b • Exemple : *Le tableau ci-dessous donne la répartition de la population active occupée (ayant effectivement un emploi) selon la CSP (catégorie socioprofessionnelle)*

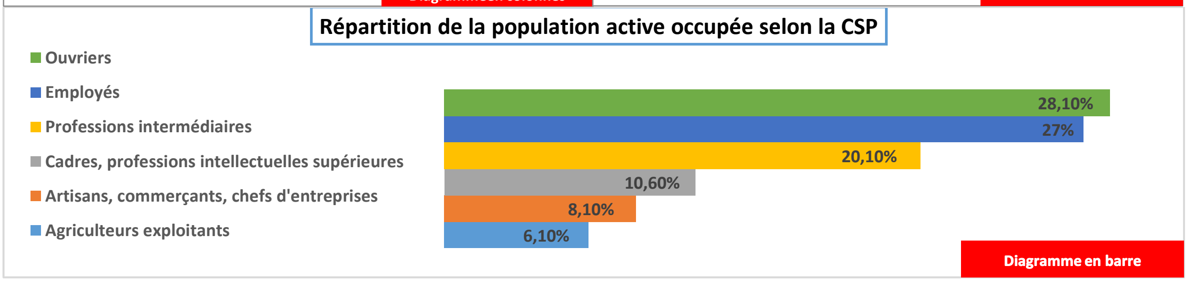
CSP (Modalités): M_i	Effectifs en milliers : n_i	Fréquences (%) : f_i
M_1: Agriculteurs exploitants	1312	6,1%
M_2: Artisans, commerçants, chefs d'entreprises	1739	8,1%
M_3: Cadres, professions intellectuelles supérieures	2267	10,6%
M_4: Professions intermédiaires	4327	20,1%
M_5: Employés	5815	27%
M_6: Ouvriers	6049	28,1%
$\sum$	21509	100%

$\mathcal{B}-3$ • *Représentations graphiques :*

Les représentations graphiques que l'on rencontre avec les variables qualitatives sont assez nombreuses. Les trois plus courantes, qui sont aussi les plus appropriées, sont :
☑ ***Le diagramme en colonnes*** ☑ ***Le diagramme en barre*** ☑ ***Le diagramme en secteurs***







$\mathcal{B}-4$ • *Mode :* ***On appelle mode d'une variable qualitative la ou les modalités ayant le plus grand effectif ou la plus grande proportion.***

Variables quantitatives discrètes

$\mathcal{C}-1$ • *Introduction :* ***On appelle variable quantitative discrète une variable quantitative ne prenant que des valeurs isolées (souvent entières et rarement décimales).***

Le nombre de valeurs distinctes d'une telle variable est habituellement assez faible.

Citons, par exemple, le nombre d'enfants dans une population de familles, le nombre d'années d'études après le bac dans une population d'étudiants...

$\mathcal{C}-2$ • *Organisation des données :*

On étudie une variable discrète X à r modalités dans une population de taille n.

Modalités x_i	Effectifs n_i	Effectifs cumulés croissants: $N_i^{\nearrow} = N_i$	Effectifs cumulés décroissants: $N_i^{\searrow}$	Fréquences f_i	Fréquences cumulées croissantes: $F_i^{\nearrow} = F_i$	Fréquences cumulées décroissants: $F_i^{\searrow}$
x_1	n_1	$N_1 = n_1$	$N_1^{\searrow} = n = \sum_{i=1}^{r} n_i$	$f_1 = \frac{n_1}{n}$	$F_1 = f_1$	$F_1^{\searrow} = 1 = \sum_{i=1}^{r} f_i$
x_2	n_2	$N_2 = n_1 + n_2$	$N_2^{\searrow} = n - N_1 = N_1^{\searrow} - n_1$	$f_2 = \frac{n_2}{n}$	$F_2 = f_1 + f_2$	$F_2^{\searrow} = 1 - F_1 = F_1^{\searrow} - f_1$
x_3	n_3	$N_3 = n_1 + n_2 + n_3$	$N_3^{\searrow} = n - N_2 = N_2^{\searrow} - n_2$	$f_3 = \frac{n_3}{n}$	$F_3 = f_1 + f_2 + f_3$	$N_3^{\searrow} = 1 - F_2 = F_2^{\searrow} - f_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_r	n_r	$N_r = n = \sum_{i=1}^{r} n_i$	$N_r^{\searrow} = n - N_{r-1} = N_{r-1}^{\searrow} - n_{r-1}$	$f_r = \frac{n_r}{n}$	$F_r = 1 = \sum_{i=1}^{r} f_i$	$F_r^{\searrow} = 1 - F_{r-1} = F_{r-1}^{\searrow} - f_{r-1}$
$\sum$	n			1		

$$n_i = Card(X = x_i) \quad N_i^{\nearrow} = N_i = \sum_{k=1}^{i} n_k = Card(X \leq x_i)$$

$$N_i^{\searrow} = Card(X \geq x_i) = n - Card(X \leq x_{i-1}) = n - N_{i-1} = N_{i-1}^{\searrow} - n_{i-1}$$

$$f_i = P(X = x_i) = \frac{n_i}{n} \quad F_i^{\nearrow} = F_i = \sum_{k=1}^{i} f_k = P(X \leq x_i) = \frac{N_i}{n}$$

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

$$F_i^{\searrow} = P(X \geq x_i) = 1 - P(X \leq x_{i-1}) = 1 - F_{i-1} = F_{i-1}^{\searrow} - f_{i-1} = \frac{N_i^{\searrow}}{n}$$

C– 3 • Représentations graphiques usuelles :

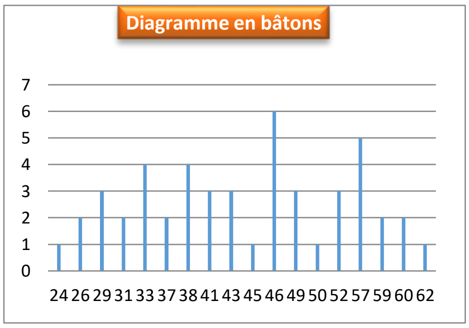
Pour une variable discrète, on rencontre essentiellement deux sortes de représentations graphiques qui sont, en fait, complémentaires : le diagramme en bâtons et le diagramme cumulatif (en escaliers).

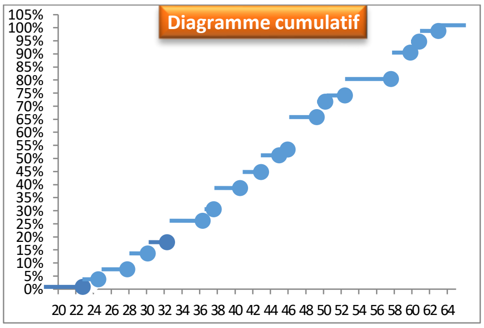
a • Exemple : *On a noté l'âge (arrondi à l'année près) des 48 salariés d'une entreprise.*

Les donnéessont listées : 43 29 57 45 50 29 37 59 46 31 46 24 33 38 49 31 62 60 52 38 38 26 41 52 60 49 52 41 38 26 37 59 57 41 29 33 33 43 46 57 46 33 46 49 57 57 46 43

Soit le tableau statistique correspondant avec des valeurs observées, effectifs, effectifs cumulés, fréquences et fréquences cumulées.

x_i	n_i	N_i	f_i (%)	F_i (%)
24	1	1	2,08	2,08
26	2	3	4,17	6,25
29	3	6	6,25	12,50
31	2	8	4,17	16,67
33	4	12	8,33	25,00
37	2	14	4,17	29,17
38	4	18	8,33	37,50
41	3	21	6,25	43,75
43	3	24	6,25	50,00
45	1	25	2,08	52,08
46	6	31	12,50	64,58
49	3	34	6,25	70,83
50	1	35	2,08	72,91
52	3	38	6,25	79,16
57	5	43	10,42	89,58
59	2	45	4,17	93,75
60	2	47	4,17	97,92
62	1	48	2,08	100
$\sum$	48		100	





C– 4 • Notion de quantile et applications :

a • Définition : *On a vu que la fréquence cumulée F_i ($0 \leq F_i \leq 1$) donne la proportion d'observations inférieures ou égales à x_i.*

Une approche complémentaire consiste à se donner, a priori, une valeur αcomprise entre

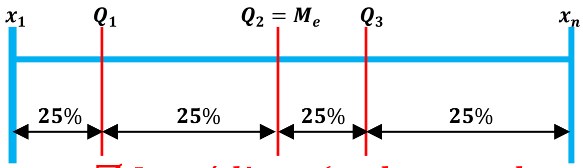
BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

$\boldsymbol{0}$ *et* $\boldsymbol{1}$, *et à rechercher* x_α, *valeur telle qu'une proportion* α *des observations lui sont inférieures ou égales (autrement dit,* x_α *vérifie* $F(x_\alpha) \cong \alpha$).

La valeur x_α *(qui n'est pas nécessairement unique) est appelée quantile (ou fractile) d'ordre* α *de la série. Les quantiles les plus utilisés sont associés à certaines valeurs particulières de* α.

Autrement dit, le quantile d'ordre α, *noté* x_α, *est tel que la proportion des observations qui lui sont inférieures ou égales vaut* α : $(F(x_\alpha) = P(X \leq x_\alpha) = \alpha)$, *tandis que la proportion des observations qui lui sont supérieures vaut* $1 - \alpha$: $(P(X > x_\alpha) = 1 - \alpha)$

✤ • La médiane et les quartiles :



☑ La médiane (ou le second quartile) $M_e = Q_2$:

La médiane est le quantile d'ordre $\frac{1}{2}$. *Elle partage donc la série des observations en deux ensembles d'effectifs égaux:* $(F(M_e) = P(X \leq M_e) = 50\%)$

• *Si le nombre d'observations n est impair, on note* $x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}$, *l'observation numéro* $\left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right)$. *La médiane est exactement la valeur* $x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}$: $\left(M_e = x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}\right)$.

• *Si le nombre d'observations n est pair, on note* $x_{\left\lfloor \frac{n}{2} \right\rfloor}$, *l'observation numéro* $\frac{n}{2}$ *et* $x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}$, *la* $\left(\left\lfloor \frac{n}{2} \right\rfloor + 1\right)$ *ème observation*.

Alors la médiane est n'importe quelle valeur comprise dans l'intervalle médian: $\left[x_{\left\lfloor \frac{n}{2} \right\rfloor}; x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}\right]$

Où $\left\lfloor \frac{n}{2} \right\rfloor$ *est la partie entière du réel* $\frac{n}{2}$

Par convention, on prend le milieu de cet intervalle : $M_e = \dfrac{x_{\left\lfloor \frac{n}{2} \right\rfloor} + x_{\left\lfloor \frac{n}{2} \right\rfloor + 1}}{2}$

☑ Le premier quartile (Q_1) et le troisième quartile (Q_3):

Le premier quartile est le quantile d'ordre $\frac{1}{4}$, *le troisième quartile celui d'ordre* $\frac{3}{4}$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

On voit donc que 25 % *des observations sont inférieures ou égales au premier quartile* :

$(F(Q_1) = P(X \leq Q_1) = 25\%)$, *tandis que* 75 % *lui sont supérieures* : $(P(X > Q_1) = 75\%)$

Pour le troisième quartile, les proportions s'inversent : 75 % *des valeurs lui sont inférieures ou égales* : $(F(Q_3) = P(X \leq Q_3) = 75\%)$, *tandis que* 25 % *lui sont supérieures* : $(P(X > Q_3) = 25\%)$

- *Si le nombre d'observations n est* **divisible par 4**, *alors* : $Q_1 = x_{\left[\frac{n}{4}\right]}$ et $Q_3 = x_{\left[\frac{3n}{4}\right]}$

- *Si le nombre d'observations n* **n'est pas divisible par 4**, *alors* : $\begin{cases} Q_1 = x_{\left[\frac{n}{4}\right]+1} \\ \text{et} \\ Q_3 = x_{\left[\frac{3n}{4}\right]+1} \end{cases}$

Où $\left[\frac{n}{4}\right]$ *et* $\left[\frac{3n}{4}\right]$ *sont respectivement les parties entières des réels* $\frac{n}{4}$ *et* $\frac{3n}{4}$

c • Les autres quantiles : *Les déciles et les centiles sont également d'usage relativement courant. Il existe* 9 *déciles qui partagent l'ensemble des observations en* 10 *parties d'égale importance (chacune contient* 10 % *des observations)*

$(F(D_1) = P(X \leq D_1) = 10\%$ et $P(X > D_1) = 90\%)$, $(F(D_2) = P(X \leq D_2) = 20\%$ et $P(X > D_2) = 80\%)$, …

, $(F(D_9) = P(X \leq D_9) = 90\%$ et $P(X > D_9) = 10\%)$ *et* 99 *centiles qui la partagent de même en* 100 *parties d'effectifs égaux* : $(F(\alpha_1) = P(X \leq \alpha_1) = 1\%$ et $P(X > \alpha_1) = 99\%)$,

$(F(\alpha_2) = P(X \leq \alpha_2) = 2\%$ et $P(X > \alpha_2) = 98\%)$, … , $(F(\alpha_{99}) = P(X \leq \alpha_{99}) = 99\%$ et $P(X > \alpha_{99}) = 1\%)$

C – 5 • Principaux paramètres de position (ou de tendance centrale) :

Si l'on doit donner un indicateur de tendance centrale unique pour caractériser le centre d'une série d'observations, on doit choisir entre la moyenne et la médiane. On retiendra que la moyenne est l'indicateur le plus naturel, le plus connu, et donc le plus utilisé dans la pratique.

De plus, la moyenne est définie par une formule mathématique donnant un résultat sans ambiguïté, ce qui n'est pas le cas de la médiane. Par contre, il faut noter que la signification de la moyenne peut être faussée par quelques valeurs très grandes ou très petites par rapport à la plupart des observations, ce qui n'est pas le cas de la médiane.

On préfère donc cette dernière lorsqu'on rencontre la situation évoquée ci-dessus, en *particulier dans le cas de séries très dissymétriques.*

a • Le mode : *On appelle mode d'une variable quantitative discrète la ou les valeur(s) ayant le plus grand effectif ou la plus grande fréquence.*

Le mode correspond aussi à la ou aux valeur(s) ayant la plus grande hauteur dans la représentation graphique de la variable.

b • Les Moyennes arithmétique, géométrique, quadratiques et harmoniques :

☑ ***La moyenne arithmétique :*** *La moyenne arithmétique est un résumé numérique et correspond au centre de gravité de la distribution.*

Elle est exprimée dans la même unité que la variable : $\overline{X} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i = \sum_{i=1}^{r} f_i x_i$

☑ ***La moyenne géométrique de r valeurs positives x_i :***

$$\overline{G} = \sqrt[n]{\prod_{i=1}^{r} x_i^{n_i}} = \left[\prod_{i=1}^{r} x_i^{n_i}\right]^{\frac{1}{n}} = \prod_{i=1}^{r} x_i^{f_i} = \exp\left[\sum_{i=1}^{r} f_i \ln(x_i)\right], car \ \ln(\overline{G}) = \sum_{i=1}^{r} f_i \ln(x_i)$$

☑ ***La moyenne quadratique :*** $\overline{Q} = \sqrt{\sum_{i=1}^{r} f_i x_i^2}$ *ou encore* $\overline{Q}^2 = \sum_{i=1}^{r} f_i x_i^2$

☑ ***La moyenne harmonique de r valeurs non nulles x_i :***

$$\overline{H} = \left[\sum_{i=1}^{r} \frac{f_i}{x_i}\right]^{-1} \text{ ou encore } \frac{1}{\overline{H}} = \sum_{i=1}^{r} \frac{f_i}{x_i}$$

La relation suivante sera toujours vérifiée : $\overline{H} \leq \overline{G} \leq \overline{X} \leq \overline{Q}$

C-6 • Principaux paramètres de dispersion :

Les paramètres de dispersion servent à préciser la variabilité de la série, c'est-à-dire à résumer l'éloignement de l'ensemble des observations par rapport à leur tendance centrale. On trouve diverses caractéristiques de dispersion, certaines étant plus courantes que d'autres.

a • L'étendue : *On appelle étendue la différence entre la plus grande valeur et la plus petite valeur prise par la variable :* $E = x_r - x_1$

b • ***L'étendue interquartile*** : ***Écart entre le troisième et le premier quartile, il contient au moins 50% des observations centrales de la distribution, il n'est pas influencé par les valeurs extrêmes*** : $EIQ = Q_3 - Q_1$

c • ***Écart absolu moyen par rapport à la médiane*** : $e_{M_e} = \sum_{i=1}^{r} f_i |x_i - M_e|$

d • ***Écart absolu moyen par rapport à la moyenne*** : $e_{\overline{X}} = \sum_{i=1}^{r} f_i |x_i - \overline{X}|$

e • ***Écart-type ou écart quadratique moyen*** : ***On appelle variance ou fluctuation, la moyenne arithmétique des carrés des écarts des résultats observés à leur moyenne*** :

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \overline{X})^2 = \sum_{i=1}^{r} f_i (x_i - \overline{X})^2$$

L'écart-type mesure la dispersion des données autour de la moyenne: $S_x = \sqrt{\sum_{i=1}^{r} f_i (x_i - \overline{X})^2}$

☑ ***Propriétés*** :

☞ $S_x^2 \geq 0$ ***et*** $S_x \geq 0$

☞ ***L'écart-type est exprimé dans la même unité que la variable.***

☞ ***Plus l'écart-type est petit, plus les données individuelles sont regroupées autour de la moyenne. Plus il est grand, plus les données individuelles sont dispersées autour de la moyenne.***

☞ ***Théorème de König-Huygens*** : $S_x^2 = \overline{X^2} - \overline{X}^2 = \overline{Q}^2 - \overline{X}^2 = \sum_{i=1}^{r} f_i x_i^2 - \overline{X}^2$

☞ $S_{ax+b}^2 = a^2 S_x^2$

☞ $S_{ax+b} = |a| S_x$

☞ ***Variance d'échantillonnage*** : $S^2 = \frac{1}{n-1}\sum_{i=1}^{r} n_i (x_i - \overline{X})^2 = \frac{n}{n-1} S_x^2$

Lorsque la série est un échantillon issu d'une population et que l'on s'intéresse aux caractéristiques de cette population via l'échantillon (inférence), on utilise plutôt S^2 ***qui est un meilleur estimateur de la variance théorique de la population*** $(V(X) = \sigma^2)$.

Dès lors que la taille n de la série est assez grande, $S^2 \approx S_x^2$

☞ *Si* $S_x^2 = S_x = 0$ *, alors toutes les données sont égales et égales à la moyenne*: $\forall i, x_i = \overline{X}$

☞ **Identité :** $\dfrac{1}{n}\sum_{i=1}^{r}(x_i - \overline{X})^2 = \dfrac{1}{n}\sum_{i=1}^{r}(x_i - a)^2 - (\overline{X} - a)^2$

f • Variable centrée réduite :

Une variable centrée et réduite est une variable dont la moyenne est nulle et l'écart-type vaut 1. *Une variable centrée et réduite s'exprime toujours sans unité.*

Pour centrer et réduire la variable X, *on fait le changement de variable :* $y_i = \dfrac{x_i - \overline{X}}{S_x}$

Alors, on vérifie que : $\overline{Y} = 0$ *et* $S_y = 1$

g • Le coefficient de variation : $C_V = \dfrac{S_x}{\overline{X}} \times 100$

☞ *Le* C_V *permet d'apprécier la représentativité de la moyenne par rapport à l'ensemble des observations. Il donne une bonne idée du degré d'homogénéité d'une série. Il faut qu'il soit le plus faible possible* ($< 15\%$ *en pratique*).

☞ *Le* C_V *est un paramètre sans dimension. On l'exprime généralement en pourcentage*

☞ *Le* C_V *sert à comparer deux distributions qui ne sont pas exprimées dans la même unité, ou dans le cas des distributions dont les moyennes sont très différentes.*

☞ *Un coefficient de variation plus élevé correspond à une distribution plus dispersée*

h • Intervalle de variation : *Soit* α *une proportion donnée* $(0 < \alpha < 1)$.

L'intervalle de variation au risque α *ou deniveau* $1-\alpha$ *contient une proportion* $1-\alpha$ *d'observations; de plus les données qui sont à l'extérieur de cet intervalle (en proportion* α*) se répartissent également: il y en a autant à "gauche" qu'à "droite", en proportion* $\dfrac{\alpha}{2}$.

On écrit donc l'intervalle de variation: où $x_{\frac{\alpha}{2}}$ *est le quantile où* $x_{\alpha/2}$ *est le quantile d'ordre* $\alpha/2$: $\left(F\left(x_{\alpha/2}\right) = \alpha/2\right)$ *et* $x_{1-\alpha/2}$ *le quantile d'ordre* $1 - \alpha/2$: $\left(F\left(x_{1-\alpha/2}\right) = 1 - \alpha/2\right)$

i • Les moments empiriques :

☑ Moments empiriques par rapport à a d'ordre s :

Soient $x_1, x_2, \dots, x_r$ *une série statistique et* $a \in \mathbb{R}$, *le moments empiriques par rapport à* a *d'ordre* s (*où* s *supposé entier positif*) *est définie par :* $\dfrac{1}{n}\sum_{i=1}^{r} n_i(x_i - a)^s$

☑ ***Moments empiriques par rapport à l'origine (ou non centrés) d'ordre k :***

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{r} x_i^k \text{, avec : } \overline{m}_1 = \overline{X}$$

☑ ***Moments empiriques centrés d'ordre k :***

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{r} n_i(x_i - \overline{X})^k \text{, avec : } \overline{\mu}_1 = 0 \text{ et } \overline{\mu}_2 = S_x^2$$

j • ***Moyennes et variances dans des groupes :*** ***upposons que les n observations soient réparties dans deux groupes G_A et G_B. Les n_A premières observations sont dans le groupe G_A et les n_B dernières observations sont dansle groupe G_B, avec la relation $n_A + n_B = n$. On suppose quela série statistique contient d'abord les unités de G_A puis les unités de G_B:*** $\underbrace{x_1, x_2, \dots, x_{n_A}}_{\text{Observations de } G_A}, \underbrace{x_{n_A+1}, x_{n_A+2}, \dots, x_{n-1}, x_n}_{\text{Observations de } G_B}$

On définit les moyennes des deux groupes: La moyenne du premier groupe: $\overline{x}_A = \frac{1}{n_A}\sum_{i=1}^{n_A} x_i$

et du deuxième groupe : $\overline{x}_B = \frac{1}{n_B}\sum_{i=n_A+1}^{n} x_i$ ***. La moyenne générale est une moyenne pondérée par la taille des groupes des moyennes des deux groupes.***

En effet: $\overline{x} = \frac{1}{n}\left(\sum_{i=1}^{n_A} x_i + \sum_{i=n_A+1}^{n} x_i\right) = \frac{1}{n}(n_A\overline{x}_A + n_B\overline{x}_B)$

On peut également définir les variances des deux groupes: La variance du premier groupe : $S_A^2 = 1/n_A \sum_{i=1}^{n_A}(x_i - \overline{x}_A)^2$ ***et du deuxième groupe :*** $S_B^2 = \frac{1}{n_B}\sum_{i=n_A+1}^{n}(x_i - \overline{x}_B)^2$

☑ ***Théorème de Huygens :*** ***La variance totale, définie par :*** $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ ***se décompose de la manière suivante :***

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \underbrace{\frac{n_A S_A^2 + n_B S_B^2}{n}}_{\text{Variance intra-groupes}} + \underbrace{\frac{n_A(\overline{x}_A - \overline{x})^2 + n_B(\overline{x}_B - \overline{x})^2}{n}}_{\text{Variance inter-groupes}}$$

☞ *Démonstration :*

$$S_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\left(\sum_{i=1}^{n_A}(x_i-\bar{x})^2+\sum_{i=n_A+1}^{n}(x_i-\bar{x})^2\right)$$

$$Or:\sum_{i=1}^{n_A}(x_i-\bar{x})^2=\sum_{i=1}^{n_A}(x_i-\bar{x}_A+\bar{x}_A-\bar{x})^2=\sum_{i=1}^{n_A}(x_i-\bar{x}_A)^2+\sum_{i=1}^{n_A}(\bar{x}_A-\bar{x})^2+2\underbrace{\sum_{i=1}^{n_A}(x_i-\bar{x}_A)(\bar{x}_A-\bar{x})}_{=0}$$

$$\sum_{i=1}^{n_A}(x_i-\bar{x})^2=\sum_{i=1}^{n_A}(x_i-\bar{x}_A+\bar{x}_A-\bar{x})^2=\sum_{i=1}^{n_A}(x_i-\bar{x}_A)^2+\sum_{i=1}^{n_A}(\bar{x}_A-\bar{x})^2+2\underbrace{\sum_{i=1}^{n_A}(x_i-\bar{x}_A)(\bar{x}_A-\bar{x})}_{=0}$$

$$\sum_{i=1}^{n_A}(x_i-\bar{x})^2=n_AS_A^2+n_A(\bar{x}_A-\bar{x})^2$$

On a évidemment la même relation dans le groupe G_B: $\sum_{i=n_A+1}^{n}(x_i-\bar{x})^2=n_BS_B^2+n_B(\bar{x}_B-\bar{x})^2$

En revenant à l'expression $\left(S_x^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\left(\sum_{i=1}^{n_A}(x_i-\bar{x})^2+\sum_{i=n_A+1}^{n}(x_i-\bar{x})^2\right)\right)$,

on obtient : $S_x^2=\frac{1}{n}\left(n_AS_A^2+n_A(\bar{x}_A-\bar{x})^2+n_BS_B^2+n_B(\bar{x}_B-\bar{x})^2\right)$

$$=\frac{n_AS_A^2+n_BS_B^2}{n}+\frac{n_A(\bar{x}_A-\bar{x})^2+n_B(\bar{x}_B-\bar{x})^2}{n}$$

Exercice 1 :

ÉNONCÉ

Montrer que la quantité $\frac{1}{n}\sum_{i=1}^{n}(x_i - t)^2$ *est minimale si* $t = \bar{x}$

Corrigé

$$\varphi(t) = \frac{1}{n}\sum_{i=1}^{n}(x_i - t)^2 \Rightarrow \frac{d\varphi(t)}{dt} = \varphi'(t) = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - t)^2\right]' = \frac{1}{n}\sum_{i=1}^{n}[(x_i - t)^2]'$$

$$\frac{d\varphi(t)}{dt} = \varphi'(t) = \frac{1}{n}\sum_{i=1}^{n}2(x_i - t)'(x_i - t) = -\frac{2}{n}\sum_{i=1}^{n}(x_i - t) \Rightarrow \frac{d^2\varphi(t)}{dt^2} = \varphi''(t) = -\frac{2}{n}\sum_{i=1}^{n}(x_i - t)'$$

$$\frac{d^2\varphi(t)}{dt^2} = \varphi''(t) = -\frac{2}{n}\sum_{i=1}^{n}-1 = 2$$

$$par\ la\ suite, \left(t_0\ minimise\ \frac{1}{n}\sum_{i=1}^{n}(x_i - t)^2\right) \Leftrightarrow \begin{cases}\varphi'(t_0) = 0\\ \varphi''(t_0) \geq 0\end{cases} \Leftrightarrow \begin{cases}-\frac{2}{n}\sum_{i=1}^{n}(x_i - t) = 0\\ 2 \geq 0\ , évidente\end{cases}$$

$$\Leftrightarrow \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}t_0 = 0 \Leftrightarrow n\bar{x} - nt_0 = 0 \Leftrightarrow t_0 = \bar{x}$$

Exercice 2 :

ÉNONCÉ

Montrer que La fonction $g(t) = \sum_{i=1}^{n}|x_i - t|$ *est minimale si t est la médiane de* $x_1, x_2, \dots, x_n$.

1)
   a) *Pour n pair soit l'observation* : $x_1 = -5, x_2 = -2, x_3 = 0, x_4 = 1, x_5 = 6, x_6 = 9$
   Avec une représentation graphique de $g(t)$, *vérifier que toute valeur de l'intervalle* $[0, 1]$ *convienne à la médiane.*
   b) *Pour n impair soit l'observation* : $x_1 = -5, x_2 = -2, x_3 = 0, x_4 = 1, x_5 = 6$
   Avec une représentation graphique de $g(t)$, *vérifier que la valeur médiane est* $M_e = 0$
2) *Généraliser ces résultats en distinguant les cas suivant la parité de la taille n de l'échantillon.*

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

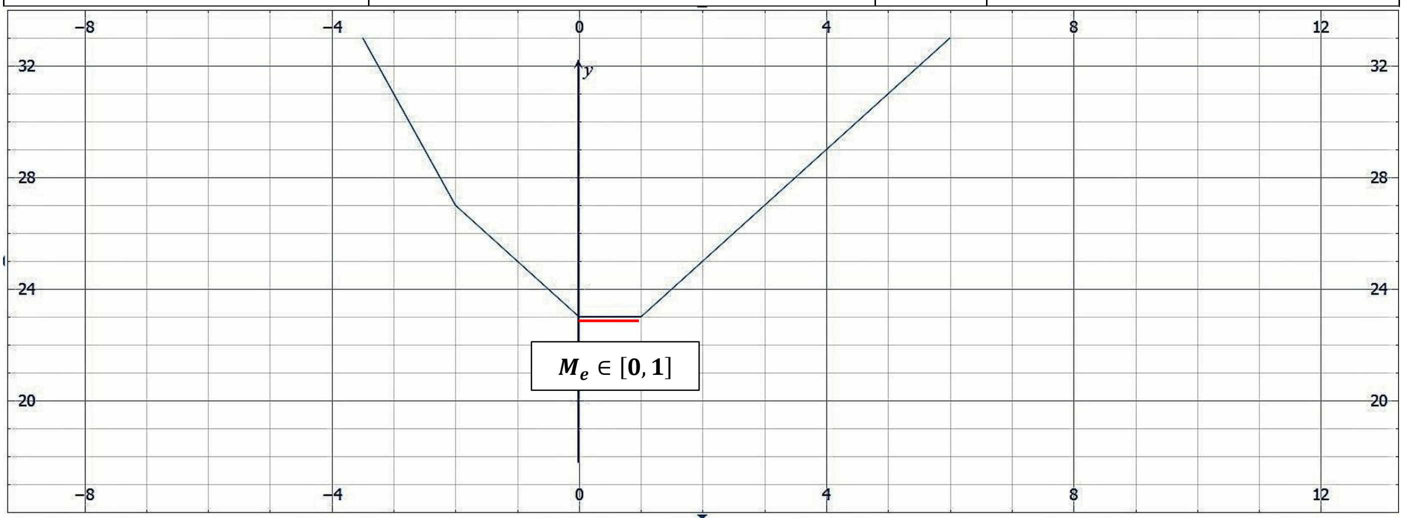
Corrigé

1)

a) $g(t)=\sum_{i=1}^{6}|x_i-t| = |-5-t|+|-2-t|+|0-t|+|1-t|+|6-t|+|9-t|$

$g(t)=|t+5|+|t+2|+|t|+|t-1|+|t-6|+|t-9|$

t	$-\infty$; -5	-5 ; -2	-2 ; 0	0 ; 1	1 ; 6	6 ; 9	9 ; $+\infty$
$\|t+5\|$	$-t-5$	$t+5$	$t+5$	$t+5$	$t+5$	$t+5$	$t+5$
$\|t+2\|$	$-t-2$	$-t-2$	$t+2$	$t+2$	$t+2$	$t+2$	$t+2$
$\|t\|$	$-t$	$-t$	$-t$	t	t	t	t
$\|t-1\|$	$1-t$	$1-t$	$1-t$	$1-t$	$t-1$	$t-1$	$t-1$
$\|t-6\|$	$6-t$	$6-t$	$6-t$	$6-t$	$6-t$	$t-6$	$t-6$
$\|t-9\|$	$9-t$	$9-t$	$9-t$	$9-t$	$9-t$	$t-a$	$t-9$
$g(t)$	$-6t+9$	$-4t+19$	$-2t+23$	23	$2t+21$	$4t+9$	$6t-9$
Sens de variation de g	$\searrow$	$\searrow$	$\searrow$	$\rightarrow$	$\nearrow$	$\nearrow$	$\nearrow$



Toute valeur de l'intervalle $[0,1]$ ***convient à la médiane.***

$$g(t)=\sum_{i=1}^{6}|x_i-t| \ \textit{est minimale lorsque } t\in[0,1] \textit{ est la médiane de } \{x_1,x_2,x_3,x_4,x_5,x_6\}$$

ou encore $M_e \in \left[x_{\left[\frac{n}{2}\right]};\ x_{\left[\frac{n}{2}\right]+1}\right]$, ***avec,*** $x_{\left[\frac{n}{2}\right]} = x_{\left[\frac{6}{2}\right]} = x_3 = 0$ ***et*** $x_{\left[\frac{n}{2}\right]+1} = x_{\left[\frac{6}{2}\right]+1} = x_4 = 1$

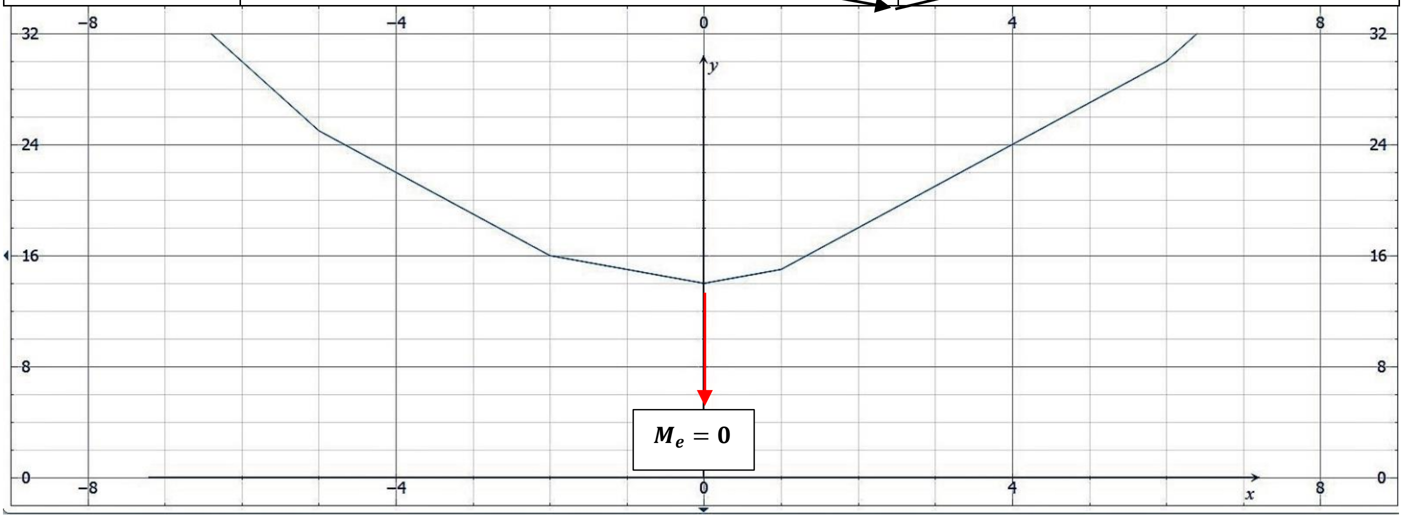
Par convention, on prend le milieu de cet intervalle : $M_e = \dfrac{x_{\left[\frac{n}{2}\right]}+x_{\left[\frac{n}{2}\right]+1}}{2} = \dfrac{x_3+x_4}{2} = \dfrac{1}{2}$

b) $g(t)=\sum_{i=1}^{5}|x_i-t| = |-5-t|+|-2-t|+|0-t|+|1-t|+|6-t|$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

$g(t) = |t+5| + |t+2| + |t| + |t-1| + |t-6|$

t	$-\infty \qquad -5$	-2	0	1	6	$+\infty$
$\lvert t+5\rvert$	$-t-5$	$t+5$	$t+5$	$t+5$	$t+5$	$t+5$
$\lvert t+2\rvert$	$-t-2$	$-t-2$	$t+2$	$t+2$	$t+2$	$t+2$
$\lvert t\rvert$	$-t$	$-t$	$-t$	t	t	t
$\lvert t-1\rvert$	$1-t$	$1-t$	$1-t$	$1-t$	$t-1$	$t-1$
$\lvert t-6\rvert$	$6-t$	$6-t$	$6-t$	$6-t$	$6-t$	$t-6$
$g(t)$	$-5t$	$-3t+10$	$-t+14$	$t+14$	$3t+12$	$5t$
Sens de variation de g	$\searrow$			$\nearrow$		



$La\ valeur\ qui\ réalise\ un\ minimum\ de\ \ g(t) = \sum_{i=1}^{5} |x_i - t|\ convient\ à\ la\ médiane\ et\ elle$

$est\ unique\ dans\ ce\ cas\ :\ M_e = 0\ est\ la\ médiane\ de\ \{x_1, x_2, x_3, x_4, x_5\}$

2) $Maintenant, on\ va\ généraliser\ les\ résultats\ obtenus\ à\ travers\ ces\ deux\ exemples\ :$

$$\forall t \in [x_r, x_{r+1}[\, , g(t) = \sum_{i=1}^{r} |x_i - t| + \sum_{i=r+1}^{n} |x_i - t|\ \ , ceci\ pour\ , r \in \{1, 2, \dots, n-1\}$$

$$or, x_r \leq t < x_{r+1} \Rightarrow \begin{cases} |x_i - t| = t - x_i\, , \forall 1 \leq i \leq r \\ |x_i - t| = x_i - t\, , \forall r+1 \leq i \leq n \end{cases}$$

$$Ainsi, g(t) = \sum_{i=1}^{r} (t - x_i) + \sum_{i=r+1}^{n} (x_i - t) = \underbrace{\sum_{i=1}^{r} t}_{rt} - \sum_{i=1}^{r} x_i + \sum_{i=r+1}^{n} x_i - \underbrace{\sum_{i=r+1}^{n} t}_{(n-r)t}$$

$$\forall t\in[x_r,x_{r+1}[\;:\;g(t)=(2r-n)t+\underbrace{\left[\sum_{i=r+1}^{n}x_i-\sum_{i=1}^{r}x_i\right]}_{s}=(2r-n)t+s$$

g est donc une fonction affine par morceaux, décroissante tant que , $2r-n<0$,

autrement dit , $r<\frac{n}{2}$ *et croissante quand* $2r-n>0$ *c.-à-d.* $r>\frac{n}{2}$.

