Cours
Statistiques bi-variées

K. BELATTAR,

Département Informatique - Université d’Alger 1

Introduction aux statistiques bi-variées

Objectif:

- Etudier sur une même population de N individus deux variables (caractères) différentes.
- Analyser la relation entre deux variables.

Exemples:

- Taille et poids
- Taux de chaumage et sexes
- Effets et doses des médicaments,
- …etc.

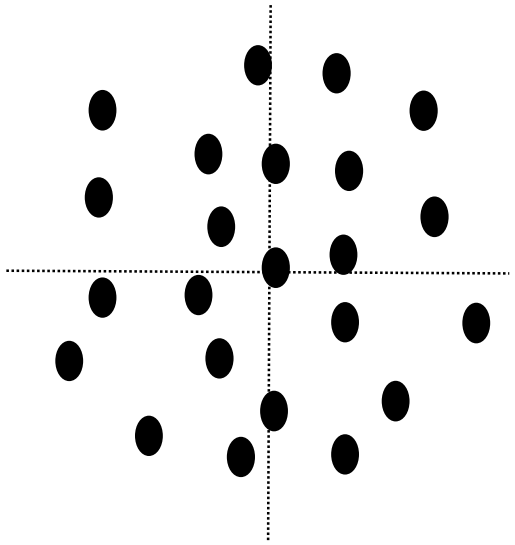
Introduction aux statistiques bi-variées

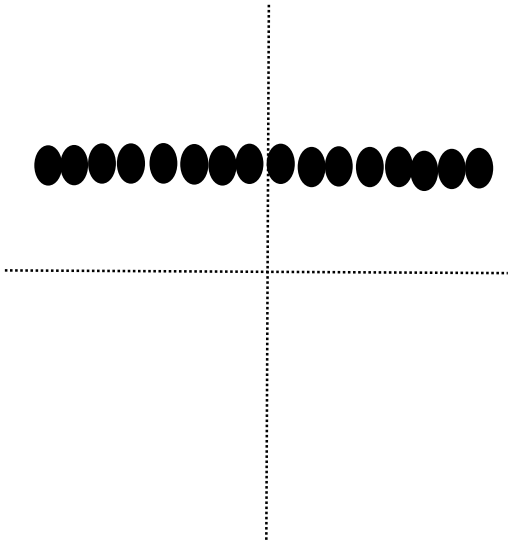
Analyser une relation entre deux variables $\boldsymbol{X}$ $\boldsymbol{et}$ $\boldsymbol{Y}$**:**

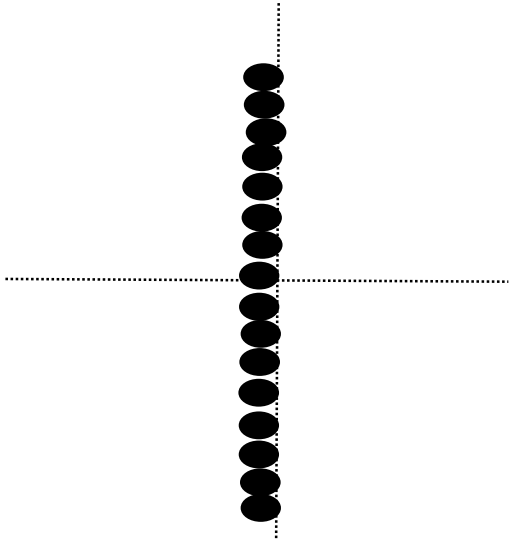
- ✓ **Déterminer s'il existe une relation** entre les deux variables,
- ✓ **Caractériser la relation** entre les deux variables :
  - Forte vs faible (**selon l'intensité**)
  - Linéaire vs non linéaire (**selon la forme**)
  - Positive vs négative (**selon la direction**)
- ✓ **Tester** si la liaison est **statistiquement significative (tests d'hypothèse)**,
- ✓ **Valider la relation** identifiée (**absence de biais**).

