Les caractéristiques de dispersion:

Variance et écart-type:

noté $S^2 = \sum_{i=1}^{n} n_i(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^2 - \bar{X}^2$

L'écart-type noté S est défini par : $S = \sqrt{S^2}$*. Il permet de mesurer la dispersion des observations autour de la moyenne* $\bar{X}$*. Un écart-type plus faible exprime une plus faible dispersion et donc une plus forte concentration autour de X.*

Ecart arithmétique moyen:

L' Ecart arithmétique moyen, noté e est défini par:

$$e = \frac{1}{n}\sum_{i=1}^{n} n_i |x_i - \bar{X}|$$

Moments:

Les moments d'ordre r sont définis par:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} n_i \; x^r$$

Les moments centrés d'ordre r sont définis par:

$\dot{\rho}_r = \frac{1}{n}\sum_{i=1}^{n} n_i(x_i - \bar{x})^r$

Le moment d'ordre 1 se confond avec la moyenne $\bar{X}$ *et le moment centré d'ordre 2 confond avec la variance* S^2

Caractéristiques de forme:

Le coefficient d'asymétrie de Fisher est défini par: $\gamma_1 = \frac{\dot{\rho}_3}{S^3}$

Le coefficient d'aplatissement de Fisher est défini par: $\gamma_2 = \frac{\dot{\rho}_4}{S^4}$

Lorsque la statistique et symétrique, on a: $\gamma_1 = 0$*. Pour une distribution normale réduite on a:* $\gamma_2 = 3$

Statistique descriptive à deux dimensions

Dans cette partie nous intéressons à deux caractères, non pas isolement ; mais nous essayerons de mettre en évidence les relations qui existent entre ces deux caractères.

La préoccupation principale sera la mesure de l'importance du lien entre deux caractères d'une population par exemple existe-t-il une certaine relation la taille d'un individu et son poids, le prix d'une maison et sa surface.

Les distributions des fréquences:

La série statistique double formé par les n couples (x_i, y_i) de valeurs observées se présente sous la forme:

X	X_1	X_2	…….	X_n
Y	Y_1	Y_2	…….	Y_n

Ou X et Y sont les deux caractères étudiés.

Exemple 1

Le tableau suivant représente une série statistique de 12 mesures:

Tableau 1

X	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
Y	*1.5*	*2.5*	*2*	*2*	*3*	*4*	*4*	*3.5*	*4.5*	*4*	*5*	*5.5*

Lorsque l'effectif total n est élevé et que les couples d'observations (x_i, y_i) correspond à une fréquence n_{ij} (ou ce qui revient au même à une fréquence $n_{ij} = \frac{n_{ij}}{n}$ on représente la distribution en un tableau à double entrée:

Y / X	Y_1	Y_2	……	Y_j	……	Y_q	Total
X_1	n_{11}	n_{12}	……	n_{1j}	……	n_{1q}	$n_{1.}$
X_2	n_{21}	n_{22}	……	n_{2j}	……	n_{ij}	$n_{2.}$
. .			……			……	. .
X_i	n_{i1}	n_{i2}	……	n_{ij}	……	n_{ij}	$n_{i.}$
. .			……				. .
X_p	n_{p1}	n_{p2}	……	n_{pj}	……	n_{ij}	$n_{p.}$
Total	$n_{.1}$	$n_{.2}$	……	$n_{.j}$	……	$n_{.q}$	n

Si le nombre de lignes ou de colonnes et trop grand, on regroupe en classes les valeurs observées, les valeurs x_i et y_i représentent les centres de classes. (Comme on déjà fa fait pour une série statistique à une dimension).

Les distributions marginales de X et Y sont définies respectivement par:

$$n_{i.} = \sum_{j=1}^{q} n_{ij} \qquad et \qquad n_{.j} = \sum_{i=1}^{p} n_{ij}$$

On définit aussi les fréquences marginales:

$$f_{i.} = \frac{n_{i.}}{n} \qquad e \qquad t \quad f_{.j} = \frac{n_{.j}}{n}$$

On dit que X et Y sont statistiquement indépendantes si : $f_{ij} = f_{i.} f_{.j}$

Exemple 2

Le tableau suivant donne les notes en mathématiques (X) et en physique (Y) obtenues par un groupe de 100 étudiants.