Plus précisément, on distingue 2 cas :

☑ *si n est pair* : $n=2p$ $(ou\ p=\frac{n}{2})\Rightarrow\forall t\in[x_p,x_{p+1}[\,,g(t)=\underbrace{(2p-n)}_{0}t+\underbrace{\left[\sum_{i=p+1}^{n}x_i-\sum_{i=1}^{p}x_i\right]}_{s}=s$

g(t) est constante sur $[x_p,x_{p+1}[$ *et elle est minimale sur cet intervalle. On reconnaît là,*

la médiane , $M_e\in[x_p,x_{p+1}[$ *avec :* $x_p=x_{\lfloor\frac{n}{2}\rfloor}$ *et* $x_{p+1}=x_{\lfloor\frac{n}{2}\rfloor+1}$

Par convention, on prend le milieu de cet intervalle : $M_e=\dfrac{x_{\lfloor\frac{n}{2}\rfloor}+x_{\lfloor\frac{n}{2}\rfloor+1}}{2}$

☑ *si n est impair* : $n=2p+1\Rightarrow 2r-n\neq 0$

Or $\forall t\in[x_r,x_{r+1}[\;:\;g(t)=(2r-n)t+s$

g étant une fonction affine par morceaux, donc, décroissante tant que , $2r-n<0$,

autrement dit , $r<\frac{n}{2}$ $\left(ou\ encore\ r<p+\frac{1}{2}\right)$ *et croissante quand* $2r-n>0$ *c.-à-d.* $r>\frac{n}{2}$

$\left(ou\ encore\ r>p+\frac{1}{2}\right)$

Ainsi, g est décroissante puis croissante avec un minimum unique pour

$t=x_{\lfloor\frac{n}{2}\rfloor+1}=x_{\lfloor\frac{2p+1}{2}\rfloor+1}=x_{p+1},$

la médiane est alors unique : $M_e=x_{\lfloor\frac{n}{2}\rfloor+1}=x_{\lfloor\frac{2p+1}{2}\rfloor+1}=x_{p+1}$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

Institut de Financement du Développement du Maghreb Arabe
CONCOURS DE RECRUTEMENT DE LA XXVII ème PROMOTION
JUILLET 2007

Exercice 3 : *(8 points :1+1+1+1+1+1+1+1)*

ÉNONCÉ

On dispose d'un ensemble E constitué de deux sous-ensembles A et B de n et m entreprises appartenant à deux secteurs d'activité distincts. Les observations relatives aux chiffres d'affaires de ces entreprises sont respectivement notées $X_1, X_2, \dots, X_n$ *et* $X_{n+1}, X_{n+2}, \dots, X_{n+m}$

1)
   a) *Rappeler l'expression de* M_A*, la moyenne du chiffre d'affaire des entreprises de la sous-population A, en fonction de* $X_1, X_2, \dots, X_n$.
   b) *Même question pour* M_B*, la moyenne du chiffre d'affaire des entreprises de la sous-population B en fonction de* $X_{n+1}, X_{n+2}, \dots, X_{n+m}$
   c) *En déduire l'expression de la moyenne générale M de l'ensemble des entreprises en fonction des moyennes* M_A *et* M_B*. Interpréter ce résultat.*
   d) *Exprimer les différences* $M - M_A$ *et* $M - M_B$ *en fonction de l'expression* $M_A - M_B$
   e) *Comparer M à* M_A *et à* M_B *si l'on suppose que* $M_A = M_B$
2) *On admet dans cette question que* $M_A = M_B$*. Exprimer la variance V de la variable X sur toute la population en fonction des variances de la même variable sur les deux sous-ensembles A et B, notées* V_A *et* V_B
3) *Reprendre la question précédente si l'on abandonne l'hypothèse d'égalité entre* M_A *et* M_B
4) *Application numérique : Retrouver le résultat de la question 3) pour le cas où* $n = m = 5$ *avec les valeurs suivantes de la variable X :*

Sous-ensemble A	$X_1 = 2$	$X_2 = 6$	$X_3 = 4$	$X_4 = 5$	$X_5 = 3$
Sous-ensemble B	$X_6 = 11$	$X_7 = 19$	$X_8 = 20$	$X_9 = 15$	$X_{10} = 25$

Corrigé

1)
   a) $M_A = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$
   b) $M_B = \frac{1}{(n+m)-(n+1)+1}\sum_{i=n+1}^{n+m} X_i = \frac{1}{m}\sum_{i=n+1}^{n+m} X_i = \frac{1}{m}(X_{n+1} + X_{n+2} + \cdots + X_{n+m})$
   c) $M = \frac{1}{n+m}\sum_{i=1}^{n+m} X_i = \frac{1}{n+m}\left[\underbrace{\sum_{i=1}^{n} X_i}_{n\,M_A} + \underbrace{\sum_{i=n+1}^{n+m} X_i}_{mM_B}\right] \Rightarrow \boxed{M = \frac{n\,M_A + mM_B}{n+m}}$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

M *représente la moyenne arithmétique pondérée des deux moyennes arithmétiques simples* M_A *et* M_B *affectée des poids* n *et* m

d)

$$\cdot\ M - M_A = \frac{n\,M_A + mM_B}{n+m} - M_A = \frac{n\,M_A + mM_B - (n+m)\,M_A}{n+m} = \frac{mM_B - m\,M_A}{n+m}$$

$$\boxed{M - M_A = -\left(\frac{m}{n+m}\right)(M_A - M_B)}$$

$$\cdot\ M - M_B = \frac{n\,M_A + mM_B}{n+m} - M_B = \frac{n\,M_A + mM_B - (n+m)\,M_B}{n+m} = \frac{nM_A - n\,M_B}{n+m}$$

$$\boxed{M - M_B = \left(\frac{n}{n+m}\right)(M_A - M_B)}$$

e) Si $M_A = M_B$

$$M - M_A = 0 \Leftrightarrow M = M_A \text{ Et } M - M_B = 0 \Leftrightarrow M = M_B$$

$D'où$ $\boxed{M_A = M_B \Rightarrow M = M_A = M_B}$

2) *Pour* $M = M_A = M_B$

$$V = \frac{1}{n+m}\sum_{i=1}^{n+m}(X_i - M)^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}\underbrace{(X_i - M)^2}_{(X_i-M_A)^2} + \sum_{i=n+1}^{n+m}\underbrace{(X_i - M)^2}_{(X_i-M_B)^2}\right]$$

$$V = \frac{1}{n+m}\left[\sum_{i=1}^{n}(X_i - M_A)^2 + \sum_{i=n+1}^{n+m}(X_i - M_B)^2\right]$$

$$Or, V_A = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_A)^2 \Rightarrow \sum_{i=1}^{n}(X_i - M_A)^2 = nV_A \; et \; V_B = \frac{1}{m}\sum_{i=n+1}^{n+m}(X_i - M_B)^2 \Rightarrow \sum_{i=n+1}^{n+m}(X_i - M_B)^2 = mV_B$$

$$Donc\quad V = \frac{1}{n+m}[nV_A + mV_B] \Rightarrow \boxed{V = \frac{nV_A + mV_B}{n+m}}$$

Sous l'hypothèse d'égalité des moyennes des deux sous-ensemble A et B, la variance V sur toute la population n'est autre que la moyenne pondérée des variances partielles.

La variance inter-groupes est nulle : $\dfrac{n_A(M_A - M)^2 + n_B(M_B - M)^2}{n} = 0$

3) *Pour* $M_A \neq M_B$

$$V = \frac{1}{n+m}\sum_{i=1}^{n+m}(X_i - M)^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}(X_i - M)^2 + \sum_{i=n+1}^{n+m}(X_i - M)^2\right]$$

$$= \frac{1}{n+m}\left[\sum_{i=1}^{n}[(X_i - M_A) - (M - M_A)]^2 + \sum_{i=n+1}^{n+m}[(X_i - M_B) - (M - M_B)]^2\right]$$

Avec,

$$\cdot \sum_{i=1}^{n}[(X_i - M_A) - (M - M_A)]^2 = \sum_{i=1}^{n}[(X_i - M_A)^2 + (M - M_A)^2 - 2(M - M_A)(X_i - M_A)]$$

$$= \underbrace{\sum_{i=1}^{n}(X_i - M_A)^2}_{nV_A} + \underbrace{\sum_{i=1}^{n}(M - M_A)^2}_{n(M-M_A)^2} - 2(M - M_A)\underbrace{\sum_{i=1}^{n}(X_i - M_A)}_{0}$$

$$\sum_{i=1}^{n}[(X_i - M_A) - (M - M_A)]^2 = nV_A + n(M - M_A)^2$$

$$\cdot \sum_{i=n+1}^{n+m}[(X_i - M_B) - (M - M_B)]^2 = \sum_{i=n+1}^{n+m}[(X_i - M_B)^2 + (M - M_B)^2 - 2(M - M_B)(X_i - M_B)]$$

$$= \underbrace{\sum_{i=n+1}^{n+m}(X_i - M_B)^2}_{mV_B} + \underbrace{\sum_{i=n+1}^{n+m}(M - M_B)^2}_{m(M-M_B)^2} - 2(M - M_B)\underbrace{\sum_{i=n+1}^{n+m}(X_i - M_B)}_{0}$$

$$\sum_{i=n+1}^{n+m}[(X_i - M_B) - (M - M_B)]^2 = mV_B + m(M - M_B)^2$$

$$Or\ V = \frac{1}{n+m}\left[\underbrace{\sum_{i=1}^{n}[(X_i - M_A) - (M - M_A)]^2}_{nV_A+n(M-M_A)^2} + \underbrace{\sum_{i=n+1}^{n+m}[(X_i - M_B) - (M - M_B)]^2}_{mV_B+m(M-M_B)^2}\right]$$

$$Par\ la\ suite : \boxed{V = \underbrace{\frac{nV_A + mV_B}{n+m}}_{\text{Variance intra-groupes}} + \underbrace{\frac{n(M - M_A)^2 + m(M - M_B)^2}{n+m}}_{\text{Variance inter-groupes}}}$$

Et comme on a les expressions de $M - M_A$ *et de* $M - M_B$ *en fonction de* $M_A - M_B$ *, on pourra ainsi exprimer la variance totale* V *en fonction de* n, m, V_A, V_B, M_A *et* M_A

$$On\ a :\ M - M_A = -\left(\frac{m}{n+m}\right)(M_A - M_B) \Rightarrow (M - M_A)^2 = \frac{m^2}{(n+m)^2}(M_A - M_B)^2$$

$$et\ M - M_B = \left(\frac{n}{n+m}\right)(M_A - M_B) \Rightarrow (M - M_B)^2 = \frac{n^2}{(n+m)^2}(M_A - M_B)^2$$

$$Ainsi : V = \frac{nV_A + mV_B}{n+m} + \frac{n(M - M_A)^2 + m(M - M_B)^2}{n+m}$$

$$= \frac{nV_A + mV_B}{n+m} + \frac{nm^2(M_A - M_B)^2 + mn^2(M_A - M_B)^2}{(n+m)^3} = \frac{nV_A + mV_B}{n+m} + \frac{nm(n+m)(M_A - M_B)^2}{(n+m)^3}$$

$$D'\text{où}, une\ autre\ expression\ de\ V : \boxed{V = \frac{nV_A + mV_B}{n+m} + \frac{nm(M_A - M_B)^2}{(n+m)^2}}$$

4)

Sous-ensemble A	$X_i = x_i\ ;\ i = 1,2,\dots,5$	$X_i - M_A$	$(X_i - M_A)^2$
Enterprise 1	$X_1 = 2$	-2	4
Enterprise 2	$X_2 = 6$	2	4
Enterprise 3	$X_3 = 4$	0	0
Enterprise 4	$X_4 = 5$	1	1
Enterprise 5	$X_5 = 3$	-1	1
$\sum$	$\sum_{i=1}^{5} X_i = 20$	$\sum_{i=1}^{5}(X_i - M_A) = 0$	$\sum_{i=1}^{5}(X_i - M_A)^2 = 10$

· $M_A = \frac{1}{5}\sum_{i=1}^{5} X_i = \frac{1}{5} \times 20 \Rightarrow \boxed{M_A = 4}$ $et\ V_A = \frac{1}{5}\sum_{i=1}^{5}(X_i - M_A)^2 = \frac{1}{5} \times 10 \Rightarrow \boxed{V_A = 2}$

Sous-ensemble B	$X_i = x_i\ ;\ i = 6,7,\dots,10$	$X_i - M_B$	$(X_i - M_B)^2$
Enterprise 6	$X_6 = 11$	-7	49
Enterprise 7	$X_7 = 19$	1	1
Enterprise 8	$X_8 = 20$	2	4
Enterprise 9	$X_9 = 15$	-3	9
Enterprise 10	$X_{10} = 25$	7	49
$\sum$	$\sum_{i=6}^{10} X_i = 90$	$\sum_{i=6}^{10}(X_i - M_B) = 0$	$\sum_{i=6}^{10}(X_i - M_B)^2 = 112$

· $M_B = \frac{1}{5}\sum_{i=6}^{10} X_i = \frac{1}{5} \times 90 \Rightarrow \boxed{M_B = 18}$

· $\begin{cases} M_A = 4 \\ M_B = 18 \end{cases} \Rightarrow M = \frac{n\,M_A + mM_B}{n+m} = \frac{5\,M_A + 5M_B}{5+5} = \frac{M_A + M_B}{2} = \frac{4+18}{2} \Rightarrow \boxed{M = 11}$

· $V_B = \frac{1}{5}\sum_{i=6}^{10}(X_i - M_B)^2 = \frac{1}{5} \times 112 \Rightarrow \boxed{V_B = \frac{112}{5}}$

· $V = \frac{nV_A + mV_B}{n+m} + \frac{nm(M_A - M_B)^2}{(n+m)^2} = \frac{(5\times 2) + \left(5 \times \frac{112}{5}\right)}{5+5} + \frac{5 \times 5 \times (4-18)^2}{(5+5)^2} \underbrace{\frac{2 + \frac{112}{5}}{2}}_{\frac{61}{5}} + \frac{14^2}{4}$

$$\boxed{V = 61,2}$$

Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Ingénieurs
Samedi 5 Janvier 2013

Exercice 4 (4 points = 3+1) :

ÉNONCÉ

Considérons la distribution statistique suivante où les valeurs y_1 et $y_2 > 0$ de la variable y sont inconnues :

y_i	n_i (*effectifs*)
9	1
y_1	2
11	1
y_2	1

1) *En admettant que la moyenne arithmétique des valeurs de la variable y vaut 6,2 et que la variance de y est égale 10,56 , calculer les valeurs de y_1 et y_2*
2) *Calculer la moyenne géométrique (G) et la moyenne harmonique (H) des valeurs de la variable y pour $y_1 = 3$ et $y_2 = 5$*

Corrigé

1) *Sachant que la moyenne arithmétique des valeurs de la variable y vaut 6,2 ; on a alors* :

$$\overline{Y} = 6,2 \Rightarrow \frac{1}{n}\sum_{i=1}^{4} n_i y_i = 6,2 \Rightarrow \frac{9+2y_1+11+y_2}{5} = 6,2 \Rightarrow 2y_1 + y_2 = 11 \quad (1)$$

Par définition la variance est égale à : $S_y^2 = \frac{1}{n}\sum_{i=1}^{4} n_i(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{4} n_i y_i^2 - \overline{y}^2$

$$\sum_{i=1}^{4} n_i y_i^2 = (9)^2 + 2(y_1)^2 + (11)^2 + (y_2)^2 = 202 + 2y_1^2 + y_2^2$$

Puisque $S_y^2 = 10,56 \Rightarrow \frac{202+2y_1^2+y_2^2}{5} - (6,2)^2 = 10,56 \Rightarrow 2y_1^2 + y_2^2 = 43 \quad (2)$

En exprimant y_1 en fonction de y_2 à partir de (1), le résultat (2) s'écrit: $3y_1^2 - 22y_1 + 39 = 0$

La résolution de cette équation donne deux valeurs possibles pour y_1 : $y_1 = 3$ *ou* $y_1 = \frac{13}{3}$

En tenant compte de (1), *on a également deux valeurs possibles de* y_2 : $y_2 = 5$ *ou* $y_2 = \frac{7}{3}$

2)

y_i	n_i	f_i	$\frac{f_i}{y_i}$	$f_i \ln y_i$	$f_i y_i^2$
9	1	0,2	0.022	0.439	16.2
3	2	0,4	0.133	0.439	3.6
11	1	0,2	0.018	0.48	24.2
5	1	0,2	0.04	0.322	5
$\sum$	5	1	0,213	1,68	49

· *Moyenne Géométrique* : $G = \sqrt[n]{\prod_{i=1}^{4} y_i^{n_i}}$

ou encore $\ln(G) = \frac{1}{n}\sum_{i=1}^{4} n_i \ln(y_i) = \sum_{i=1}^{4} f_i \ln(y_i) = 1,68 \Rightarrow G = e^{1,68} \Rightarrow \boxed{G = 5,366}$

· *Moyenne Quadratique* : $Q = \sqrt{\sum_{i=1}^{4} f_i y_i^2} \Rightarrow \boxed{Q = 7}$

· *Moyenne Harmonique* : $H = \frac{n}{\sum_{i=1}^{4} \frac{n_i}{y_i}} \Rightarrow \frac{1}{H} = \sum_{i=1}^{4} \frac{f_i}{y_i} = 0,213 \Rightarrow \boxed{H = 4,695}$

La relation suivante sera toujours vérifiée : $\underbrace{H}_{4,695} \leq \underbrace{G}_{5,366} \leq \underbrace{\overline{Y}}_{6,2} \leq \underbrace{Q}_{7}$

Variables quantitatives continues

D-1 • Généralités :

Une variable quantitative continue est à valeurs réelles. Elle prend un trop grand nombre de valeurs pour qu'on puisse toutes les recenser.

Ce qui nous pousse, dans ce cas, à regrouper les individus par classes. On décompose l'ensemble des valeurs possibles en une partition d'intervalles.

En général, les deux raisons principales qui peuvent amener à considérer comme continue une variable quantitative sont le grand nombre d'observations distinctes (un traitement en discret serait, dans ce cas, peu commode) et le caractère "sensible" d'une variable (lors d'une enquête, il est moins gênant de demander à des individus leur classe de salaire que leur salaire précis ; même chose pour l'âge). Deux exemples de variables quantitatives *fréquemment considérées comme continues sont ainsi le revenu et l'âge (pour un groupe d'individus).*

a • Choix du nombre de classes : *Le regroupement en classe présente une part de subjectivité. En effet aucune loi ni théorème ne permet de déterminer le nombre de classe à utiliser. Ce nombre doit être ni trop grand (en général ≤ 20) ni trop petit (en général ≥ 5), car tout regroupement entraine une perte d'information.*

On utilise une règle empirique, dite la règle de Sturge, par laquelle on choisit un nombre de classe de même amplitude proche de : $u = 1 + \frac{10}{3}\log_{10}(n) = 1 + \frac{10\ln(n)}{3\ln(10)} \cong r$

où n est le nombre d'observations et r l'entier le plus proche de u

b • Choix de la longueur des classes : *On définit l'étendue E d'une série statistique par la différence entre la plus grande valeur et la plus petite valeur de X :* $E = x_{max} - x_{min}$

On choisit alors une longueur h telle que $h = \frac{E}{r} \cong a$, *on arrondit h, par excès à l'entier le plus proche qui sera l'amplitude a de chaque classe*

c • Choix des limites des classes : *Il est souhaitable que les limites des classes comportent une décimale de plus que les observations.*

Dans la pratique le choix du nombre de classes et souvent guidé par le bon sens et la pratique. Un nombre de classe variant entre 10 et 20 semble être une bonne chose

D-2 • Organisation des données :

Comme dans le cas discret, le tableau statistique permet de présenter de manière synthétique les observations d'une variable quantitative continue. Par contre, les graphiques changent dans ce cas : la répartition des observations est représentée au moyen d'un histogramme, tandis que leur cumul est maintenant représenté au moyen de la courbe cumulative. Enfin, les caractéristiques numériques qui résument ces variables sont les mêmes que dans le cas discret, mais leur calcul nécessite quelques adaptations.

On utilise alors un tableau statistique analogue à celui vu dans la section précédente, en

disposant maintenant dans la première colonne les classes rangées par ordre croissant.

Les notions d'effectifs, de fréquences (ou pourcentages), d'effectifs cumulés et de fréquences (ou pourcentages) cumulées sont définies de la même façon que dans le cas discret.

Modalités (classes) : $[e_i, e_{i+1}[$	Amplitude : $a_i = e_{i+1} - e_i$	Centres des classes : $c_i = \frac{e_i + e_{i+1}}{2} = e_i + \frac{a_i}{2}$	Effectifs : n_i	Fréquences : $f_i = \frac{n_i}{n}$	Densité de fréquence : $d_i = \frac{f_i}{a_i}$	Effectifs cumulés croissants : $N_i^{\nearrow} = N_i = \sum_{k=1}^{i} n_k$	Effectifs cumulés décroissants : $N_i^{\searrow} = \begin{cases} n, si\ j = 1 \\ n - N_{i-1} = N_{i-1}^{\searrow} - n_{i-1}, si\ i \geq 2 \end{cases}$	Fréquences cumulées croissantes : $F_i^{\nearrow} = F_i = \sum_{k=1}^{i} f_k$	Fréquencescumulées décroissantes : $F_i^{\searrow} = \begin{cases} 1, si\ i = 1 \\ 1 - F_{i-1} = F_{i-1}^{\searrow} - f_{i-1}, si\ i \geq 2 \end{cases}$
$[e_1, e_2[$	a_1	c_1	n_1	f_1	d_1	N_1	$N_1^{\searrow} = n$	F_1	$F_1^{\searrow} = 1$
$[e_2, e_3[$	a_2	c_2	n_2	f_2	d_2	N_2	$N_2^{\searrow}$	F_2	$F_2^{\searrow}$
$[e_3, e_4[$	a_3	c_3	n_3	f_3	d_3	N_3	$N_3^{\searrow}$	F_3	$F_3^{\searrow}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[e_r, e_{r+1}[$	a_r	c_r	n_r	f_r	d_r	$N_r = n$	$N_r^{\searrow}$	$F_r = 1$	$F_r^{\searrow}$
$\sum$			n	1					

☞ ***Amplitude*** : $a_i = e_{i+1} - e_i$ ☞ ***Centres des classes*** : $c_i = \frac{e_i + e_{i+1}}{2} = e_i + \frac{a_i}{2}$

☞ ***Densité de fréquence*** : $d_i = \frac{f_i}{a_i}$ ☞ ***Densité d'effectifs*** : $d_i' = \frac{n_i}{a_i}$

☞ $n_i = Card(\ X \in [e_i, e_{i+1}[\ )$ ☞ $N_i^{\nearrow} = N_i = \sum_{k=1}^{i} n_k = Card(X < e_{i+1}) = Card(X \leq e_i)$

☞ $N_i^{\searrow} = Card(X \geq e_i) = n - Card(X < e_i) = n - Card(X \leq e_{i-1}) = n - N_{i-1} = N_{i-1}^{\searrow} - n_{i-1}$

☞ $f_i = P(X \in [e_i, e_{i+1}[) = \frac{n_i}{n}$ ☞ $F_i^{\nearrow} = F_i = \sum_{k=1}^{i} f_k = P(X < e_i) = P(X \leq e_i) = \frac{N_i}{n}$

☞ $P(a < X \leq b) = P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

☞ $F_i^{\searrow} = P(X \geq e_i) = 1 - P(X < e_{i-1}) = 1 - P(X \leq e_{i-1}) = 1 - F_{i-1} = F_{i-1}^{\searrow} - f_{i-1} = \frac{N_i^{\searrow}}{n}$

$\mathcal{D}$– 3 • *Représentations graphiques* :

Les deux graphiques usuels remplaçant respectivement dans ce cas le diagramme en bâtons et le diagramme cumulatif sont l'histogramme et la courbe cumulative.

a • L'histogramme : ***La représentation graphique des effectifs (resp. fréquences) d'une distribution d'une variable statistique continue s'appelle histogramme.***

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

La population d'une classe est représentée par un rectangle dont la surface est proportionnelle à son effectif (resp. fréquence).

☞ *si les amplitudes des classes sont identiques, la hauteur du rectangle est proportionnelle à l'effectif (resp. Fréquence).*

☞ *Lorsque les classes sont d'amplitudes inégales, il faut procéder à un calcul des effectifs par intervalle élémentaire choisi pour assurer la proportionnalité des airs des rectangles aux effectifs, on prend généralement l'amplitude la plus faible comme amplitude de référence. Si l'on désigne par n_i l'effectif de la classe i, a_i son amplitude et a_0 l'amplitude de référence, l'effectif corrigé n_i^c est donné par* $n_i^c = d_i' \times a_0 = \frac{n_i}{a_i} \times a_0$

Ou en utilisant les fréquences, on obtient la fréquence corrigée $f_i^c = d_i \times a_0 = \frac{f_i}{a_i} \times a_0$

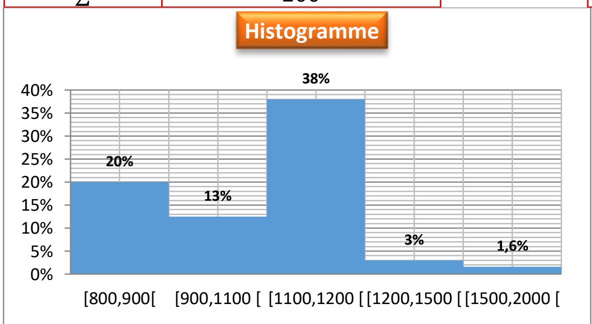
☞ ***Remarque*** : $\sum_{i=1}^{r} n_i^c \neq n$ et $\sum_{i=1}^{r} f_i^c \neq 1$

☞ *On peut tracer le polygone des fréquences (resp. effectifs) en joignant par des segments de droite les milieux des côtés supérieurs des rectangles de l'histogramme. Le polygone des fréquences permet ainsi d'évaluer visuellement le poids de chaque classe représenté par son centre.*

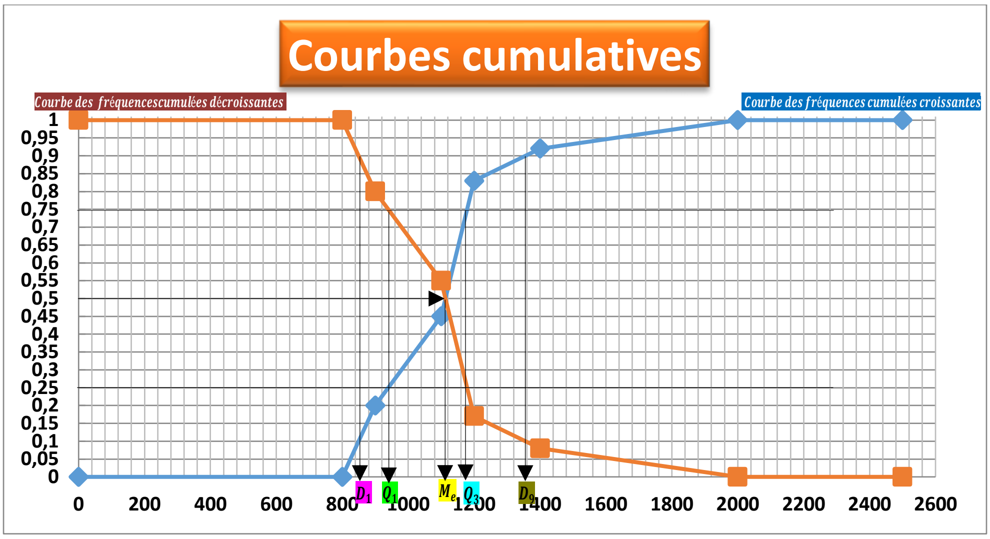
☑ ***Exemple*** :

Le tableau suivant donne la répartition des employés d'une entreprise selon le salaire :

Salaire en D $[e_i, e_{i+1}[$	Nombre d'employés : n_i	a_i	c_i	f_i	$d_i = \frac{f_i}{a_i}$ (%)	$f_i^c = d_i \times a_0$ $a_0 = 100$	$F_i^{\nearrow}$	$F_i^{\searrow}$
[800, 900[	40	100	850	0,2	0,2%	20%	0,2	1
[900, 1100[	50	200	1000	0,25	0,125%	12,5%	0,45	0,8
[1100, 1200[	76	100	1150	0,38	0,38%	38%	0,83	0,55
[1200, 1400[	18	300	1350	0,09	0,03%	3%	0,92	0,17
[1400, 2000[	16	500	1750	0,08	0,016%	1,6%	1	0,08
$\sum$	200			1				



b • La courbe cumulative : *Pour chaque valeur v qui est une borne de classe, on associe un point de coordonnées $(v, F(v))$. On joint les points consécutifs par un segment. On termine en prolongeant en 0 et en 1 aux deux extrêmes.*



D-4 • Principaux paramètres de position (ou de tendance centrale) :

a • Classes modales et modes : *La classe modale est la classe ayant la plus grande densité de fréquence* $\left(d_i = \dfrac{f_i}{a_i}\right)$. *Graphiquement, c'est la classe correspondant au rectangle le plus haut dans l'histogramme.*

☞ *Il peut y avoir une ou plusieurs classes modales, on choisit la classe ayant la plus grande densité. Il existe une formule empirique permettant de calculer le mode :*

Soit $[e_m, e_{m+1}[$ *, la classe modale ayant une amplitude* a_m, *un effectif* n_m *et une fréquence* f_m

notons $[e_{m-1}, e_m[$ *la classe qui la précède ayant pour effectif* n_{m-1} *et une fréquence* f_{m-1}

et $[e_{m-1}, e_m[$ *la classe qui la succède ayant pour effectif* n_{m+1} *et une fréquence* f_{m+1}

ainsi : $$M_o = e_m + a_m\left[\frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})}\right] = e_m + a_m\left[\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})}\right]$$

b • Les Moyennes arithmétique, géométrique, quadratiques et harmoniques :

☑ ***La moyenne arithmétique :*** *La moyenne arithmétique est un résumé numérique et correspond au centre de gravité de la distribution. Elle est exprimée dans la même unité que la variable* $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{r} n_i c_i = \sum_{i=1}^{r} f_i c_i$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

☑ ***La moyenne géométrique*** :

$$\overline{G} = \sqrt[n]{\prod_{i=1}^{r} c_i^{n_i}} = \left[\prod_{i=1}^{r} c_i^{n_i}\right]^{\frac{1}{n}} = \prod_{i=1}^{r} c_i^{f_i} = \exp\left[\sum_{i=1}^{r} f_i \ln(c_i)\right], car \ \ln(\overline{G}) = \sum_{i=1}^{r} f_i \ln(c_i)$$

☑ ***La moyenne quadratique*** : $\overline{Q} = \sqrt{\sum_{i=1}^{r} f_i c_i^2}$ *ou encore* $\overline{Q}^2 = \sum_{i=1}^{r} f_i c_i^2$

☑ ***La moyenne harmonique*** : $\overline{H} = \left[\sum_{i=1}^{r} \frac{f_i}{c_i}\right]^{-1}$ *ou encore* $\frac{1}{\overline{H}} = \sum_{i=1}^{r} \frac{f_i}{c_i}$

La relation suivante sera toujours vérifiée : $\overline{H} \leq \overline{G} \leq \overline{X} \leq \overline{Q}$

c • *Quantiles et applications* :

☑ ***Interpolation linéaire*** : ***On note*** $F(x) = P(X \leq x). \forall x \in [a, b[\ \subset [x_{min}, x_{max}]$

on a: $F(x) = F(a) + \left[(F(b) - F(a))\left(\frac{x-a}{b-a}\right)\right]$

• ***Exemple*** :

Salaire en D $[e_i, e_{i+1}[$	*Nombre d'employés* : n_i	a_i	c_i	f_i	$d_i = \frac{f_i}{a_i}$ (%)	$f_i^c = d_i \times a_0$ $a_0 = 100$	$F_i^{\nearrow}$	$F_i^{\searrow}$
$[800, 900[$	40	100	850	0,2	0,2%	20%	0,2	1
$[900, 1100[$	50	200	1000	0,25	0,125%	12,5%	0,45	0,8
$[1100, 1200[$	76	100	1150	0,38	0,38%	38%	0,83	0,55
$[1200, 1400[$	18	300	1350	0,09	0,03%	3%	0,92	0,17
$[1400, 2000[$	16	500	1750	0,08	0,016%	1,6%	1	0,08
$\sum$	200			1				

On se propose de calculer la proportion d'employés touchant au plus 1120 D ?

