- **Distribution à deux variables**

1. Corrélation :

- **Distribution à deux variables, notion de corrélation :**

Dans les cours précédents, les distributions statistiques étudiées portaient sur un seul caractère, une seule variable, que l'on observait sur chacune des unités statistiques composant la population.

Or il est tout à fait possible de se trouver face d'une population telle qu'on puisse étudier deux caractères différents sur une même unité statistique.

Par exemple :

Unité statistique	Variable1 : désignée par x	Variable2 : désignée par y
Enfants d'une école	Taille	Poids
Enfants d'une école	Taille	Moyenne arithmétique de notes obtenues au cours de l'année scolaire
Mariages célébrés en Algérie en 2011	Age de l'époux au moment du mariage	Age de l'épouse au moment du mariage
Année	Nombre de spectateurs payants dans les stades Algériens	Prix moyens des places dans les stades Algériens
Mois	Tonnage transporté par une entreprise des transports routiers	Consommation de carburants dans la même entreprise

Une population statistique ayant été étudiée à un double point de vue, le tableau statistique traduisant les observations faites se présentera de la façon suivante :

Variable1 X_i	Variable2 Y_i	Effectifs
X_1	Y_1	1
X_2	Y_2	1
.	.	
.	.	
.	.	1
X_i	Y_i	1
XN	YN	1

Ce qui signifiera que les N unités statistiques observées :

Une unité représente à la fois la mesure x_1 de la variable 1, et la mesure y_1 de la variable 2

Une unité représente à la fois la mesure x_2 de la variable 1, et la mesure y_2 de la variable 2

Et ainsi de suite

Si cette présentation est adoptée la colonne des effectifs dans laquelle tous les effectifs sont égaux à 1, pourra être supprimée.

Sur le tableau ainsi obtenu on pourra s'attacher à l'étude des variations simultanées (croissance ou décroissance) des deux ou variables x et y. il est intuitif que pourront se présenter les situations suivantes :

a) Les variations des deux variables n'ont aucun lien entre elles ; par exemple, les tailles respectives des N enfants de l'école et les moyennes arithmétiques respectives des notes obtenus par ces N enfants au cours de l'année scolaire.
On dira dans ce cas que les deux variables sont indépendantes

b) Les deux variables son liées l'un à l'autre de façon telle que la connaissance de la mesure, pour chaque unité statistique, de l'une des deux variables, entraine la connaisse exacte, pour la même unité de l'autre variable. Par exemple le revenu annuel déclaré par chacun des N contribuables d'un pays, et le montant de l'impôt sur le revenu payé par chacun de ces N contribuables.
On dira alors que les deux variables sont en liaison fonctionnelle.

c) Sans être liés rigoureusement, les deux variables sont en dépendance, plus ou moins marquée. Leurs valeurs varient dans le même sens(les deux variables croissent ensemble, ou décroissent ensemble).
On dira alors que les deux variables sont en corrélation, positive ou négative suivant le cas.
Exemple :
   - Le poids et la taille des enfants d'une école sont deux variables en corrélation positive.
   - Le nombre de spectateurs payants dans les stades, et le prix des places sont deux variables qui a priori sont en corrélation négative (nous disons ici « a priori » car d'autres variables pourraient intervenir, par exemple la variable « pouvoir d'achat »).

C'est sur la situation c, c'est-à-dire sur la corrélation entre deux variables que nous allons concentrés.

A. Mesure de la corrélation. Coefficient de dépendance :

x_i	y_i	Différence sur x_i : $x_{i+1} - x_i$	Différence sur y_i : $y_{i+1} - y_i$	Produits des différences	
				-	+
16	20				
18	24	18 – 16 = +2	24 – 20 =+4		8
23	28	23 – 18 = +5	28 – 24= +4		20
24	22	24 – 23 = +1	22 – 28 = -6	-6	
28	32	28 – 24 = +4	32 – 22= +10		40
29	28	29 – 28 = +1	28 – 32 = - 4	-4	
26	32	26 – 29 = -3	32 – 28 = +4	-12	
31	36	31 – 26 = +5	36 – 32 = +4		20
32	41	32 – 31 = +1	41 – 36 = +5		5
34	41	34 – 32 = +2	41 – 41 = 0		0
			Total	-22	93

Coefficient de dépendance = $\frac{\textbf{sommes des produits positifs} - /\textbf{somme des produits négatifs}/}{\textbf{somme des produits positifs} + /\textbf{somme des produits négatifs}/} = \frac{93-22}{93+22} = \frac{71}{115} = +0.62$

La marche à suivre pour calculer le coefficient de dépendance est clairement indiquer sur le tableau de calcul.