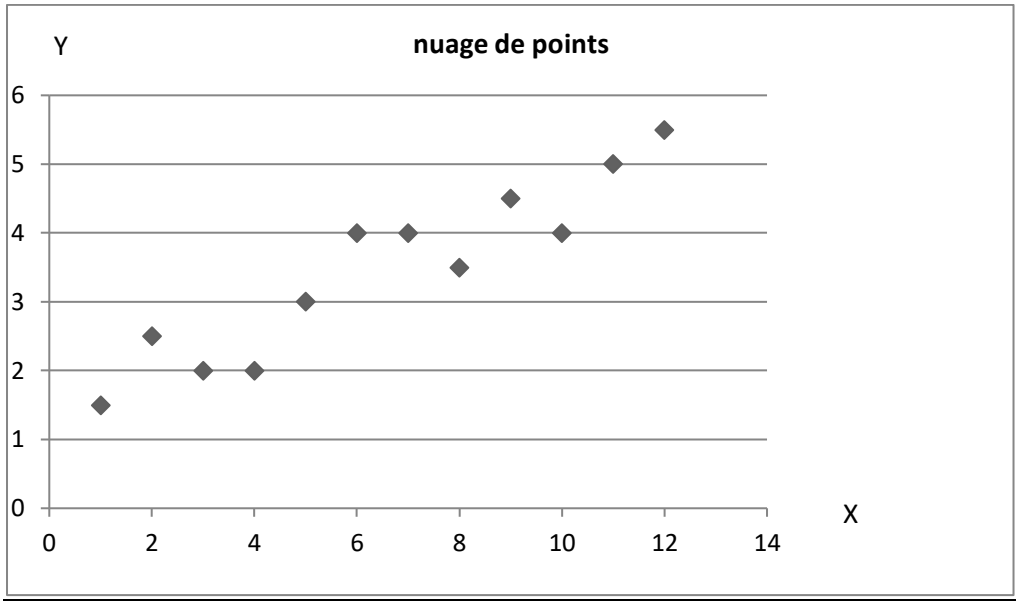
Tableau 2

Y / *X*	***4≤ Y ≤6***	***6≤ Y ≤8***	***8≤ Y ≤10***	***10≤ Y ≤12***	***12≤ Y ≤14***	***14≤ Y ≤16***	***Total***
4≤ Y ≤6	**2**	**4**	**9**	**9**			**24**
6≤ Y ≤8	1	7	**9**	**18**	**3**	**1**	**39**
8≤ Y ≤10			***1***	***8***	***1***	***1***	***11***
10≤ Y ≤12		**1**	***1***	***10***	***7***	**5**	**24**
12≤ Y ≤14					***1***		**1**
14≤ Y ≤16						**1**	**1**
Total	**3**	**12**	**20**	**45**	**12**	**8**	**100**

Représentation graphique:

Si chaque couple n'a pas qu'une seule observation, on représente la distribution sous la forme d'un nuage de points dans IR^2.

Fig-1



Covariance:

Lorsqu'i n'ya qu'une seule observation par couple la covariance Cov(X , Y) est définie par:

$$S_{XY}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})$$

Et le coefficient de corrélation r ou r_{XY} est définie par:

$$r = \frac{S_{XY}^2}{S_X S_Y}$$

Ou $S_X^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{X}^2$ *et* $S_Y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{Y}^2$

Si n_{ij} *est l'effectif observé de* (x_i , y_i), *on a:*

$$S_{XY}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\,(x_i - \overline{X})(y_i - \overline{Y})$$

Et

$$r = \frac{S_{XY}^2}{S_X S_Y}$$

Ou $S_X^2 = \frac{1}{n}\sum_{i=1}^{p} x_i^2 - \overline{X}^2$ *et* $S_Y^2 = \frac{1}{n}\sum_{j=1}^{q} y_i^2 - \overline{Y}^2$

En reprenant les données du tableau 2(exemple 2) on calcul la moyenne et l'écart-type pour chacune des variables ***X*** *et* ***Y*** *ainsi que le coefficient de corrélation. Pour chaque classe nous choisirons un centre de classe.*