Autrement dit $P(X \leq 1120) = F(1120)$

Or $1120 \in \left[\underbrace{1100}_{a}, \underbrace{1200}_{b}\right[$, ***avec*** $F(1100) = P(X \leq 1100) = 0,45$ ***et*** $F(1200) = P(X \leq 1200) = 0,83$

Par interpolation linéaire on obtient :

$$F(1120) = F(1100) + \left[(F(1200) - F(1100))\left(\frac{1120-1100}{1200-1120}\right)\right] = 0,45 + \left[(0,83 - 0,45) \times \left(\frac{20}{80}\right)\right] = 0,545$$

☑ ***La médiane (ou le second quartile)*** $M_e = Q_2$:

La médiane est le quantile d'ordre $\frac{1}{2}$. ***Elle partage donc la série des observations en deux ensembles d'effectifs égaux***: $(F(M_e) = P(X \leq M_e) = P(X > M_e) = 50\%)$

• *On cherche l'indice i tel que* $F_{i-1} \leq 0,5$ (*resp.* $N_{i-1} \leq \frac{n}{2}$) *et* $F_i > 0,5$(*resp.* $N_i > \frac{n}{2}$)

Par interpolation linéaire on obtient : $M_e = e_i + a_i\left(\frac{0,5 - F_{i-1}}{F_i - F_{i-1}}\right) = e_i + a_i\left(\frac{n/2 - N_{i-1}}{N_i - N_{i-1}}\right)$

avec e_i *la borne inférieure de la classe correspondante à* F_i *et* a_i *l'amplitude de cette même classe.*

• **Graphiquement** *la médiane n'est autre que l'abscisse du point d'intersection entre la courbe des fréquences cumulées croissantes et celle des fréquences cumulées décroissantes. Autrement dit l'abscisse du point de la courbe des fréquences cumulées croissantes (resp. décroissantes) dont l'ordonnée est égale à* $0,5$

☑ **Le premier quartile (Q_1) et le troisième quartile (Q_3):**

Le premier quartile est le quantile d'ordre $1/4$, *le troisième quartile celui d'ordre* $3/4$

On voit donc que $25\ \%$ *des observations sont inférieures ou égales au premier quartile*: ($F(Q_1) = P(X \leq Q_1) = 25\%$), *tandis que* $75\ \%$ *lui sont supérieures* : ($P(X > Q_1) = 75\%$)

Pour le troisième quartile, les proportions s'inversent : $75\ \%$ *des valeurs lui sont inférieures ou égales* : ($F(Q_3) = P(X \leq Q_3) = 75\%$), *tandis que* $25\ \%$ *lui sont supérieures* : ($P(X > Q_3) = 25\%$)

• **Le premier quartile Q_1** : *On cherche l'indice i tel que* $F_{i-1} \leq 0,25$ (*resp.* $N_{i-1} \leq n/4$) *et* $F_i > 0,25$(*resp.* $N_i > n/4$)

Par interpolation linéaire on obtient : $Q_1 = e_i + a_i\left(\frac{0,25 - F_{i-1}}{F_i - F_{i-1}}\right) = e_i + a_i\left(\frac{n/4 - N_{i-1}}{N_i - N_{i-1}}\right)$

avec e_i *la borne inférieure de la classe correspondante à* F_i *et* a_i *l'amplitude de cette même classe.*

• **Le troisième quartile Q_3** : *On cherche l'indice i tel que* $F_{i-1} \leq 0,75$ (*resp.* $N_{i-1} \leq 3n/4$) *et* $F_i > 0,75$(*resp.* $N_i > 3n/4$)

Par interpolation linéaire on obtient : $Q_3 = e_i + a_i\left(\frac{0,75 - F_{i-1}}{F_i - F_{i-1}}\right) = e_i + a_i\left(\frac{3n/4 - N_{i-1}}{N_i - N_{i-1}}\right)$

avec e_i *la borne inférieure de la classe correspondante à* F_i *et* a_i *l'amplitude de cette même classe.*

• **Détermination graphique de Q_1**: *le premier quartile Q_1 est l'abscisse du du point de la courbe des fréquences cumulées croissantes dont l'ordonnée est égale à $0,25$ ou encore l'abscisse du point de la courbe des fréquences cumulées décroissantes dont l'ordonnée est égale à $0,75$*

• **Détermination graphique de Q_3**: *le troisième quartile Q_3 est l'abscisse du du point de la courbe des fréquences cumulées croissantes dont l'ordonnée est égale à $0,75$ ou encore l'abscisse du point de la courbe des fréquences cumulées décroissantes dont l'ordonnée est égale à $0,25$*

☑ **Les autres quantiles**

Les déciles et les centiles sont également d'usage relativement courant. Il existe 9 déciles qui partagent l'ensemble des observations en 10 parties d'égale importance (chacune contient 10 % des observations) $(F(D_1) = P(X \leq D_1) = 10\%\ et\ P(X > D_1) = 90\%)$, $(F(D_2) = P(X \leq D_2) = 20\%\ et\ P(X > D_2) = 80\%)$, …, $(F(D_9) = P(X \leq D_9) = 90\%\ et\ P(X > D_9) = 10\%)$ *et 99 centiles qui la partagent de même en 100 parties d'effectifs égaux :*

$(F(\alpha_1) = P(X \leq \alpha_1) = 1\%\ et\ P(X > \alpha_1) = 99\%)$, $(F(\alpha_2) = P(X \leq \alpha_2) = 2\%\ et\ P(X > \alpha_2) = 98\%)$, …, $(F(\alpha_{99}) = P(X \leq \alpha_{99}) = 99\%\ et\ P(X > \alpha_{99}) = 1\%)$

• **Le premier décile D_1** : *On cherche l'indice i tel que* $F_{i-1} \leq 0,1$ $(resp. N_{i-1} \leq n/10)$ *et* $F_i > 0,1 (resp. N_i > n/10)$. *Par interpolation linéaire on obtient :*

$$D_1 = e_i + a_i\left(\frac{0,1 - F_{i-1}}{F_i - F_{i-1}}\right) = e_i + a_i\left(\frac{n/10 - N_{i-1}}{N_i - N_{i-1}}\right)$$

avec e_i la borne inférieure de la classe correspondante à F_i et a_i l'amplitude de cette même classe.

• **Le neuvième décile D_9** : *On cherche l'indice i tel que* $F_{i-1} \leq 0,9$ $(resp. N_{i-1} \leq \frac{9n}{10})$ *et* $F_i > 0,9 (resp. N_i > 9n/10)$ *Par interpolation linéaire on obtient :*

$$D_9 = e_i + a_i\left(\frac{0,9 - F_{i-1}}{F_i - F_{i-1}}\right) = e_i + a_i\left(\frac{9n/10 - N_{i-1}}{N_i - N_{i-1}}\right)$$

avec e_i la borne inférieure de la classe correspondante à F_i et a_i l'amplitude de cette même classe.

• **Détermination graphique de D_1**: *le premier décile D_1 est l'abscisse du du point de la courbe des fréquences cumulées croissantes dont l'ordonnée est égale à $0,1$*

ou encore l'abscisse du point de la courbe des fréquences cumulées ***décroissantes*** *dont l'ordonnée est égale à* **0,9**

• ***Détermination graphique de D_9*** : *le neuvième décile D_9 est l'abscisse du du point de la courbe des fréquences cumulées* ***croissantes*** *dont l'ordonnée est égale à* **0,9** *ou encore l'abscisse du point de la courbe des fréquences cumulées* ***décroissantes*** *dont l'ordonnée est égale à* **0,1**

D– 5 • Paramètres de dispersion :

a • L'étendue : *On appelle étendue la différence entre la plus grande valeur et la plus petite valeur prise par la variable* : $E = e_{max} - e_{min}$

b • L'étendue interquartile : *Écart entre le troisième et le premier quartile, il contient au moins 50% des observations centrales de la distribution, il n'est pas influencé par les valeurs extrêmes* : $EIQ = Q_3 - Q_1$

c • Écart absolu moyen par rapport à la médiane : $e_{M_e} = \sum_{i=1}^{r} f_i |c_i - M_e|$

d • Écart absolu moyen par rapport à la moyenne : $e_{\overline{X}} = \sum_{i=1}^{r} f_i |c_i - \overline{X}|$

e • Écart-type ou écart quadratique moyen : *On appelle variance ou fluctuation, arithmétique des carrés des écarts des résultats observés à leur moyenne* :

$S_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_i (c_i - \overline{X})^2 = \sum_{i=1}^{r} f_i (c_i - \overline{X})^2$ *L'écart-type mesure la dispersion des données autour de la moyenne* : $S_x = \sqrt{\sum_{i=1}^{r} f_i (c_i - \overline{X})^2}$

☑ ***Propriétés*** :

☞ $S_x^2 \geq 0$ *et* $S_x \geq 0$ ☞ *L'écart-type est exprimé dans la même unité que la variable.*

☞ *Plus l'écart-type est petit, plus les données individuelles sont regroupées autour de la moyenne. Plus il est grand, plus les données individuelles sont dispersées autour de la moyenne.*

☞ **Théorème de König-Huygens :** $S_x^2 = \overline{X^2} - \overline{X}^2 = \overline{Q}^2 - \overline{X}^2 = \sum_{i=1}^{r} f_i c_i^2 - \overline{X}^2$

☞ $S_{ax+b}^2 = a^2 S_x^2$

☞ $S_{ax+b} = |a| S_x$

☞ **Variance d'échantillonnage :** $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{r} n_i (c_i - \overline{X})^2 = \dfrac{n}{n-1} S_x^2$

Lorsque la série est un échantillon issu d'une population et que l'on s'intéresse aux caractéristiques de cette population via l'échantillon (inférence), on utilise plutôt S^2 qui est un meilleur estimateur de la variance théorique de la population $(V(X) = \sigma^2)$.

Dès lors que la taille n de la série est assez grande, $S^2 \approx S_x^2$

☞ *Si* $S_x^2 = S_x = 0$ *, alors toutes les données sont égales et égales à la moyenne:* $\forall i, x_i = \overline{X}$

☞ **Identité :** $\dfrac{1}{n}\sum_{i=1}^{r}(c_i - \overline{X})^2 = \dfrac{1}{n}\sum_{i=1}^{r}(c_i - a)^2 - (\overline{X} - a)^2$

f • Les moments empiriques :

☑ **Moments empiriques par rapport à a d'ordre k :**

Soient $x_1, x_2, \dots, x_r$ une série statistique et $a \in \mathbb{R}$, le moments empiriques par rapport à a d'ordre k (où s supposé entier positif) est définie par : $\dfrac{1}{n}\sum_{i=1}^{r} n_i (c_i - a)^k = \sum_{i=1}^{r} f_i (c_i - a)^k$

☑ **Moments empiriques par rapport à l'origine (ou non centrés) d'ordre k :**

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{r} n_i c_i^k = \sum_{i=1}^{r} f_i c_i^k \text{ , avec : } \overline{m}_1 = \overline{X}$$

☑ **Moments empiriques centrés d'ordre k :**

$$\overline{\mu}_k = \frac{1}{n}\sum_{i=1}^{r} n_i (c_i - \overline{X})^k = \sum_{i=1}^{r} f_i (c_i - \overline{X})^k \text{ , avec : } \overline{\mu}_1 = 0 \text{ et } \overline{\mu}_2 = S_x^2$$

g • L'inégalité de Bienaymé-Tchebychev :

Pour toute population de moyenne $\overline{X}$ et d'écart-type S_x : $P(\overline{X} - \lambda s \le X \le \overline{X} + \lambda s) \ge 1 - \dfrac{1}{\lambda^2}$ *, pour tout* $\lambda > 1$

$P(\overline{X} - 2s \le X \le \overline{X} + 2s) \ge 3/4$ et $P(\overline{X} - 3s \le X \le \overline{X} + 3s) \ge 8/9$

D– 6 • Paramètres de forme :

En plus de l'étude de la tendance et de la dispersion, il est intéressant d'étudier la forme de la courbe d'une distribution, mise en évidence par la représentation graphique. Les paramètres de forme caractérisent la dissymétrie et l'aplatissement.

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

a • Asymétrie d'une distribution :

☑ ***Coefficient d'asymétrie de Pearson*** ***est basé sur une comparaison de la et du mode, et est standardisé par l'écart-type*** : $\mathcal{P}_1 = \dfrac{\overline{X} - M_e}{S_x}$

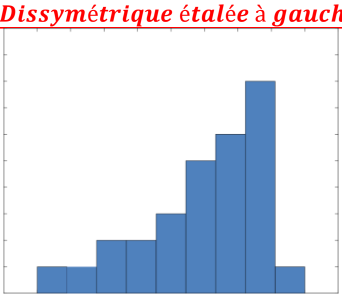
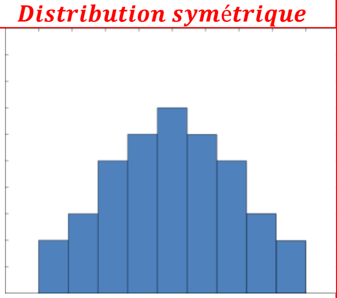
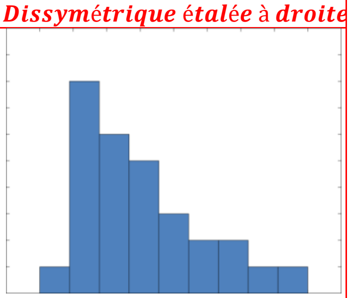
☑ ***Coefficient de Yule & Kendall*** : ***est basé sur les positions des 3 quartiles (1^er^ quartile, médiane et 3^ème^ quartile), et est normalisé par la distance interquartile***:

$$\mathcal{Y}_k = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

☑ ***Coefficient d'asymétrie de Fisher*** :

• ***Pour les moments centrés d'ordre impair, on constate que la somme des écarts positifs compense celle des écarts négatifs donc ces moments sont nuls.***
La distribution est alors symétrique.

• ***Lorsque la distribution est dissymétrique à gauche les*** $(c_i - \overline{X})$ ***négatifs sont plus nombreux, mais petits en valeur absolue, tandis que les*** $(c_i - \overline{X})$ ***positifs sont moins nombreux mais plus grand.***
Un autre moyen pour mesurer l'asymétrie d'une distribution sera le coefficient d'asymétrie de Fisher : $\gamma_1 = \dfrac{\overline{\mu}_3}{S_x^3} = \dfrac{\sum_{i=1}^{r} f_i (c_i - \overline{X})^3}{\left(\sqrt{\sum_{i=1}^{r} f_i (c_i - \overline{X})^2}\right)^3}$

Dissymétrique étalée à gauche	Distribution symétrique	Dissymétrique étalée à droite	
			
$\overline{X} < M_e < M_o$	$\overline{X} = M_e = M_o$	$\overline{X} > M_e > M_o$	Comparaison Mode-Moyenne- Médiane
$\mathcal{P}_1 = \dfrac{\overline{X} - M_e}{S_x} < 0$	$\mathcal{P}_1 = \dfrac{\overline{X} - M_e}{S_x} = 0$	$\mathcal{P}_1 = \dfrac{\overline{X} - M_e}{S_x} > 0$	Coefficient d'asymétrie de Pearson
$\mathcal{Y}_k = \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} < 0$	$\mathcal{Y}_k = \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 0$	$\mathcal{Y}_k = \dfrac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} > 0$	Coefficient de Yule & Kendall
$\gamma_1 = \dfrac{\overline{\mu}_3}{S_x^3} < 0$	$\gamma_1 = \dfrac{\overline{\mu}_3}{S_x^3} = 0$	$\gamma_1 = \dfrac{\overline{\mu}_3}{S_x^3} > 0$	Coefficient d'asymétrie de Fisher

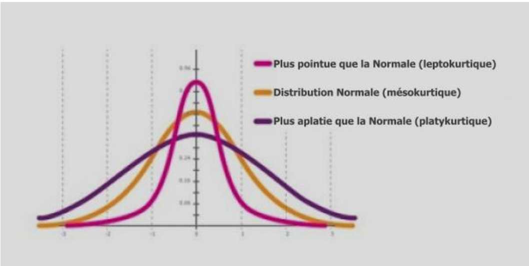
Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

b • *Paramètre d'aplatissement (kurtosis) :*

L'aplatissement est mesuré par le coefficient d'aplatissement de Pearson : $\mathcal{P}_2 = \dfrac{\overline{\mu}_4}{S_x^4}$

ou le coefficient d'aplatissement de Fisher : $\gamma_2 = \mathcal{P}_2 - 3 = \dfrac{\overline{\mu}_4}{S_x^4} - 3$



Plus pointue que la Normale (*leptokurtique*)	*Distribution Normale* (*mésokurtique*)	*Plus aplatie que la Normale* (*platykurtique*)
$\gamma_2 = \dfrac{\overline{\mu}_4}{S_x^4} - 3 > 0$	$\gamma_2 = \dfrac{\overline{\mu}_4}{S_x^4} - 3 = 0$	$\gamma_2 = \dfrac{\overline{\mu}_4}{S_x^4} - 3 < 0$

D-7 • Paramètres de concentration (mesures de l'inégalité) :

a • Introduction :

Des indicateurs particuliers ont été développés pour mesurer les inégalités des revenus ou les inégalités de patrimoine. On considère qu'une société est parfaitement égalitaire si tous les individus reçoivent le même revenu. La situation théorique la plus inégalitaire est la situation où un individu perçoit la totalité des revenus, et les autre individus n'ont aucun revenu.

b • Définitions et détermination algébrique :

☑ ***Valeurs globales :***

On appelle valeurs globales d'une distribution statistique continue les produits $(n_i c_i)_{1 \le i \le r}$
(***Exp: les salaires par classe***) ***où*** n_i ***: les effectifs par classe ;*** c_i ***: les centres de classe***

☑ ***Valeurs globales totales :***

$$VGT = \sum_{i=1}^{r} n_i c_i = n\overline{X}$$ (***Exp: la masse totale des salaires***)

☑ ***Valeurs globales relatives : On appelle valeurs globales relatives***
(***VGR***) ***de la distribution statistique les rapports définies par :*** $VGR_i = q_i = \dfrac{n_i c_i}{\sum_{i=1}^{r} n_i c_i}$,
qu'on l'exprimesouvent en pourcentage.

(***Exp: les salaires par classe en pourcentage de revenu total***)

☑ ***Valeurs globales relatives cumulées croissantes :***

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

Ce sont les quantités $Q_i = \sum_{k=1}^{i} q_k$. *On a toujours* : $0 \leq Q_i \leq F_i \leq 1$

(*Exp*: *le pourcentage de la masse totale des salaires par les classes*)

☑ **Médiale** : *On cherche l'indice i tel que* $Q_{i-1} \leq 0,5$ *et* $Q_i > 0,5$. *Par interpolation linéaire on obtient*: $Ml_e = e_i + a_i\left(\frac{0,5 - Q_{i-1}}{Q_i - Q_{i-1}}\right)$ *avec* e_i *la borne inférieure de la classe correspondante à* Q_i *et* a_i *l'amplitude de cette même classe.*

On a toujours: $Ml_e \geq M_e$

☑ **Exemple** : *Pour mieux comprendre la notion de concentration, considérons la distribution groupée des salaires d'une entreprise de* 250 *personnes* :

Classe des salaires mensuels	c_i	n_i	Valeurs globales $n_i c_i$	f_i	F_i	Valeurs globales relatives $VGR_i = q_i = \frac{n_i c_i}{VGT}$	valeurs globales relatives cumulées croissantes : Q_i
[1200 ; 1600[	1400	25	35000	0,1	0,1	0,056	0,056
[1600 ; 2000[	1800	32	57600	0,128	0,228	0,092	0,147
[2000 ; 2400[	2200	54	118800	0,216	0,444	0,189	0,336
[2400 ; 2800[	2600	60	156000	0,24	0,684	0,248	0,584
[2800 ; 3200[	3000	44	132000	0,176	0,86	0,210	0,794
[3200 ; 3600[	3400	19	64600	0,076	0,936	0,103	0,896
[3600 ; 4000[	3800	10	38000	0,04	0,976	0,060	0,957
[4000 ; 4600[	4300	4	17200	0,016	0,992	0,027	0,984
[4600 ; 5400[	5000	2	10000	0,008	1	0,016	1
∑		250	$VGT = 629200$	1		1	

- *La masse totale des salaires* : $VGT = 629200$
- *Prenons par exemple la classe des salaires* : [3600 ; 4000[

☞ $f_7 = 4\%$. *Interprétation* : 4% *des Salariés* (*qui sont au nombre de* 10) *touchent entre* 3600 *et* 4000

☞ $q_7 = 6\%$. *Interprétation* : *Les salariés qui ont un salaire dont le montant est compris entre* 3600 *et* 4000 € *représentent ensemble, globalement, un pourcentage de la masse salariale totale égal à* : 6 %.

☞ $F_7 = 97,6\%$. *Interprétation* : 97,6% *des salariés touchent moins de* 4000

☞ $Q_7 = 95,7\%$. *Interprétation* : *Les salariés qui ont un salaire dont le montant est inférieur à* 4000 € *représentent ensemble, globalement, un pourcentage de la masse salariale totale égale à* : 89,63 %

☞ $\binom{F_7}{Q_7} = \binom{97,6\%}{95,7\%}$. *Interprétation* : 97,6% *des salariés représentent* 95,7% *de l'ensembledes salaires*

☞ $M_e = e_i + a_i\left(\frac{0,5 - F_{i-1}}{F_i - F_{i-1}}\right) = 2400 + \left[400 \times \left(\frac{0,5 - 0,444}{0,684 - 0,444}\right)\right] = 2493,333$

Interprétation : 50% *des salariés rouchent moins de* 2493,333

☞ $Ml_e = e_i + a_i\left(\frac{0,5 - Q_{i-1}}{Q_i - Q_{i-1}}\right) = 2400 + \left[400 \times \left(\frac{0,5 - 0,336}{0,584 - 0,336}\right)\right] = 2664,516$

Interprétation : *Les salariés qui ont un salaire inférieur à* 2664 € *représentent ensemble la moitié de la masse salariale globale. Évidemment, ceux qui gagnent plus de*

2664 € représentent ensemble l'autre moitié de la masse salariale globale

c • Courbe de concentration (ou de Lorenz) :

La courbe de Lorenz, est une représentation graphique permettant de visualiser la distribution d'une variable (actif, patrimoine, revenu, etc.) au sein d'une population.

Plus précisément, la courbe de Lorenz relie les points (F_i, Q_i) pour $i = 1, \dots, r$.

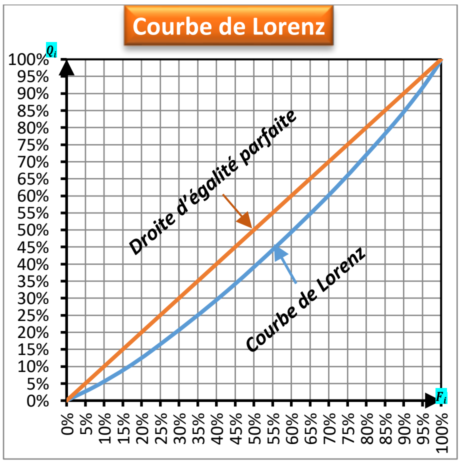
En abscisse, on a donc une proportion d'individus classés par ordre de revenu, et en ordonnée la proportion du revenu total reçu par ces individus.

Cette courbe permet de visualiser la répartition des revenus, des patrimoines, des terres agricoles … et donc permet de comprendre l'économie, voire la politique …

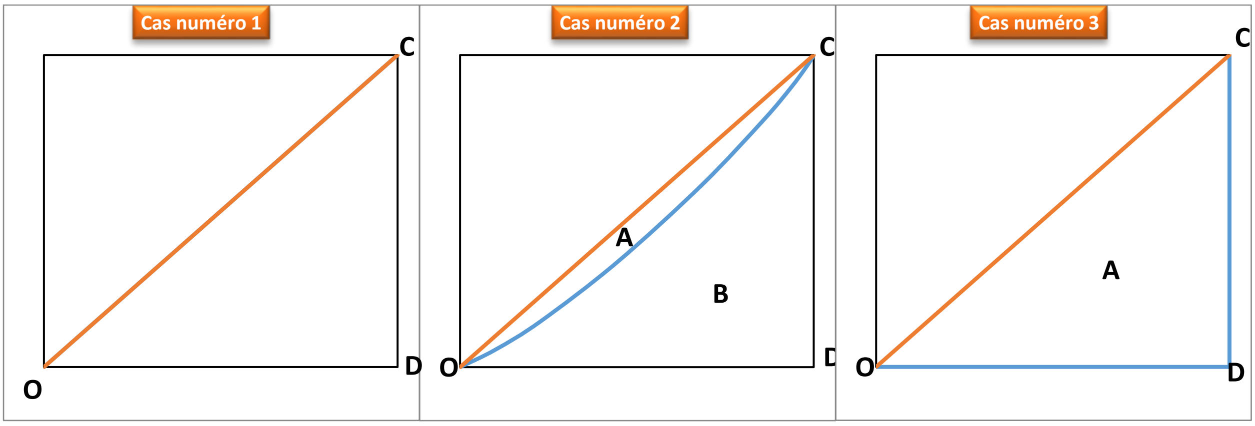
La courbe de Lorenz s'inscrit donc dans un carré. Pour apprécier l'inégalité, on doit comparer cette courbe avec la droite d'égalité parfaite qui correspond à la diagonale .

De façon générale, plus une courbe de Lorenz se rapproche de la droite d'égalité parfaite et plus la répartition de la masse considérée au sein de la population est égalitaire.

En effet, dans ce cas, la masse (des salaires, de la richesse, du revenu, etc.) est peu concentrée sur quelques uns. Inversement, plus une courbe de Lorenz s'éloigne de la droite d'égalité parfaite et plus la répartition de la masse considérée au sein de la population est inégalitaire car la masse (des salaires, de la richesse, du revenu) est alors concentrée sur un petit nombre d'unités statistiques.



3 cas typiques, dont les deux cas limites sont représentés par les graphiques ci-dessous :



Cas numéro 1 : Égalité parfaite	*Cas numéro 2 : Inégalité modérée*	*Cas numéro 3 : Inégalité totale*
La courbe de Lorenz se confond avec la droite (OC) d'égalité parfaite. Chaque individu de la population possède la même part de la masse totale	*La courbe de Lorenz partage le triangle OCD en 2 surfaces. Plus la surface A augmente aux dépends de la surface B et plus l'inégalité augmente*	*La courbe de Lorenz Est donnée OCD. La surface A occupe tout le triangle OCD et la surface B a disparue. C'est le cas rhétorique où un seul individu possède 100% de la masse totale et les autres rien*

d • Indice de concentration (ou Indice de GINI) :

☑ ***Définition : Le coefficient de GINI est une mesure de l'inégalité associée à la courbe de Lorenz. Il est donné par la formule :***

$$I_G = \frac{A}{A+B} = \frac{\text{aire de la surface de concentration}}{\text{aire du triangle } ODC}$$

Où A représente la surface comprise entre la courbe de Lorenz et la droite d'égalité parfaite et B représente la surface située sous la droite d'égalité parfaite moins la surface A. Le meilleur indicateur visuel de cette formule est le cas numéro 2 du tableau ci-avant. Le coefficient de Giniest compris entre zéro et 1. En cas d'égalité parfaite, il est égal à zéro (car $A = 0$). En cas d'inégalité totale il est égal à 1, car $B = 0$.

Par conséquent, à mesure que I_G augmente de zéro à 1, l'inégalité de la répartition augmente. Le coefficient de GINI permet ainsi de faire de nombreuses comparaisons.

Sachant que la courbe de Lorenz est inscrite dans un carré de 1×1, on voit que la surface $A + B$ est égale à la moitié de cette surface. On a donc : $A + B = \frac{1}{2} \Rightarrow I_G = \frac{A}{1/2} = 2A$

comme : $A + B = \frac{1}{2} \Rightarrow A = \frac{1}{2} - B$: $I_G = \frac{A}{A+B} = 2A = 2\left(\frac{1}{2} - B\right) = 1 - 2B$

☑ ***Détermination*:**

• ***Méthode graphique*** : ***On trace la courbe de Lorenz sur un papier millimétré***

$(1mm = 1\%)$ ***puis on évalue l'aire A de concentration exprimé en*** cm^2. ***Alors*** : $I_G = 2A$

• ***Méthode des triangles*** : $I_G = \sum_{i=1}^{r-1}[(F_iQ_{i+1}) - (F_{i+1}Q_i)] = \sum_{i=1}^{r-1}\begin{vmatrix} F_i & Q_i \\ F_{i+1} & Q_{i+1} \end{vmatrix}$

Exemple :

F_i	Q_i	$(F_iQ_{i+1}) - (F_{i+1}Q_i)$
0,1	0,056	$(F_1Q_2) - (F_2Q_1) = 0,1932\%$
0,228	0,147	$(F_2Q_3) - (F_3Q_2) = 1,134\%$
0,444	0,336	$(F_3Q_4) - (F_4Q_3) = 2,9472\%$
0,684	0,584	$(F_4Q_5) - (F_5Q_4) = 4,0856\%$
0,86	0,794	$(F_5Q_6) - (F_6Q_5) = 2,7376\%$
0,936	0,896	$(F_6Q_7) - (F_7Q_6) = 2,1256\%$
0,976	0,957	$(F_7Q_8) - (F_8Q_7) = 1,104\%$
0,992	0,984	$(F_8Q_9) - (F_9Q_8) = 0,8\%$
1	1	$I_G = 15,1272\%$

$$I_G = \underbrace{\begin{vmatrix} 0,1 & 0,056 \\ 0,228 & 0,147 \end{vmatrix}}_{0,1932\%} + \underbrace{\begin{vmatrix} 0,228 & 0,147 \\ 0,444 & 0,336 \end{vmatrix}}_{1,134\%} + \underbrace{\begin{vmatrix} 0,444 & 0,336 \\ 0,684 & 0,584 \end{vmatrix}}_{2,9472\%}$$

$$+ \underbrace{\begin{vmatrix} 0,684 & 0,584 \\ 0,86 & 0,794 \end{vmatrix}}_{4,0856\%} + \underbrace{\begin{vmatrix} 0,86 & 0,794 \\ 0,936 & 0,896 \end{vmatrix}}_{2,7376\%} + \underbrace{\begin{vmatrix} 0,936 & 0,896 \\ 0,976 & 0,957 \end{vmatrix}}_{2,1256\%}$$

$$+ \underbrace{\begin{vmatrix} 0,976 & 0,957 \\ 0,992 & 0,984 \end{vmatrix}}_{1,104\%} + \underbrace{\begin{vmatrix} 0,992 & 0,984 \\ 1 & 1 \end{vmatrix}}_{0,8\%} = 15,1272\%$$

• ***Méthode des trapèzes*** : $I_G = 1 - \left[f_1Q_1 + \sum_{i=1}^{r} f_i(Q_{i-1} + Q_i)\right]$

Ou encore : $I_G = 1 - \left[\sum_{i=1}^{r} f_i(Q_{i-1} + Q_i)\right]$; ***avec*** $Q_0 = 0$

Exemple :

f_i	Q_i	$Q_{i-1} + Q_i$	$f_i(Q_{i-1} + Q_i)$
0,1	0,056	$Q_0 + Q_1 = Q_1 = 0,056$	0,56%
0,128	0,147	$Q_1 + Q_2 = 0,203$	2,5984%
0,216	0,336	$Q_2 + Q_3 = 0,483$	10,4328%
0,24	0,584	$Q_3 + Q_4 = 0,92$	22,08%
0,176	0,794	$Q_4 + Q_5 = 1,378$	24,2528%
0,076	0,896	$Q_5 + Q_6 = 1,69$	12,844%
0,04	0,957	$Q_6 + Q_7 = 1,853$	7,412%
0,016	0,984	$Q_7 + Q_8 = 1,941$	3,1056%
0,008	1	$Q_8 + Q_9 = 1,984$	1,5872%
$\sum$			84,8728%
			$I_G = 100\% - 84,8728\% = 15,1272\%$

• ***Méthode de la différence moyenne*** : $I_G = \frac{1}{2n(n-1)\bar{X}}\sum_{i=1}^{r}\sum_{j=1}^{r} n_in_j|c_i - c_j|$

• ***Méthode par intégration de la fonction de concentration*** (g) :

$$I_G = 1 - 2\int_0^1 g(x)dx = 2\int_0^1 (x - g(x))dx$$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Ingénieurs
Lundi 17 Octobre 2011

Exercice 5 (4 points = 1+1+1+1) :

ÉNONCÉ

Considérons la distribution statistique suivante :

Classes	$20-30$	$30-40$	$40-z$	$z-70$	$70-100$	100 *et plus*
Effectifs : n_i	100	140	125	200	180	55

1) *Sachant que la médiane de cette distribution est égale à* $56,8$ *calculer* z
2) *Supposons que la moyenne arithmétique vaut* $60,5$. *On notera* y *le centre de la classe dont la borne inférieure est égale à* 100 :
   a) *Calculer* y *en utilisant la valeur de* z *calculée à la question* 1)
   b) *En déduire la valeur de la borne supérieure de la classe de borne inférieure égale à* 100

Corrigé

1) *Détermination de la valeur de* z :

La moitié de l'effectif total est égale à $\frac{n}{2} = 400$. *Ce nombre est compris entre*

$N_{i-1} = 365$ *et* $N_i = 565$, *ce qui signifie que la médiane est comprise entre* z *et* 70.

Classes: $[e_i, e_{i+1}[$	*Amplitude* : a_i	*Effectifs* : n_i	*Effectifs cumulés* $\nearrow$: N_i
$[20,30[$	10	100	100
$[30,40[$	10	140	240
$[40,z[$	$z-40$	125	365
$[z,70[$	$70-z$	200	565
$[70,100[$	30	180	745
$[100,t[$	$t-100$	55	800
$\sum$		800	

Or, la médiane est égale à $56,8$ *ce qui implique* :

$$M_e = z + (70-z)\left(\frac{\frac{n}{2} - N_{i-1}}{N_i - N_{i-1}}\right) \Leftrightarrow 56,8 = z + (70-z)\left(\frac{400-365}{565-365}\right) \Leftrightarrow 56,8 = z + \frac{7(70-z)}{40}$$

$$\Leftrightarrow \frac{33z}{40} = 56,8 - \frac{49}{4} \Leftrightarrow z = \frac{891}{20} \times \frac{40}{33} \; d'où : \boxed{z = 54}$$

2)

a) $\textit{Calcul de y centre de la classe de borne supérieure égale à } 100$:

$\textit{Calcul des centres de classes } c_i \textit{ et des } c_i n_i$:

$Classes: [e_i, e_{i+1}[$	$Amplitude : a_i$	$Centres\ de\ classes\ c_i$	n_i	$c_i n_i$
$[20, 30[$	10	25	100	2500
$[30, 40[$	10	35	140	4900
$[40, 54[$	14	47	125	5875
$[54, 70[$	16	62	200	12400
$[70, 100[$	30	85	180	15300
$[100, t[$	$t - 100$	y	55	$55y$
$\sum$			800	$40975 + 55y$

$\textit{Moyenne arithmétique}$: $\bar{X} = \frac{1}{n}\sum_{i=}^{6} c_i n_i$

$$donc, 60,5 = \frac{1}{800}(40975 + 55y) \Leftrightarrow \frac{55}{800}y = 60,5 - \frac{40975}{800} \Leftrightarrow y = \frac{297}{32} \times \frac{800}{55}, d'où: \boxed{y = 135}$$

b) $\textit{Calcul de la borne sup de la classe de borne inférieure ou égale à } 100$:

$\textit{Centre de la classe}$: $[100, t[\textit{ est } y = 135$

$\textit{Par ailleurs}$: $\frac{100 + t}{2} = 135 \Rightarrow \boxed{t = 170}$

Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Ingénieurs
Samedi 21 Septembre 2013

Exercice 6 *(4 points = 1 + 1 + 2) :*

ÉNONCÉ

$\textit{La répartition des salaires d'une entreprise est donnée par le tableau suivant}$:

$Salaires$	$Nombre\ de\ salariés$
[0-1000[	0
[1000-1400[	100
[1400-1800[	150
[1800-2200[	40
[2200-3000[	10

1) $\textit{Calculer le salaire moyen ainsi que l'écart-type de cette distribution}$
2) $\textit{Calculer la médiale et la médiane. Analyser}$

3) *Définir et tracer la courbe de Lorenz*

Corrigé

1)

- *Amplitude* : a_i
- *Nombre de salariés* : n_i
- *Centres des classes* : c_i
- *Valeurs globales relatives* : $q_i = \frac{n_i c_i}{\sum n_i c_i}$
- *Fréquences relatives* : $f_i = \frac{n_i}{\sum n_i}$
- *Fréquences cumulées croissantes* : F_i
- *Valeurs globales cumulées croissantes* : Q_i

Salaires	a_i	n_i	c_i	$n_i c_i$	$n_i c_i^2$	q_i	f_i	F_i	Q_i
[0-1000[	1000	0	500	0	0	0	0	0	0
[1000-1400[	400	100	1200	120000	144000000	0,26	0,33	0,33	0,26
[1400-1800[	400	150	1600	240000	384000000	0,52	0,50	0,83	0,77
[1800-2200[	400	40	2000	80000	160000000	0,17	0,13	0,97	0,94
[2200-3000[	800	10	2600	26000	67600000	0,06	0,03	1	1
$\sum$		300		466000	755600000	1	1		

- *Le salaire moyen* : $\bar{X} = \frac{1}{n}\sum_{i=1}^{N} n_i c_i = \frac{466000}{300} \Rightarrow \boxed{\bar{X} = 1553,33}$

- $S_X^2 = \frac{1}{n}\sum_{i=1}^{n} n_i c_i^2 - \bar{X}^2 = \frac{755600000}{300} - (1553,33)^2 \Rightarrow \boxed{S_X^2 = 105832,58} \Rightarrow \boxed{S_X = 325,32}$

2)

- $Q_2 \leq 0,5$ *et* $Q_3 > 0,5$, *avec* $Q_2 = 0,26$ *et* $Q_3 = 0,77$

$$Ml_e = e_3 + a_3\left(\frac{0,5 - Q_2}{Q_3 - Q_2}\right) \quad avec\ e_3 = 1400\ et\ a_3 = 400$$

$$Mle = 1400 + \left[\left(\frac{0,5 - 0,26}{0,77 - 0,26}\right) \times 400\right] = 1400 + \left[\left(\frac{0,24}{0,51}\right) \times 400\right] = 1400 + 188 \Rightarrow \boxed{Mle = 1588}$$

La médiale implique que 50% *de la masse salariale est versée aux salariés gagnant moins que* 1588

- $F_2 \leq 0,5$ *et* $F_3 > 0,5$, *avec* $F_2 = 0,33$ *et* $F_3 = 0,83$

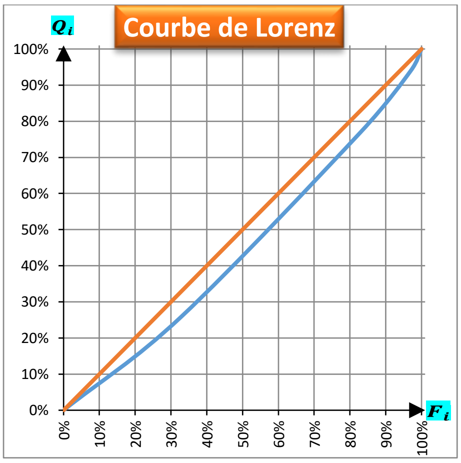
$$M_e = e_3 + a_3\left(\frac{0,5 - F_2}{F_3 - F_2}\right) \quad avec\ e_3 = 1400, (F(1400) = P(X \leq 1400) = F_2 = 0,26) et\ a_3 = 400$$

$$Me = 1400 + \left[\left(\frac{0,5 - 0,33}{0,83 - 0,33}\right) \times 400\right] = 1400 + \left[\left(\frac{0,17}{0,5}\right) \times 400\right] = 1400 + 136 \Rightarrow \boxed{Me = 1536}$$

Le salaire médian signifie que 50% *des salariés gagnant moins de* 1536

Ainsi, 50% *des salariés gagnent moins de* 50% *de la masse salariale* $(Ml_e \geq M_e)$

3) *La courbe de Lorenz, pour la tracer il faut mettre en abscisse la fréquence cumulée croissante de la série* (F_i) *et en ordonnée les valeurs globales cumuléescroissantes*(Q_i). *Lorsqu'on trace cette courbe, le centre d'intérêt est la distance entre la première bissectrice et la courbe de Lorenz. Ainsi plus l'aire comprise entre les deux est importante, plus il y a des inégalités (ou plus la concentration est importante)*



Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Ingénieurs
Jeudi 29 octobre 2015

Exercice 7 ***(4 points =0,5+0,5+0,5+2+0,5) :***

ÉNONCÉ

Le tableau suivant fournit la répartition par tranche d'âge des agriculteurs exerçantdans des exploitations agricoles d'une région donnée.