Type de relations (selon l'intensité)

Nuage de points

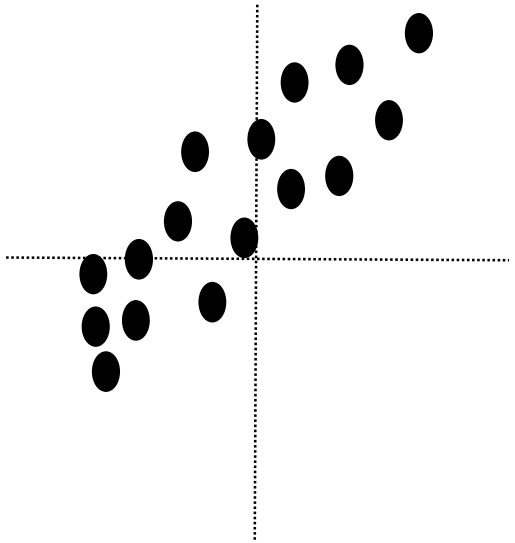




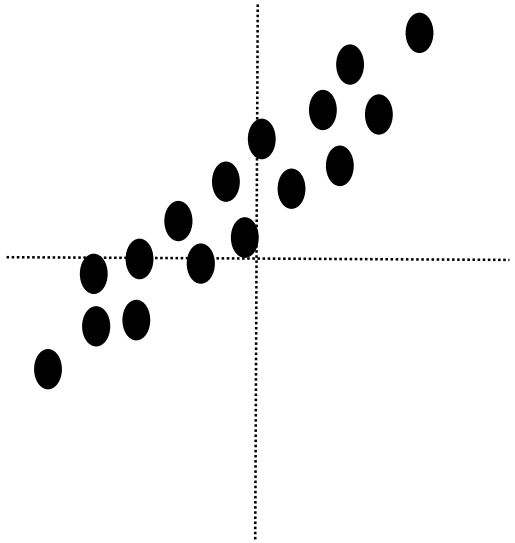


Absence d'une relation

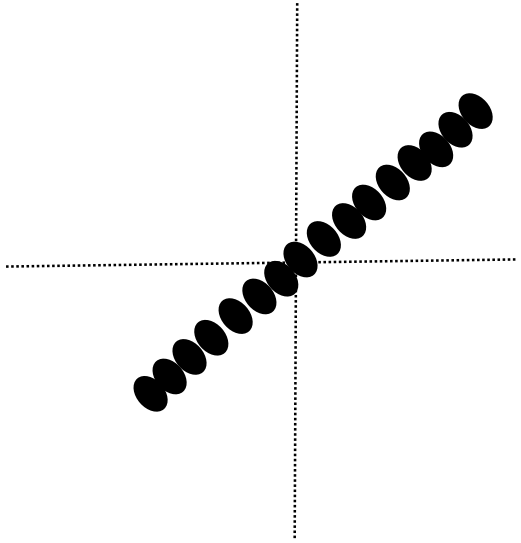
Type de relations (selon l'intensité)



