Université Hassiba BenBouali Chlef
Faculté des Sciences de la Nature et de la Vie
Département de Biotechnologie

Série statistique à deux variables

Niveau: 2ème année

Pr : KEHAILI ABDELKADER

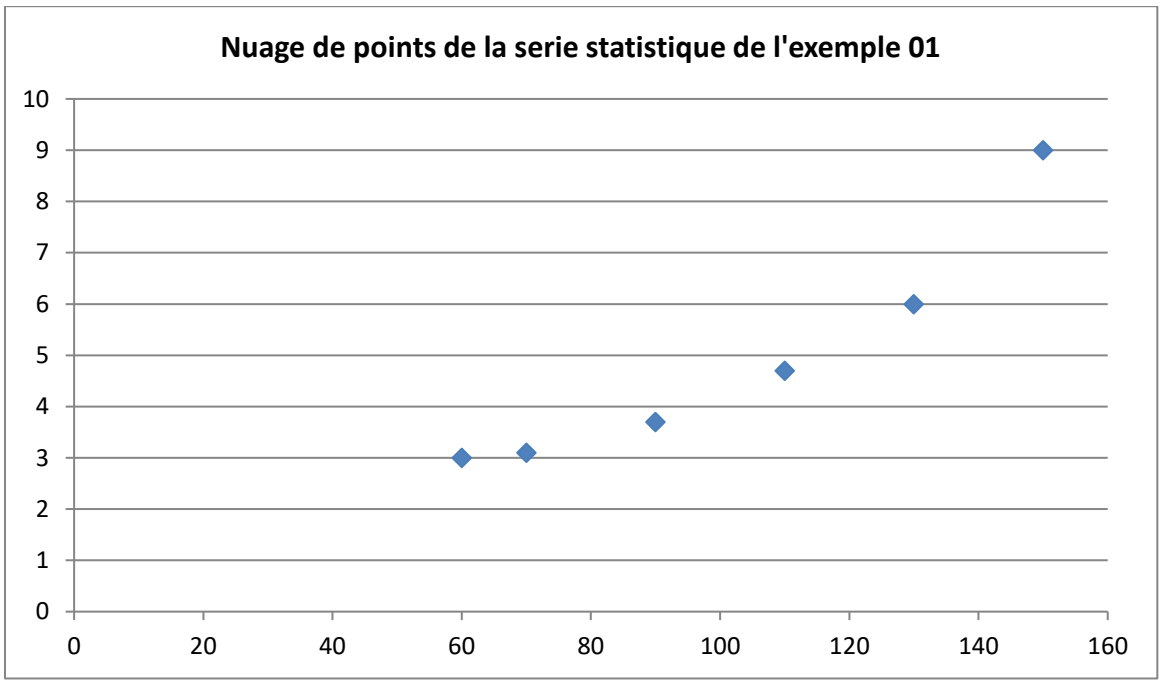
Définition : On appelle série statistique à deux variables (ou série statistique doubles) une série statistique à deux caractères sont étudiés simultanément.

Exemple 01:

On a relevé, pour un modèle de voiture, la consommation en carburant (en L/100 km) pour différentes vitesse (en km/h) sur le cinquième rapport :

Vitesse x_i (en km/h)	60	70	90	110	130	150
Consommation y_i (en L/100 km)	3	3.1	3.7	4.7	6	9

Définition : Dans un repère orthogonal, l'ensemble des points M_i de coordonnées $(x_i,\ y_i)$ constitue le nuage de points associé à la série statistique à deux variables.



Point moyen

Définition : le Point moyen d'un nuage de points G de coordonnées $(\bar{x},\ \bar{y})$où :

$\bar{x}$ représente la moyenne des x_i :

$$\bar{x} = \frac{\boldsymbol{x_1 + x_2 + \cdots + x_n}}{\boldsymbol{n}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i$$

$\bar{y}$ représente la moyenne des y_i :

$$\bar{y} = \frac{\boldsymbol{y_1 + y_2 + \cdots + y_n}}{\boldsymbol{N}} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_i$$

Exemple : le Point moyen de l'exemple 01 est

$$\bar{x} = \frac{\mathbf{60 + 70 + 90 + 110 + 130 + 150}}{\mathbf{6}} = \mathbf{101.66}$$

$$\bar{y} = \frac{\mathbf{3 + 3.1 + 3.7 + 4.7 + 6 + 9}}{\mathbf{6}} = \mathbf{4.91}$$