Moins de 25 *ans*	*e* 25 à 29 *ans*	*De* 30 à 39 *ans*	*De* 40 à 49 *ans*	*De* 50 à 59 *ans*	*Au moins* 60 *ans*
580 *exploitations*	2162 *exploitants*	8063 *exploitants*	9569 *exploitants*	10660 *exploitants*	15913 *exploitants*

1) *Définir la population étudiée, l'individu, le caractère et la modalité.*
2) *Élaborer le tableau statistique de cette série : fréquence, fréquence cumulées, croissance et décroissance. On retiendra* 20 *ans et* 70 *ans come âges minimale et maximale.*
3) *Quelle est la proportion des agriculteurs qui ont : au moins* 40 *ans ? Moins de* 30 *ans ? Entre* 25 *et* 60 *ans ?*

4) *Tracer le graphique des fréquences cumulées croissantes et décroissantes puis déterminer par calcul la médiane* (Me) *les quartiles* Q_1 *et* Q_3. *Donner une estimation graphique des déciles* D_1 *et* D_9. *Placer les points sur le graphique.*

5) *Quelle est la propension des agriculteurs qui ont entre 35 et 65 ans (détermination graphique) ?*

Corrigé

1)
- ***Population*** : ***exploitations agricoles***
- ***Individu*** : ***une exploitation***
- ***Caractère*** : ***âge de l'agriculteur***
- ***Modalité*** : ***classe d'âge***

2)

Classes	Amplitudes	Effectifs	Fréquences	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
$[20, 25[$	5	580	1,2%	1,2%	100%
$[25, 30[$	5	2162	4,6%	5,8%	98,8%
$[30, 40[$	10	8063	17,2%	23,0%	94,2%
$[40, 50[$	10	9569	20,4%	43,4%	77%
$[50, 60[$	10	10660	22,7%	66,1%	56,6%
$[60, 70[$	10	15913	33,9%	100%	33,9%
$\sum$		46947	100%		

3)
- $P(X \geq 40) = 100\% - P(X \leq 40) = 100\% - 23,0\% = 77\%$

77% ***ont au moins 40 ans***
- $P(X < 30) = P(X \leq 30) = 5,8\%$

$5,8\%$ ***ont moins de 30 ans***
- $P(X < 60) = P(X \leq 60) = 66,1\%$

moins de 60 ans : $66,1\%$
- $P(X < 25) = P(X \leq 25) = 1,2\%$

moins de 25 ans : $1,2\%$
- $P(25 \leq X \leq 60) = F(60) - F(25) = P(X \leq 60) - P(X \leq 25) = 66,1\% - 1,2\% = 64,9\%$

entre 25 et 60 ans : $64,9\%$

4)
- $F_4 \leq 50\%$ *et* $F_5 > 50\%$, *avec* $F_4 = P(X \leq 50) = 43,4\%$ *et* $F_5 = P(X \leq 60) = 66,1\%$

$\Rightarrow M_e \in [50, 60[$

$$M_e = e_5 + a_5\left(\frac{50\% - F_4}{F_5 - F_4}\right) \quad avec\ e_5 = 50\ et\ a_5 = 10$$

$$M_e = 50 + \left(10 \times \left(\frac{50\% - 43,4\%}{66,1\% - 43,4\%}\right)\right) \Rightarrow \boxed{Me = 52,9\ ans \approx 53\ ans}$$

- $F_3 \leq 25\%$ *et* $F_4 > 25\%$, *avec* $F_3 = P(X \leq 40) = 23\%$ *et* $F_4 = P(X \leq 50) = 43,4\%$

$\Rightarrow Q_1 \in [40, 50[$

$$Q_1 = e_4 + a_4\left(\frac{25\% - F_3}{F_4 - F_3}\right) \quad avec\ e_4 = 40\ et\ a_4 = 10$$

$Q_1 = 40 + \left(10 \times \left(\frac{25\% - 23\%}{43,4\% - 23\%}\right)\right) \Rightarrow \boxed{Q_1 \ = \ 40,98 \ ans \ \approx \ 41 \ ans}$

· $F_5 \leq 75\%$ et $F_6 > 75\%$, avec $F_5 = P(X \leq 60) = 66,1\%$ et $F_6 = P(X \leq 70) = 100\%$

$\Rightarrow Q_3 \in [60, 70[$

$Q_3 = e_4 + a_4\left(\frac{75\% - F_5}{F_6 - F_5}\right)$ avec $e_6 = 60$ et $a_6 = 10$

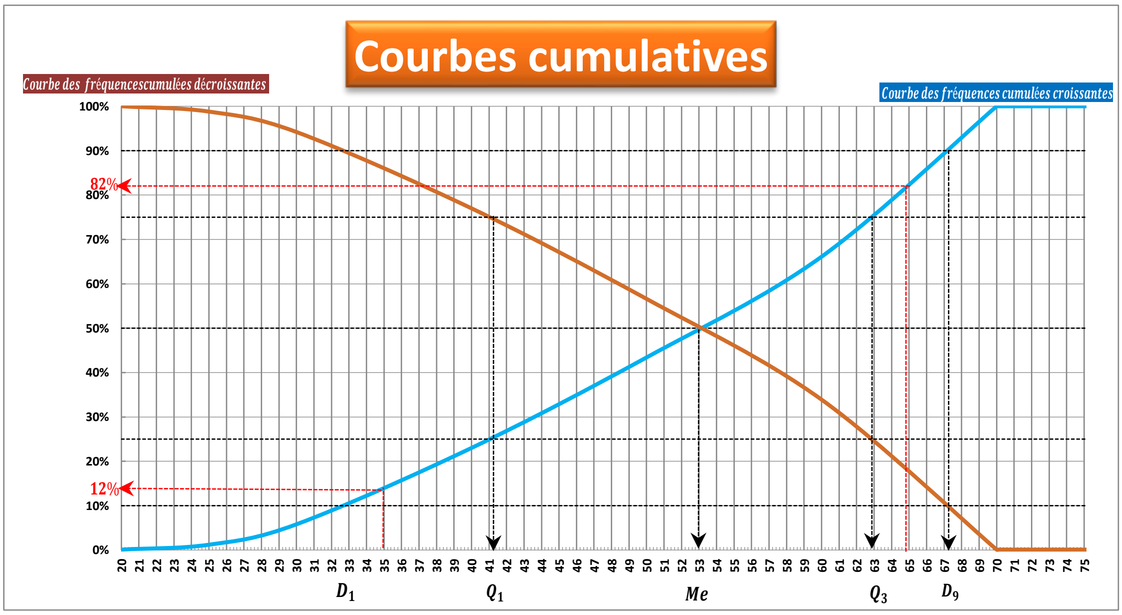
$Q_3 = 60 + \left(10 \times \left(\frac{75\% - 66,1\%}{100\% - 66,1\%}\right)\right) \Rightarrow \boxed{Q_3 \ = \ 62,63 \ ans \ \approx \ 63 \ ans}$

· *Détermination graphique de D_1 : le 1^{er} décile D_1 est l'abscisse du point de la courbe des fréquences cumulées croissantes dont l'ordonnée est égale à 0,1 ou encore l'abscisse du point de la courbe des fréquences cumulées décroissantes dont l'ordonnée est égale à 0,9.*

$\boxed{D_1 = 32,5 \ ans}$

· *Détermination graphique de D_9 : le $9^{ème}$ décile D_9 est l'abscisse du point de la courbe des fréquences cumulées croissantes dont l'ordonnée est égale à 0,9 ou encore l'abscisse du point de la courbe des fréquences cumulées décroissantes dont l'ordonnée est égale à 0,1*

$\boxed{D_9 = 67,5 \ ans}$



5) *Graphiquement on peut déterminer que* :

· *Les moins de 65 ans représentent 82% de la population* : $P(X \leq 65) = F(65) = 82\%$

· *Les moins de 35 ans représentent 12% de la population* : $P(X \leq 35) = F(35) = 12\%$

- *les entre 35 et 64 ans représentent 60% de la population*:

$P(35 \leq X \leq 65) = F(635) - F(35) = 60\%$

Exercice 8

ÉNONCÉ

La série suivante représente le nombre de pièces non-conformes par jour, dans une entreprise :

122-111-154-98-93-67-134-167-123-142-132-151-127-119-137-130-127-135-187
165-161-151-143-148-132-127-99-132-139-100-136-132-127-167-118-116-83-77.

1) ***Classer ces données dans des intervalles de même amplitude.***
2) ***Tracer le diagramme différentiel et le diagramme intégral.***
3) ***Calculer la médiane et déterminer la classe modale.***
4) ***Calculer les moyennes arithmétique, géométrique, harmonique, et quadratique.***
5) ***Étudier l'asymétrie et la forme de cette série.***
6) ***Calculer le coefficient de variation.***

Corrigé

1) ***Rangeons les données par ordre croissant*** : **67-77-83-93-98-99-100-111-116-118-119-122-123-127-127-127-127-130-132132-132-132-134-135-136-137-139-142-143-148-151-151-154-161-165-167-**

167-187

Utilisons la règle empirique de Sturges pour déterminer le nombre de classes :

$$u = 1 + \frac{10}{3} log_{10}(n) = 1 + \frac{10}{3} log_{10}(38) \cong 6, \text{le nombre de classes sera} : r = 6$$

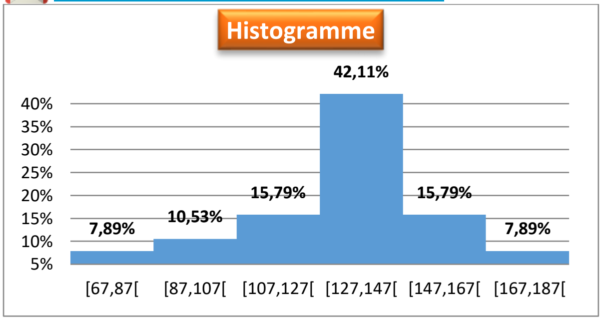
D'autre part l'étendue : $E = x_{max} - x_{min} = 178 - 67 = 120$

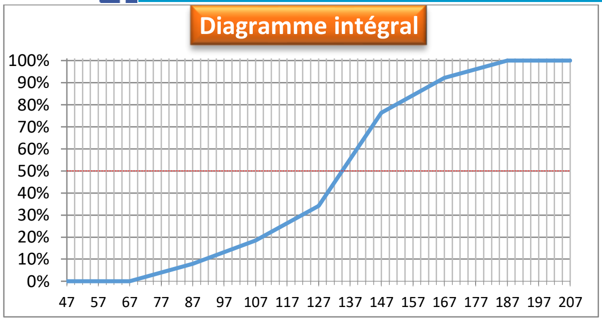
Par la suite l'amplitude de chaque classe sera : $h = \frac{E}{r} = \frac{120}{6} = 20 \Rightarrow a = 20(\text{ l'amplitude })$

D'où la présentation des données sous forme d'un tableau statistique :

$[e_i, e_{i+1}[$	a_i	c_i	n_i	f_i	F_i	$f_i c_i$	$f_i c_i^2$	$\frac{f_i}{c_i}$	$f_i \ln(c_i)$	$c_i - \overline{X}$	$f_i(c_i - \overline{X})^3$	$f_i(c_i - \overline{X})^4$
[67, 87[	20	77	3	7,89%	7,89%	6,0753	467,7981	$10,2468.10^{-4}$	$34,2726.10^{-2}$	−54,21	−12569 ,79	10946,05
[87, 107[	20	97	4	10,53%	18,42%	10,2141	990,7677	$10,8557.10^{-4}$	$48,1717.10^{-2}$	−34,21	−4216,07	144233,92
[107, 127[	20	117	6	15,79%	34,21%	18,4743	2161,4931	$13,4957.10^{-4}$	$75,1947.10^{-2}$	−14,21	−453,12	6439,06
[127, 147[	20	137	16	42,11%	76,32%	57,6907	7903,6259	$30,7372.10^{-4}$	$207,1804.10^{-2}$	5,79	81,72	473,09
[147, 167[	20	157	6	15,79%	92,11%	24,7903	3892,0771	$10,0573.10^{-4}$	$79,8381.10^{-2}$	25,79	2708,38	69847,69
[167, 187[	20	177	3	7,89%	100%	13,9653	2471,8581	$4,4576.10^{-4}$	$40,8398.10^{-2}$	45,79	7574,85	346848,33
$\sum$			38	100%		131,21	17887,62	$79,85030.10^{-4}$	$485,4973.10^{-2}$		−6874,03	578788,14

2)





· $F_3 \leq 50\%$ *et* $F_4 > 50\%$, *avec* $F_3 = P(X \leq 127) = 34,21\%$ *et* $F_4 = P(X \leq 147) = 76,32\%$

$\Rightarrow M_e \in [127, 147[$

$$M_e = e_4 + a_4\left(\frac{50\% - F_3}{F_4 - F_3}\right) \quad avec\ e_4 = 127\ et\ a_4 = 20$$

$$M_e = 127 + \left(20 \times \left(\frac{50\% - 34,21\%}{76,32\% - 34,21\%}\right)\right) \Rightarrow \boxed{Me\ =\ 134,5}$$

· $[e_m, e_{m+1}[= [127, 147[$ *étant la classe modale d'amplitude* $a_m = 20$ *possédant l'effectif le plus élevé* $n_m = 16$, $[e_{m-1}, e_m[= [107, 127[$ *la classe qui la précède d'effectif* $n_{m-1} = 6$, $[e_{m-1}, e_m[= [147, 167[$ *la classe qui la succède d'effectif* $n_{m+1} = 6$

$$ainsi:\ M_o = e_m + a_m\left[\frac{n_m - n_{m-1}}{(n_m - n_{m-1}) + (n_m - n_{m+1})}\right] = 127 + \left(20 \times \left(\frac{16-6}{(16-6)+(16-6)}\right)\right)$$

$\Rightarrow \boxed{M_o\ =\ 137}$

4)

· *Moyenne arithmétique* : $\bar{X} = \sum_{i=1}^{6} f_i c_i \Rightarrow \boxed{\bar{X}\ =\ 131,21}$

· *Moyenne géométrique* : $\bar{G} = \sqrt[38]{\prod_{i=1}^{6} c_i^{\,n_i}} = \left[\prod_{i=1}^{6} c_i^{\,n_i}\right]^{\frac{1}{38}} = \prod_{i=1}^{6} c_i^{\,f_i} = \exp\left[\sum_{i=1}^{6} f_i \ln(c_i)\right]$,

car $\ln(\bar{G}) = \sum_{i=1}^{6} f_i \ln(c_i) = 485,4973.10^{-2} \Rightarrow \bar{G} = e^{485,4973.10^{-2}} \Rightarrow \boxed{\bar{G}\ =\ 128,38}$

· *Moyenne quadratique* : $\bar{Q} = \sqrt{\sum_{i=1}^{6} f_i c_i^2} = \sqrt{17887,62} \Rightarrow \boxed{\bar{Q}\ =\ 133,74}$

· *Moyenne harmonique* : $\frac{1}{\bar{H}} = \sum_{i=1}^{6} \frac{f_i}{c_i} = 79,85030.10^{-4} \Rightarrow \bar{H} = \left[\sum_{i=1}^{r} \frac{f_i}{c_i}\right]^{-1} = \frac{1}{79,85030.10^{-4}}$

$\Rightarrow \boxed{\bar{H}\ =\ 125,23}$

La relation suivante sera toujours vérifiée : $\underbrace{\overline{H}}_{125,23} \leq \underbrace{\overline{G}}_{128,38} \leq \underbrace{\overline{X}}_{131,21} \leq \underbrace{\overline{Q}}_{133,74}$

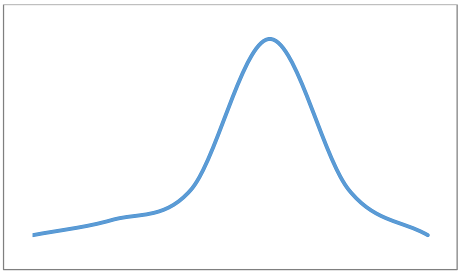
5)

☑ ***Étude de l'asymétrie :***

Calculons le coefficient de dissymétrie de Fisher : $\gamma_1 = \dfrac{\overline{\mu}_3}{S_x^3} = \dfrac{\sum_{i=1}^{6} f_i (c_i - \overline{X})^3}{\left(\sqrt{\sum_{i=1}^{6} f_i (c_i - \overline{X})^2}\right)^3}$

$$or\ , S_x^2 = \sum_{i=1}^{6} f_i (c_i - \overline{X})^2 = \sum_{i=1}^{6} f_i c_i^2 - \overline{X}^2 = 17887,62 - 131,21^2 = 671,56 \Rightarrow S_x = 25,91$$

$\gamma_1 = \dfrac{-6874,03}{25,91^3} = -0,395 < 0$. $\boxed{\textit{En effet la distribution est dissymétrique étalée à gauche}}$



☑ ***Étude de l'aplatissement :***

Calculons le coefficient de Fisher : $\gamma_2 = \dfrac{\overline{\mu}_4}{S_x^4} - 3 = \dfrac{\sum_{i=1}^{6} f_i (c_i - \overline{X})^4}{\left(\sum_{i=1}^{6} f_i (c_i - \overline{X})^2\right)^2} - 3 = \dfrac{578788,14}{671,56^2} - 3$

$\gamma_2 = -1,717 < 0.$ $\boxed{\textit{la distribution est plus aplatie que la Normale (platykurtique)}}$

6)

Coefficient de variation : $C_V = \dfrac{S_x}{\overline{X}} \times 100 = \dfrac{25,91}{131,21} \times 100 \Rightarrow \boxed{C_V = 19,75\%}$

BEN AHMED MOHSEN
omega.center.cp@gmail.com
Téléphone: (+216) 97 619191 / 54 619191
https://web.facebook.com/OMEGACENTER2014

Série statistique à deux variables quantitatives

Présentation des données

$\mathcal{A}$– 1 • Tableau des données ponctuelles :

Soient deux caractères X et Y définis sur une même population d'effectif total n.

Les couples $(x_i, y_i)_{1 \le i \le n}$ *constituent une série statistique à deux variables.*

(*On dit aussi deux dimensions.*)

Il s'agit d'un tableau à trois colonnes (ou trois lignes) du type :

Observation n°	Valeur de X	Valeur de Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
i	x_i	y_i
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

$\mathcal{A}$– 2 • Tableau à double entrée (ou tableau de contingence) :

Lorsqu'un certain nombre d'observations sont identiques, il est préférable de présenter les données dans un tableau à double entrée. On reporte les (r) valeurs distinctes de X en lignes et les (s) valeurs distinctes de Y en colonnes.

À l'intersection de la $i^{ème}$ *ligne et de la* $j^{ème}$ *colonne, on reporte l'effectif* $n_{i,j}$ *correspondant à l'observation conjointe de* $X = x_i$ *et* $Y = y_j$.

☞ *l'effectif jointe* $n_{i,j}$ *correspondant à l'observation conjointe de* $X = x_i$ *et* $Y = y_j$

☞ *la fréquence jointe* $f_{i,j}$ *correspondant à l'observation conjointe de* $X = x_i$ *et* $Y = y_j$

☞ *A partir de ce tableau il est possible de retrouver la description des séries statistiques X et Y "seules" (distributions marginales). C'est à dire : on peut retrouver l'effectif marginal correspondant à*

☑ *la situation* $X = x_i$: $n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$ *et la fréquence marginale* $f_{i\bullet} = \sum_{j=1}^{s} f_{ij} = \dfrac{n_{i\bullet}}{n}$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

☑ $la\ situation\ Y = y_j :\ n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$ $et\ la\ fréquence\ marginale\ f_{\bullet j} = \sum_{i=1}^{r} f_{ij} = \frac{n_{\bullet j}}{n}$

☞ $l'effectif\ total\ sera:\ n = n_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}$, $on\ a\ aussi: \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} f_{ij} = 1$

On synthétise les données de la distribution jointe du couple (X, Y) par un tableau à double entrée appelé tableau de contingence

		Distribution conditionnelle de $(X\|Y = y_2)$						
$X \backslash Y$	y_1	y_2	…	y_j	…	y_s	Distribution marginale de X	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1\bullet} = \sum_{j=1}^{s} n_{1j}$	
	f_{11}	f_{12}	…	f_{1j}	…	f_{1s}	$f_{1\bullet} = \sum_{j=1}^{s} f_{1j}$	
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2\bullet} = \sum_{j=1}^{s} n_{2j}$	Distribution conditionnelle de $(Y\|X = x_2)$
	f_{21}	f_{22}	…	f_{2j}	…	f_{2s}	$f_{2\bullet} = \sum_{j=1}^{s} f_{2j}$	
⋮	⋮	⋮	⋮		⋮	⋮	⋮	
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$	
	f_{i1}	f_{i2}	…	f_{ij}	…	f_{is}	$f_{i\bullet} = \sum_{j=1}^{s} f_{ij}$	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r\bullet} = \sum_{j=1}^{s} n_{rj}$	
	f_{r1}	f_{r2}	…	f_{rj}	…	f_{rs}	$f_{r\bullet} = \sum_{j=1}^{s} f_{rj}$	
Distribution marginale de Y	$n_{\bullet 1} = \sum_{i=1}^{r} n_{i1}$	$n_{\bullet 2} = \sum_{i=1}^{r} n_{i2}$	…	$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$	…	$n_{\bullet s} = \sum_{i=1}^{r} n_{is}$	$n = n_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}$	
	$f_{\bullet 1} = \sum_{i=1}^{r} f_{i1}$	$f_{\bullet 2} = \sum_{i=1}^{r} f_{i2}$	…	$f_{\bullet j} = \sum_{i=1}^{r} f_{ij}$	…	$f_{\bullet s} = \sum_{i=1}^{r} f_{is}$	$1 = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} f_{ij}$	

A – 3 • Distributions conditionnelles :

Le principe des distributions conditionnelles est de décrire le comportement de l'une des deux variables quand l'autre a une valeur donnée.

☞ ***A la ligne "i" du tableau de contingence, on lit la distribution conditionnelle de la variable Y sachant que la variable X prend la modalité x_i ; elle est notée distribution***

conditionnelle de $(Y|X = x_i)$

☑ *La fréquence conditionnelle de* $(Y|X = x_i)$ *est* : $f_{j/i} = f_j^i = \dfrac{n_{ij}}{n_{i\bullet}} = \dfrac{f_{ij}}{f_{i\bullet}}$

☑ *Il y a r distributions conditionnelles de* $(Y|X = x_i)$

Distributions conditionnelles de $(Y\|X = x_i)$	n_{ij}	$f_{j/i}$
y_1	n_{i1}	$f_{1/i}$
y_2	n_{i2}	$f_{2/i}$
$\vdots$	$\vdots$	$\vdots$
y_j	n_{ij}	$f_{j/i}$
$\vdots$	$\vdots$	$\vdots$
y_s	n_{is}	$f_{s/i}$
$\sum$	$n_{i\bullet}$	1

☞ *A la colonne "j" du tableau de contingence, on lit la distribution conditionnelle de la variable X sachant que la variable Y prend la modalité* y_j ; *elle est notée* ***distribution conditionnelle de*** $(X|Y = y_j)$

☑ *La fréquence conditionnelle de* $(X|Y = y_j)$ *est* : $f_{i/j} = f_i^j = \dfrac{n_{ij}}{n_{\bullet j}} = \dfrac{f_{ij}}{f_{\bullet j}}$

☑ *Il y a s distributions conditionnelles de* $(X|Y = y_j)$

Distributions conditionnelles de $(X\|Y = y_j)$	n_{ij}	$f_{i/j}$
x_1	n_{1j}	f_{1j}
x_2	n_{2j}	f_{2j}
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_{ij}
$\vdots$	$\vdots$	$\vdots$
x_r	n_{rj}	f_{rj}
$\sum$	$n_{\bullet j}$	1

☞ *Remarque*: $f_{ij} = \dfrac{n_{ij}}{n} = f_{i\bullet}f_{j/i} = f_{\bullet j}f_{i/j}$

$\mathcal{A}$– 4 • Indépendance statistique :

Le caractère X est dit indépendant du caractère Y, si toutes les distributions conditionnelles de X sont identiques. Elles sont alors égales à la distribution marginale de X. l'indépendance est une relation réciproque :

X indépendante de Y $\Leftrightarrow$ *Y indépendante de X*

☞ *Si X et Y sont indépendants, la relation entre les fréquences devient* :

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

$\mathcal{A}$– 5 • Valeurs typiques :

a • Distributions marginales :

☑ *Moyennes*:

☞ $\bar{X} = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}x_i = \sum_{i=1}^{r} f_{i\bullet}x_i$ ☞ $\bar{Y} = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}y_j = \sum_{j=1}^{s} f_{\bullet j}y_j$

☞ *Le point* $(\bar{X}, \bar{Y})$ *est appelé centre de gravité de la distribution à deux dimensions*

☑ ***Variances***:

☞ $S_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(x_i - \bar{X})^2 = \sum_{i=1}^{r} f_{i\bullet}(x_i - \bar{X})^2 = \sum_{i=1}^{r} f_{i\bullet}x_i^2 - \bar{X}^2$

☞ $S_y^2 = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}(y_j - \bar{Y})^2 = \sum_{j=1}^{s} f_{\bullet j}(y_j - \bar{Y})^2 = \sum_{j=1}^{s} f_{\bullet j}y_j^2 - \bar{Y}^2$

b • Distributions conditionnelles :

☑ ***Moyennes***:

☞ *On appelle moyenne conditionnelle de* $(X|Y = y_j)$ *et on note* :

$$\bar{X}_j = \frac{1}{n_{\bullet j}}\sum_{i=1}^{r} n_{ij}x_i = \sum_{i=1}^{r} f_{i/j}x_i$$

☞ *On appelle moyenne conditionnelle de* $(Y|X = x_i)$ *et on note* :

$$\bar{Y}_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{s} n_{ij}y_j = \sum_{j=1}^{s} f_{j/i}y_j$$

☑ ***Variances***:

☞ *On appelle moyenne conditionnelle de* $(X|Y = y_j)$ *et on note* :

$$S_j^2 = V_j(X|Y = y_j) = V_j(X) = \frac{1}{n_{\bullet j}}\sum_{i=1}^{r} n_{ij}(x_i - \bar{X}_j)^2 = \sum_{i=1}^{r} f_{i/j}(x_i - \bar{X}_j)^2$$

☞ *On appelle moyenne conditionnelle de* $(Y|X = x_i)$ *et on note* :

$$S_i^2 = V_i(Y|X = x_i) = V_i(Y) = \frac{1}{n_{i\bullet}}\sum_{j=1}^{s} n_{ij}(y_j - \bar{Y}_i)^2 = \sum_{j=1}^{s} f_{j/i}(y_j - \bar{Y}_i)^2$$

c • Relations entre les caractéristiques marginales et conditionnelles :

☑ ***Relation entre les moyennes*** :

La moyenne marginale est égale à la moyenne des moyennes conditionnelles, pondérées par les effectifs marginaux :

☞*La moyenne de X dans la population totale est la moyenne des « s » moyennes des X dans les distributions conditionnelles :* $\bar{X} = \frac{1}{n}\sum_{j=1}^{s} n_{.j}\bar{X}_j = \sum_{j=1}^{s} f_{.j}\bar{X}_j$

• ***Preuve :***

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{r} \underbrace{n_{i.}}_{\sum_{j=1}^{s} n_{ij}} x_i = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}x_i = \frac{1}{n}\sum_{j=1}^{s}\underbrace{\sum_{i=1}^{r} n_{ij}x_i}_{n_{.j}\bar{X}_j} = \frac{1}{n}\sum_{j=1}^{s} n_{.j}\bar{X}_j = \sum_{j=1}^{s} f_{.j}\bar{X}_j$$

☞*La moyenne de Y dans la population totale est la moyenne des « r » moyennes des Y dans les distributions conditionnelles :* $\bar{Y} = \frac{1}{n}\sum_{i=1}^{r} n_{i.}\bar{Y}_i = \sum_{i=1}^{r} f_{i.}\bar{Y}_i$

• ***Preuve :***

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{s} \underbrace{n_{.j}}_{\sum_{i=1}^{r} n_{ij}} y_j = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}y_j = \frac{1}{n}\sum_{i=1}^{r}\underbrace{\sum_{j=1}^{s} n_{ij}y_j}_{n_{i.}\bar{Y}_i} = \frac{1}{n}\sum_{i=1}^{r} n_{i.}\bar{Y}_i = \sum_{i=1}^{r} f_{i.}\bar{Y}_i$$

☑ ***Relation entre les variances :***

☞*La variance marginale de X est égale à la somme de la moyenne des « s » variances conditionnelles et de la variance des « s » moyennes conditionnelles :* $S_x^2 = \overline{\left(V_j(X)\right)} + S_{\bar{X}_j}^2$

$$S_x^2 = \frac{1}{n}\sum_{j=1}^{s} n_{.j}V_j(X) + \frac{1}{n}\sum_{j=1}^{s} n_{.j}\left(\bar{X}_j - \bar{X}\right)^2 = \sum_{j=1}^{s} f_{.j}V_j(X) + \sum_{j=1}^{s} f_{.j}\left(\bar{X}_j - \bar{X}\right)^2 = \sum_{j=1}^{s} f_{.j}V_j(X) + S_{\bar{X}_j}^2$$

• ***Preuve :***

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i.}(x_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i.}\left[(x_i - \bar{X}_j) - (\bar{X} - \bar{X}_j)\right]^2$$

$$= \frac{1}{n}\sum_{i=1}^{r} n_{i.}\left[(x_i - \bar{X}_j)^2 - 2(x_i - \bar{X}_j)(\bar{X} - \bar{X}_j) + (\bar{X} - \bar{X}_j)^2\right]$$

$$= \frac{1}{n}\sum_{i=1}^{r} \underbrace{n_{i.}}_{\sum_{j=1}^{s} n_{ij}} (x_i - \bar{X}_j)^2 - \frac{2}{n}\sum_{i=1}^{r} \underbrace{n_{i.}}_{\sum_{j=1}^{s} n_{ij}} (x_i - \bar{X}_j)(\bar{X} - \bar{X}_j) + \frac{1}{n}\sum_{i=1}^{r} \underbrace{n_{i.}}_{\sum_{j=1}^{s} n_{ij}} (\bar{X} - \bar{X}_j)^2$$

$$= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - \bar{X}_j)^2 - \frac{2}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(x_i - \bar{X}_j)(\bar{X} - \bar{X}_j) + \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}(\bar{X} - \bar{X}_j)^2$$

$$S_x^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)^2}_{\mathcal{A}} - \underbrace{\frac{2}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)(\overline{X} - \overline{X}_j)}_{\mathcal{B}} + \underbrace{\frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(\overline{X} - \overline{X}_j)^2}_{\mathcal{C}}$$

$$\bullet \mathcal{B} = \frac{2}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)(\overline{X} - \overline{X}_j) = \frac{2}{n}\sum_{j=1}^{s}\left[(\overline{X} - \overline{X}_j)\left(\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)\right)\right]$$

$$= \frac{2}{n}\sum_{j=1}^{s}\left[(\overline{X} - \overline{X}_j)\left(\sum_{i=1}^{r} n_{ij}x_i - \sum_{i=1}^{r} n_{ij}\overline{X}_j\right)\right] = \frac{2}{n}\sum_{j=1}^{s}\left[(\overline{X} - \overline{X}_j)\left(\sum_{i=1}^{r} n_{ij}x_i - \overline{X}_j \underbrace{\sum_{i=1}^{r} n_{ij}}_{n_{\bullet j}}\right)\right]$$

$$= \underbrace{\frac{2}{n}\sum_{j=1}^{s}\left[(\overline{X} - \overline{X}_j)\underbrace{\left(\underbrace{\sum_{i=1}^{r} n_{ij}x_i}_{n_{\bullet j}\overline{X}_j} - n_{\bullet j}\overline{X}_j\right)}_{0}\right]}_{0} \Rightarrow \boxed{\mathcal{B} = 0}$$

$$\bullet \mathcal{A} = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)^2 = \frac{1}{n}\sum_{j=1}^{s}\left[\underbrace{\left(\sum_{i=1}^{r} n_{ij}(x_i - \overline{X}_j)^2\right)}_{n_{\bullet j}V_j(X)}\right] \Rightarrow \boxed{\mathcal{A} = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}V_j(X)}$$

$$\bullet \mathcal{C} = \frac{1}{n}\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(\overline{X} - \overline{X}_j)^2 = \frac{1}{n}\sum_{j=1}^{s}\left[\left(\sum_{i=1}^{r} n_{ij}(\overline{X} - \overline{X}_j)^2\right)\right] = \frac{1}{n}\sum_{j=1}^{s}(\overline{X} - \overline{X}_j)^2\left[\underbrace{\left(\sum_{i=1}^{r} n_{ij}\right)}_{n_{\bullet j}}\right]$$

$$\boxed{\mathcal{C} = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}(\overline{X} - \overline{X}_j)^2 = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}(\overline{X}_j - \overline{X})^2}$$

$$En\ effet:\ S_x^2 = \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}V_j(X) + \frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}(\overline{X}_j - \overline{X})^2$$

☞ ***La variance marginale de Y est égale à la somme de la moyenne des « r » variances conditionnelles et de la variance des « r » moyennes conditionnelles :*** $S_y^2 = \overline{(V_\iota(Y))} + S_{\overline{Y}_i}^2$

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}V_i(Y) + \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(\overline{Y}_i - \overline{Y})^2 = \sum_{i=1}^{r} f_{i\bullet}V_i(Y) + \sum_{i=1}^{r} f_{i\bullet}(\overline{Y}_i - \overline{Y})^2 = \sum_{i=1}^{r} f_{i\bullet}V_i(Y) + S_{\overline{Y}_i}^2$$

$\mathcal{A}$– 6 • *Les moments* :

a • Moments empiriques par rapport à l'origine (ou non centrés) d'ordre k et l :

$$\overline{m}_{k,l}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}x_i^k y_j^l=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i^k y_j^l$$

☞ $\overline{m}_{0,0}=1$

☞ $\overline{m}_{1,0}=\overline{X}$

• ***Preuve*** : $\overline{m}_{1,0}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}x_i^1 y_j^0=\frac{1}{n}\sum_{i=1}^{r}x_i\underbrace{\left(\sum_{j=1}^{s}n_{ij}\right)}_{n_{i\bullet}}=\overline{X}=\frac{1}{n}\sum_{i=1}^{r}n_{i\bullet}x_i=\sum_{i=1}^{r}f_{i\bullet}x_i$

☞ $\overline{m}_{0,1}=\overline{Y}$

• ***Preuve*** : $\overline{m}_{0,1}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}x_i^0 y_j^1=\frac{1}{n}\sum_{i=1}^{r}y_j\underbrace{\left(\sum_{j=1}^{s}n_{ij}\right)}_{n_{\bullet j}}=\overline{Y}=\frac{1}{n}\sum_{j=1}^{s}n_{\bullet j}y_j=\sum_{j=1}^{s}f_{\bullet j}y_j$

b • Moments empiriques centrés d'ordre k et l :

$$\overline{\mu}_{k,l}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}(x_i-\overline{X})^k\left(y_j-\overline{Y}\right)^l=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}(x_i-\overline{X})^k\left(y_j-\overline{Y}\right)^l$$

☞ $\overline{\mu}_{0,0}=1$ ☞ $\overline{\mu}_{1,0}=0$ ☞ $\overline{\mu}_{0,1}=0$

☞ $\overline{\mu}_{1,1}=S_{x,y}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}(x_i-\overline{X})\left(y_j-\overline{Y}\right)=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}(x_i-\overline{X})\left(y_j-\overline{Y}\right)=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i y_j-\overline{X}\overline{Y}$

• ***Preuve*** : $\overline{\mu}_{1,1}=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}(x_i-\overline{X})\left(y_j-\overline{Y}\right)=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}\left(x_i y_j-\overline{X}y_j-\overline{Y}x_i+\overline{X}\overline{Y}\right)$

$$=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i y_j-\overline{X}\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}y_j}_{\overline{Y}}-\overline{Y}\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i}_{\overline{X}}+\overline{X}\overline{Y}\underbrace{\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}}_{1}$$

$$=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i y_j-\overline{X}\overline{Y}-\overline{X}\overline{Y}+\overline{X}\overline{Y}=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_i y_j-\overline{X}\overline{Y}$$

A– 7 • Covariance, Corrélation :

a • Covariance :

La notion de covariance est « homogène » à celle de la variance des séries à une dimension, donc c'est une généralisation de bidimensionnelle la notion de variance. C'est un indice

rendant compte numériquement de la manière dont les deux variables considérées varient simultanément. La covariance de X et Y est le nombre réel défini par :

$$S_{x,y}=\bar{\mu}_{1,1}=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}(x_i-\bar{X})(y_j-\bar{Y})=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}(x_i-\bar{X})(y_j-\bar{Y})=\sum_{i=1}^{r}\sum_{j=1}^{s}f_{ij}x_iy_j-\bar{X}\bar{Y}$$
$$=\overline{(XY)}-\bar{X}\bar{Y}$$

☑ **Propriétés :**

☞ $S_{x,y}=S_{y,x}$ ☞ $S_{x,x}=S_x^2$ et $S_{y,y}=S_y^2$ ☞ **Inégalité de Cauchy-Schwarz :** $S_{x,y}^2\le S_x^2S_y^2$

☞ $S_{x+y}^2=S_x^2+S_y^2+2S_{x,y}$ ☞ $S_{x-y}^2=S_x^2+S_y^2-2S_{x,y}$ ☞ $S_{\alpha x+\beta y}^2=\alpha^2S_x^2+\beta^2S_y^2+2\alpha\beta S_{x,y}$

☞ *La covariance peut prendre des valeurs positives, négatives ou nulles :* $S_{x,y}\in\mathbb{R}$

☞ *La covariance dépend des unités de X et de Y*

☞ **Remarque**: *Dans le cas de la variance, on doit passer à l'écart-type pour avoir un indicateur interprétable ; dans celui de la covariance, il faudra passer au coefficient de corrélation linéaire.*

b • **Le coefficient de corrélation linéaire :**

Il est clair que la covariance dépend des unités de mesure dans lesquelles sont exprimées les variables considérées. En ce sens, ce n'est pas un indice de liaison « intrinsèque ».
C'est la raison pour laquelle on définit le coefficient de corrélation linéaire (souvent appelé coefficient de Pearson, plus rarement de Bravais-Pearson), rapport entre la covariance et le produit des écarts-types. Ce coefficient caractérise, de façon intrinsèque, la liaison linéaire entre les deux variables considérées.
En particulier, il ne dépend pas des unités de mesure des deux variables.