Relation faible

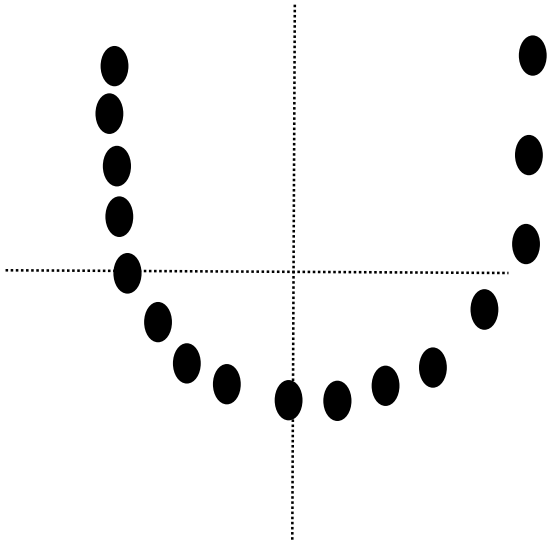


Relation modérée

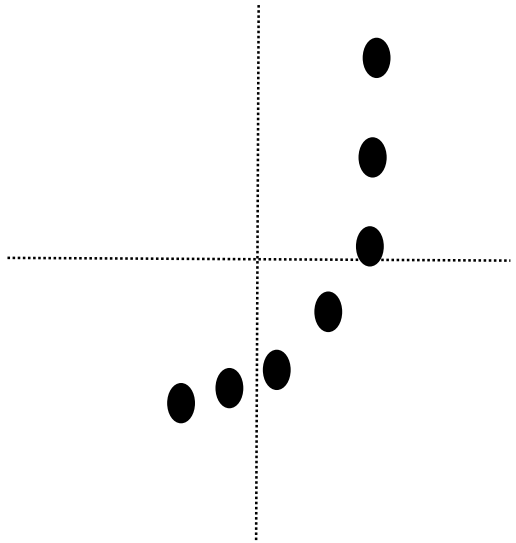


Relation forte

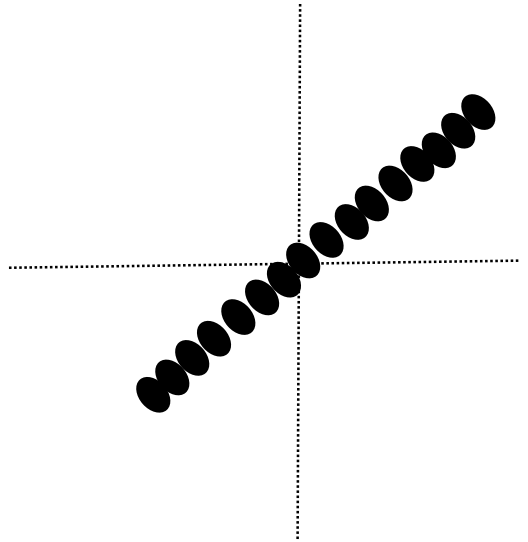
Type de relations (selon la forme)



Relation non linéaire et non monotone

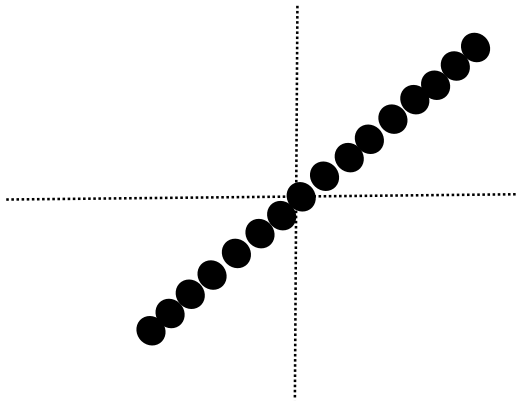


Relation non linéaire et monotone

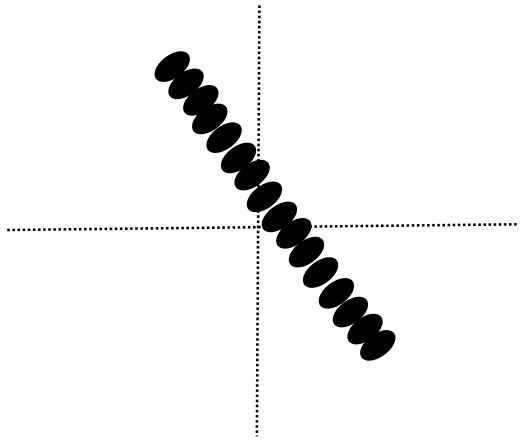


Relation linéaire (toujours monotone)

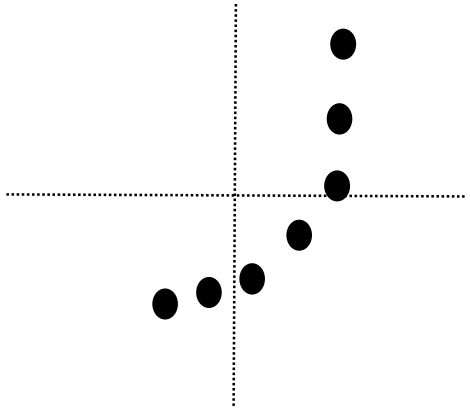
Type de relations (selon la direction)



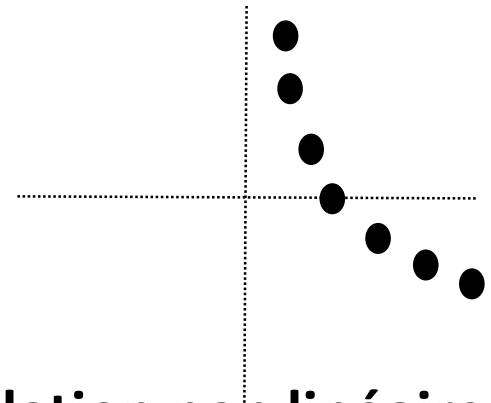
Relation positive (et linéaire)



Relation négative (et linéaire)



Relation non linéaire positive



Relation non linéaire et négative

Séries statiques à deux variables

- Considérons deux variables (quantitatives, qualitatives, une quantitative et l'autre qualitative) $\boldsymbol{X}$ et $\boldsymbol{Y}$ définies sur la **même population** d'effectif total $\boldsymbol{N}$.