Interprétons le résultat obtenu :

Les différences calculées sur x et y montrent dans quel sens (positif ou négatif) varient les mesures des deux variables quand on passe d'une valeur à la suivante.

Si sur une même ligne les différences sont de même signe – donc variation dans le même sens pour les deux variables – leur produit sera positif. Il sera négatif dans le cas contraire.

Si tous les produits sont positifs ; la somme des produits négatifs est évidemment nulle. Le coefficient de dépendance sera égale a +1. La dépendance entre les deux variables sera positive et très serrée, les deux variables varient toujours dans le même sens.

Si les produits de différences sont tous négatifs, la somme des produits positifs est nulle, le coefficient de dépendance est égale a – 1. La dépendance entre les deux variables x et y sera encore très serrée, les deux variables varient toujours dans le même sens.

Nous venons d'envisager des cas extrêmes, - produits des différences tous positifs, ou tous négatifs – qui montrent que le coefficient de dépendance peut varier de -1 à +1.

Dans le cas ou la somme des produits négatifs et la somme des produits positifs seraient égale en valeur absolue, le coefficient de dépendance serait égale à zéro.

Dans le calcul que nous venons de faire, le coefficient de dépendance, égal à 0.62 est la marque d'une bonne corrélation positive.

B. **<u>Mesure de la corrélation, calcul pratique du coefficient de corrélation linéaire :</u>**

Dans le cas ou le nuage de points qui permet de croire à l'existence d'une corrélation entre les deux variables x et y prend une forme allongée telle que les points qui le constituent paraissent s'être regroupés au voisinage d'une droite, un coefficient de corrélation linéaire, que nous désignerons par r, peut être calculé comme suit :

Faisons d'abord les changements de variables :

$$\boldsymbol{X_1 = \frac{x_1 - \bar{x}}{\sigma_x}} \qquad\qquad \boldsymbol{Y_1 = \frac{y_1 - \bar{y}}{\sigma_y}}$$

$\bar{x}$ Désignant la moyenne arithmétique des N valeurs ; $\bar{x} = \frac{\sum x_i}{N}$

$\bar{y}$ Désignant la moyenne arithmétique des N valeurs ; $\bar{y} = \frac{\sum y_i}{N}$

$\boldsymbol{\sigma_x}$ Désignant l'écart type des N valeurs de x ; $\boldsymbol{\sigma_x^2 = \frac{\sum (x_i - \bar{x})^2}{N}}$

$\boldsymbol{\sigma_y}$ Désignant l'écart type des N valeurs de y ; $\boldsymbol{\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{N}}$

Les nouvelles variables

Corrélation

1. **Mesure de la corrélation, calcul pratique du coefficient de corrélation linéaire :**

Dans le cas ou le nuage de points qui permet de croire à l'existence d'une corrélation entre les deux variables x et y prend une forme allongée telle que les points qui le constituent paraissent s'être regroupés au voisinage d'une droite.

Un coefficient de corrélation linéaire, que nous désignerons par r, peut être calculé comme suit :

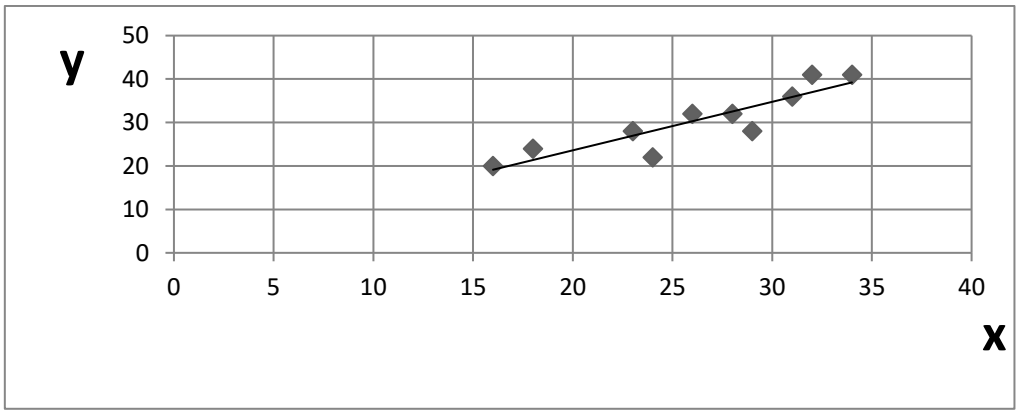
$$\boldsymbol{r} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} \quad \textbf{- Et :} \ (-1) \leq r \leq +1$$