Sa définition est donc la suivante : $Corr(x,x)=r_{x,y}=\dfrac{S_{x,y}}{S_xS_y}$

☑ **Propriétés :**

☞ $r_{x,y}=S_{\left(\frac{x-\bar{X}}{S_x};\frac{y-\bar{Y}}{S_y}\right)}=Cov_{empirique}\left(\dfrac{x-\bar{X}}{S_x};\dfrac{y-\bar{Y}}{S_y}\right)$

☞ $r_{x,y}=r_{y,x}$

☞ *En s'appuyant sur l'inégalité de Cauchy-Schwarz, on obtient :* $-1\le r_{x,y}\le 1$

☞ $r_{x,x} = 1$

☞ $r_{x,ax+b} = \begin{cases} 1, si\ a > 0 \\ -1, si\ a < 0 \end{cases}$

☑ ***Le coefficient de corrélation linéaire/interprétation :***

☞ *Le signe du coefficient indique le sens de la liaison. Ainsi, une valeur positive indique que les deux variables ont tendance à varier dans le même sens (sur une population de ménages, penser aux revenus, variable X, et aux dépenses vestimentaires, variable Y). Au contraire, une valeur négative du coefficient de corrélation linéaire indique que les deux variables ont tendance à varier en sens opposés (toujours sur une population de ménages, penser maintenant aux dépenses totales, variable X, et à l'épargne, variable Y).*

$r_{x,y} > 0 \Rightarrow$ *les deux variables x et y ont tendance à varier dans le même sens*

$r_{x,y} < 0 \Rightarrow$ *les deux variables x et y ont tendance à varier en sens opposés*

☞ *La valeur absolue du coefficient indique l'intensité de la liaison. Plus cette valeur absolue est proche de 1, plus la liaison est forte ; au contraire, plus elle est proche de 0 et plus la liaison est faible. Ainsi, un coefficient de 0,9 indique une liaison très forte ; un coefficient de 0,5 indique une liaison moyenne ; un coefficient de 0,1 indique une liaison très faible.*

$|r_{x,y}| \cong 1 \Rightarrow$ *la liaison entre les deux variables x et y est forte*

$|r_{x,y}| \cong 0 \Rightarrow$ *la liaison entre les deux variables x et y est faible*

$|r_{x,y}| \cong 0,5 \Rightarrow$ *la liaison entre les deux variables x et y est moyenne*

☞ $|r_{x,y}| = 1$ *correspondent à :* $Y = aX + b$ *et* $X = cY + d$

il existe, donc une liaison linéaire entre X et Y, bien entendu, un tel cas ne se rencontre en général pas avec des données réelles.

☞ *En pratique, lorsque* $|r_{x,y}| > 0,8$ *, alors la liaison linéaireest considérée comme forte.*

☞ *X et Y indépendantes* $\Rightarrow r_{x,y} = 0$

Ajustement analythique

B– 1 • Nuage de points :

a • Introduction :

Soit (X,Y) une série statisrique à deux caractères. L'ensembledes points du plan $(O,\vec{i},\vec{j})$ de coordonnées $(x_i,y_i)_{1\leq i\leq n}$ s'appellele nuage de points représentant la série statistique (X,Y) Lorsque des points se superposent, on ajoute entre parenthèses leur effectif n_{ij} sur la représentation graphique du nuage.

Une fois le nuage dessin´e, on peut essayer de trouver une fonction f telle que la courbe d'équation $y = f(x)$"passe le plus près possible" des points du nuage.

C'est le problème de l'ajustement.

b • Régression linéaire entre deux variables :

Lorsque deux variables quantitatives sont correctement corrélées $|r_{x,y}| \cong 1$ et que l'on peut considérer, a priori, que l'une (nous supposerons qu'il s'agit de X)est cause de l'autre (il s'agira donc de Y), il est alors assez naturel de chercher une fonction de X approchant Y, « le mieux possible » en un certain sens. La méthode statistique permettant de trouver une telle fonction s'appelle la régression de Y sur X.

Pour pouvoir mettre en œuvre une régression, il est au préalable nécessaire d'une part de définir un ensemble de fonctions dans lequel on va chercher « la meilleure », d'autre part de préciser le sens (mathématique)que l'on donne aux expressions telles que « le mieux possible » ou encore « la meilleure ».

Si l'on choisit pour ensemble de fonctions celui des fonctions affines (du type $f(X) = aX + b$), on parle alors de régression linéaire (parce que le graphe d'une telle fonction est une droite). C'est le choix que l'on fait le plus fréquemment dans la pratique et c'est celui que nous ferons ici. Pour donner un sens mathématique à l'expression « le mieux possible », on utilise en général le critère appelé des moindres carrés car il consiste à minimiser une somme de carrés, ou parfois le critère de Mayer.

c • Critère de Mayer :

Effectuons un « agrandissement » d'une portion de la droite. Considérons le point « observation » de coordonnées (x_i, y_i). Il existe sur la droite d'ajustement un point qui n'est généralement pas une observation, de même abscisse x_i et dont l'ordonnée est :

$y'_i = ax_i + b$, *avec* $i = 1, 2, \ldots, n$

Soit $\delta_i = y_i - y'_i$: $\begin{cases} \text{si } \delta_i > 0 \Rightarrow \text{le point } (x_i, y'_i) \text{ se trouve au-dessus de la droite} \\ \text{si } \delta_i < 0 \Rightarrow \text{le point } (x_i, y'_i) \text{ se trouve en dessous de la droite} \end{cases}$

Le critère de Mayer consiste à imposer : $\displaystyle\sum_{i=1}^{n} \delta_i = 0$

Comme $\delta_i = y_i - y'_i = y_i - ax_i - b$; $(i = 1, 2, \ldots, n)$, *on a :*

$$\sum_{i=1}^{n} \delta_i = 0 \Leftrightarrow \sum_{i=1}^{n} (y_i - ax_i - b) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i - nb - a\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} y_i - b - a\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = 0$$

$$\Leftrightarrow \overline{Y} - b - a\overline{X} = 0 \text{ d'où } \boxed{\overline{Y} = b + a\overline{X}}$$

Le critère de Mayer conduit à imposer cette condition. Or cette équation implique que la droite passe par le point $G(\overline{X}, \overline{Y})$ appelé centre de gravité de l'ensemble des points observés,

ainsi : $a = \dfrac{y'_i - \overline{Y}}{x_i - \overline{X}}$

d • Méthode de Mayer :

Divisons l'ensemble des observations en deux parties (I et II) comprenant p et q observations tels que $p + q = n$ et $|p - q| \le 1$: $\displaystyle\sum_{i=1}^{p} \delta_i = 0$ *(dans I) et :* $\displaystyle\sum_{i=p+1}^{n} \delta_i = 0$ *(dans II)*

Ainsi on a deux équations faisant intervenir les centres de gravité $G_I(\overline{X}_I, \overline{Y}_I)$ et $G_{II}(\overline{X}_{II}, \overline{Y}_{II})$ des deux régions I et II : $\begin{cases} \overline{Y}_I = b + a\overline{X}_I \\ \overline{Y}_{II} = b + a\overline{X}_{II} \end{cases}$

d'où l'équation de la dtoite $(G_I, G_{II}) : y - \overline{Y}_I = \left(\dfrac{\overline{Y}_{II} - \overline{Y}_I}{\overline{X}_{II} - \overline{X}_I}\right)(x - \overline{X}_I)$ *, avec* $G(\overline{X}, \overline{Y}) \in (G_I, G_{II})$

☞ **Exemple :** *On mesure le poids Y et la taille X de 20 individus :*

x_i	155	162	157	170	164	162	169	170	178	173	180	175	173	175	179	175	180	185	189	187
y_i	60	61	64	67	68	69	70	70	72	73	75	76	78	80	85	90	96	96	98	101

On divise l'ensemble des 20 observations en deux parties (I et II) comprenant chacune 10

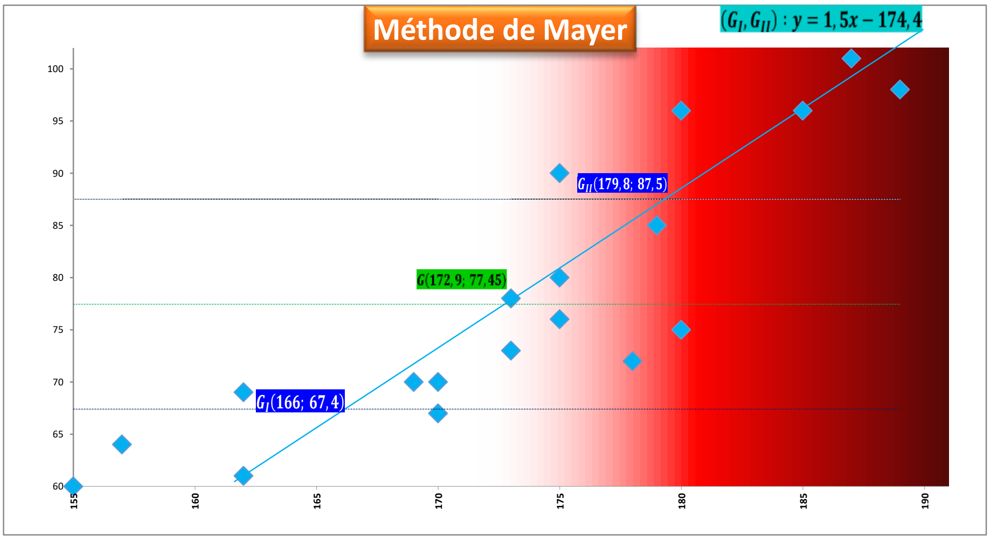
On obtient respectivement : $\overline{X} = 172,9$; $\overline{Y} = 77,45$; $\overline{X}_I = 166$; $\overline{X}_{II} = 179,8$; $\overline{Y}_I = 67,4$

et $\overline{Y}_{II} = 87,5$. *Par la suite les points* $G(172,9;\ 77,45)$; $G_I(166;\ 67,4)$ *et* $G_{II}(179,8;\ 87,5)$

l'équation de la dtoite(G_I, G_{II}) *obtenue par la méthode de Mayer sera donc* :

$$(G_I, G_{II}) : y - \overline{Y}_I = \left(\frac{\overline{Y}_{II} - \overline{Y}_I}{\overline{X}_{II} - \overline{X}_I}\right)(x - \overline{X}_I) \Leftrightarrow y - 67,4 = \left(\frac{87,5 - 67,4}{179,8 - 166}\right)(x - 166)$$

$$(G_I, G_{II}) : y = 1,5x - 174,4$$



$\mathcal{B}$– 2 • *Méthodes des moindres carrés* :

a • *Introduction* :

Si l'on applique la fonction $a + bX$ *à la valeur* x_i *de la variable* X *observée sur l'individu « i », on obtient*$(a + bx_i)$ *La différence entre cette valeur et celle qu'elle est censée approcher,* y_i*, vaut* $y_i - (a + bx_i)$ *Elle représente l'erreur commise en approchant* y_i *par* $a + bx_i$.

Pour obtenir l'erreur globale commise sur l'ensemble de l'échantillon, il faut ensuite faire la somme de l'ensemble de ces quantités. Comme dans la définition de la variance, il est nécessaire au préalable de les prendre soit en valeur absolue soit au carré, pour éviter que les erreurs positives ne compensent les erreurs négatives. L'utilisation des carrés étant nettement plus commode au niveau des calculs, c'est eux que l'on utilise en général et c'est ainsi que l'on obtient le critère des moindres carrés.

Pour mémoire, on notera que $[y_i - (a + bx_i)]$ *représente, dans le nuage de points associé*

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

aux observations, la distance verticale du point figurant « i » à la droite d'équation

$Y = a + bX$ ***et c'est aussi l'erreur que l'on commet en utilisant la droite de régression***

pour prédire y_i ***à partir de*** x_i

Les « différences » $[y_i - (a + bx_i)]$ ***peuvent être positifs ou négatifs.***

b • Détermination des coefficients de la régression linéaire :

Pour déterminer la valeur des coefficients a et b on utilise le principe des moindres carrés qui consiste à chercher la droite qui minimise la somme des carrés des « différences » $[y_i - (a + bx_i)]$:

$$\min_{a,b} \sum_{i=1}^{n} \mathcal{M}(a,b) = \min_{a,b} \underbrace{\sum_{i=1}^{n} (y_i - a - bx_i)^2}_{\mathcal{M}(a,b)}$$

• Condition nécessaire :

$$\begin{cases} \partial\mathcal{M}(a,b)/\partial a = 0 \\ \partial\mathcal{M}(a,b)/\partial b = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} n\bar{Y} - na - nb\bar{X} = 0 \\ na\bar{X} + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \Leftrightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ n(\bar{Y} - b\bar{X})\bar{X} + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ b\sum_{i=1}^{n} x_i^2 - nb\bar{X}^2 = \sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y} \end{cases} \Leftrightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ b\left(\sum_{i=1}^{n} x_i^2 - n\bar{X}^2\right) = \sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y} \end{cases}$$

$$\Leftrightarrow \begin{cases} a = \bar{Y} - b\bar{X} \\ b = \dfrac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{X}^2} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{X})^2} \end{cases}$$

• Condition suffisante :

$$\partial^2\mathcal{M}/\partial a^2\,(a,b) = \frac{\partial\left(-2\sum_{i=1}^{n}(y_i - a - bx_i)\right)}{\partial a} = -2\sum_{i=1}^{n} -1 = 2n$$

$$\partial^2\mathcal{M}/\partial b^2\,(a,b) = \frac{\partial\left(-2\sum_{i=1}^{n}(y_i - a - bx_i)x_i\right)}{\partial b} = -2\sum_{i=1}^{n} -x_i^2 = 2\sum_{i=1}^{n} x_i^2$$

$$\partial^2\mathcal{M}/\partial b\partial a\,(a,b) = \frac{\partial^2\mathcal{M}}{\partial a\partial b}(a,b) = \frac{\partial\left(-2\sum_{i=1}^{n}(y_i - a - bx_i)\right)}{\partial b} = -2\sum_{i=1}^{n} -x_i = 2\sum_{i=1}^{n} x_i$$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

$$|H(\mathcal{M}(a,b))| = \begin{vmatrix} \frac{\partial^2 \mathcal{M}}{\partial a^2}(a,b) & \frac{\partial^2 \mathcal{M}}{\partial b \partial a}(a,b) \\ \frac{\partial^2 \mathcal{M}}{\partial a \partial b}(a,b) & \frac{\partial^2 \mathcal{M}}{\partial b^2}(a,b) \end{vmatrix} = \begin{vmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{vmatrix}$$

$$|H(\mathcal{M}(a,b))| = 4n\sum_{i=1}^{n} x_i^2 - 4\left(\sum_{i=1}^{n} x_i\right)^2 = 4n\sum_{i=1}^{n} x_i^2 - 4(n\overline{X})^2 = 4n^2\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2\right)$$

$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2$ *étant la variance empirique de* $(x_1, \dots, x_n)$

$S_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2$ *étant la variance empirique de* $(y_1, \dots, y_n)$

$S_{x,y} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y}$ *étant la covariance empirique*

$|H(\mathcal{M}(a,b))| = 4n^2 S_x^2 > 0 \Rightarrow (a,b)$ *est un minimum local*

$$\begin{cases} a = \overline{Y} - b\overline{X} \\ b = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{X}^2} = \dfrac{\sum_{i=1}^{n} (x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{X})^2} = \dfrac{S_{x,y}}{S_x^2} \end{cases}$$

La droite est : $y = a + bx = \overline{Y} - \frac{S_{x,y}}{S_x^2}\overline{X} + \frac{S_{x,y}}{S_x^2}x$ *ou* $y - \overline{Y} = \frac{S_{x,y}}{S_x^2}(x - \overline{X})$ *ou* $\frac{y - \overline{Y}}{S_x} = r_{x,y}\left(\frac{x - \overline{X}}{S_y}\right)$

c • Remarques :

☞ *Dans le cas où* $S_{x,y} = 0$ *on obtient dans la formule ci-dessus* : $Y = \overline{Y}$ *donc* X *ne joue aucun rôle, il n'y a pas de relations entre* Y *et* X.

C'est pourquoi nous disons que X *et* Y *sont non corrélées lorsque* $S_{x,y} = 0$.

☞ *La droite de régression de* y *en* x *n'est pas la même que la droite de régression de* x *en* y.

☞ *Le point* $(\overline{X}, \overline{Y})$ *appartient à la droite de régression*: $\overline{Y} = a + b\overline{X}$

☞ *Les « différences »* $[y_i - (a + bx_i)]$ *représentent la partie inexpliquée des* y_i *par la droite de régression*

☞ $\frac{1}{n}\sum_{i=1}^{n}[y_i - (a + bx_i)] = 0$

☞ $\sum_{i=1}^{n} x_i[y_i - (a + bx_i)] = 0$

☞ $b = \frac{S_y}{S_x} r_{x,y}$; b et $r_{x,y}$ auront toujours le même signe

☞ Tous les points du nuage sont alignés $\Leftrightarrow |r_{x,y}| = 1$

☞ Tous les points du nuage sont presque alignés $\Leftrightarrow |r_{x,y}| \cong 1$

d • Régression linéaire de X sur Y :

Dans ce qui précède on a cherché à exprimer Y en fonction de X : $(Y = a + bX)$, régression linéaire de Y sur X. Nous pouvons aussi chercher une relation linéaire du type : $(X = a' + b'Y)$, régression linéaire de X sur Y. Les résultats précédents se généralisent sans difficultés :

$$\begin{cases} a' = \overline{X} - b'\overline{Y} \\ b' = \dfrac{\sum_{i=1}^{n} x_i y_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} y_i^2 - n\overline{Y}^2} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(y_i - \overline{Y})^2} = \dfrac{S_{x,y}}{S_y^2} \end{cases}$$

La droite de régression est: $x = a' + b'y = \overline{X} - \frac{S_{x,y}}{S_y^2}\overline{Y} + \frac{S_{x,y}}{S_y^2} y$ ou encore $x - \overline{X} = \frac{S_{x,y}}{S_y^2}(y - \overline{Y})$

☞ $bb' = r_{x,y}^2$

☞ $r_{x,y} = b\frac{S_x}{S_y} = b'\frac{S_y}{S_x}$

☞ $r_{(a+bx, a'+b'y)} = \frac{bb'}{|bb'|} r_{x,y}$

☑ **Droites de régressions :** Considérons les deux droites de régressions qu'on peut déterminer par l'ajustement : la droite D de Y en X et la droite D' de X en Y

☞ D et D' confondues $\Longrightarrow$ il existe dans ce cas une relation linéaire exacte entre x et y

☞ $D \perp D' \Longrightarrow \begin{cases} \text{L'indépendance entre } x \text{ et } y \\ \text{ou} \\ \text{L'absence de corrélation réciproque entre } x \text{ et } y \\ \text{ou} \\ \sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = 0 \ (\text{c.-à-d. } S_{x,y} = 0) \end{cases}$

☞ Les droites D et D' sont distinctes et non perpendiculaires. C'est le cas général. Les deux droites ont alors des pentes respectives b et b' de même signe que celui de

$\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) \Longrightarrow \begin{cases} \text{s'il existe une relation linéaire exacte alors } bb' = 1 \\ \text{si les deux caractères sont indépendants, alors } bb' = 0 \end{cases}$

e • Autre cas d'ajustement :

La méthode des moindres carrés permet de trouver la « meilleure » relation affine entre Y et X. Cependant, il se peut que le nuage de points ne représente pas une droite mais une parabole. Dans ce cas, la relation entre Y et X est peut être du type $Y \cong a + bX^2$ *. Nous allons voir comment trouver de bonnes valeurs pour a et b dans ce cas.*

Nous allons utiliser la méthode des moindres carrés et un changement de variables.

L'idée est la suivante: nous allons étudier un nouveau couple de séries statistiques (U,V).

U et V sont définies de la manière suivante: $U = X^2$ *et* $V = Y^2$. *Cela signifie que l'on a* $u_i = x_i^2$ *et* $v_i = y_i^2$. *Puisque* $Y \cong a + bX^2$, *il en découle* $V \cong a + bU$.

Nous nous sommes donc ramenés à un cas de régression linéaire de V en U. Nous pouvons donc calculer a et b à l'aide de la méthode des moindres carrés.

f • *Changements de variables usuelles :*

Équation d'une courbe	Changement de variables	Équation de droite
$Y = \alpha X^n$	$U = \ln X$ et $V = \ln Y$	$V = nU + \ln \alpha$
$Y = a + bX^n$	$U = X^n$ et $V = Y$	$V = a + bU$
$Y = \alpha e^{\beta X}$	$U = X$ et $V = \ln Y$	$V = \beta U + \ln \alpha$
$Y = a + b\ln X$	$U = \ln X$ et $V = Y$	$V = a + bU$
$Y = a + \frac{b}{X}$	$U = \frac{1}{X}$ et $V = Y$	$V = a + bU$

g • *Ajustement et corrélation :*

☞ *Si le coefficient de corrélation est positif, les points sont alignés le long d'une droite croissante.*

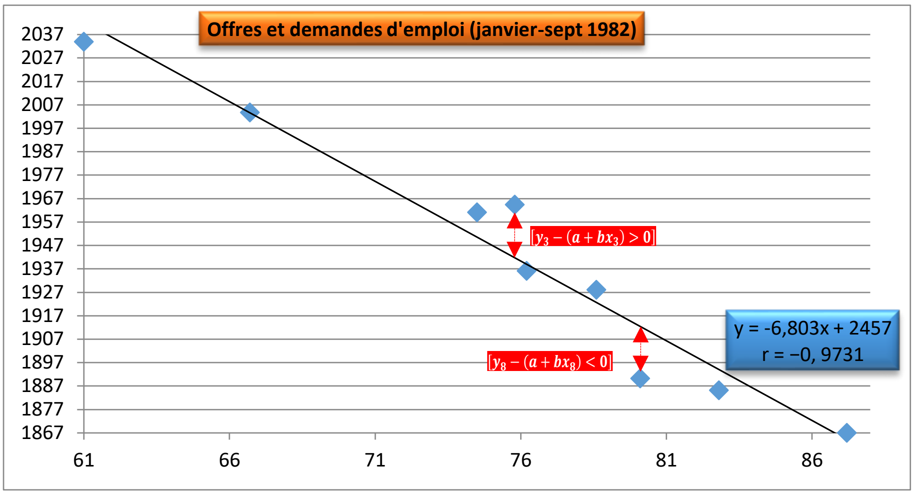
☞ *Si le coefficient de corrélation est négatif, les points sont alignés le long d'une droite décroissante.*

☞ *Si le coefficient de corrélation est nul ou proche de zéro, il n'y a pas de dépendance linéaire. On peut cependant avoir une dépendance non-linéaire avec un coefficient de corrélation nul.*

h • *Exemple :*

On donne pour les 9 premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant des emplois durables à plein temps) et de demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles, à la recherche d'un emploi *durable à plein temps). Les nombres sont exprimés en milliers.*

Offres : X	61	66,7	75,8	78,6	82,8	87,2	76,2	80,1	74,5
Demandes : Y	2034	2003,8	1964,5	1928,2	1885,3	1867,1	1936,2	1890,3	1961,2



$\mathcal{B}$– 3 • Analyse de la variance :

a • Somme des carrés totale :

On appelle somme des carrés totale la quantité : $SCT = \sum_{i=1}^{n}(y_i - \bar{Y})^2 = nS_y^2$

Elle indique la variabilité totale de Y.

La variance totale peut alors être définie par : $VT = \frac{SCT}{n} = S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2$

$\mathcal{b}$ • Somme des carrés expliqués ou de la régression:

On appelle somme des carrés expliqués la quantité : $SCE = \sum_{i=1}^{n}\left((a + bx_i) - \bar{Y}\right)^2$

Elle indique la variation de Y due à sa régression linéaire sur X

La variance expliquée peut alors être définie par : $VE = \frac{SCE}{n} = \frac{1}{n}\sum_{i=1}^{n}\left((a + bx_i) - \bar{Y}\right)^2$

c • Somme des carrés des résidus (ou résiduelle) :

On appelle somme des carrés résiduelle la quantité : $SCR = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2$

Elle indique la variabilité de Y non expliquée par le modèle.

La variance résiduelle peut alors être définie par : $VR = \dfrac{SCR}{n} = \dfrac{1}{n}\sum_{i=1}^{n}[y_i - (a + bx_i)]^2$

d • L'équation d'analyse de la variance :

$SCT = SCE + SCR$ *ou encore* $VT = VE + VR$

☑ **Preuve :**

$$SCT = \sum_{i=1}^{n}(y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left[(y_i - (a + bx_i)) - (\bar{Y} - (a + bx_i))\right]^2$$

$$= \underbrace{\sum_{i=1}^{n}\left(y_i - (a + bx_i)\right)^2}_{SCR} + \underbrace{\sum_{i=1}^{n}\left((a + bx_i) - \bar{Y}\right)^2}_{SCE} - 2\sum_{i=1}^{n}(y_i - (a + bx_i))(\bar{Y} - (a + bx_i))$$

$$SCT = SCR + SCE - 2\sum_{i=1}^{n}(y_i - (a + bx_i))(\bar{Y} - (a + bx_i))$$

$$Or \sum_{i=1}^{n}\left(y_i - \left(\underbrace{a}_{\bar{Y} - b\bar{X}} + bx_i\right)\right)\left(\bar{Y} - \left(\underbrace{a}_{\bar{Y} - b\bar{X}} + bx_i\right)\right) = \sum_{i=1}^{n}(y_i - (\bar{Y} - b\bar{X} + bx_i))(\bar{Y} - (\bar{Y} - b\bar{X} + bx_i))$$

$$= \sum_{i=1}^{n}((y_i - \bar{Y}) - b(x_i - \bar{X}))(-b(x_i - \bar{X})) = -b\underbrace{\sum_{i=1}^{n}(y_i - \bar{Y})(x_i - \bar{X})}_{nS_{x,y}} + b^2\underbrace{\sum_{i=1}^{n}(x_i - \bar{X})^2}_{nS_x^2}$$

$$= nb\left[bS_x^2 - S_{x,y}\right] = nb\underbrace{\left[\left(\frac{S_{x,y}}{S_x^2}\right)S_x^2 - S_{x,y}\right]}_{0}$$

$$D'où \sum_{i=1}^{n}(y_i - (a + bx_i))(\bar{Y} - (a + bx_i)) = 0 \text{ et } SCT = SCE + SCR$$

e • Décomposition de la variance/Coefficient de détermination :

La part de variance de Y expliquée par le modèle est toujours traduite par le coefficient de détermination : $r_{x,y}^2 = \dfrac{SCE}{SCT} = \dfrac{VE}{VT} = 1 - \dfrac{SCR}{SCT} = 1 - \dfrac{VR}{VT} \in [0, 1]$

☑ **Preuve :**

$$VE = \frac{1}{n}\sum_{i=1}^{n}\left((a + bx_i) - \bar{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left((\bar{Y} - b\bar{X} + bx_i) - \bar{Y}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(\bar{Y} + b(x_i - \bar{X}) - \bar{Y})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(b(x_i-\bar{X})\right)^2 = b^2\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{X})^2\right) = \left(\frac{S_{x,y}}{S_x^2}\right)^2 S_x^2 = \left(\frac{S_{x,y}^2}{S_x^4}\right)S_x^2 = \frac{S_{x,y}^2}{S_x^2} = \left(\frac{S_{x,y}^2}{S_y^2 S_x^2}\right)S_y^2$$

$$VE = r_{x,y}^2 S_y^2 = r_{x,y}^2 VT \Leftrightarrow r_{x,y}^2 = \frac{VE}{VT} = \frac{\frac{SCE}{n}}{\frac{SCT}{n}} = \frac{SCE}{SCT} = \frac{SCT-SCR}{SCT} 1 - \frac{SCR}{SCT}$$

☞ $SCR = SCT(1 - r_{x,y}^2)$

☞ $VR = VT(1 - r_{x,y}^2)$

☞ $SCE = nb^2S_x^2 = nbS_{x,y}$

☞ $VE = b^2S_x^2 = bS_{x,y}$

☞ *Plus le $r_{x,y}^2$ est proche de 1, meilleur est l'ajustement, la connaissance des valeurs de X permet de deviner avec précision celles de Y.*

☞ *Plus le $r_{x,y}^2$ est proche de 0, mauvais est l'ajustement, X n'apporte pas d'informations utiles sur Y.*

Au meilleur des cas	*Au pire des cas*
$SCR = 0$	$SCE = 0$
$SCT = SCE$	$SCT = SCR$
$r_{x,y}^2 = 1$	$r_{x,y}^2 = 0$
Le modèle est parfait, la droite de régression passe par tous les points du nuage.	*Le modèle est mauvais, la meilleure prédiction de Y est sa propre moyenne.*

B-4 • Rapports de corrélation :

Dans certain cas les deux variables X et Y sont liées de manière non-linéaire, ou si l'une était une variable qualitative, comment mesurer l'intensité de leur liaison ?

a • Définition :

Dans certain cas il est possible de se ramener au cas linéaire, en transformant les variables (utilisation du ln ...) si non on définit un rapports de corrélation :

☞ *de Y en X :* $\eta_{Y/X}^2 = \frac{\frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(\bar{Y}_i - \bar{Y})^2}{S_y^2}$

☞ *de X en Y :* $\eta_{X/Y}^2 = \frac{\frac{1}{n}\sum_{j=1}^{s} n_{\bullet j}(\bar{X}_j - \bar{X})^2}{S_x^2}$

b • Propriétés :

☞ $\eta^2_{Y/X} \neq \eta^2_{X/Y}$

☞ $0 \leq r^2_{x,y} \leq \eta^2_{Y/X} \leq 1 \; et \; 0 \leq r^2_{x,y} \leq \eta^2_{X/Y} \leq 1$

☞ $\eta^2_{Y/X} = 0 \Longrightarrow$ *les deux vatiables sont indépendantes*

☞ $\eta^2_{Y/X} = 1 \Longrightarrow$ *les observations sont concentrées sur la courbe de régression*

Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Économistes et Gestionnaires
Samedi 5 Janvier 2013

Exercice 9 (5 points = 1+2+1+1) :

ÉNONCÉ

Nous disposons des données statistiques suivantes ($n = 10$ *observations*)*relatives* à *deux variables X et Y conformément au tableau* n^o 4 *qui suit* :

	X	Y	X^2	Y^2	XY
	16	20	256	400	320
	18	24	324	576	432
	23	28	529	784	644
	24	22	576	484	528
	28	32	784	1024	896
	29	28	841	784	812
	26	32	676	1224	832
	31	36	961	1296	1116
	32	41	1024	1681	1312
	34	41	1156	1681	1394
$\sum$	261	304	7127	9934	8286

Questions :

1) *Calculer les valeurs moyennes des variables X et Y*
2) *Calculer la variance et l'écart type de chaque variable*
3) *Calculer* $Cov(X,Y)$
4) *Calculer le coefficient de corrélation linéaire noté R*

Corrigé

1) · $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{261}{10} = 26,1$ · $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{304}{10} = 30,4$

2) · $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = \frac{7127}{10} - (26,1)^2 = 31,49 \Rightarrow S_x = \sqrt{31,49} = 5,611$

$$S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2 = \frac{9934}{10} - (30,4)^2 = 49,24 \Rightarrow S_y = \sqrt{49,24} = 7,017$$

3) $S_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{X}\overline{Y} = \frac{8286}{10} - (26,1 \times 30,4) = 35,16$

4) $r_{x,y} = \frac{S_{x,y}}{S_x S_y} = \frac{35,16}{5,611 \times 7,017} = 0,89$

Ecole Nationale d'Administration
Concours d'Entrée au Cycle Supérieur (Économie&Gestion)
Candidats Économistes et Gestionnaires
Jeudi 29 octobre 2015

Exercice 10 *(5 points = 2+1+1+0,5+0,5) :*

ÉNONCÉ

Une entreprise envisage la fabrication d'un nouveau produit (x). Elle étudie la demande pour ce produit, afin de déterminer le prix de vente qui lui permettra de maximiser la recette.

Dans le tableau suivant figurent les résultats d'une enquête réalisée pour déterminer la demande (d) de ce nouveau produit en fonction de son prix de vente $p(x)$ en DT.