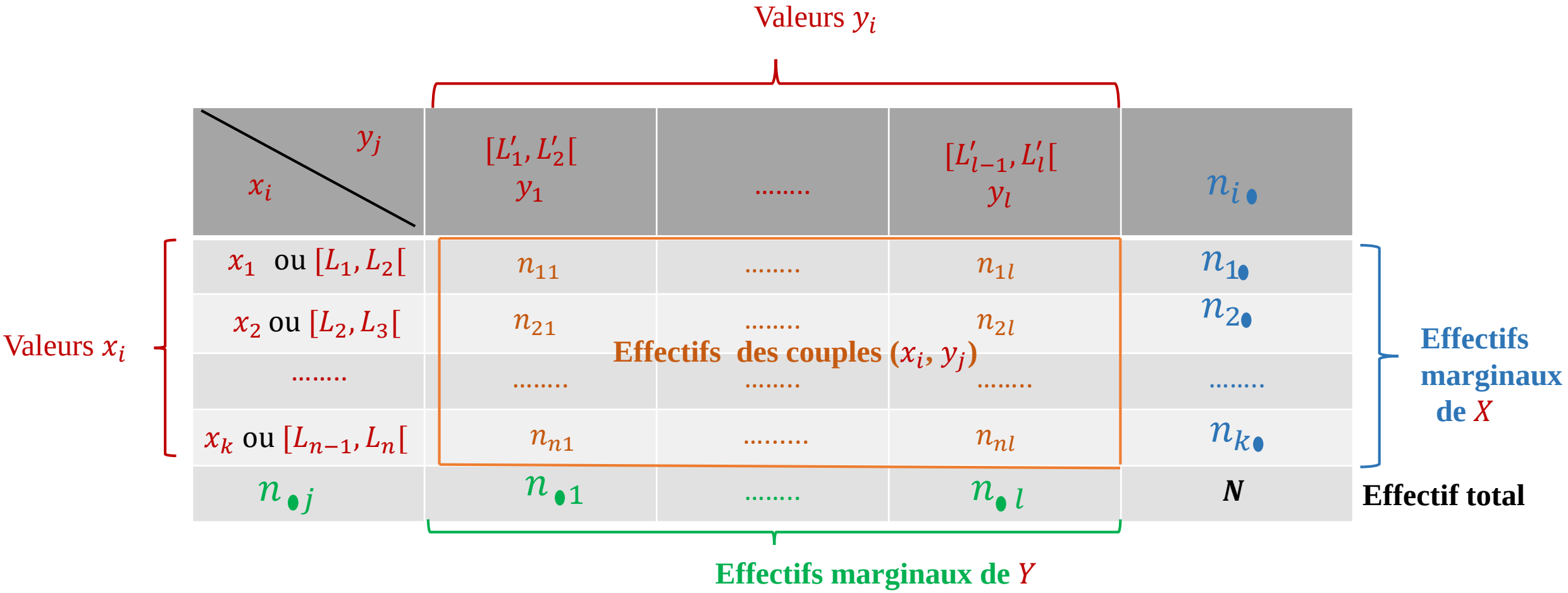
$$N = \sum_i \sum_j n_{ij}$$

$\boldsymbol{n_{ij}}$: effectif conjoint des valeurs $\boldsymbol{x_i, y_j}$ pour lesquels $\boldsymbol{X}$ prend la valeur $\boldsymbol{x_i}$ et $\boldsymbol{Y}$ prend la valeur $\boldsymbol{y_j}$.