Donc le point moyen $\boldsymbol{G(101.66,\ 4.91)}$**.**

Ajustement affine par la méthode des moindres carrés

Définition : On appelle covariance de x et de y le nombre

$$\boldsymbol{cov(x,y)} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (x_i - \bar{x})\,(y_i - \bar{y}) = \left(\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i\, y_i \right) - \bar{x}\bar{y}$$

Rappel : La variance de caractère x est :

$$\boldsymbol{V(x)} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (x_i - \bar{x})^{\mathbf{2}} = \left(\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} {x_i}^{\mathbf{2}} \right) - \bar{x}^2 = \boldsymbol{cov(x,x)}$$

La variance de caractère y est :

$$\boldsymbol{V(y)} = \frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (y_i - \bar{y})^{\mathbf{2}} = \left(\frac{\mathbf{1}}{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} {y_i}^{\mathbf{2}} \right) - \bar{y}^2 = \boldsymbol{cov(y,y)}$$

Elle est utilisée pour le calcul de l'écart type : $\sigma(x) = \sqrt{\boldsymbol{V(x)}}$, $\sigma(y) = \sqrt{\boldsymbol{V(y)}}$.

Exemple : Calculer dans l'exemple 01 $\boldsymbol{cov(x,y)}$, $\boldsymbol{cov(x,x)}, \boldsymbol{cov(y,y)}, \sigma(x), \sigma(y)$.

On a

							Somme
x_i	60	70	90	110	130	150	
y_i	3	3.1	3.7	4.7	6	9	
$x_i y_i$	180	217	333	517	780	1350	3377
$x_i^{\mathbf{2}}$	3600	4900	8100	12100	16900	22500	68100
$y_i^{\mathbf{2}}$	9	9.61	13.69	22.09	36	81	171.39

$\bar{x} = \mathbf{101.66}, \qquad \bar{y} = \mathbf{4.91.}$

$\boldsymbol{cov(x,y)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i\, y_i\right) - \bar{x}\bar{y} = \frac{3377}{6} - 499.15 = 63.68.$

$\boldsymbol{V(x)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i^{\mathbf{2}}\right) - \bar{x}^2 = \frac{\mathbf{68100}}{\mathbf{6}} - (\mathbf{101.66})^{\mathbf{2}} = \mathbf{1015.2444.}$

$\boldsymbol{V(y)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_i^{\mathbf{2}}\right) - \bar{y}^2 = \frac{\mathbf{171.39}}{\mathbf{6}} - (\mathbf{4.91})^{\mathbf{2}} = \mathbf{4.4569.}$

$\sigma(x) = \sqrt{\boldsymbol{V(x)}} = \sqrt{\mathbf{1015.2444}} = 31.86.$

$\sigma(y) = \sqrt{\boldsymbol{V(y)}} = \sqrt{\mathbf{4.4569}} = 2.11.$

Théorème : Lors d'un ajustement affine par la méthode des moindres carrés.

1. La droite de régression D de Y en X a pour équation $D(Y/X)$ $Y = aX + b$ où :

$$a = \frac{\boldsymbol{cov(x,y)}}{\boldsymbol{V(x)}}$$

Passe par le point moyen du nuage $G(\bar{x},\ \bar{y})$ c'est-à-dire vérifié $\bar{Y} = a\bar{X} + b$, donc $b = \bar{Y} - a\bar{X}$.

2. La droite de régression D de X en Y a pour équation $D(X/Y)$ $X = a'Y + b'$ où :

$$a' = \frac{\boldsymbol{cov(x,y)}}{\boldsymbol{V(y)}}$$

Passe par le point moyen du nuage $G(\bar{x},\ \bar{y})$ c'est-à-dire vérifié $\bar{X} = a'\bar{Y} + b'$, donc $b' = \bar{X} - a'\bar{Y}$.

Exemple : Calculer dans l'exemple 01 La droite de régression D de Y en X

On a

$\bar{x} = \mathbf{101.66}$, $\bar{y} = \mathbf{4.91}$, $\boldsymbol{cov(x,y)} = 63.68$, $\boldsymbol{V(x)} = \mathbf{1015.2444}$, $\boldsymbol{V(y)} = \mathbf{4.4569.}$

1. $D(Y/X)\ \ Y = aX + b$

$a = \frac{\boldsymbol{cov(x,y)}}{\boldsymbol{V(x)}} = 0.0627, \quad b = \bar{Y} - a\bar{X} = -1.46$.