Tableau 3

Y / ***X***	***5***	***7***	***9***	***11***	***13***	***15***	***Total*** $n_{i.}$	$n_{i.}\ x_i$	$n_{i.}\ x_i^2$	$x_i \sum_i n_{ij} y_j$
5	**2**	**4**	**9**	**9**			**24**	**120**	**600**	**1090**
7	1	7	**9**	**18**	**3**	**1**	**39**	**273**	**1911**	**2709**
9			***1***	***8***	***1***	***1***	***11***	***99***	***891***	***1125***
11		**1**	***1***	***10***	***7***	**5**	**24**	**264**	**2904**	**3212**
13					***1***		**1**	**13**	**169**	**169**
15						**1**	**1**	**15**	**225**	**225**
Total $n_{.j}$	**3**	**12**	**20**	**45**	**12**	**8**	**100**	**784**	**6700**	**8530**
$n_{.j}\ y_i$	**15**	**84**	**180**	**495**	**156**	**120**	**1050**			
$n_{.j}\ y_i^2$	**75**	**588**	**1620**	**5445**	**2028**	**1800**	**11556**			
$Y_j \sum_i n_{ij} x_i$	**85**	**560**	**1152**	**3883**	**1560**	**1290**	**8530**			

$$S_{XY}^2 = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\,(x_i - \overline{X})(y_i - \overline{Y})$$

$$= \frac{1}{n}\sum_{i.j} n_{ij}x_i y_j - \left(\frac{1}{n}\sum_{i.j} n_{ij}x_i\right)\left(\frac{1}{n}\sum_{i.j} n_{ij}y_j\right)$$

$S_{XY}^2 = \frac{1}{100}8530 - \frac{1}{100}\ 784\ \frac{1}{100}1050$=*2.99*

$S_X^2 = \frac{1}{100}6700 - (7.84)^2$=*5.53*

$S_y^2 = \frac{1}{100}11556 - (10.5)^2$=*5.31* *et donc*

$$r = \frac{S_{XY}^2}{S_X S_Y} = \frac{2.98}{\sqrt{5.53}\sqrt{5.31}} = 0.55$$

le coefficient de corrélation r *est indépendant du choix des unités. Or remarque que:*

- *I* r *I≤ 1*
- *Si les deux variables* ***X*** *et* ***Y*** *sont indépendantes entre elles, alors* $r = 0$*. La réciproque n'est pas vraie.*
- *Si les 2 caractères varient en général dans le même sens,* $r\ est\ positif$.
- *Si les 2 caractères varient en général en sens contraire,* $r\ est\ négatif$.

Ajustement linéaire:

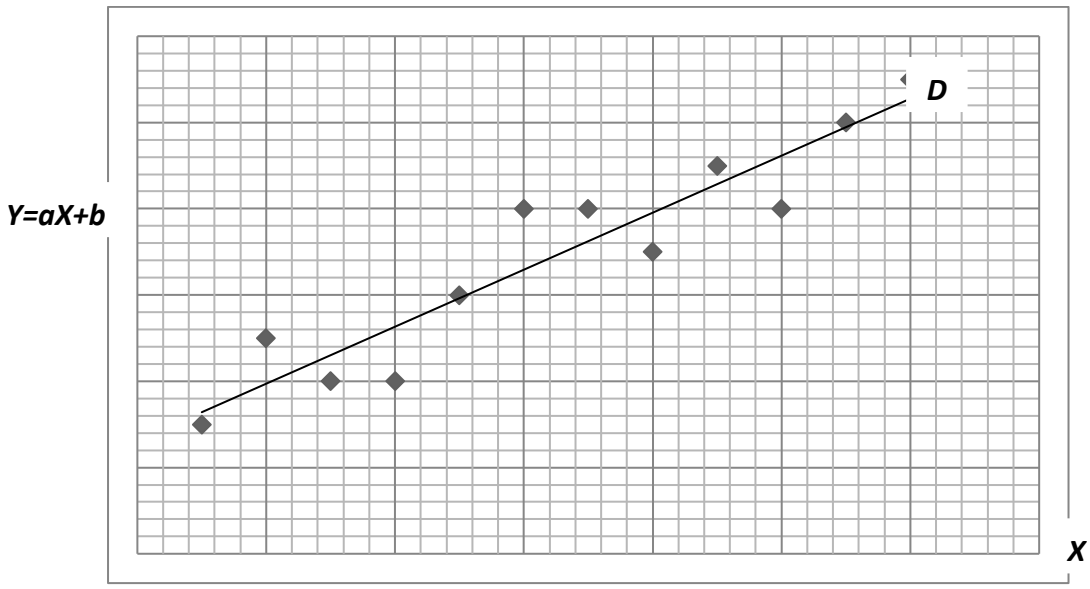
Dans une statistique double, on est amené à répondre à la question de savoir quelle relation peut exister entre les deux caractères étudies. Il s'agit de trouver, si elle existe, la relation entre X et Y d'une façon précise.