$p(x)$	200	250	300	350	450	500
$d(x)$	550	430	400	310	260	210

1) *Représenter graphiquement le nuage de points. Déterminer l'équation de la droite de Mayer. Placer cette droite sur le graphique.*
2) *Déterminer l'ajustement linéaire de d en fonction de $p(x)$ de la forme : $d(x) = a + bp(x)$ par la méthode des moindres carrées.*
   Calculer le coefficient de corrélation et le coefficient de détermination.
3) *On cherche maintenant à déterminer un ajustement de $d(x)$ en fonction de $p(x)$ de la forme : $d(x) = b[p(x)]^a$.*
   Déterminer a et b. On ramènera à un ajustement linéaire en posant $V = Ln\,(d(x))$; $B = Ln\,(b)$ et $U = Ln(p(x))$.
   Calculer le coefficient de corrélation entre U et V puis le coefficient de détermination. Interpréter ce dernier coefficient.
4) *Lequel des deux ajustements semble le plus judicieux*
5) *Estimer la demande, si le prix de vente est fixé à **400 DT**.*

Corrigé

1) *On sépare le nuage de points en deux groupes de 3 points chacun. On cherche le point*

moyen (Centre de gravité de chaque groupe)

Groupe 1 :

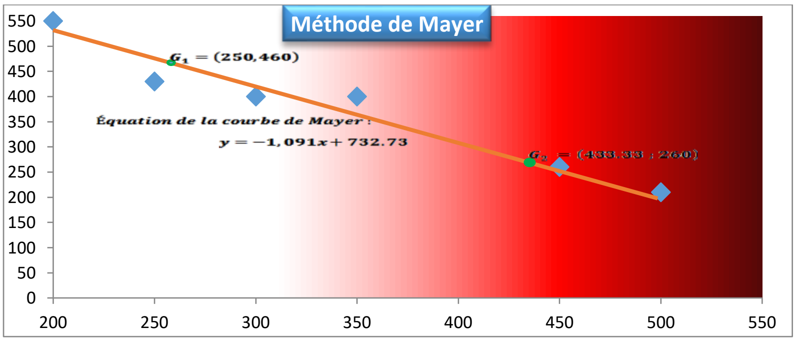
$p(x)$	200	250	300	$\begin{cases}\overline{X}_I = \dfrac{200+250+300}{3} = 250 \\ \overline{Y}_I = \dfrac{550+430+400}{3} = 460\end{cases} \Rightarrow G_1 = (250\,;460)$
$d(x)$	550	430	400	

Groupe 2 :

$p(x)$	350	450	500	$\begin{cases}\overline{X}_{II} = \dfrac{350+450+500}{3} = 433,33 \\ \overline{Y}_{II} = \dfrac{310+260+210}{3} = 260\end{cases} \Rightarrow G_2 = (433,33\,;260)$
$d(x)$	310	260	210	

L'équation de la de la droite de Mayer qui doit passer par les points G_1 et G_2:

$$(G_I, G_{II}): y - \overline{Y}_I = \left(\frac{\overline{Y}_{II} - \overline{Y}_I}{\overline{X}_{II} - \overline{X}_I}\right)(x - \overline{X}_I) \Leftrightarrow y = 460 + \frac{-200}{183,33}(x - 250) \Leftrightarrow y = -1,091x + 732,73$$



2)

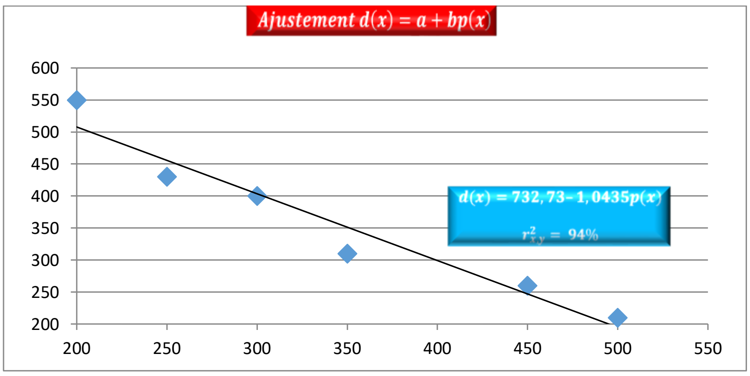
Ajustement linéaire par la méthode des moindres carrés, dont l'expression est sous

la forme : $d(x) = a + bp(x)$, ***avec,*** $b = \dfrac{\sum_{i=1}^{6}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{6}(x_i - \overline{X})^2} = -1,0435$ ***et*** $a = \overline{Y} - b\overline{X} = 732,73$

Le calcul du coefficient de corrélation avec la formule: $r_{x,y} = \dfrac{\sum_{i=1}^{6}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{6}(x_i - \overline{X})^2 \sum_{i=1}^{6}(y_i - \overline{Y})^2}}$

Donne le résultat : $r_{x,y} = -0,97$ ***ainsi x et y varient au sens contraire et*** $r_{x,y}^2 = 94\%$

c'est-à-dire 94% des variations de y sont expliquées par la droite : $d(x) = a + bp(x)$



3) $y = bx^a$ avec $ln(y) = ln(b) + a\,ln(x)$ on pose $u = ln(x)$; $v = ln(y)$ et $B = ln(b)$
on obtient la forme linéaire : $v = a\,u + B$

$u = Ln(p(x))$	5,298	5,521	5,704	5,858	6,109	6,215
$v = Ln(d(x))$	6,31	6,064	5,991	5,737	5,561	5,347

$\bar{U} = 5,784$; $\bar{V} = 5,835$; $S_u = 0,318$; $S_v = 0,323$

$\sum_{i=1}^{6}(u_i - \bar{U})^2 = 0,608$; $\sum_{i=1}^{6}(u_i - \bar{U})^2 = 0,608$; $\sum_{i=1}^{6} u_i v_i = 201,893$; $\sum_{i=1}^{6}(u_i - \bar{U})(v_i - \bar{V}) = \sum_{i=1}^{6} u_i v_i - n\bar{U}\bar{V} = -0,61$

$\sum_{i=1}^{6}(u_i - \bar{U})^2 = nS_u^2 = 0,608$; $\sum_{i=1}^{6}(v_i - \bar{V})^2 = nS_v^2 = 0,625$; $a = \frac{\sum_{i=1}^{6}(u_i - \bar{U})(v_i - \bar{V})}{\sum_{i=1}^{6}(u_i - \bar{U})^2} = -1,0032$

$B = \bar{V} - b\bar{U} = 11,6376$; $B = ln(b) \Rightarrow b = e^B = 113278,806$; $r = \frac{S_{u,v}}{S_u S_v} = -0.9891$ et $r^2 = 97.8\%$

On aura : $v = -1,0032\,u + 11,6376$ et $y = 113278,806x^{-1,0032}$

$r^2 = 97.8\% \Rightarrow 97,8\%$ des variations de y sont expliquées par la fonction de puissance :

puissance : $y = bx^a$

4) On remarque que le dernier coefficient de détermination associé à la relation de puissance est plus élevé que celui de la relation linéaire : $y = ax + b$

Prévision pour $x = 400$: Il est recommandé d'utiliser la relation de puissance

$y = 113278,8x-1,0032 \Rightarrow y = 113278,806 \times (400^{-1,0032}) = 278$

Exercice 11 :

ÉNONCÉ

Une société veut vendre des machines destinées à certaines entreprises.

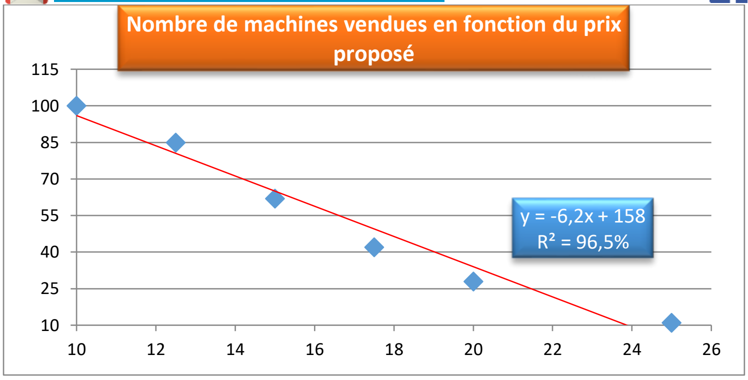
Le prix de vente minimal est fixé à 10 000 euros. Le nombre prévisible y de machines vendues, est fonction du prix proposé, en millier d'euros, x. Une enquête auprès des clients potentiels a donné les résultats suivants :

x_i	10	12,5	15	17,5	20	25
y_i	100	85	62	42	28	11

1) Représenter les six points du nuage.
2) On pose $z_i = \ln\left(\frac{y_i}{x_i - 6}\right)$. Donner les valeurs de z_i arrondies au millièmele le plus proche.
3) Donner une équation de la droite de régression de z en x ; les coefficients seront arrondis au millième le plus proche.
4) En déduire une expression approchée de y de la forme : $y = \alpha(x-6)e^{\beta x}$

Corrigé

1)



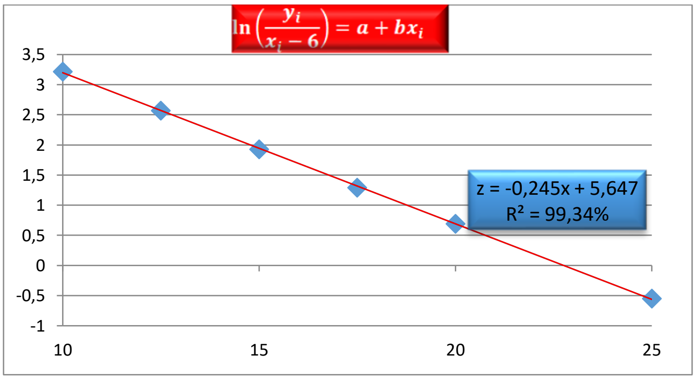
2)

x_i	10	$12,5$	15	$17,5$	20	25
y_i	100	85	62	42	28	11
$z_i = \ln\left(\frac{y_i}{x_i - 6}\right)$	$3,219$	$2,571$	$1,93$	$1,295$	$0,693$	$-0,547$

3) *On veut ajuster z en x sous la forme : $z = a + bx$. En utilisant la MCO , on obtient :*

$$\begin{cases} a = \overline{Z} - b\overline{X} = 5,647 \\ b = \dfrac{S_{x,z}}{S_x^2} = -0,245 \end{cases} \quad avec\ \overline{X} = \frac{1}{6}\sum_{i=1}^{6} x_i = 16,667 ; \overline{Z} = \frac{1}{6}\sum_{i=1}^{6} z_i = 1,527 ; \sum_{i=1}^{n} x_i^2 = 1812,5 ;$$

$$S_x^2 = \frac{1}{6}\sum_{i=1}^{n} x_i^2 - \overline{X}^2 = 24,306 ; \sum_{i=1}^{n} x_i z_i = 121,525 ; S_{x,z} = \frac{1}{6}\sum_{i=1}^{n} x_i z_i - \overline{X}\overline{Z} = -5,943 \Rightarrow z = 5,647 - 0,245x$$



4) $$\begin{cases} z = \ln\left(\frac{y}{x-6}\right) \\ z = a + bx \end{cases} \Rightarrow \ln\left(\frac{y}{x-6}\right) = a + bx \Leftrightarrow \frac{y}{x-6} = e^{(a+bx)} \Leftrightarrow y = (x-6)e^{(a+bx)}$$

$$\Leftrightarrow y = \underbrace{e^a}_{\alpha}(x-6)e^{\overbrace{b}^{\beta}x}, or\ a = 5,647\ et\ b = -0,245\ donc\ \alpha = e^{5,647} = 283,44\ et\ \beta = -0,245$$

Conclusion : $y = 283,44(x-6)e^{-0,245x}$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

Série statistique à deux variables qualitatives

Présentation des données

A– 1 • Introduction :

Lorsqu'on étudie simultanément deux variables qualitatives, il est commode de présenter les données sous forme d'une table de contingence, synthèse des observations selon les modalités des variables qu'elles ont présentées.

À partir de cette table, on définit la notion de profil, dont on se sert pour réaliser un diagramme de profils faisant bien apparaître la liaison entre les deux variables, lorsqu'il en existe une.

Pour quantifier cette liaison, l'indicateur fondamental est le khi-deux. Toutefois, comme il n'est pas d'usage commode dans la pratique, on introduit encore les indicateurs phi-deux, T de Tschuprow et C de Cramér, liés au khi-deux. Les deux derniers sont compris entre 0 et 1, et sont d'autant plus grands que la liaison est forte, ce qui facilite leur interprétation.

A– 2 • Données observées et Tableau de contingence :

a • Données observées :

Si les deux variables X et Y sont qualitatives, alors les données observées sont une suite de couples de variables $(x_1, y_1), \dots (x_i, y_j), \dots, (x_r, y_s)$ *chacune des deux variables prend comme valeurs des modalités qualitatives.*

b • Définition des profils :

On appelle $l^{\text{ième}}$ *profil-ligne l'ensemble des fréquences de la variable Y conditionnelles à la modalité* x_l *de X (c'est-à-dire définies au sein de la sous-population* C_l *de C associée à cette modalité). Il s'agit donc des quantités :* $\left\{\frac{n_{l1}}{n_{l\bullet}}, \dots, \frac{n_{lj}}{n_{l\bullet}}, \dots, \frac{n_{ls}}{n_{l\bullet}}\right\}$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

On définit de façon analogue le $h^{ième}$ *profil-colonne :* $\left\{\frac{n_{1h}}{n_{\bullet h}}, \dots, \frac{n_{ih}}{n_{\bullet h}}, \dots, \frac{n_{rh}}{n_{\bullet h}}\right\}$

c • Tableau de contingence :

Les données observées peuvent être regroupées sous la forme d'un tableau de contingence

Modalités de Y / Modalités de X	y_1	y_2	…	y_j	…	y_s	Σ
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1\bullet} = \sum_{j=1}^{s} n_{1j}$
	f_{11}	f_{12}	…	f_{1j}	…	f_{1s}	$f_{1\bullet} = \sum_{j=1}^{s} f_{1j}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2\bullet} = \sum_{j=1}^{s} n_{2j}$
	f_{21}	f_{22}	…	f_{2j}	…	f_{2s}	$f_{2\bullet} = \sum_{j=1}^{s} f_{2j}$
⋮	⋮	⋮	⋮		⋮	⋮	⋮
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{is}	$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$
	f_{i1}	f_{i2}	…	f_{ij}	…	f_{is}	$f_{i\bullet} = \sum_{j=1}^{s} f_{ij}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r\bullet} = \sum_{j=1}^{s} n_{rj}$
	f_{r1}	f_{r2}	…	f_{rj}	…	f_{rs}	$f_{r\bullet} = \sum_{j=1}^{s} f_{rj}$
Σ	$n_{\bullet 1} = \sum_{i=1}^{r} n_{i1}$	$n_{\bullet 2} = \sum_{i=1}^{r} n_{i2}$	…	$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$	…	$n_{\bullet s} = \sum_{i=1}^{r} n_{is}$	$n = n_{\bullet\bullet} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}$
	$f_{\bullet 1} = \sum_{i=1}^{r} f_{i1}$	$f_{\bullet 2} = \sum_{i=1}^{r} f_{i2}$	…	$f_{\bullet j} = \sum_{i=1}^{r} f_{ij}$	…	$f_{\bullet s} = \sum_{i=1}^{r} f_{is}$	$1 = \sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = \sum_{j=1}^{s}\sum_{i=1}^{r} f_{ij}$

☑ **Remarque :** *Toutes les notations de la distribution bidimensionnelle quantitatives restent inchangées y compris celles des distributions conditionnelles*

$\mathcal{A}$– 3 • Les représentations graphiques :

On peut envisager, dans le cas de l'étude simultanée de deux variables qualitatives, d'adapter les graphiques présentés dans le cas unidimensionnel : on découpe chaque partie (colonne, partie de barre ou secteur) représentant une modalité de l'une des variables selon les effectifs des modalités de l'autre. Mais, de façon générale, il est plus approprié de réaliser des graphiques représentant des quantités très utiles dans ce cas, que l'on appelle les profils.

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

a • Exemple : ***On s'intéresse à une éventuelle relation entre le sexe de 200 personnes et la couleur des yeux.***

Tableau des effectifs : n_{ij}

	Bleu	*Vert*	*Marron*	*Total*
Homme	10	50	20	80
Femme	20	60	40	120
Total	30	110	60	200

Tableau des fréquences : f_{ij}

	Bleu	*Vert*	*Marron*	*Total*
Homme	0,05	0,25	0,1	0,4
Femme	0,1	0,3	0,2	0,6
Total	0,15	0,55	0,3	1

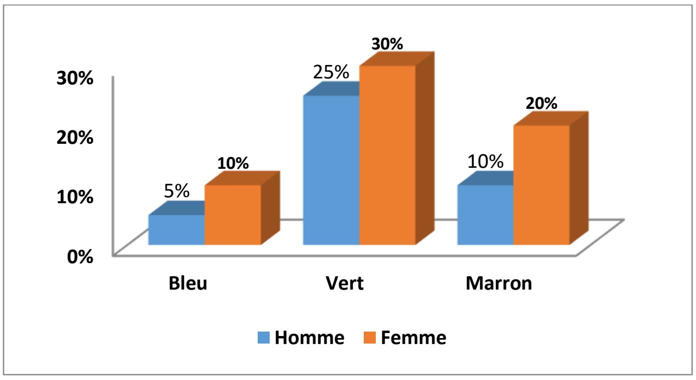


Tableau des profils lignes

	Bleu	*Vert*	*Marron*	*Total*
Homme	0,13	0,63	0,25	1
Femme	0,17	0,5	0,33	1
Total	0,15	0,55	0,3	1

Tableau des profils colonnes

	Bleu	*Vert*	*Marron*	*Total*
Homme	0,33	0,45	0,33	0,4
Femme	0,67	0,55	0,67	0,6
Total	1	1	1	1

Les indices de liaison : le khi-deux et ses dérivés

B–1 • Effectifs théoriques et khi-deux :

On cherche souvent une interaction entre des lignes et des colonnes, un lien entre les variables. Pour mettre en évidence ce lien, on construit un tableau d'effectifs théoriques qui représente la situation où les variables ne sont pas liées (indépendance). Ces effectifs théoriques sont construits de la manière suivante : $n^*_{ij} = \dfrac{n_{i\bullet} n_{\bullet j}}{n_{ij}}$

On peut établir l'équivalence des trois propriétés suivantes :

☞ $n^*_{ij} = \dfrac{n_{i\bullet} n_{\bullet j}}{n_{ij}}$ $\Leftrightarrow$ ***tous les profils-lignes sont égaux*** $\Leftrightarrow$ ***tous les profils-colonnes sont égaux***

☞ ***Les effectifs observés*** n_{ij} ***ont les mêmes marges que les effectifs théoriques*** n^*_{ij}

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

☞ *Enfin, les écarts à l'indépendance sont définis par :* $e_{ij} = n_{ij} - n_{ij}^*$

☞ *La dépendance du tableau se mesure au moyen du khi-deux défini par :*

$$\chi_{obs}^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-n_{ij}^*)^2}{n_{ij}^*} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{e_{ij}^2}{n_{ij}^*} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij}-\frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\left[\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i.}n_{.j}}\right)-1\right]$$

☞ *Le khi-deux peut être normalisé pour ne plus dépendre du nombre d'observations.*

On définit le phi-deux par : $\phi^2 = \frac{\chi_{obs}^2}{n}$ *Le* ϕ^2 *ne dépend plus du nombre d'observations.*

Il est possible de montrer que : $\phi^2 \leq \min(r-1, s-1)$

☞ *Le V de Cramer est définit par :* $V = \sqrt{\frac{\phi^2}{\min(r-1,s-1)}} = \sqrt{\frac{\chi_{obs}^2/n}{\min(r-1,s-1)}}$

Le V de Cramer est compris entre 0 *et* 1. *Il ne dépend ni de la taille de l'échantillon ni de la taille du tableau. Si* $V \cong 0$, *les deux variables sont indépendantes. Si* $V = 1$, *il existe une relation fonctionnelle entre les variables, ce qui signifie que chaque ligne et chaque colonne du tableau de contingence ne contiennent qu'un seul effectif différent de* 0 (*il faut que le tableau ait le même nombre de lignes que de colonnes*).

a • Exemple :

Tableau des effectifs théoriques : $n_{ij}^* = \frac{n_{i.}n_{.j}}{n_{ij}}$

	Bleu	*Vert*	*Marron*	*Total*
Homme	12	44	24	80
Femme	18	66	36	120
Total	30	110	60	200

Tableau des écarts à l'indépendance : $e_{ij} = n_{ij} - n_{ij}^*$

	Bleu	*Vert*	*Marron*	*Total*
Homme	−2	6	−4	0
Femme	2	−6	4	0
Total	0	0	0	0

Tableau des $\frac{e_{ij}^2}{n_{ij}^*}$

	Bleu	*Vert*	*Marron*	*Total*
Homme	0,33	0,82	0,67	1,82
Femme	0,22	0,55	0,44	1,21
Total	0,55	1,36	1,11	3,03

☞ *Le khi-deux observé vaut :* $\chi_{obs}^2 = 3,03$

☞ *Le phi-deux observé vaut :* $\phi^2 = 0,01515$

☞ *Comme le tableau a deux lignes et trois colonnes, donc,* $\min(r-1, s-1) = \min(1,2) = 1$

Le V de Cramer est égal à : $V = \sqrt{\frac{\phi^2}{1}} = \sqrt{0,01515} = 0,123$

La dépendance entre les deux variables est très faible.

Une variable quantitative et une qualitative

Présentation des données

$\mathcal{A}-1$ • Introduction :

Si X est la variable qualitative à r modalités, elle définit une partition de l'ensemble des observations en r « classes ». La classe courante, notée C_i $(i = 1, \ldots, r)$ contient les individus ayant présenté la modalité x_i de X. On peut alors définir moyenne et variance partielles de la variable quantitative Y au sein de chaque classe C_i . La façon dont les moyennes partielles varient donne une première idée de la liaison entre X et Y.

Enfin, une idée encore plus précise sur cette liaison est donnée par le rapport de corrélation, indicateur compris entre 0 et 1 et d'autant plus grand que la liaison est forte.

$\mathcal{A}-2$ • Les données :

Nous disposons toujours ici de deux variables mais, maintenant, l'une est quantitative et l'autre qualitative.

La variable qualitative est X, supposée à r modalités notées: $x_1, \ldots, x_i \ldots, x_r$

La variable quantitative est Y, de moyenne $\overline{Y}$ et de variance S_y^2. On peut ainsi repartir l'ensemble des individus observés en r parties, ou sous-ensembles, en fonction de la modalité de X présentée par chaque individu. Ainsi, nous noterons C_i l'ensemble des individus de l'échantillon ayant présenté la modalité x_i de X ; on obtient ainsi ce que l'on appelle une partition en r classes (on parle de partition lorsque chaque individu présente une modalité et une seule de la variable X). Nous noterons $n_{1\bullet}, \ldots, n_{i\bullet} \ldots, n_{r\bullet}$ les effectifs des différentes classes (avec toujours $n = \sum_{i=1}^{r} n_{i\bullet}$ est le nombre total d'individus observés).

Par exemple, avec la variable sexe, on définit deux classes : C_1 pour les hommes et C_2 pour les femmes.

$X \backslash Y$	y_1	y_2	...	y_j	...	y_s	
Modalité x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1\bullet} = \sum_{j=1}^{s} n_{1j}$
Modalité x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2\bullet} = \sum_{j=1}^{s} n_{2j}$
⋮	⋮	⋮	⋮		⋮	⋮	⋮
Modalité x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i\bullet} = \sum_{j=1}^{s} n_{ij}$
⋮	⋮	⋮	⋮		⋮	⋮	⋮
Modalité x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r\bullet} = \sum_{j=1}^{s} n_{rj}$
	$n_{\bullet 1} = \sum_{i=1}^{r} n_{i1}$	$n_{\bullet 2} = \sum_{i=1}^{r} n_{i2}$	...	$n_{\bullet j} = \sum_{i=1}^{r} n_{ij}$	...	$n_{\bullet s} = \sum_{i=1}^{r} n_{is}$	n

On peut alors définir la moyenne et la variance partielles de Y sur chaque classe C_i de la partition; nous les noterons respectivement $\overline{Y}_i$ et S_i^2 :

$$\overline{Y}_i = \frac{1}{n_{i\bullet}} \sum_{j=1}^{s} n_{ij} y_j \text{ et } S_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{s} n_{ij} \left(y_j - \overline{Y}_i\right)^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{s} n_{ij} y_j^2 - \overline{Y}_i^2$$

$(Y \mid Classe\ C_i)$	n_{ij}	$n_{ij} y_j$	$n_{ij} y_j^2$
y_1	n_{i1}	$n_{i1} y_1$	$n_{i1} y_1^2$
y_2	n_{i2}	$n_{i2} y_2$	$n_{i2} y_2^2$
⋮	⋮		
y_j	n_{ij}	$n_{ij} y_j$	$n_{ij} y_j^2$
⋮	⋮		
y_s	n_{is}	$n_{is} y_s$	$n_{is} y_s^2$
Σ	$n_{i\bullet}$	$\sum_{j=1}^{s} n_{ij} y_j = n_{i\bullet} \overline{Y}_i$	$\sum_{j=1}^{s} n_{ij} y_j^2$

Formules de décomposition-Rapport de corrélation

$\mathcal{B}$– 1 • *Formules :*

Ces formules sont nécessaires pour définir un indice de liaison entre les deux variables. Elles indiquent comment se décomposent la moyenne et la variance globales $\overline{Y}$ et S_y^2 de Y en fonction de leurs valeurs partielles $\overline{Y}_i$ et S_i^2 définies sur la partition ou la classe C_i de la variable qualitative X

☞***La moyenne globale de Y , $(\overline{Y})$ est la moyenne des « r » moyennes partielles $(\overline{Y}_i)_{1 \le i \le r}$***

: $\overline{Y} = \frac{1}{n} \sum_{i=1}^{r} n_{i\bullet} \overline{Y}_i$

☞ ***La variance globale de Y est égale à la somme de la moyenne des « r » variances*** *partielles noté* $S_R^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}S_i^2$ *est appelée variance résiduelle; (on parle encore de variance intra-classes, ou à l'intérieur des classes.) et de la variance des « r » moyennes partielles :* $S_E^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(\overline{Y}_i - \overline{Y})^2$ *appelée variance expliquée par la partition ou la classe* C_i *de la variable qualitative X; (on l'appelle aussi variance inter-classes, ou entre les classes) ou encore variance de la variable quantitative Y expliquée par la variable qualitative X :*

$$S_y^2 = \underbrace{\frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}S_i^2}_{S_R^2} + \underbrace{\frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(\overline{Y}_i - \overline{Y})^2}_{S_E^2} = \overline{S_i^2} + S_{\overline{Y}_i}^2 = S_R^2 + S_E^2$$

☑ ***Interprétation :***

La variance expliquée, S_E^2*, représente ce que serait la variance de Y si, dans chaque classe* C_i *de la partition définie par X, Y était constante et valait* $\overline{Y}_i$ *.*

De son côté, la variance résiduelle S_R^2 *représente ce qu'il reste comme variation de Y, en moyenne, dans chaque classe. Ainsi, plus* S_E^2 *est grande par rapport à* S_R^2 *plus les deux variables X et Y sont liées.*

B– 2 • Rapport de corrélation empirique :

a • Idée générale : *Le rapport de corrélation empirique offre une mesure simple de la liaison entre une variable qualitative et une variable quantitative. Considérons que la variable qualitative possède r modalités. On obtient alors une partition naturelle de notre échantillon de données en r groupes : chaque individu appartient au groupe naturellement défini par la modalité de la variable qualitative qu'il présente*

b • Aspects calculatoires : *Le théorème de Huygens (p92) est essentiel puisqu'il permet de comprendre que la variabilité totale dans un échantillon est la somme de la contribution des variations à l'intérieur des groupes et entre les groupes.*

Le poids relatif des variances intra et inter est déterminant pour comprendre la structure et la pertinence d'un découpage en groupes : si la variance intra-groupes est nettement plus élevée que la variance inter-groupes, on est dans un cas où les groupes *sont en moyenne assez semblables entre eux, mais où chacun d'eux abrite en son sein une énorme variabilité inter-individuelle. Les groupes sont donc vraisemblablement mal définis, et ne correspondent pas à une réalité (physique, biologique, sociale, ...) bien définie. À l'inverse, si la variabilité inter-groupes est nettement plus élevée que la variabilité intra, nous sommes alors en présence de groupes bien différenciés les uns des autres et bien homogènes en leur sein : le découpage en groupes est pertinent et correspond à une réalité concrète.*

L'idée du rapport de corrélation est tout simplement de mesurer le poids de la contribution de la variance inter-groupes dans la variance totale (ce qui revient donc à mesurer le poids relatif de la variance inter et de la variance intra)

☑ **Définition :** *Soient Y une variable quantitative et X une variable qualitative à « r » modalités. Ces deux variables sont mesurées sur « n » individus, et on suppose que chacune des « r » modalités de X est présente sur au moins deux individus. Les individus sont alors naturellement répartis en « r » groupes correspondant aux « r » modalités de X. Le rapport de corrélation de Y en X, noté $\eta^2_{Y/X}$, est le rapport de la variance inter sur la variance totale :*

$$\eta^2_{Y/X} = \frac{\text{Variance inter-classes}}{\text{Variance totale}} = \frac{S_E^2}{S_y^2} = \frac{\text{Variance inter-classes}}{\text{Variance inter-classes} + \text{Variance intra-classes}} = \frac{S_E^2}{S_E^2 + S_R^2}$$

c • Propriétés :

☞ $\eta^2_{Y/X}$ *n'est pas symétrique cette propriété est évidente, compte-tenu que X et Y ne sont pas de même nature :* $\eta^2_{Y/X} \neq \eta^2_{X/Y}$

☞ $0 \leq \eta^2_{Y/X} \leq 1$

☞ $\eta^2_{Y/X} = 1 \Leftrightarrow S_R^2 = 0 \Leftrightarrow S_i^2 = \frac{1}{n_{i\bullet}} \sum_{j=1}^{s} n_{ij}\left(y_j - \overline{Y}_i\right)^2 = 0, \forall i \Leftrightarrow y_j = \overline{Y}_i, \forall i$

d'après la définition de S_R^2 (la somme des carrés de ces quantités est nulle, donc chacune de ces quantités est nulle) ; par conséquent, Y est constante sur chaque classe C_i (puisque sa variance est nulle sur chacune de ces classes) ; dans un tel cas, la connaissance de X

$(donc\ de\ la\ classe\ \mathcal{C}_i$ à $laquelle\ appartient\ chaque\ individu) est\ suffisante\ pour\ connaître$

$Y\ (qui\ vaut\ \overline{Y}_i):\ il\ y\ a\ liaison\ totale\ entre\ X\ et\ Y$.

$$\eta^2_{Y/X} = 0 \Leftrightarrow S^2_E = 0 \Leftrightarrow S^2_E = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}(\overline{Y}_i - \overline{Y})^2 = 0, \forall i \Leftrightarrow \overline{Y}_i = \overline{Y}, \forall i$$

$En\ moyenne, X\ n'a\ aucune\ influence\ sur\ Y\ (puisque\ la\ valeur\ moyenne\ de\ Y\ est\ la\ même,$

$quelle\ que\ soit\ la\ modalité\ de\ X):\ il\ n'y\ a\ pas\ de\ liaison\ entre\ les\ deux\ variables.$

$On\ retiendra\ que\ plus\ \eta^2_{Y/X}\ est\ grand, plus\ la\ liaison\ entre\ X\ et\ Y\ est\ forte$

$d \bullet Test\ de\ significativité$:

$On\ admettra\ que\ sous\ l'hypothèse\ H_0: \eta^2_{Y/X} = 0\ (nullité\ du\ rapport\ de\ corrélation)$

$, la\ quantité\ F = \dfrac{(n-r)\eta^2_{Y/X}}{(r-1)\eta^2_{Y/X}} \sim \mathcal{F}\big((r-1),(n-r)\big)$, $(Loi\ de\ Fisher$ à $(r-1) et\ (n-r)\ degrés$

$de\ liberté). Pour\ déterminer\ si\ la\ valeur\ \eta^2_{Y/X}\ est\ significativement\ différente\ de\ 0$

$(avec\ un\ risque\ d'erreur\ \alpha = 5\%), il\ suffit\ donc\ de\ comparer\ la\ statistique\ de\ test\ F\ au$

$quantile\ d'ordre\ 0.95\ de\ la\ loi\ de\ Fisher$ à $(r-1) et\ (n-r)\ degrés\ de\ liberté$.

Institut de Financement du Développement du Maghreb Arabe
CONCOURS DE RECRUTEMENT DE LA XXXVIIème PROMOTION(BANQUE)
AOÛT 2017

Exercice 12 *(6 points = 1 point par question) :*

ÉNONCÉ

$La\ répartition\ statistique\ d'un\ ensemble\ n = 1000\ chèques\ (mille\ chèques) selon\ le$

$montant, noté\ X, et\ la\ région\ d'émission, notée\ R\ est\ comme\ suit$:

X \ Région	R_1	R_2
10	150	50
20	A	200
40	300	B

1) $Déterminer\ la\ relation\ entre\ A\ et\ B$
2) $Trouver\ A\ et\ B\ sachant\ que\ le\ montant\ moyen\ de\ tous\ ces\ chèques\ est\ égal$ à 28

3) *Déterminer la distribution statistique de la variable X*
4) *En déduire la variance de X*
5) *Calculer le montant moyen des chèques pour chacune des deux régions R_1 et R_2*
6) *Sur la base des calculs précédents, les caractéristiques des chèques «les montants » et les « régions d'émission » sont-elles indépendantes ? Justifier votre réponse.*

Corrigé

1) *Notons n_{ij} l'effectif conjoint qui correspond au couple $(x_i, R_j), (i,j) \in \{1,2,3\} \times \{1,2\}$,*

$n_{i\bullet}$ l'effectif marginal correspondant à x_i, $n_{\bullet j}$ l'effectif marginal correspondant à R_j et

$n = 1000$ l'effectif total

$X \backslash R$	R_1	R_2	Σ
$x_1 = 10$	$n_{11} = 150$	$n_{12} = 50$	$n_{1\bullet} = 200$
$x_2 = 20$	$n_{21} = A$	$n_{22} = 200$	$n_{2\bullet} = A + 200$
$x_3 = 40$	$n_{31} = 300$	$n_{32} = B$	$n_{3\bullet} = B + 300$
Σ	$n_{\bullet 1} = A + 450$	$n_{\bullet 2} = B + 250$	$n = 1000$

$$Or\ n = \sum_{i=1}^{3}\sum_{j=1}^{2} n_{ij} = \sum_{j=1}^{2}\sum_{i=1}^{3} n_{ij} = \sum_{i=1}^{3} n_{i\bullet} = \sum_{j=1}^{2} n_{\bullet j} \Leftrightarrow (A+450)+(B+250) = 1000$$

$$d'où\ \boxed{B = 300 - A}\quad (1)$$

2) $\bar{X} = \frac{1}{n}\sum_{i=1}^{3} n_{i\bullet}.x_i \Leftrightarrow 28 = \frac{n_{1\bullet}.x_1 + n_{2\bullet}.x_2 + n_{3\bullet}.x_3}{n}$

$$\Leftrightarrow 28 = \frac{(200 \times 10) + [(A+200)\times 20] + [(B+300)\times 40]}{1000} \Leftrightarrow \boxed{A + 2B = 500}\quad (2)$$

Avec (1) et (2), on obtient le systéme : $\begin{cases} B = 300 - A \\ A + 2B = 500 \end{cases} \Leftrightarrow \begin{cases} B = 300 - A \\ A + 600 - 2A = 500 \end{cases}$

Finalement, on trouve : $\boxed{A = 100}$ *et* $\boxed{B = 200}$

3) *Distribution statistique de la variable X :*

X	$n_{i\bullet}$	$f_{i\bullet}$	$f_{i\bullet}.x_i$	$f_{i\bullet}.x_i^2$
$x_1 = 10$	200	0,2	2	20
$x_2 = 20$	300	0,3	6	120
$x_3 = 40$	500	0,5	20	800
Σ	1000	1	$\bar{X} = 28$	$\sum_{i=1}^{3} f_{i\bullet}.x_i^2 = 940$

4) $S_x^2 = \frac{1}{n}\sum_{i=1}^{r} n_{i\bullet}.(x_i - \bar{X})^2 = \sum_{i=1}^{r} f_{i\bullet}.(x_i - \bar{X})^2 = \sum_{i=1}^{r} f_{i\bullet}.x_i^2 - \bar{X}^2 = 940 - 28^2 \Leftrightarrow \boxed{S_x^2 = 156}$

5) *Notons $\bar{X}_j$, $(j = 1,2)$ les moyenne partielles de X sur chaque région R_j de la partition*

$(X\|R_1)$	n_{i1}	$n_{i1}x_i$	$n_{i1}x_i^2$
$x_1=10$	$n_{11}=150$	1500	15000
$x_2=20$	$n_{21}=100$	2000	40000
$x_3=40$	$n_{31}=300$	12000	480000
$\sum$	$n_{\cdot 1}=550$	15500	535000

$(X\|R_2)$	n_{i2}	$n_{i2}x_i$	$n_{i2}x_i^2$
$x_1=10$	$n_{12}=50$	500	5000
$x_2=20$	$n_{22}=200$	4000	80000
$x_3=40$	$n_{32}=200$	8000	320000
$\sum$	$n_{\cdot 2}=450$	12500	405000

$\cdot\ \bar{X}_1=\frac{1}{n_{\cdot 1}}\sum_{i=1}^{3}n_{i1}x_i=\frac{15500}{550}=28,182$ $\cdot\ \bar{X}_2=\frac{1}{n_{\cdot 2}}\sum_{i=1}^{3}n_{i2}x_i=\frac{12500}{450}=27,778$

6)

▪ $1^{ère}$ ***Méthode :***

☑ ***Calculons la variance résiduelle*** $S_R^2=\frac{1}{n}\sum_{j=1}^{2}n_{\cdot j}S_j^2$ ***, variance intra-classes, ou à l'intérieur des régions., où*** S_j^2 $(j=1,2)$ ***les variances partielles de X sur chaque région*** R_j ***de la partition***

$$\cdot\ S_1^2=\frac{1}{n_{\cdot 1}}\sum_{i=1}^{3}n_{i1}(x_i-\bar{X}_1)^2=\frac{1}{n_{\cdot 1}}\sum_{i=1}^{3}n_{i1}x_i^2-\bar{X}_1^2=\frac{535000}{550}-(28,182)^2=178,512$$

$$\cdot\ S_2^2=\frac{1}{n_{\cdot 2}}\sum_{i=1}^{3}n_{i2}(x_i-\bar{X}_2)^2=\frac{1}{n_{\cdot 2}}\sum_{i=1}^{3}n_{i2}x_i^2-\bar{X}_2^2=\frac{405000}{450}-(27,778)^2=128,395$$

$$\cdot\ S_R^2=\frac{1}{n}\sum_{j=1}^{2}n_{\cdot j}S_j^2=\frac{1}{n}(n_{\cdot 1}S_1^2+n_{\cdot 2}S_2^2)=\frac{1}{1000}\Big[\underbrace{(550\times 178,512)}_{98181,818}+\underbrace{(450\times 128,395)}_{57777,778}\Big]=155,96$$

☑ ***Calculons la variance expliquée*** $S_E^2=\frac{1}{n}\sum_{j=1}^{2}n_{\cdot j}(\bar{X}_j-\bar{X})^2$ ***, variance inter-classes, ou entre les régions***

$$\cdot\ S_E^2=\frac{1}{n}\sum_{j=1}^{2}n_{\cdot j}(\bar{X}_j-\bar{X})^2=\frac{1}{n}[n_{\cdot 1}(\bar{X}_1-\bar{X})^2+n_{\cdot 2}(\bar{X}_2-\bar{X})^2]$$

$$S_E^2=\frac{1}{1000}\Big[\underbrace{(550\times(28,182-28)^2)}_{18,182}+\underbrace{(450\times(27,778-28)^2)}_{22,222}\Big]=0,04$$

On vérifie bien que S_x^2 ***, la variance globale de la variable X est la somme de la variance résiduelle et la variance expliquée :*** $\underbrace{S_x^2}_{156}=\underbrace{S_R^2}_{155,96}+\underbrace{S_E^2}_{0,04}$

☑ ***Calculons Le rapport de corrélation empirique de X en R*** :

$$\eta^2_{X/Y} = \frac{S_E^2}{S_x^2} = \frac{S_E^2}{S_E^2 + S_R^2} = \frac{0,04}{156} = 0,026\% \cong 0\%$$

La répartition des montants des chèques est indépendante des régions d'émission, on pourra dire que la répartition sur des régions d'émission n'est pas pertinente et elle est vraisemblablement mal définie

▪ **2ème *Méthode (plus recommandée)*** :

$$\overline{X}_1 = 28,182 \,; \overline{X}_2 = 27,778 \text{ et } \overline{X} = 28 \Rightarrow \overline{X}_1 \cong \overline{X}_2 \cong \overline{X} \Rightarrow S_E^2 = \frac{1}{n}\sum_{j=1}^{2} n_{.j}\left(\overline{X}_j - \overline{X}\right)^2 \cong 0 \Rightarrow \eta^2_{X/Y} \cong 0\%$$

En moyenne, R_j (régions d'émission) n'a aucune influence sur X (puisque la valeur moyenne de X est la même, quelle que soit la région d'émission: il n'y a pas de liaison entre les deux variables.