Donc $D(Y/X)\ \ Y = aX + b = 0.0627X - 1.46$

2. $D(X/Y)\ \ X = a'Y + b'$

$a' = \frac{\boldsymbol{cov(x,y)}}{\boldsymbol{V(y)}} = 14.287, \ \ b' = \bar{X} - a'\bar{Y} = 31.51$

Donc $D(X/Y)\ \ X = a'Y + b' = 14.287Y + 31.51$

Coefficient de corrélation linéaire

Définition : le Coefficient de corrélation linéaire d'une série statistique à deux variables x et y est le nombre r défini par :

$$\boldsymbol{r} = \frac{\boldsymbol{cov(x,y)}}{\sqrt{\boldsymbol{V(x)}}\sqrt{\boldsymbol{V(y)}}} = \frac{\boldsymbol{cov(x,y)}}{\sigma(x)\sigma(\boldsymbol{y})}$$

Remarque :

1. $-1 \leq r \leq 1.$
2. Si $\mathrm{r} = 1$ ou $\mathrm{r} = -1$ alors il ya une corrélation positive ou négative parfaite entre X et Y et les points $(\mathrm{x_i},\ \mathrm{y_i})$ sont tous sur la droite de régression.
   Une corrélation positive c'est-à-dire une augmentation de X entraîne une augmentation de Y.
   Une corrélation négative c'est-à-dire une augmentation de X entraîne une diminution de Y ou le contraire.
3. Si $\mathrm{r} = 0$ alors il n'ya pas de corrélation entre X et Y et les points $(\mathrm{x_i},\ \mathrm{y_i})$ sont dispersés au hasard.
4. Si $0 < r < 1$ alors il y a une corrélation positive faible, moyenne ou forte entre X et Y.
5. Si $-1 < r < 0$ alors il y a une corrélation négative faible, moyenne ou forte entre X et Y.

Exemple : Calculer dans l'exemple 01 le Coefficient de corrélation linéaire.

On a $\boldsymbol{cov(x,y)} = 63.68$, $\sigma(x) = 31.86,\ \sigma(y) = 2.11.$

Donc

$\boldsymbol{r} = \frac{\boldsymbol{cov(x,y)}}{\sigma(x)\sigma(\boldsymbol{y})} =$**0.947**

alors il y a une corrélation positive forte entre X et Y

Coefficient de détermination

Définition : le Coefficient de détermination d'une série statistique à deux variables x et y est le nombre $\boldsymbol{R^2}$défini par :

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{\mathrm{Y}})^2}{\sum_{i=1}^{n}(\mathrm{y_i} - \overline{\mathrm{Y}})^2}$$

Tel que $\widehat{y_i} = ax_i + b$

Exemple : Calculer dans l'exemple 01 le Coefficient de détermination.

							Somme
$\mathrm{x_i}$	60	70	90	110	130	150	
$\mathrm{y_i}$	3	3.1	3.7	4.7	6	9	
$(\mathrm{y_i} - \overline{\mathrm{Y}})^2$	3.65	3.27	1.46	0.044	1.188	16.72	26.332
$\widehat{\mathrm{y_i}} = \mathrm{ax_i} + \mathrm{b}$	2.3	2.92	4.18	5.43	6.69	7.94	
$(\widehat{\mathrm{y_i}} - \overline{\mathrm{Y}})^2$	6.8	3.92	0.53	0.28	3.17	9.21	23.92

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y_i} - \overline{\mathrm{Y}})^2}{\sum_{i=1}^{n}(\mathrm{y_i} - \overline{\mathrm{Y}})^2} = \frac{23.92}{26.332} = 0.9.$$