Le nuage de points de la figure 1 suggère de chercher une relation de type ***Y=aX+b*** *, on cherche donc la droite (dite droite de régression) qui passe au* ***plus prés*** *des différents points du nuages. La méthode des moindres carrés permet de déterminer parmi toutes les droites, celle qui approche* ***le plus*** *le nuage de points c'est-à-dire de trouver une droite D d'équation* ***Y=aX+b*** *telle que:*

$\sum = \sum_{i=1}^{n}(y_i - ax_i - b)^2$ *Soit minimale.*

Fig-2

Y



Le minimum est déterminé en annulant les dérivées partielles par rapport à a et à b on donc

$$\begin{cases} \text{bn} + a \sum_{i=1}^{n} \text{x}_\text{i} = \sum_{i=1}^{n} \text{y}_\text{i} = \\ \text{b} \sum_{i=1}^{n} \text{x}_\text{i} + a \sum_{i=1}^{n} \text{x}_\text{i}^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

dites 2auqtions normales

On déduit de la première équation $\bar{\boldsymbol{Y}} = \boldsymbol{a}\,\bar{\boldsymbol{X}} + \boldsymbol{b}$.

Et en déduire que $a = \frac{\mathbf{cov(X,Y)}}{\mathbf{S_X^2}}$ = $\frac{\mathbf{S_{XY}^2}}{\mathbf{S_X^2}}$

La droite de régression de Y en X s'écrit donc : ***Y=aX+b*** avec:

$\boldsymbol{a} = \frac{\mathbf{cov(X,Y)}}{\mathbf{S_X^2}}$ = $\frac{\mathbf{S_{XY}^2}}{\mathbf{S_X^2}}$ et $\bar{\boldsymbol{Y}} = \boldsymbol{a}\,\bar{\boldsymbol{X}} + \boldsymbol{b}$

En reprenant les données du tableau 1, on détermine la droite de régression de Y en X et le coefficient de corrélation $\boldsymbol{r}$.

Tableau 4

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
1	1.5	1	1.5	2.25
2	2.5	4	5	6.25
3	2	9	6	4
4	2	16	8	4
5	3	25	15	9
6	4	36	24	16
7	4	49	28	16
8	3.5	64	28	12.25
9	4.5	81	40.5	20.25
10	4	100	40	16
11	5	121	55	25
12	5.5	144	66	30.25
$\sum x_i = 78$	$\sum y_i = 41.5$	$\sum x_i^2 = 650$	$\sum x_i y_i = 317$	$\sum y_i^2 = 161.25$

On a : $\begin{cases} 12\ b + 78\ a = 41.5 \\ 78\ b + 650\ a = 317 \end{cases}$ d'où a = 0.33 et b = 1.31

L'équation de la droite s'écrit donc. ***Y = 0.33 X + 1.31***

Comme $\bar{X} = 6.5$ et $\bar{Y} = 3.46$ on a Cov(X,Y) = $\frac{1}{12}$ 317 - 6.5(3.46) = 3.94

$S_X^2 = \frac{1}{12}650 - (65)^2$ *= 3.45*

$S_Y^2 = \frac{1}{12}16.25 - (3.46)^2$ *= 1.22*

d'où $r = \frac{S_{XY}^2}{S_X S_Y} = \frac{3.94}{3.45(1.22)} = 0.94$