Les indices statistiques

Taux de croissance et indices élémentaires

$\mathcal{A}$– 1 • Calcul des taux de croissance :

a • Temps continu : *Soit $y = f(t)$, une variable dont l'évolution au cours du temps est décrite par la fonction f.*

☑ **Taux de croissance d'une variable :** *La dérivée de la variable y par rapport au temps est par convention notée : $y' = \dfrac{\partial f(t)}{\partial t}$ et indique la variation instantanée de y en t.*

Le taux de croissance de cette fonction, noté $\mathcal{T}_y$ est défini comme le rapport de cette variation temporelle à la valeur de la fonction f en un instant t du temps, soit :

$$\mathcal{T}_y = \frac{y'}{y} = \frac{\partial f(t)/\partial t}{f(t)}$$

☑ **Taux de croissance d'un produit de variables :** *Dans la plupart des applications économiques, la fonction f s'écrit souvent comme le produit de plusieurs variables qui dépendent ou pas du temps: $f(t) = kh(t)^{\alpha} l(t)^{\beta}$ (1) . Pour calculer le taux de croissance de y en fonction du taux de croissance de ces autres variables, on utilise la propriété suivante :* $\mathcal{T}_y = \dfrac{y'}{y} = \dfrac{\partial \ln f(t)/\partial t}{f(t)}$

Autrement dit, le taux de croissance d'une variable qui dépend du temps n'est rien d'autre que la dérivée du logarithme de cette variable par rapport au temps.

Utilisons cette propriété pour calculer le taux de croissance de f en fonction du taux de croissance des variables du membre de droite de l'équation (1). On commence par calculer le logarithme de y : $\ln y = \ln\big(f(t)\big) = \ln k + \alpha \ln\big(h(t)\big) + \beta \ln\big(l(t)\big)$

Dérivons cette expression par rapport au temps :

$$\frac{\partial \ln(f(t))}{\partial t} = \frac{\partial \ln k}{\partial t} + \alpha \frac{\partial \ln(h(t))}{\partial t} + \beta \frac{\partial \ln(l(t))}{\partial t} \text{ , soit : } \mathcal{T}_y = \alpha \mathcal{T}_h + \beta \mathcal{T}_l$$

où $\mathcal{T}_h$ et $\mathcal{T}_l$ désignent les taux de croissance des variables dont l'évolution au cours du temps est décrite par les fonctions h et l respectivement.

☑ **Propriété importante :** *Si une variable y croît au taux constant a, alors :*

$y = f(t) = e^{at} f(0)$

La démonstration de cette propriété est une application directe de la méthode de résolution des équations différentielles linéaires du premier ordre.

La solution générale d'une équation du type : $y' = ay$ est $y(t) = ke^{at}$.

Pour trouver la valeur de k, il suffit d'annuler cette équation, ce qui donne : $k = y(0)$.

• Temps discret : *Dans de nombreux modèles, on ne considère pas l'évolution d'une variable y en chaque point du temps, mais plutôt à intervalles réguliers: $t, t+1, t+2$, etc.*

☑ **Taux de croissance d'une variable :** *Dans ce cas, on note y_t la valeur de la variable y à l'instant t. Son taux de croissance $\mathcal{T}_y$ entre t et $t+1$ est donné par la formule :* $\mathcal{T}_y = \frac{y_{t+1} - y_t}{y_t} = \frac{\Delta y_t}{y_t} = \frac{y_{t+1}}{y_t} - 1$

☑ **Taux de croissance d'un produit de variables :** *Supposons que la variable y soit elle-même le produit d'autres variables k (constante au cours du temps), h_t et l_t (variablesau cours du temps) :* $y_t = k h_t^{\alpha} l_t^{\beta}$

Pour calculer le taux de croissance de y en fonction du taux de croissance de ces autres variables, on utilise l'approximation suivante : $\mathcal{T}_y = \frac{y_{t+1} - y_t}{y_t} = \frac{y_{t+1}}{y_t} - 1 \approx \ln y_{t+1} - \ln y_t$

puisque si y_{t+1} est proche de y_t alors $\frac{y_{t+1}}{y_t} \to 1$ et $\underbrace{\ln\left(\frac{y_{t+1}}{y_t}\right)}_{\ln y_{t+1} - \ln y_t} \sim \underbrace{\frac{y_{t+1}}{y_t} - 1}_{\mathcal{T}_y}$

Cette approximation est une application directe de la propriété : $\left(\ln(x) \sim_1 (x-1)\right)$

En utilisant cette approximation et en notant $(\mathcal{T}_h \approx \ln(h_{t+1}) - \ln(h_t))$ *et* $(\mathcal{T}_l \approx \ln(l_{t+1}) - \ln(l_t))$ *les taux de croissance respectifs de h_t et l_t, on a :*

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

$$\begin{cases} \mathcal{T}_y \approx \ln y_{t+1} - \ln y_t \\ \ln y_{t+1} = \ln\left(k h_{t+1}^{\alpha} l_{t+1}^{\beta}\right) = \ln(k) + \alpha \ln(h_{t+1}) + \beta \ln(l_{t+1}) \\ ln\, y_t = ln\left(k h_t^{\alpha} l_t^{\beta}\right) = ln(k) + \alpha\, ln(h_t) + \beta\, ln(l_t) \end{cases}$$

d'où : $\mathcal{T}_y \approx \ln y_{t+1} - \ln y_t = \alpha \mathcal{T}_h + \beta \mathcal{T}_l$

☞ $Si\ y_t = h_t l_t\ , alors\ \mathcal{T}_y = (1 + \mathcal{T}_h)(1 + \mathcal{T}_l) - 1\ ou\ encore\ \mathcal{T}_y \approx \mathcal{T}_h + \mathcal{T}_l$

☞ $Si\ y_t = \dfrac{h_t}{l_t}\ , alors\ \mathcal{T}_y = \dfrac{1 + \mathcal{T}_h}{1 + \mathcal{T}_l} - 1\ ou\ encore\ \mathcal{T}_y \approx \mathcal{T}_h - \mathcal{T}_l$

☑ ***Propriété importante*** *Si une variable y en temps discret croît au taux constant,* $\mathcal{T}_y = r$, *alors quantité* y_t *est augmentée de* $\mathcal{T}_y y_t = r y_t$ *et devient :*

$y_{t+1} = y_t + r y_t = (1 + r) y_t$ *sa valeur à la date t est donnée par la formule:*

$y_{t+1} = \left(1 + \mathcal{T}_y\right) y_t = (1 + r) y_t$

Le terme $(1 + r)$ *est appelé le coefficient multiplicateur associé au taux de variation r.*

☞ ***Remarque*** : *le nombre r est positif dans le cas d'une grandeur qui croît etnégatif dans le cas d'une grandeur qui décroît.*

Appliquer (t) *fois de suite un taux de variation r, revient à multiplier* (t) *fois par le coefficient multiplicateur* $1 + r$ *. Cela revient finalement à multiplier par* $(1 + r)^t$

Si on note y_t *la valeur au temps* (t) *d'une quantité qui augmente de r%, on a la formule :*

$y_t = y_0 (1 + r)^t$

c • Taux global et taux moyen : *Lorsqu'une grandeur varie sur une période de t années, on peut calculer son taux de croissance global sur cette période.*

Si elle passe de la valeur y_0 *à la valeur* y_t *, ce taux global est de :* $\mathcal{T}_G = \dfrac{y_t - y_0}{y_0} = \dfrac{y_t}{y_0} - 1$

Le rapport $\dfrac{y_t}{y_0}$ *apparaît donc comme le coefficient multiplicateur sur la période globale puisque :* $\dfrac{y_t}{y_0} = 1 + \mathcal{T}_G$

☑ ***Définition (taux moyen)***: *Le taux d'accroissement annuel moyen est le taux qui donnerait le même taux global au bout de la même période de t années.*

On le calcule selon la formule : $\mathcal{T}_M = \sqrt[t]{y_t / y_0} - 1 = \sqrt[t]{1 + \mathcal{T}_G} - 1$

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

$\mathcal{A}-2$ • Indices élementaires :

a • Définition : *L'indice est le rapport de deux valeurs différentes d'une même variable.*

La notion d'indice permet de mesurer l'évolution d'une grandeur dans deux "contextes" différents : en général, on s'intéresse à une grandeur à deux dates différentes, mais cela peut aussi correspondre à l'observation de cette grandeur dans deux régions différentes. Dans le premier cas, on parle d'indice temporel ou chronologique, dans le second cas d'un indice spatial ou régional.

Soit g_t la valeur d'une grandeur G à la date t et $g_{t'}$ sa valeur à l'époque t'.

On définit l'indice élémentaire par l'expression : $i_{t/t'} = \dfrac{g_t}{g_{t'}} = (1 + \mathcal{T}_G)$

Les indices sont souvent exprimés (comme les pourcentages) sur une base de 100.

Dans ce cas, on écrit : $I_{t/t'} = 100 \times \dfrac{g_t}{g_{t'}} = 100 \times i_{t/t'}$

☑ **Remarques :**

☞ *Dans les études d'indices, on choisit en général une date initiale ou date de référence qualifiée de temps zéro" et on établit les autres indices par rapport à cette date.*

L'indice valeur 100 à la date initiale est : $I_{t/0} = 100 \times \dfrac{g_t}{g_0} = 100 \times i_{t/0} = 100 \times (1 + \mathcal{T}_G)$

☞ *Le choix de la base est arbitraire et on pourrait aussi bien prendre une base 1 ou une base 1000. La plupart des formules qui vont suivre seront souvent définies à un multiple de la base près, en général à un multiple de 100 près.*

☞ **Pourcentage de variation :** $PV_{t/0} = 100 \times \mathcal{T}_G = 100 \times \left(\dfrac{g_t - g_0}{g_0}\right) = I_{t/0} - 100$

☞ **Taux d'accroissemententre les dates t et t + 1 d'une variable y :**

$$\mathcal{T}_y = \frac{y_{t+1} - y_t}{y_t} = \frac{\Delta y_t}{y_t} = \frac{y_{t+1}}{y_t} - 1 = i_{t+1/t}(y) - 1$$

$\mathcal{b}$ • Indices particuliers (Indices de consommation):

☑ **Indice des prix :** $I_{t/0}(p) = 100 \times \dfrac{p_t}{p_0} = 100 \times i_{t/0}(p)$

☑ **Indice des quantités** : $I_{t/0}(q) = 100 \times \frac{q_t}{q_0} = 100 \times i_{t/0}(q)$

☑ **Indice des valeurs** : $I_{t/0}(v) = 100 \times \frac{v_t}{v_0} = 100 \times i_{t/0}(v)$

c • Propriétés des indices élementaires :

☑ **Identité** : $\begin{cases} i_{t/t} = \frac{g_t}{g_t} = 1 \\ I_{t/t} = 100 \times \frac{g_t}{g_t} = 100 \end{cases}$

☑ **Homogénéité** : *L'indice est indépendant des unités de mesure*

☑ **Réversibilité** : $\begin{cases} i_{t/0} = \frac{1}{i_{0/t}} \\ I_{t/0} = \frac{10^4}{I_{0/t}} \end{cases}$

☑ **Transitivité (Transférabilité ou enchaînable)** : $\begin{cases} i_{t/u} \times i_{u/v} = i_{t/v} \\ I_{t/u} \times I_{u/v} = 100 \times I_{t/v} \end{cases}$

Grâce à cette propriété, pour comparer deux grandeurs à deux dates différentes, il suffit de faire le quotient de leurs indices à ces deux dates : $I_{t/u} = 100 \times \frac{I_{t/v}}{I_{u/v}}$

en particulier pour $v = 0$: $I_{t/u} = 100 \times \frac{I_{t/0}}{I_{u/0}}$

☑ **Circularité** : $\begin{cases} i_{0/t} \times i_{t/u} \times i_{u/0} = 1 \\ I_{0/t} \times I_{t/u} \times I_{u/0} = 10^6 \end{cases}$

☑ **Multiplication** : *Soit p et q les valeurs de deux grandeurs dont le produit (pq) a un sens, l'indice élémentaire jouit de la propriété de multiplication :*

$$\begin{cases} i_{t/0}(pq) = i_{t/0}(p)i_{t/0}(q) \\ 100 \times I_{t/0}(pq) = I_{t/0}(p)I_{t/0}(q) \end{cases}$$

Cette propriete peut se concevoir pour plus de deux grandeurs

☑ **Remarques** : $(v = p \times q)$

$valeur = prix \times quantité \Leftrightarrow dépense = prix \times volume$

Exercice 13 *(7 points = 1 point par question) :*

ÉNONCÉ

Considérons deux variables quantitatives X_t et Y_t observées à des périodes successives. On note RX_t et RY_t respectivement les taux de croissance entre la période $t-1$ et la période t de X_t et Y_t. On s'intéresse à la somme et le produit de ces deux variables :
$S_t = X_t + Y_t$ et $P_t = X_t Y_t$

1)
   i. *Rappeler l'expression de RX_t en fonction X_t et de X_{t-1}. Exprimer le rapport $\dfrac{X_t}{X_{t-1}}$ en fonction de RX_t*
   ii. *Quelle serait l'expression de X_t dans le cas particulier où RX_t est une constante c indépendante du temps ?*

2)
   i. *Exprimer RP_t le taux de croissance du produit P_t en fonction de RX_t et RY_t*
   ii. *interpréter le résultat précédent si l'on considère X_t comme un prix et Y_t comme une quantité*
   iii. *En déduire de la question 2)-i. le taux de croissance de X_t^2 en fonction de RX_t*

3)
   i. *Exprimer RS_t le taux de croissance de la somme S_t en fonction de RX_t et RY_t et de X_{t-1} et Y_{t-1}*
   ii. *Que devient le résultat précèdent si l'on a une somme de trois variables ? Justifier vos propos.*

Corrigé

1)
   i.

$$RX_t = \frac{\Delta X_t}{X_{t-1}} = \frac{X_t - X_{t-1}}{X_{t-1}} = \frac{X_t}{X_{t-1}} - 1 \Leftrightarrow \boxed{\frac{X_t}{X_{t-1}} = 1 + RX_t}$$

De même, on vérifie bien que : $\boxed{\dfrac{Y_t}{Y_{t-1}} = 1 + RY_t}$

   ii. *Si X_t augmente constamment (indépendamment du temps) de $c\%$ $(RX_t = c)$*

$On\ obtient, alors : \frac{X_t}{X_{t-1}} = 1 + RX_t = 1 + c \Leftrightarrow X_t = (1+c)X_{t-1}$

$On\ suppose\ que\ X_t \neq 0$ c-à-d. $RX_t \neq -1\ ou\ encore\ c \neq -1\ et\ notons\ X_0\ la\ valeur\ de\ la$

$variable\ X$ à $la\ date\ initiale($ $de\ r$éfé$rence)$.

$Ainsi\ la\ relation\ de\ r$é$currence\ X_t = (1+c)X_{t-1}\ d$é$finie\ une\ suite\ g$é$om$é$trique\ de\ raison$

$non\ nulle\ (1+c)\ et\ de\ premier\ terme\ \ X_0\ et\ dont\ le\ terme\ g$é$n$é$ral\ nous\ donne$

$l'expression\ de\ X_t\ :\ \boxed{X_t = X_0(1+c)^t = X_0 e^{t\ln(1+c)}}$

2)

i. $$RP_t = \frac{\Delta P_t}{P_{t-1}} = \frac{P_t - P_{t-1}}{P_{t-1}} = \frac{P_t}{P_{t-1}} - 1 = \frac{X_tY_t}{X_{t-1}Y_{t-1}} - 1 = \underbrace{\left(\frac{X_t}{X_{t-1}}\right)}_{1+RX_t}\underbrace{\left(\frac{Y_t}{Y_{t-1}}\right)}_{1+RY_t} - 1$$

$D'o$ù $\boxed{(1+RP_t) = (1+RX_t)(1+RY_t) Ou\ encore: RP_t = RX_t + RY_t + RX_tRY_t}$

ii. $Notons\ respectivement : \frac{X_t}{X_{t-1}} = i_{t/t-1}(p)\ et\ \frac{Y_t}{Y_{t-1}} = i_{t/t-1}(q)$, $les\ indices\ des$ $prix\ de\ la\ grandeur\ X\ etles\ indices\ des\ prix\ de\ la\ grandeur\ X\ et\ des\ quantit$é$s\ de\ la$ $grandeur\ Y\ entre\ les\ dates\ t-1\ et\ t$.

$$Or\ \frac{X_t}{X_{t-1}} = 1 + RX_t\ et\ \frac{Y_t}{Y_{t-1}} = 1 + RY_t \Rightarrow \begin{cases} i_{t/t-1}(p) = 1 + RX_t \\ i_{t/t-1}(q) = 1 + RY_t \end{cases}$$

$D'autre\ part, P_t = X_tY_t\ donc\ la\ grandeur\ P\ n'est\ autre\ que\ le\ volume\ (ou\ la\ valeur)d'un$ $bien\ dont\ le\ prix\ est\ repr$é$sent$é $par\ la\ grandeur\ X\ et\ la\ quantit$é $par\ la\ grandeur\ Y$

$$On\ aura\ aussi\ \frac{P_t}{P_{t-1}} = i_{t/t-1}(v), \frac{P_t}{P_{t-1}} = 1 + PX_t\ et\ \text{é}videmment\ i_{t/t-1}(v) = 1 + RP_t$$

$En\ effet\ (1+RP_t) = (1+RX_t)(1+RY_t) \Leftrightarrow \boxed{i_{t/t-1}(v) = i_{t/t-1}(p) \times i_{t/t-1}(q)}$

$Elle\ d$é$finit\ l'une\ des\ propri$é$t$é$s\ des\ indices$ élé$mentaires\ (multiplication)tant\ que\ le$ $produit\ P_t = X_tY_t$ à $un\ sens\ (volume\ )$

iii. $On\ a :\ 1 + RP_t = (1+RX_t)(1+RY_t)\ or\ si\ X_t = Y_t$, $alors\ P_t = X_t^2$

$On\ obtient\ par\ la\ suite : 1 + RX_t^2 = (1+RX_t)^2 \Leftrightarrow 1 + RX_t^2 = 1 + (RX_t)^2 + 2RX_t$

$D'o$ù: $\boxed{RX_t^2 = (RX_t)^2 + 2RX_t}$

3)

i. $$RS_t = \frac{\Delta S_t}{S_{t-1}} = \frac{S_t - S_{t-1}}{S_{t-1}} = \frac{S_t}{S_{t-1}} - 1 = \frac{X_t + Y_t}{X_{t-1} + Y_{t-1}} - 1$$

$$Or \begin{cases} \dfrac{X_t}{X_{t-1}} = 1 + RX_t \\ \dfrac{Y_t}{Y_{t-1}} = 1 + RY_t \end{cases} \Leftrightarrow \begin{cases} X_t = (1 + RX_t)X_{t-1} \\ Y_t = (1 + RY_t)Y_{t-1} \end{cases} \Leftrightarrow RS_t = \frac{(1 + RX_t)X_{t-1} + (1 + RY_t)Y_{t-1}}{X_{t-1} + Y_{t-1}} - 1$$

$$RS_t = \frac{X_{t-1} + Y_{t-1} + RX_tX_{t-1} + RY_tY_{t-1}}{X_{t-1} + Y_{t-1}} - 1 = 1 + \frac{RX_tX_{t-1} + RY_tY_{t-1}}{X_{t-1} + Y_{t-1}} - 1$$

$$D'où \boxed{RS_t = \left(\frac{X_{t-1}}{X_{t-1} + Y_{t-1}}\right) RX_t + \left(\frac{Y_{t-1}}{X_{t-1} + Y_{t-1}}\right) RY_t\ ;}$$

$\boxed{\textit{Le taux de croissance de la somme est la moyenne pondérée des deux taux de croissance.}}$

ii. *Étant donné trois variables quantitatives* X_t, Y_t, Z_t *et leur somme* $S_t = X_t + Y_t + Z_t$

On se propose d'exprimer $RS_t = \dfrac{\Delta S_t}{S_{t-1}} = \dfrac{S_t - S_{t-1}}{S_{t-1}} = \dfrac{S_t}{S_{t-1}} - 1 = \dfrac{X_t + Y_t + Z_t}{X_{t-1} + Y_{t-1} + Z_{t-1}} - 1$

en fonction de $RX_t, RY_t, RZ_t, X_{t-1}, Y_{t-1}$ *et* Z_{t-1} :

$$Or \begin{cases} X_t/X_{t-1} = 1 + RX_t \\ Y_t/Y_{t-1} = 1 + RY_t \\ Z_t/Z_{t-1} = 1 + RZ_t \end{cases} \Leftrightarrow \begin{cases} X_t = (1 + RX_t)X_{t-1} \\ Y_t = (1 + RY_t)Y_{t-1} \\ Z_t = (1 + RZ_t)Z_{t-1} \end{cases}$$

$$\Leftrightarrow RS_t = \frac{(1 + RX_t)X_{t-1} + (1 + RY_t)Y_{t-1} + (1 + RY_t)Y_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}} - 1$$

$$\Leftrightarrow RS_t = \frac{X_{t-1} + Y_{t-1} + Z_{t-1} + RX_tX_{t-1} + RY_tY_{t-1} + RZ_tZ_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}} - 1$$

$$\Leftrightarrow RS_t = 1 + \frac{RX_tX_{t-1} + RY_tY_{t-1} + RZ_tZ_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}} - 1$$

$$D'où \boxed{RS_t = \left(\frac{X_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}}\right) RX_t + \left(\frac{Y_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}}\right) RY_t + \left(\frac{Z_{t-1}}{X_{t-1} + Y_{t-1} + Z_{t-1}}\right) RZ_t}$$

$\boxed{\textit{Le taux de croissance de la somme est la moyenne pondérée des deux taux de croissance.}}$

Exercice 14 :

ÉNONCÉ

Une entreprise fabrique deux biens en quantités q_1 *et* q_2 *qu'elle vend au prix respectifs de* p_1 *et* p_2 *. On sait que la recette procurée par le bien 2 est le double de celle procurée par le bien 1.*

1) *Au cours d'une année, les prix* p_1 *et* p_2 *augmentent respectivement de* $1,5\%$ *et de* 2% *tandis que les ventes des deux biens augmentent de* 5% *et* 3% *respectivement. Calculer le taux de variation de la recette totale de l'entreprise.*
2) *Quel est le rapport, au bout d'un an, entre les recettes des deux biens ?*
3) *Au cours de l'année suivante, les ventes du bien 2 ont chuté de* 10% *tandis que le prix* p_2 *augmentait de* 5%. *En même temps, les ventes du bien 1 ont progressé de* 10%. *Quelle variation du prix* p_1 *ferait en sorte que la recette totale reste inchangée ?*
4) *Quel est, au bout des deux années, le rapport entre les recettes des deux biens ?*

Corrigé

1) *Notons* r_1 *et* r_2 *les recettes procurées par les deux biens. La recette totale au départ est de* $r = r_1 + r_2$, *elle dépend des quantités vendues et du prix de vente de chaque bien :*

$$\begin{cases} r_1 = p_1 q_1 \\ r_2 = p_2 q_2 \end{cases}$$

Soient p'_1, p'_2, q'_1, q'_2 *les prix et des quantités des deux biens au bout d'un an ;* $\mathcal{T}_{p_1}, \mathcal{T}_{p_2}, \mathcal{T}_{q_1}$ *et* $\mathcal{T}_{q_2}$ *les taux de croissance respectifs .*

$$On\ a : \mathcal{T}_{p_1} = \frac{p'_1 - p_1}{p_1} = \frac{p'_1}{p_1} - 1 \Leftrightarrow \frac{p'_1}{p_1} = 1 + \mathcal{T}_{p_1} \Leftrightarrow p'_1 = (1 + \mathcal{T}_{p_1})p_1 = (100\% + 1,5\%)p_1$$

$$p'_1 = 1,015p_1 \ et \ q'_1 = (1 + \mathcal{T}_{q_1})p_1 = (100\% + 5\%)q_1 = 1,05q_1$$

$$p'_2 = (1 + \mathcal{T}_{p_2})p_2 = (100\% + 2\%)p_2 = 1,02p_2 \ et \ q'_2 = (1 + \mathcal{T}_{q_2})q_2 = (100\% + 3\%)q_2 = 1,03q_2$$

On obtient ainsi les recettes des deux biens au bout d'un an :

$$\begin{cases} r'_1 = p'_1 q'_1 = 1,015p_1 \times 1,05q_1 = 1,06575p_1q_1 = 1,06575r_1 \\ r'_2 = p'_2 q'_2 = 1,02p_2 \times 1,03q_2 = 1,0506p_2q_2 = 1,0506r_2 \end{cases}$$

La nouvelle recette est : $r' = r'_1 + r'_2 = 1,06575r_1 + 1,0506r_2$

Le taux de variation de la recette totale de l'entreprise est :

$$\mathcal{T}_r = \frac{r' - r}{r} = \frac{(1,06575r_1 + 1,0506r_2) - (r_1 + r_2)}{r_1 + r_2} = \frac{0,06575r_1 + 0,0506r_2}{r_1 + r_2}$$

Or la recette procurée par le bien 2 est le double de celle procurée par le bien 1 $\Rightarrow r_2 = 2r_1$

$$Par\ la\ suite, \mathcal{T}_r = \frac{0,06575r_1 + (0,0506 \times 2r_1)}{r_1 + 2r_1} = \frac{0,16695r_1}{3r_1} = 0,05565$$

$$\boxed{\mathcal{T}_r \approx 5,57\%}$$

2) *Soit* γ' *le nouveau rapport entre les recettes des deux biens et* $\gamma = \frac{r_2}{r_1} = 2$,

$le\ rapport\ au\ départ :\ \gamma' = \dfrac{r'_2}{r'_1} = \dfrac{1,0506r_2}{1,06575r_1} = 0,9858\dfrac{r_2}{r_1} = 0,9858\gamma = 1,9716$

Le rapport a un peu baissé : il est passé de 2 *à* $1,9716$ *, soit une baisse de* $1,42\%$

3) *Au cours de l'année suivante :*

· *Les ventes du bien 2 ont chuté de* 10% : $q''_2 = (1 + \mathcal{T}'_{q_2})q'_2 = (100\% - 10\%)q'_2 = 0,9q'_2$

· *Les prix du bien 2 augmentait de* 5%: $p''_2 = (1 + \mathcal{T}'_{p_2})p'_2 = (100\% + 5\%)p'_2 = 1,05p'_2$

· *Les ventes du bien 1 ont progressé de* 10% : $q''_1 = (1 + \mathcal{T}'_{q_1})q'_1 = (100\% + 10\%)q'_1 = 1,1q'_1$

· *Les nouvelles recettes des deux biens :*

$$\begin{cases} r''_1 = p''_1q''_1 = (1 + \mathcal{T}'_{p_1})p'_1 \times 1,1q'_1 = 1,1(1 + \mathcal{T}'_{p_1})p'_1q'_1 = 1,1(1 + \mathcal{T}'_{p_1})r'_1 \\ r''_2 = p''_2q''_2 = 1,05p'_2 \times 0,9q'_2 = 0,945p'_2q'_2 = 0,945r'_2 \end{cases}$$

La nouvelle recette est : $r'' = r''_1 + r''_2 = 1,1(1 + \mathcal{T}'_{p_1})r'_1 + 0,945r'_2$

On a vu à la question précédente que $\gamma' = \dfrac{r'_2}{r'_1} = 1,9716 \Rightarrow r'_2 = 1,9716r'_1$

Ce qui donne : $r'' = 1,1(1 + \mathcal{T}'_{p_1})r'_1 + 0,945r'_2 = 1,1(1 + \mathcal{T}'_{p_1})r'_1 + (0,945 \times 1,9716r'_1)$

$r'' = 1,1\mathcal{T}'_{p_1}r'_1 + 1,1r'_1 + 1,863162r'_1 = 1,1\mathcal{T}'_{p_1}r'_1 + 2,96316r'_1$

$Or\ r'_1 = 1,06575r_1 \Rightarrow r'' = (1,1 \times 1,06575\mathcal{T}'_{p_1}r_1) + (2,96316 \times 1,06575r_1)$

$r'_1 = 1,172325\mathcal{T}'_{p_1}r_1 + 3,15798777r_1 = (1,172325\mathcal{T}'_{p_1} + 3,15798777)r_1$

$D'autrepart, r' = r'_1 + r'_2 = 1,06575r_1 + 1,0506\underbrace{r_2}_{2r_1} = 3,16695r_1$

La recette totale reste inchangée $\Leftrightarrow r' = r'' \Leftrightarrow 3,16695r_1 = (1,172325\mathcal{T}'_{p_1} + 3,15798777)r_1$

$\Leftrightarrow 3,16695 = 1,172325\mathcal{T}'_{p_1} + 3,15798777 \Leftrightarrow \mathcal{T}'_{p_1} = \dfrac{3,16695 - 3,15798777}{1,172325} = 0,0076$

$\boxed{\mathcal{T}'_{p_1} = 0,76\%}$

4) *Soit* γ'' *le rapport entre les recettes des deux biens au bout des deux années*

$$\gamma'' = \frac{r''_2}{r''_1} = \frac{0,945r'_2}{1,1(1 + \mathcal{T}'_{p_1})r'_1} = \frac{0,945r'_2}{1,1 \times (100\% + 0,76\%)r'_1} = \frac{0,945r'_2}{1,10836r'_1} = 0,8526\underbrace{(r'_2/r'_1)}_{\gamma'=1,9716}$$

$$\boxed{\gamma'' = \frac{r''_2}{r''_1} = 1,68}$$

Indices synthétiques simples

Un indice est dit synthétique lorsqu'il porte sur plusieurs grandeurs de même nature à la fois.