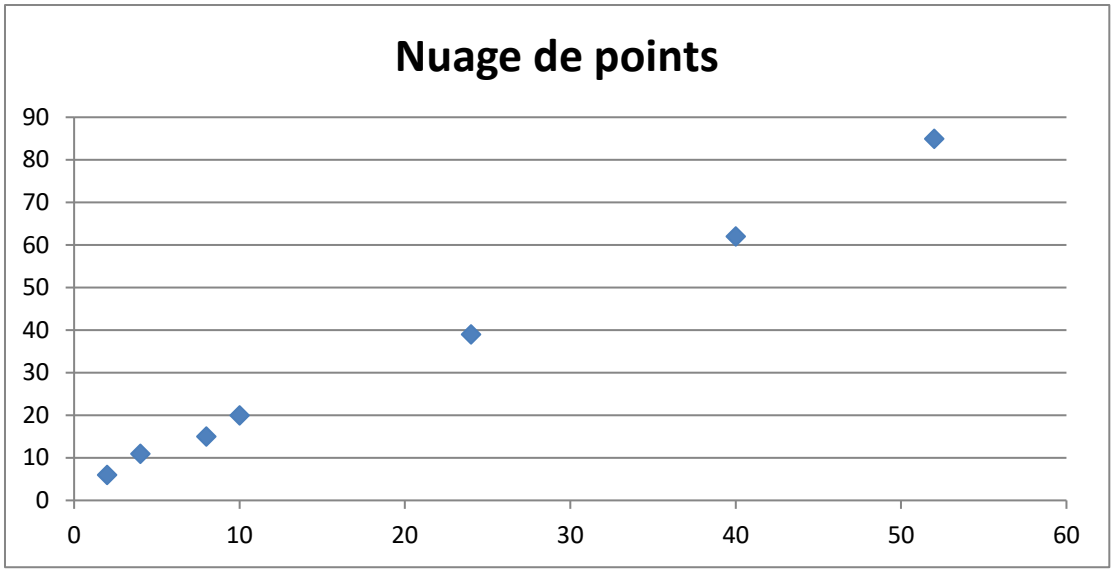
Exercice : Dans la série statistique suivante, X représente le nombre de jours d'exposition au soleil d'une feuille et Y le nombre de stomates aérifères au millimètre carré:

X	2	4	8	10	24	40	52
y	6	11	15	20	39	62	85

1. Tracer le nuage des points.
2. Calculer le coefficient de corrélation linéaire entre X et Y. Conclusion?
3. Déterminer l'équation de la droite de régression de Y en fonction de X.
4. Si on expose au soleil une feuille 15 jours; quel est le nombre de stomates aérifères peut-on prévoir ?

Corrigé type

1. **Le nuage des points**



2. Le coefficient de corrélation linéaire entre X et Y.

								Somme
x_i	2	4	8	10	24	40	52	140
y_i	6	11	15	20	39	62	85	238
$x_i y_i$	12	44	120	200	936	2480	4420	8212
x_i^2	4	16	64	100	576	1600	2704	5064
y_i^2	36	121	225	400	1521	3844	7225	13372

On a

$\bar{x} = \frac{140}{7} = \mathbf{20}$ $\bar{y} = \frac{\mathbf{238}}{7} = \mathbf{34.}$

$\boldsymbol{cov(x,y)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i\, y_i\right) - \bar{x}\bar{y} = \frac{8212}{7} - 680 = 493.14.$

$\boldsymbol{V(x)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} x_i^{\mathbf{2}}\right) - \bar{x}^2 = \frac{\mathbf{5064}}{\mathbf{7}} - (\mathbf{20})^{\mathbf{2}} = \mathbf{323.43.}$

$\boldsymbol{V(y)} = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_i^{\mathbf{2}}\right) - \bar{y}^2 = \frac{\mathbf{13372}}{\mathbf{7}} - (\mathbf{34})^{\mathbf{2}} = \mathbf{754.28.}$

$\sigma(x) = \sqrt{\boldsymbol{V(x)}} = \sqrt{\mathbf{323.43}} = 17.98.$

$\sigma(y) = \sqrt{\boldsymbol{V(y)}} = \sqrt{\mathbf{754.28}} = 27.46.$

Donc le coefficient de corrélation linéaire entre X et Y.

$\boldsymbol{r} = \frac{\boldsymbol{cov(x,y)}}{\sigma(x)\sigma(\boldsymbol{y})} = 0.99$

alors il y a une corrélation positive forte entre X et Y

3. L'équation de la droite de régression de Y en fonction de X.

$D(Y/X)\ Y = aX + b$

$a = \frac{cov(x,y)}{V(x)} = 1.52, \quad b = \bar{Y} - a\bar{X} = 3.6$.

Alors $D(Y/X)\ Y = 1.52\,X + 3.6$

4. le nombre de stomates aérifères

On a $Y = 1.52\,X + 3.6 = 1.52 \times 15 + 3.6 = 26.4$

								Somme
x_i	2	4	8	10	24	40	52	140
y_i	6	11	15	20	39	62	85	238
$(\mathrm{y_i} - \overline{\mathrm{Y}})^2$	784	529	361	196	25	784	2601	
$\widehat{\mathrm{y_i}} = \mathrm{ax_i} + \mathrm{b}$	6.64	9.68	15.76	18.8	40.08	64.4	82.64	
$(\widehat{\mathrm{y_i}} - \overline{\mathrm{Y}})^2$	748.57	591.46	332.69	231.04	36.96	924.16	2365.84	

$Y = 1.52\,X + 3.6$