- Le signe de coefficient de corrélation $\boldsymbol{r}$ indique le sens positif ou négatif de la corrélation.
- La valeur absolue de $\boldsymbol{r}$ est inférieure ou égale à **1,** est la corrélation est d'autant plus serrée que cette valeur absolue est plus voisine de **1**

Exemple de calcul du coefficient de corrélation linéaire entre deux variables x et y:

x_i	16	18	23	24	28	29	26	31	32	34
y_i	20	24	28	22	32	28	32	36	41	41

La représentation graphique a montrée que l'hypothèse d'une corrélation linéaire positive pouvait être retenue (nuage aplati, allongé, de forme linéaire, de pente positive)



x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	20	-10,1	-10,4	102,01	108,16	105,04
18	24	-8,1	-6,4	65,61	40,96	51,84
23	28	-3,1	-2,4	9,61	5,76	7,44
24	22	-2,1	-8,4	4,41	70,56	17,64
28	32	1,9	1,6	3,61	2,56	3,04
29	28	2,9	-2,4	8,41	5,76	-6,96
26	32	-0,1	1,6	0,01	2,56	-0,16
31	36	4,9	5,6	24,01	31,36	27,44
32	41	5,9	10,6	34,81	112,36	62,54
34	41	7,9	10,6	62,41	112,36	83,74
261	304	0	0	314,9	492,4	351,6

- La moyenne arithmétique des x ; $\bar{x} = \frac{\sum x_i}{N} = \frac{261}{10} = 26{,}1$
- La moyenne arithmétique des y ; $\bar{y} = \frac{\sum y_i}{N} = \frac{304}{10} = 30{,}4$

- $\boldsymbol{r} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} = \frac{\mathbf{351,60}}{\sqrt{\mathbf{492,40 \times 314,90}}}$

- $\boldsymbol{r} = \frac{\mathbf{351,60}}{\sqrt{\mathbf{155\,056,76}}} = \frac{\mathbf{351,60}}{\mathbf{393,77}} = \mathbf{+0,89}$

Corrélation positive et assez serrée, le coefficient r ayant une valeur absolue voisine de 1

2. <u>Droite de régression :</u>

Lorsque de variables sont en corrélation linéaire, il arrive que l'on puisse admettre que les variations de l'une des deux variables sont causes des variations de l'autre.

Dans l'exemple précédent on peut considérer que le poids d'un enfant dépendre de son âge.

Il sera alors légitime d'essayer d'exprimer par une fonction linéaire les valeurs de la variable considérée comme conséquence, à partir des valeurs de la variable considérée comme cause.

Si la variable fonction (conséquence) est désignée par ***y***, la variable cause étant désignée par ***x***, il sera donc normal de rechercher une fonction ***y = ax + b*** qui permettra d'ajuster la variable ***y*** à partir de la variable ***x***.

La droite dont il est question est dite droite de régression, ou droite d'estimation de ***y*** à partir de ***x***.

Il suffit simplement de substituer aux deux colonnes ; « variable x_i » ; « effectif y_i » du tableau habituel le tableau à deux colonnes « variable x_i », « variable y_i ».

Les deux paramètres $\boldsymbol{a}$ et **b** sont de la fonction cherchée sont donnés par la formule suivante :

$$\boldsymbol{a} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ Et } \mathbf{b} = \bar{y} - a\bar{x}$$

<u>Exemple :</u> Reprenons le même exemple précédent.