Soit n valeurs de grandeurs de même nature : $g^1, g^2, \ldots, g^n$ *, on peut donner une même importance à ces grandeurs, on aura alors un indice synthétique simple, en leur affectant un coefficient de pondération on obtiendera alors un indice pondéré.*

B-1 • Indice de Bradstreet (ou indice des moyennes arithmétiques) :

$$B_{t/0}(g) = \frac{g_t^1 + g_t^2 + \cdots + g_t^n}{g_0^1 + g_0^2 + \cdots + g_0^n} = \frac{\sum_{k=1}^n g_t^k}{\sum_{k=1}^n g_0^k}$$

B-2 • Moyenne des indices :

Pour chaque grandeur, on peut considérer son indice élémentaire et calculer alors une moyenne simple de ces n indices.

a • Indice moyenne arithmétique : $A_{t/0}(g) = \dfrac{1}{n}\sum_{k=1}^{n}\dfrac{g_t^k}{g_0^k} = \dfrac{1}{n}\sum_{k=1}^{n} i_{t/0}(g^k)$

b • Indice moyenne harmonique : $H_{t/0}(g) = \dfrac{n}{\sum_{k=1}^n \dfrac{1}{g_t^k/g_0^k}} = \dfrac{n}{\sum_{k=1}^n \dfrac{g_0^k}{g_t^k}} = \dfrac{n}{\sum_{k=1}^n i_{0/t}(g^k)}$

c • Indice moyenne géométrique : $G_{t/0}(g) = \sqrt[n]{\prod_{k=1}^{n}\dfrac{g_t^k}{g_0^k}} = \sqrt[n]{\prod_{k=1}^{n} i_{t/0}(g^k)}$

d • Propriété importante des moyenne : *Ces indices (arithmétique, harmonique, géométriquene) vérifient généralement pas les propriétés des indices élémentaires*

☑ **Exemple :** $A_{t/u}(g) \times A_{u/v}(g) \neq A_{t/v}(g)$

B-3 • Propriétés particulières de ces moyennes :

On s'intéresse uniquement aux grandeurs prix et quantité. A titre d'exemple, considérons l'indice des prix :

a • Cas de l'indice moyenne arithmétique :

$$On\ a : A_{t/0}(p) = \frac{1}{n}\sum_{k=1}^{n}\frac{p_t^k}{p_0^k} = \frac{1}{n}\left[\frac{p_t^1}{p_0^1} + \frac{p_t^2}{p_0^2} + \cdots + \frac{p_t^n}{p_0^n}\right]$$

$$D'autre\ part : \sum_{k=1}^{n}\left(p_t^1 \times \frac{1}{p_0^k}\right) \Big/ \sum_{k=1}^{n}\left(p_0^1 \times \frac{1}{p_0^k}\right) = \frac{\left(p_t^1 \times \frac{1}{p_0^1}\right) + \left(p_t^2 \times \frac{1}{p_0^2}\right) + \cdots + \left(p_t^n \times \frac{1}{p_0^n}\right)}{\underbrace{\left(p_0^1 \times \frac{1}{p_0^1}\right) + \left(p_0^2 \times \frac{1}{p_0^2}\right) + \cdots + \left(p_0^n \times \frac{1}{p_0^n}\right)}_{n}} = \frac{1}{n}\sum_{k=1}^{n}\frac{p_t^k}{p_0^k}$$

$$D'où, \boxed{A_{t/0}(p) = \frac{1}{n}\sum_{k=1}^{n}\frac{p_t^k}{p_0^k} = \frac{\sum_{k=1}^{n}\left(p_t^k \times \frac{1}{p_0^k}\right)}{\sum_{k=1}^{n}\left(p_0^k \times \frac{1}{p_0^k}\right)}}$$

Cet indice fait intervenir en plus des coefficients de pondération, l'ensemble $\left\{\frac{1}{p_0^1}, \frac{1}{p_0^2}, \dots, \frac{1}{p_0^n}\right\}$ *où* $\left(\frac{1}{p_0^k}\right)_{1\le k\le n}$ *représente la quantité de la grandeur considérée qu'onpeut déterminer pour une valeur unitaire appelée (panier de consommation fixe).*

☑ ***L'indice*** $A_{t/0}(p)$ ***est plus sensible à des prix en hausse qu'àdes prix en baisse. En effet :*** $A_{t/0}(p) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{p_0^k + \Delta p_t^k}{p_0^k}\right) = 1 + \frac{1}{n}\sum_{k=1}^{n}\frac{\Delta p_t^k}{p_0^k}$ ***, avec*** $\Delta p_t^k = p_t^k - p_0^k$

☑ ***L'indice moyenne arithmétique est égal à 100% plus la moyenne arithmétique des accroissements relatifs*** $\left(\frac{\Delta p_t^k}{p_0^k}\right)$

b • Cas de l'indice moyenne harmonique : $\boxed{H_{t/0}(p) = \frac{n}{\sum_{k=1}^{n}\frac{p_0^k}{p_t^k}} = \frac{\sum_{k=1}^{n}\left(p_t^k \times \frac{1}{p_t^k}\right)}{\sum_{k=1}^{n}\left(p_0^k \times \frac{1}{p_t^k}\right)}}$

$H_{t/0}(p)$ ***est une généralisation de l'indice de Bradstreet, obtenue en introduisant un ensemble de coefficients de pondération*** $\left\{\frac{1}{p_t^1}, \frac{1}{p_t^2}, \dots, \frac{1}{p_t^n}\right\}$ *où* $\left(\frac{1}{p_t^k}\right)_{1\le k\le n}$ ***représente le (panier de consommation variable).***

☑ ***L'indice de moyenne harmonique,*** $H_{t/0}(p)$ ***est sensible aux prix en baisse***

c • Remarques :

☑ *On constate que :* $A_{t/0}(g) = \dfrac{1}{H_{0/t}(g)}$

☑ *Les indices H, G et A satisfont aux propriétés générales des moyennes :*

$H_{t/0}(g) \leq G_{t/0}(g) \leq A_{t/0}(g)$

Indices synthétiques pondérés

Quand on veut calculer un indice à partir de plusieurs prix, le problème devient sensiblement plus compliqué. Un indice synthétique est une grandeur d'un ensemble de biens par rapport à une année de référence. On ne peut pas construire un indice synthétique en additionnant simplement des indices simples.

Il faut, en effet, tenir compte des quantités achetées.

Généralison l'indice de Bradstreet en le pondérant à l'aide de pondérations (w) des grandeurs (g) : $B_{t/0}(g) = \dfrac{w_*^1 g_t^1 + w_*^2 g_t^2 + \cdots + w_*^n g_t^n}{w_*^1 g_0^1 + w_*^2 g_0^2 + \cdots + w_*^n g_0^n} = \dfrac{\sum_{k=1}^{n} w_*^k g_t^k}{\sum_{k=1}^{n} w_*^k g_0^k}$

Pour calculer un indice de prix de n biens de consommation étiquetés de $1, 2, \ldots, n$, on utilise la notation suivante :

- p_t^k *représente le prix du bien de consommation (i) au temps (t)*
- q_t^k *représente la quantité de bien (i) consommée au temps (t)*

Il existe deux méthodes fondamentales pour calculer les indices de prix, l'indice de Paasche et l'indice de Laspeyres.

C-1 • Indice de Laspeyres :

a • Définition : C'est un indice synthétique pondéré dont la pondération (w_0^k) dépend de l'époque de base seulement: $L_{t/0} = 100 \times \left[\dfrac{w_0^1 g_t^1 + w_0^2 g_t^2 + \cdots + w_0^n g_t^n}{w_0^1 g_0^1 + w_0^2 g_0^2 + \cdots + w_0^n g_0^n}\right] = 100 \times \left(\dfrac{\sum_{k=1}^{n} w_0^k g_t^k}{\sum_{k=1}^{n} w_0^k g_0^k}\right)$

☑ *Cas particuliers :*

- *Prix :* $L_{t/0}(p) = 100 \times \left[\dfrac{q_0^1 p_t^1 + q_0^2 p_t^2 + \cdots + q_0^n p_t^n}{q_0^1 p_0^1 + q_0^2 p_0^2 + \cdots + q_0^n p_0^n}\right] = 100 \times \left(\dfrac{\sum_{k=1}^{n} q_0^k p_t^k}{\sum_{k=1}^{n} q_0^k p_0^k}\right)$

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

On utilise pour le calculer, les quantités $\{q_0^k\}$ *du temps de référence par rapport auquel on veut calculer l'indice.*

• **Quantité :** $L_{t/0}(q) = 100 \times \left[\dfrac{p_0^1 q_t^1 + p_0^2 q_t^2 + \cdots + p_0^n q_t^n}{p_0^1 q_0^1 + p_0^2 q_0^2 + \cdots + p_0^n q_0^n}\right] = 100 \times \left(\dfrac{\sum_{k=1}^n p_0^k q_t^k}{\sum_{k=1}^n p_0^k q_0^k}\right)$

On utilise pour le calculer, les prix $\{p_0^k\}$ *du temps de référence comme pondération.*

• **Propriétés :** *Pour* $L_{t/0}(p)$ *les coefficients de pondération* $\{q_0^k\}$ *constituent un pannier de consommation fixe, représentant une quantité globale considérée à la date base* (0) :

$$\frac{L_{t/0}(p)}{100} = \frac{q_0^1 p_t^1 + q_0^2 p_t^2 + \cdots + q_0^n p_t^n}{q_0^1 p_0^1 + q_0^2 p_0^2 + \cdots + q_0^n p_0^n} = \frac{\overbrace{(p_0^1 q_0^1)}^{\alpha_1}\overbrace{\left(\frac{p_t^1}{p_0^1}\right)}^{i_{t/0}^1(p)} + \overbrace{(p_0^2 q_0^2)}^{\alpha_2}\overbrace{\left(\frac{p_t^2}{p_0^2}\right)}^{i_{t/0}^2(p)} + \cdots + \overbrace{(p_0^n q_0^n)}^{\alpha_n}\overbrace{\left(\frac{p_t^n}{p_0^n}\right)}^{i_{t/0}^n(p)}}{\underbrace{q_0^1 p_0^1}_{\alpha_1} + \underbrace{q_0^2 p_0^2}_{\alpha_2} + \cdots + \underbrace{q_0^n p_0^n}_{\alpha_n}}$$

$$\frac{L_{t/0}(p)}{100} = \frac{\alpha_1 i_{t/0}^1(p) + \alpha_2 i_{t/0}^2(p) + \cdots + \alpha_n i_{t/0}^n(p)}{\alpha_1 + \alpha_2 + \cdots + \alpha_n} = \frac{\alpha_1 i_{t/0}^1(p) + \alpha_2 i_{t/0}^2(p) + \cdots + \alpha_n i_{t/0}^n(p)}{\sum_{k=1}^n \alpha_k}$$

$$\frac{L_{t/0}(p)}{100} = \underbrace{\left(\frac{\alpha_1}{\sum_{k=1}^n \alpha_k}\right)}_{w_0^1} i_{t/0}^1(p) + \underbrace{\left(\frac{\alpha_2}{\sum_{k=1}^n \alpha_k}\right)}_{w_0^2} i_{t/0}^2(p) + \cdots + \underbrace{\left(\frac{\alpha_n}{\sum_{k=1}^n \alpha_k}\right)}_{w_0^n} i_{t/0}^n(p) = \sum_{k=1}^n w_0^k i_{t/0}^k(p)$$

L'indice de Laspeyres est une moyenne arithmétique des indices simples pondérés par les coefficients budgétaires à la date 0 ; $\{w_0^k\}$: où $w_0^k = \dfrac{p_0^k q_0^k}{\sum_{k=1}^n p_0^k q_0^k}$ *et vérifie :* $\sum_{k=0}^n w_0^k = 1$

$$L_{t/0}(p) = \sum_{k=1}^n w_0^k I_{t/0}^k(p) = 100 \times \sum_{k=1}^n w_0^k i_{t/0}^k(p) = 100 \times \sum_{k=1}^n w_0^k \left(\frac{p_t^k}{p_0^k}\right)$$

- ☑ *Le poids* $p_0^k q_0^k$ *correspondant à la recette totale du bien* (k) *au temps* 0
- ☑ *L'indice de Laspeyres ne possède ni la propriété de circularité ni de réversibilité.*
- ☑ *L'indice de Laspeyres est facile à calculer, car seules les quantités* $\{q_0^k\}$ *du temps de référence sont nécessaires pour le calculer.*
- ☑ *Les coefficients budgétaires à la date* 0 ; $\{w_0^k\}$ *sont constants dans le temps.*

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba

☑ *L'interprétation de l'indice de Laspeyres est quelque peu problématique puisque la méthode de pondération suppose que les quantités de référence ne varient pas quand les prix changent. De plus, l'indice de Laspeyres tend à perdre sa représentativité au cours du temps.*

C – 2 • Indice de Paasche :

***a* • Définition :** *C'est un indice synthétique pondéré dont la pondération (w_t^k) dépend de l'époque courante :* $P_{t/0} = 100 \times \left[\frac{w_t^1 g_t^1 + w_t^2 g_t^2 + \cdots + w_t^n g_t^n}{w_t^1 g_0^1 + w_t^2 g_0^2 + \cdots + w_t^n g_0^n}\right] = 100 \times \left(\frac{\sum_{k=1}^n w_t^k g_t^k}{\sum_{k=1}^n w_t^k g_0^k}\right)$

☑ ***Cas particuliers :***

• ***Prix :*** $P_{t/0}(p) = 100 \times \left[\frac{q_t^1 p_t^1 + q_t^2 p_t^2 + \cdots + q_t^n p_t^n}{q_t^1 p_0^1 + q_t^2 p_0^2 + \cdots + q_t^n p_0^n}\right] = 100 \times \left(\frac{\sum_{k=1}^n q_t^k p_t^k}{\sum_{k=1}^n q_t^k p_0^k}\right)$

On utilise pour le calculer, les quantités $\{q_t^k\}$ du temps courant par rapport auquel on veut calculer l'indice.

• ***Quantité*** $P_{t/0}(q) = 100 \times \left[\frac{p_t^1 q_t^1 + p_t^2 q_t^2 + \cdots + p_t^n q_t^n}{p_t^1 q_0^1 + p_t^2 q_0^2 + \cdots + p_t^n q_0^n}\right] = 100 \times \left(\frac{\sum_{k=1}^n p_t^k q_t^k}{\sum_{k=1}^n p_t^k q_0^k}\right)$

On utilise pour le calculer, les prix $\{p_0^k\}$ du temps courant par rapport auquel on veut calculer l'indice de Paasche .

***b* • Propriétés :** *Pour $P_{t/0}(p)$ les coefficients de pondération $\{q_t^k\}$ constituent un pannier de consommation variable représentant une quantité globale considérée à la date courante (t)*

$$\frac{P_{t/0}(p)}{100} = \frac{q_t^1 p_t^1 + q_t^2 p_t^2 + \cdots + q_t^n p_t^n}{q_t^1 p_0^1 + q_t^2 p_0^2 + \cdots + q_t^n p_0^n} = \frac{\overbrace{q_t^1 p_t^1}^{\beta_1} + \overbrace{q_t^2 p_t^2}^{\beta_2} + \cdots + \overbrace{q_t^n p_t^n}^{\beta_n}}{\underbrace{(p_t^1 q_t^1)}_{\beta_1}\underbrace{\left(\frac{p_0^1}{p_t^1}\right)}_{\frac{1}{i_{t/0}^1(p)}} + \underbrace{(p_t^2 q_t^2)}_{\beta_2}\underbrace{\left(\frac{p_0^2}{p_t^2}\right)}_{\frac{1}{i_{t/0}^2(p)}} + \cdots + \underbrace{(p_t^n q_t^n)}_{\beta_n}\underbrace{\left(\frac{p_0^n}{p_t^n}\right)}_{\frac{1}{i_{t/0}^n(p)}}}$$

$$\frac{P_{t/0}(p)}{100} = \frac{\beta_1 + \beta_2 + \cdots + \beta_n}{\frac{\beta_1}{i_{t/0}^1(p)} + \frac{\beta_2}{i_{t/0}^2(p)} + \cdots + \frac{\beta_n}{i_{t/0}^n(p)}} = \frac{\sum_{k=1}^n \beta_k}{\frac{\beta_1}{i_{t/0}^1(p)} + \frac{\beta_2}{i_{t/0}^2(p)} + \cdots + \frac{\beta_n}{i_{t/0}^2(p)}}$$

$$\frac{P_{t/0}(p)}{100}=\frac{1}{\overbrace{\frac{\beta_1}{\sum_{k=1}^{n}\beta_k}}^{w_t^1}\Big/ i_{t/0}^{1}(p)+\overbrace{\frac{\beta_2}{\sum_{k=1}^{n}\beta_k}}^{w_t^2}\Big/ i_{t/0}^{2}(p)+\cdots+\overbrace{\frac{\beta_n}{\sum_{k=1}^{n}\beta_k}}^{w_t^n}\Big/ i_{t/0}^{2}(p)}=\left[\sum_{k=1}^{n}\frac{w_t^k}{i_{t/0}^{k}(p)}\right]^{-1}$$

L'indice de Paasche est une moyenne harmonique des indices simples pondérée par les coefficients budgétaires à la date t ; $\{w_t^k\}$: où $w_t^k=\dfrac{p_t^k q_t^k}{\sum_{k=1}^{n}p_t^k q_t^k}$ *et vérifie* : $\sum_{k=0}^{n} w_t^k=1$

$$P_{t/0}(p)=\left[\sum_{k=1}^{n}\frac{w_t^k}{I_{t/0}^{k}(p)}\right]^{-1}=100\times\left[\sum_{k=1}^{n}\frac{w_t^k}{i_{t/0}^{k}(p)}\right]^{-1}=100\times\left[\sum_{k=1}^{n}\frac{w_t^k}{p_t^k/p_0^k}\right]^{-1}$$

- ☑ *Le poids* $p_t^k q_t^k$ *correspondant à la recette totale du bien* (k) *au temps courant* (t)
- ☑ *L'indice de Paasche ne possède ni la propriété de circularité ni de réversibilité.*
- ☑ *L'indice de Paasche est plus difficile à calculer que l'indice de Laspeyres, car on doit connaitre les quantités pour chaque valeur de* (t) *et c'est pour cette raison que l'indice de Laspeyres est le plus utilisé dans la pratique*
- ☑ *L'indice de Paasche, utilisé avec l'indice de Laspeyres, sert à déterminer une fourchette d'estimation*
- ☑ *Les poids* $\{w_t^k\}$ *(ne sont pas constants dans le temps) sont des coefficients budgétaires « artificiels » évoluant dans le temps, correspondant au coût des quantités de la période courante aux prix de la période de base.*
- ☑ *L'utilisation de l'indice de Paasche n'est pas aisée car les comparaisons interpériodes sont rendues complexes puisque les pondérations varient dans le temps.*

C – 3 • Relation entre les indices :

- $L_{t/0}(g)=\dfrac{100^2}{P_{0/t}(g)}$
- $L_{t/0}(g)\geq P_{t/0}(g)$, *(en général)*

C−4 • Autres indices :

a • Indice de Sidgwick : $S_{t/0}(g) = \dfrac{L_{t/0}(g) + P_{t/0}(g)}{2}$

b • Indice de Edgeworth : $E_{t/0}(g) = \dfrac{\sum_{k=1}^{n}(w_0^k + w_t^k)g_t^k}{\sum_{k=1}^{n}(w_0^k + w_t^k)g_0^k}$

c • Indice de Fisher : $F_{t/0}(g) = \sqrt{L_{t/0}(g)P_{t/0}(g)}$

☑ *Cet indice possède la propriété de réversibilité, il est plus utilisé que les deux précédents* : $F_{t/0}(g) = \dfrac{100^2}{F_{0/t}(g)}$

☑ *Il donne une meilleure estimation d'une hausse des prix. En effet l'indice de Laspeyres a tendance à les surévaluer, tandis que celui de Paasche à les sous-estimer.*

d • Indices synthétiques de valeur :

$$V_{t/0}(v) = B_{t/0}(v) = \frac{\sum_{k=1}^{n} w_t^k v_t^k}{\sum_{k=1}^{n} w_0^k v_0^k} \text{ où } w_*^k = \frac{p_*^k q_*^k}{\sum_{k=1}^{n} p_*^k q_*^k} \text{ et vérifie : } \sum_{k=0}^{n} w_*^k = 1;\ (* = 0, t)$$

e • Indices-chaînes : *Pour comparer des « époques distincts de l'époque de référence, si les indices sont transitifs, le problème est résolu. Dans le cas contraire, on construit des indices-chaînes. Pour cela, on considère une suite de dates* $0, 1, 2, \dots, k, \dots$ *et une suite d'indices exprimés en base* 100 *de l'année précédente* $I_{1/0}, I_{2/1}, \dots, I_{k/k-1}, \dots$

On construit une suite d'indices-chaînes, base 100 *à la date* (0) *de la manière suivante* :

$$CI_{1/0} = 100 \times i_{1/0} = I_{1/0}$$

$$CI_{2/0} = CI_{1/0} i_{2/1} = \frac{1}{100} CI_{1/0} I_{2/1}$$

$$\dots \dots \dots \dots \dots \dots \dots \dots$$

$$CI_{n/0} = CI_{n-1/0} i_{n/n-1} = \frac{1}{100} CI_{n-1/0} I_{n/n-1}$$

$$\forall t \in \{1, \dots, n\}, CI_{t/0} = \frac{1}{100^{t-1}} \prod_{k=1}^{t} I_{k/k-1} = 100 \prod_{k=1}^{t} i_{k/k-1}$$

A partir de cette relation, on peut calculer l'indice relatif à deux dates quelconques de la *suite, soit* : $\forall t,t' \in \{1,\dots,n\}\,, i_{t'/t} = \dfrac{CI_{t'/0}}{CI_{t/0}}$

☑ ***Utilisation :*** *L'indice-chaîne permet mieux que les indices de Laspeyres ou de Paasche, de suivre l'évolution d'une grandeur entre deux dates successives.*

Exercice 15 :

ÉNONCÉ

Le tableau suivant présente les prix et quantités de trois bien pendant 3 ans

Période / Bien	0		1		2	
	Prix: p_0^k	Quantités: q_0^k	Prix: p_1^k	Quantités: q_1^k	Prix: p_2^k	Quantités: q_2^k
Bien 1	100	14	150	10	200	8
Bien 2	60	10	50	12	40	14
Bien 3	160	4	140	5	140	5

1)
   a) ***Calculer de deux manières les indices de Laspeyres prix***
   b) ***Calculer de deux manières les indices de Laspeyres quantités***

2)
   a) ***Calculer de deux manières les indices de Paasche prix***
   b) ***Calculer de deux manières les indices de Paasche quantités***

3) ***Déduire les indices de Fisher prix et quantités***

Corrigé

1)

Période / Bien	0		1		2		$q_0^kp_0^k$	$q_0^kp_1^k$	$q_0^kp_2^k$	$q_1^kp_2^k$	$q_1^kp_1^k$	w_0^k	w_1^k	$\frac{p_1^k}{p_0^k}$	$\frac{p_2^k}{p_0^k}$	$\frac{p_2^k}{p_1^k}$
	p_0^k	q_0^k	p_1^k	q_1^k	p_2^k	q_2^k										
Bien 1	100	14	150	10	200	8	1400	2100	2800	2000	1500	$\frac{35}{66}$	$\frac{15}{28}$	$\frac{3}{2}$	2	$\frac{4}{3}$
Bien 2	60	10	50	12	40	14	600	500	400	480	600	$\frac{5}{22}$	$\frac{3}{14}$	$\frac{5}{6}$	$\frac{2}{3}$	$\frac{4}{5}$
Bien 3	160	4	140	5	140	5	640	560	560	700	700	$\frac{8}{33}$	$\frac{1}{4}$	$\frac{7}{8}$	$\frac{7}{8}$	1
$\sum$							2640	3160	3760	3180	2800	1	1			

Période / Bien	0		1		2		$p_0^kq_1^k$	$p_0^kq_2^k$	$p_1^kq_2^k$	$p_2^kq_2^k$	w_0^k	w_1^k	w_2^k	$\frac{q_1^k}{q_0^k}$	$\frac{q_2^k}{q_0^k}$	$\frac{q_2^k}{q_1^k}$
	p_0^k	q_0^k	p_1^k	q_1^k	p_2^k	q_2^k										
Bien 1	100	14	150	10	200	8	1000	800	1200	1600	$\frac{35}{66}$	$\frac{15}{28}$	$\frac{80}{143}$	$\frac{5}{7}$	$\frac{4}{7}$	$\frac{4}{5}$
Bien 2	60	10	50	12	40	14	720	840	700	560	$\frac{5}{22}$	$\frac{3}{14}$	$\frac{28}{143}$	$\frac{6}{5}$	$\frac{7}{5}$	$\frac{7}{6}$
Bien 3	160	4	140	5	140	5	800	800	700	700	$\frac{8}{33}$	$\frac{1}{4}$	$\frac{35}{143}$	$\frac{5}{4}$	$\frac{5}{4}$	1
$\sum$							2520	2440	2600	2860	1	1	1			

a)

☑ **$1^{ère}$ *manière* :**

$$\cdot\ L_{1/0}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_0^k p_1^k}{\sum_{k=1}^{3} q_0^k p_0^k}\right) = 100 \times \left(\frac{3160}{2640}\right) = 119,697\%$$

$$\cdot\ L_{2/0}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_0^k p_2^k}{\sum_{k=1}^{3} q_0^k p_0^k}\right) = 100 \times \left(\frac{3760}{2640}\right) = 142,424\%$$

$$\cdot\ L_{2/1}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_1^k p_2^k}{\sum_{k=1}^{3} q_1^k p_1^k}\right) = 100 \times \left(\frac{3180}{2800}\right) = 113,571\%$$

☑ $2^{\text{ème}}$ ***manière*** :

$$\cdot\ L_{1/0}(p) = \sum_{k=1}^{3} w_0^k I_{1/0}^k(p) = 100 \times \sum_{k=1}^{3} w_0^k i_{1/0}^k(p) = 100 \times \sum_{k=1}^{3} w_0^k \left(\frac{p_1^k}{p_0^k}\right)$$

$$L_{1/0}(p) = 100 \times \left[w_0^1\left(\frac{p_1^1}{p_0^1}\right) + w_0^2\left(\frac{p_1^2}{p_0^2}\right) + w_0^3\left(\frac{p_1^3}{p_0^3}\right)\right] = 100 \times \left[\left(\frac{35}{66} \times 1,5\right) + \left(\frac{5}{22} \times \frac{2}{3}\right) + \left(\frac{8}{33} \times \frac{7}{8}\right)\right]$$

$$L_{1/0}(p) = 119,697\%$$

$$\cdot\ L_{2/0}(p) = \sum_{k=1}^{3} w_0^k I_{2/0}^k(p) = 100 \times \sum_{k=1}^{3} w_0^k i_{2/0}^k(p) = 100 \times \sum_{k=1}^{3} w_0^k \left(\frac{p_2^k}{p_0^k}\right)$$

$$L_{2/0}(p) = 100 \times \left[w_0^1\left(\frac{p_2^1}{p_0^1}\right) + w_0^2\left(\frac{p_2^2}{p_0^2}\right) + w_0^3\left(\frac{p_2^3}{p_0^3}\right)\right] = 100 \times \left[\left(\frac{35}{66} \times 2\right) + \left(\frac{5}{22} \times \frac{2}{3}\right) + \left(\frac{8}{33} \times \frac{7}{8}\right)\right]$$

$$L_{2/0}(p) = 142,424\%$$

$$\cdot\ L_{2/1}(p) = \sum_{k=1}^{3} w_1^k I_{2/1}^k(p) = 100 \times \sum_{k=1}^{3} w_1^k i_{2/1}^k(p) = 100 \times \sum_{k=1}^{3} w_1^k \left(\frac{p_2^k}{p_1^k}\right)$$

$$L_{2/1}(p) = 100 \times \left[w_1^1\left(\frac{p_2^1}{p_1^1}\right) + w_1^2\left(\frac{p_2^2}{p_1^2}\right) + w_1^3\left(\frac{p_2^3}{p_1^3}\right)\right] = 100 \times \left[\left(\frac{15}{28} \times \frac{4}{3}\right) + \left(\frac{3}{14} \times \frac{4}{5}\right) + \left(\frac{1}{4} \times 1\right)\right]$$

$$L_{2/1}(p) = 113,571\%$$

b)

☑ $1^{\text{ère}}$ ***manière*** :

$$\cdot\ L_{1/0}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_0^k q_1^k}{\sum_{k=1}^{3} p_0^k q_0^k}\right) = 100 \times \left(\frac{2520}{2640}\right) = 95,455\%$$

$$\cdot\ L_{2/0}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_0^k q_2^k}{\sum_{k=1}^{3} p_0^k q_0^k}\right) = 100 \times \left(\frac{2440}{2640}\right) = 92,424\%$$

· $L_{2/1}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_1^k q_2^k}{\sum_{k=1}^{3} p_1^k q_1^k}\right) = 100 \times \left(\frac{2600}{2800}\right) = 92,857\%$

☑ $2^{\text{ème}}$ *manière* :

· $L_{1/0}(q) = \sum_{k=1}^{3} w_0^k I_{1/0}^k(q) = 100 \times \sum_{k=1}^{3} w_0^k i_{1/0}^k(q) = 100 \times \sum_{k=1}^{3} w_0^k \left(\frac{q_1^k}{q_0^k}\right)$

$$L_{1/0}(q) = 100 \times \left[w_0^1\left(\frac{q_1^1}{q_0^1}\right) + w_0^2\left(\frac{q_1^2}{q_0^2}\right) + w_0^3\left(\frac{q_1^3}{q_0^3}\right)\right] = 100 \times \left[\left(\frac{35}{66} \times \frac{5}{7}\right) + \left(\frac{5}{22} \times \frac{6}{5}\right) + \left(\frac{8}{33} \times \frac{5}{4}\right)\right]$$

$L_{1/0}(q) = 95,455\%$

· $L_{2/0}(q) = \sum_{k=1}^{3} w_0^k I_{2/0}^k(q) = 100 \times \sum_{k=1}^{3} w_0^k i_{2/0}^k(q) = 100 \times \sum_{k=1}^{3} w_0^k \left(\frac{q_2^k}{q_0^k}\right)$

$$L_{2/0}(q) = 100 \times \left[w_0^1\left(\frac{q_2^1}{q_0^1}\right) + w_0^2\left(\frac{q_2^2}{q_0^2}\right) + w_0^3\left(\frac{q_2^3}{q_0^3}\right)\right] = 100 \times \left[\left(\frac{35}{66} \times \frac{4}{7}\right) + \left(\frac{5}{22} \times \frac{7}{5}\right) + \left(\frac{8}{33} \times \frac{5}{4}\right)\right]$$

$L_{2/0}(q) = 92,424\%$

· $L_{2/1}(q) = \sum_{k=1}^{3} w_1^k I_{2/1}^k(q) = 100 \times \sum_{k=1}^{3} w_1^k i_{2/1}^k(q) = 100 \times \sum_{k=1}^{3} w_1^k \left(\frac{q_2^k}{q_1^k}\right)$

$$L_{2/1}(q) = 100 \times \left[w_1^1\left(\frac{q_2^1}{q_1^1}\right) + w_1^2\left(\frac{q_2^2}{q_1^2}\right) + w_1^3\left(\frac{q_2^3}{q_1^3}\right)\right] = 100 \times \left[\left(\frac{15}{28} \times \frac{4}{5}\right) + \left(\frac{3}{14} \times \frac{7}{6}\right) + \left(\frac{1}{4} \times 1\right)\right]$$

$L_{2/1}(q) = 92,857\%$

2)

a)

☑ $1^{\text{ère}}$ *manière* :

· $P_{1/0}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_1^k p_1^k}{\sum_{k=1}^{3} q_1^k p_0^k}\right) = 100 \times \left(\frac{2800}{2520}\right) = 111,111\%$

· $P_{2/0}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_2^k p_2^k}{\sum_{k=1}^{3} q_2^k p_0^k}\right) = 100 \times \left(\frac{2860}{2440}\right) = 117,213\%$

· $P_{2/1}(p) = 100 \times \left(\frac{\sum_{k=1}^{3} q_2^k p_2^k}{\sum_{k=1}^{3} q_2^k p_1^k}\right) = 100 \times \left(\frac{2860}{2600}\right) = 110\%$

☑ $2^{\text{ème}}$ *manière* :

· $P_{1/0}(p) = \left[\sum_{k=1}^{3} \frac{w_1^k}{I_{1/0}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_1^k}{i_{1/0}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_1^k}{p_1^k/p_0^k}\right]^{-1}$

$$P_{1/0}(p) = \frac{100}{\left[\frac{w_1^1}{p_1^1/p_0^1} + \frac{w_1^2}{p_1^2/p_0^2} + \frac{w_1^3}{p_1^3/p_0^3}\right]} = \frac{100}{\left[\frac{15/28}{3/2} + \frac{3/14}{5/6} + \frac{1/4}{7/8}\right]} = \frac{100}{9/10} = 111,111\%$$

· $P_{2/0}(p) = \left[\sum_{k=1}^{3} \frac{w_2^k}{I_{2/0}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_2^k}{i_{2/0}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_2^k}{p_2^k/p_0^k}\right]^{-1}$

$$P_{2/0}(p) = \frac{100}{\left[\frac{w_2^1}{p_2^1/p_0^1} + \frac{w_2^2}{p_2^2/p_0^2} + \frac{w_2^3}{p_2^3/p_0^3}\right]} = \frac{100}{\left[\frac{80/143}{2} + \frac{28/143}{2/3} + \frac{35/143}{7/8}\right]} = \frac{100}{122/143} = 117,213\%$$

· $P_{2/1}(p) = \left[\sum_{k=1}^{3} \frac{w_2^k}{I_{2/1}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_2^k}{i_{2/1}^k(p)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_2^k}{p_2^k/p_1^k}\right]^{-1}$

$$P_{2/1}(p) = \frac{100}{\left[\frac{w_2^1}{p_2^1/p_1^1} + \frac{w_2^2}{p_2^2/p_1^2} + \frac{w_2^3}{p_2^3/p_1^3}\right]} = \frac{100}{\left[\frac{80/143}{4/3} + \frac{28/143}{4/5} + \frac{35/143}{1}\right]} = \frac{100}{10/11} = 110\%$$

b)

☑ $1^{ère}$ ***manière*** :

· $P_{1/0}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_1^k q_1^k}{\sum_{k=1}^{3} p_1^k q_0^k}\right) = 100 \times \left(\frac{2800}{3160}\right) = 88,608\%$

· $P_{2/0}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_2^k q_2^k}{\sum_{k=1}^{3} p_2^k q_0^k}\right) = 100 \times \left(\frac{2860}{3760}\right) = 76,064\%$

· $P_{2/1}(q) = 100 \times \left(\frac{\sum_{k=1}^{3} p_2^k q_2^k}{\sum_{k=1}^{3} p_2^k q_1^k}\right) = 100 \times \left(\frac{2860}{3180}\right) = 89,937\%$

☑ $2^{ème}$ ***manière*** :

· $P_{1/0}(q) = \left[\sum_{k=1}^{3} \frac{w_1^k}{I_{1/0}^k(q)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_1^k}{i_{1/0}^k(q)}\right]^{-1} = 100 \times \left[\sum_{k=1}^{3} \frac{w_1^k}{q_1^k/q_0^k}\right]^{-1}$

$$P_{1/0}(q) = \frac{100}{\left[\frac{w_1^1}{q_1^1/q_0^1} + \frac{w_1^2}{q_1^2/q_0^2} + \frac{w_1^3}{q_1^3/q_0^3}\right]} = \frac{100}{\left[\frac{15/28}{5/7} + \frac{3/14}{6/5} + \frac{1/4}{5/4}\right]} = \frac{100}{79/70} = 88,608\%$$

$$\cdot\ P_{2/0}(q)=\left[\sum_{k=1}^{3}\frac{w_2^k}{I_{2/0}^k(q)}\right]^{-1}=100\times\left[\sum_{k=1}^{3}\frac{w_2^k}{i_{2/0}^k(q)}\right]^{-1}=100\times\left[\sum_{k=1}^{3}\frac{w_2^k}{q_2^k/q_0^k}\right]^{-1}$$

$$P_{2/0}(q)=\frac{100}{\left[\frac{w_2^1}{q_2^1/q_0^1}+\frac{w_2^2}{q_2^2/q_0^2}+\frac{w_2^3}{q_2^3/q_0^3}\right]}=\frac{100}{\left[\frac{80/143}{4/7}+\frac{28/143}{7/5}+\frac{35/143}{5/4}\right]}=\frac{100}{188/143}=76,064\%$$

$$\cdot\ P_{2/1}(q)=\left[\sum_{k=1}^{3}\frac{w_2^k}{I_{2/1}^k(q)}\right]^{-1}=100\times\left[\sum_{k=1}^{3}\frac{w_2^k}{i_{2/1}^k(q)}\right]^{-1}=100\times\left[\sum_{k=1}^{3}\frac{w_2^k}{q_2^k/q_1^k}\right]^{-1}$$

$$P_{2/1}(q)=\frac{100}{\left[\frac{w_2^1}{q_2^1/q_1^1}+\frac{w_2^2}{q_2^2/q_1^2}+\frac{w_2^3}{q_2^3/q_1^3}\right]}=\frac{100}{\left[\frac{80/143}{4/5}+\frac{28/143}{7/6}+\frac{35/143}{1}\right]}=\frac{100}{159/143}=89,937\%$$

3)

☑ $F_{t/t'}(p)=\sqrt{L_{t/t'}(p)P_{t/t'}(p)}$:

$$\cdot\ F_{1/0}(p)=\sqrt{L_{1/0}(p)P_{1/0}(p)}=\left(\sqrt{119,697\times111,111}\right)\%=115,324\%$$

$$\cdot\ F_{2/0}(p)=\sqrt{L_{2/0}(p)P_{2/0}(p)}=\left(\sqrt{142,424\times117,213}\right)\%=129,205\%$$

$$\cdot\ F_{2/1}(p)=\sqrt{L_{2/1}(p)P_{2/1}(p)}=\left(\sqrt{113,571\times110}\right)\%=111,771\%$$

☑ $F_{t/t'}(q)=\sqrt{L_{t/t'}(q)P_{t/t'}(q)}$:

$$\cdot\ F_{1/0}(q)=\sqrt{L_{1/0}(q)P_{1/0}(q)}=\left(\sqrt{95,455\times88,608}\right)\%=91,968\%$$

$$\cdot\ F_{2/0}(q)=\sqrt{L_{2/0}(q)P_{2/0}(q)}=\left(\sqrt{92,424\times76,064}\right)\%=83,846\%$$

$$\cdot\ F_{2/1}(q)=\sqrt{L_{2/1}(q)P_{2/1}(q)}=\left(\sqrt{92,857\times89,937}\right)\%=91,385\%$$

Axe ② : Table des matières

Distribution statistique à un seul caractère

Série statistique à deux variables quantitatives

Série statistique à deux variables qualitatives

Une variable quantitative et une qualitative

BEN AHMED MOHSEN Téléphone: (+216) 97 619191 / 54 619191
omega.center.cp@gmail.com https://web.facebook.com/OMEGACENTER2014

Les indices statistiques

Adresse : 2, Rue Ibn Arafa, av de l'environnement Manouba Centre, Manouba