1 COURS
Rappels sur les statistiques double

L'objectif de cette étude statistique est d'étudier sur une même population de N individus, deux caractères différents (ou modalités différentes) et de rechercher s'il existe un lien ou corrélation entre ces deux variables. Exemple de relations possibles entre les variables suivantes : taille et âge ; diabète et poids ; taux de cholestérol et régime alimentaire ; niche écologique et population ; ensoleillement et croissance végétale ; toxine et réaction métabolique ; survie et pollution ; effets et doses; organe 1 et 2 ; organe et fonction biologique ; ...

1 les séries statistique double

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(enkm/h)	60	70	90	110	130	150
Consommation y_i(enL/100km)	3	3.1	3.7	4.7	6	9

1.1.1 Nuage de points

Définition

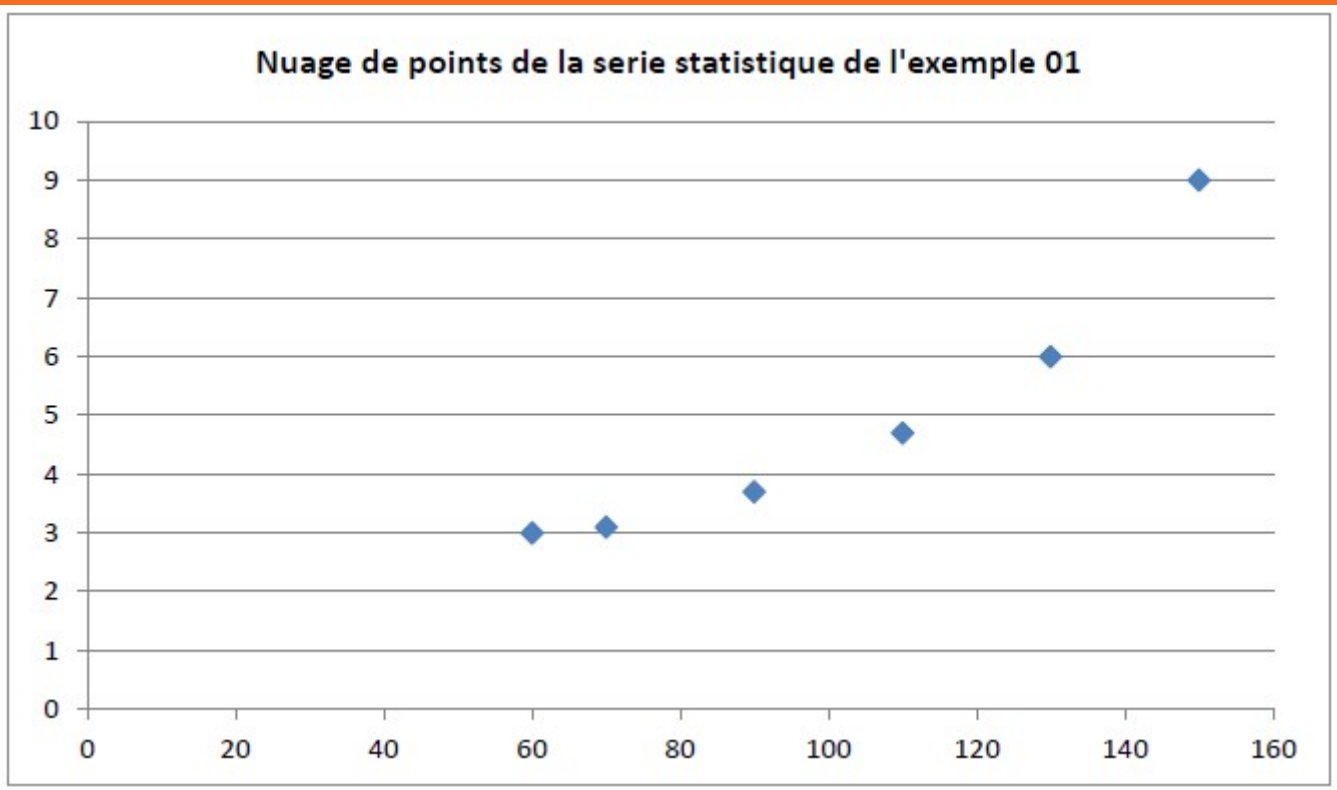
Dans un repère orthogonal, l'ensemble des points M_i de coordonnées (x_i, y_i) constitue le nuage de points associé à la série statistique à deux variables.

1.1.2 Les moyennes marginales

Définition

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

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



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

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

Exemple

les moyennes marginales de l'exemple 01 sont:

$$\bar{x} = \frac{60 + 70 + 90 + 110 + 130 + 150}{6} = 101.66$$
$$\bar{y} = \frac{3 + 3.1 + 3.7 + 4.7 + 6 + 9}{6} = 4.91$$

1.1.3 La covariance

Définition

On appelle covariance de x et de y le nombre

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

Rappel

La variance de caractère x est :

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

La variance de caractère y est :

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

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 $\text{cov}(\boldsymbol{x},\boldsymbol{y}), \boldsymbol{\text{cov}}(\boldsymbol{x},\boldsymbol{x}), \boldsymbol{cov}(\boldsymbol{y},\boldsymbol{y}), \sigma(x), \sigma(y)$. 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
${x_i}^2$	3600	4900	8100	12100	16900	22500	68100
${y_i}^2$	9	9.61	13.69	22.09	36	81	171.39

$\bar{x} = 101.66, \quad \bar{y} = 4.91$

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

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

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

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

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

Théorème

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

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

et $\quad b = \bar{Y} - a\bar{X}.$

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

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

et $\quad b' = \bar{X} - a'\bar{Y}.$

Exemple

Calculer dans l'exemple 01 La droite de régression D de Y en X On a

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

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

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

Donc $D(Y/X)\ Y = aX + b = 0.0627X - 1.46$
2. $D(X/Y) \quad X = a'Y + b'$

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

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

1.1.4 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 :

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

Remarque

1. $-1 \leq r \leq 1$.
2. Si $r = 1$ ou $r = -1$ alors il ya une corrélation positive ou négative parfaite entre X et Y et les points $(\text{x}_\text{i}, \text{y}_\text{i})$ 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 $r = 0$ alors il n'ya pas de corrélation entre X et Y et les points (x_i, 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 $\text{cov}(x, y) = 63.68, \sigma(x) = 31.86, \sigma(y) = 2.11$.
Donc

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

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