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	20	-10,1	-10,4	102,01	108,16	105,04
18	24	-8,1	-6,4	65,61	40,96	51,84
23	28	-3,1	-2,4	9,61	5,76	7,44
24	22	-2,1	-8,4	4,41	70,56	17,64
28	32	1,9	1,6	3,61	2,56	3,04
29	28	2,9	-2,4	8,41	5,76	-6,96
26	32	-0,1	1,6	0,01	2,56	-0,16
31	36	4,9	5,6	24,01	31,36	27,44
32	41	5,9	10,6	34,81	112,36	62,54
34	41	7,9	10,6	62,41	112,36	83,74
261	304	0	0	314,9	492,4	351,6

$$\boldsymbol{a} = \frac{\mathbf{351,60}}{\mathbf{314,9}} = \mathbf{1,117} \text{ Et } \mathbf{b} = \mathbf{30,4} - (\mathbf{1,117}) \times (\mathbf{26,1}) = \mathbf{1,25}$$

Equation de régression $\mathbf{y} = \mathbf{1,117}x + \mathbf{1,25}$

Par analogie on pourra chercher l'équation de la droite de régression de x par rapport à y, équation de la forme **:**

$\boldsymbol{x = a'y + b'}$ **avec :** $\boldsymbol{a' = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(y_i - \overline{y})^2}}$ **et b'** $\boldsymbol{= \overline{x} - a'\overline{y}}$

Dans le même exemple : $\boldsymbol{a' = \frac{351{,}60}{492{,}40} = 0{,}714 \; et \; b' = 26{,}1 - (0{,}714 \times 30{,}4) = 4{,}4}$

Equation de régression de x par rapport à y :

$$\boldsymbol{x = 0{,}714y + 4{,}4}$$

Exercice :

x_i	y_i	$(x_i - \overline{x})$	$(y_i - \overline{y})$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
29	22	-1,5	-12,5	2,25	156,25	18,75
35	36	4,5	1,5	20,25	2,25	6,75
36	38	5,5	3,5	30,25	12,25	19,25
41	21	10,5	-13,5	110,25	182,25	-141,75
12	21	-18,5	-13,5	342,25	182,25	249,75
27	21	-3,5	-13,5	12,25	182,25	47,25
23	29	-7,5	-5,5	56,25	30,25	41,25
12	25	-18,5	-9,5	342,25	90,25	175,75
23	41	-7,5	6,5	56,25	42,25	-48,75
44	54	13,5	19,5	182,25	380,25	263,25
45	52	14,5	17,5	210,25		
39	54	8,5	19,5	72,25	380,25	165,75
366	414	0	0	1437	1640,75	797,25

- **Exercice :**

La société anonyme par action R augmente son capital. On a relevé pendant 10 jours le cours en bourse de l'action de celui du droit de souscription.

x_i	98	94	97	98	100	102	102	104	104	101
y_i	6.5	5.4	6.1	6.4	6.9	8	7.5	7.5	7.4	7.3

- Calculer le coefficient de corrélation linéaire entre les deux variables x et y
- Donner l'équation de la droite de régression qui permet d'estimer le cours du droit de souscription à partir du cours d'action

Solution :

x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
98	6,5	-2	-0,4	4	0,16	0,8
94	5,4	-6	-1,5	36	2,25	9
97	6,1	-3	-0,8	9	0,64	2,4
98	6,4	-2	-0,5	4	0,25	1
100	6,9	0	0	0	0	0
102	8	2	1,1	4	1,21	2,2
102	7,5	2	0,6	4	0,36	1,2
104	7,5	4	0,6	16	0,36	2,4
104	7,4	4	0,5	16	0,25	2
101	7,3	1	0,4	1	0,16	0,4
1000	69	0	0	94	5,64	21,4

- La moyenne arithmétique des x ; $\bar{x} = \frac{\sum x_i}{N} = \frac{1000}{10} = 100$
- La moyenne arithmétique des y ; $\bar{y} = \frac{\sum y_i}{N} = \frac{69}{10} = 6.9$

- $\boldsymbol{r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} = \frac{21.4}{\sqrt{94 \times 5.64}} = \frac{21.4}{\sqrt{530.16}} = +0.93}$

Corrélation positive et assez serrée, le coefficient r ayant une valeur absolue voisine de 1

Equation de régression : $\boldsymbol{y = ax + b}$

$\boldsymbol{a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}}$ **Et b** $\boldsymbol{= \bar{y} - a\bar{x}}$

$\boldsymbol{a = \frac{21.4}{94} = 0.228}$ **Et b** $\boldsymbol{= 6.9 - (0.228) \times (100) = -15.9}$

Equation de régression $\boldsymbol{y = 0.228x + (-15.9) = 0.228x - 15.9}$

$$\boldsymbol{y = 0.228x - 15.9}$$