- Une série statistique bi-variée est représentée par l'ensemble des couples

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_l)\}$$

Tableau de contingence

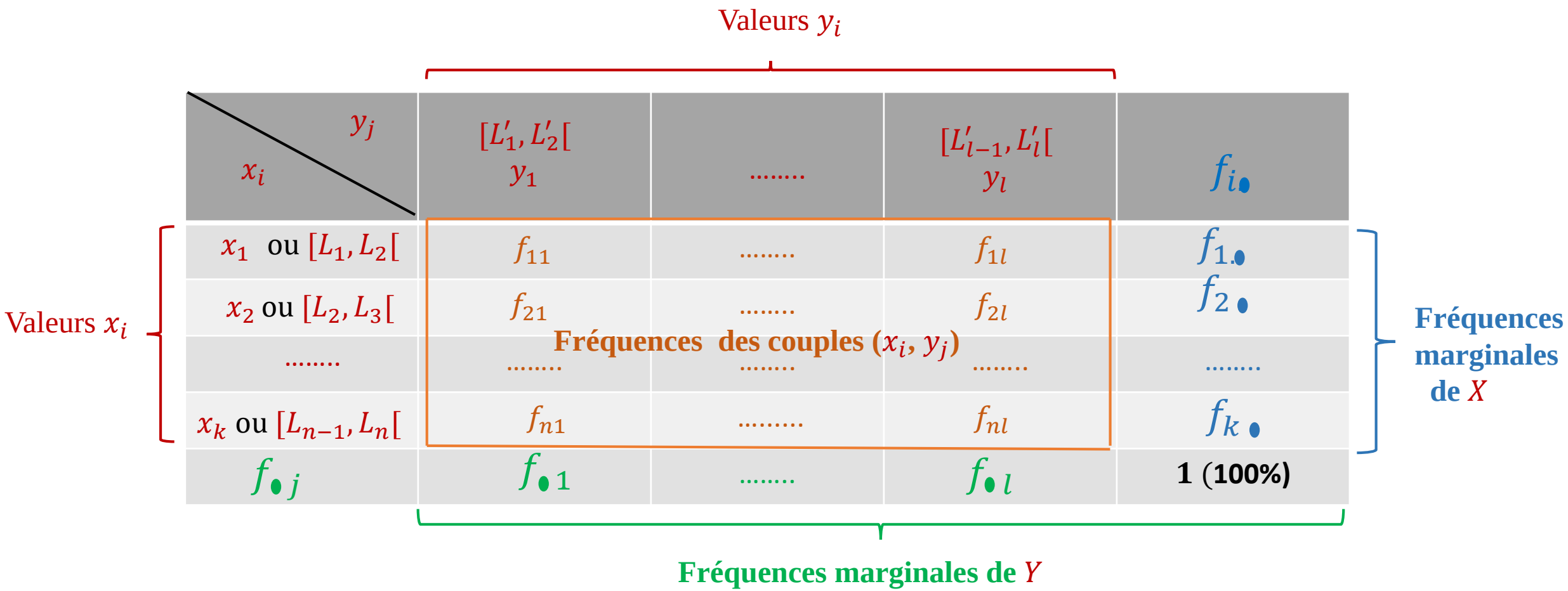


Effectif marginal de X: $n_{i.} = \sum_j n_{ij}$

Effectif marginal de Y : $n_{.j} = \sum_i n_{ij}$

Effectif des couples (x_i, y_j) : $N = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{.j}$

Tableau de contingence



Fréquence marginale de X: $f_{i.} = \sum_j f_{ij}$

Fréquence marginale de Y : $f_{.j} = \sum_i f_{ij}$

Fréquence des couples (x_i, y_j) : $\sum_i \sum_j f_{ij} = \sum_i f_{i.} = \sum_j f_{.j} = 1$ avec $f_{ij} = \frac{n_{ij}}{N}$

Tableau de contingence

Exemple :

Soit la répartition des salaires d'une entreprise selon le nombre d'enfant (X) et le salaire mensuel(Y) en 1000DA.

x_i \ y_j	[2-6[	[6-10[	[10-14[	$n_{i.}$	$f_{i.}$
1	15	8	2	25	0.42
2	13	4	1	18	0.3
3	11	3	3	17	0.28
$n_{.j}$	39	15	6	60	
$f_{.j}$	0.65	0.25	0.1		1

Distributions marginales

- **Effectif (fréquence) marginal (e) de X:** $n_{i\,.} = \sum_j n_{ij}$ $(f_{i\,.} = \sum_j f_{ij})$
- **Effectif (fréquence) marginal (e) de Y :** $n_{.j} = \sum_i n_{ij}$ $(f_{.j} = \sum_i f_{ij})$
- **Effectifs (fréquence) des couples (x_i, y_j) :**

$$\mathrm{N} = \sum_i \sum_j n_{ij} = \sum_i n_{i.} = \sum_j n_{.j} \left(\sum_i \sum_j f_{ij} = \sum_i f_{i.} = \sum_j f_{.j} = 1 \text{ avec } f_{ij} = \frac{n_{ij}}{N}\right)$$

- **Moyenne marginale (variables discrètes)**

$$\bar{X} = \frac{\sum_{i=1}^{N} n_{\,i.} x_i}{N} = \sum_{i=1}^{N} f_{i.} x_i \qquad \bar{Y} = \frac{\sum_{j=1}^{N} n_{.\mathrm{j}}\, y_j}{N} = \sum_{i=1}^{N} f_{.\mathrm{j}}\, y_j$$

- **Variance marginale (d'un échantillon)**

$$\sigma^2(X) = \frac{\sum_{i=1}^{N} n_{i.}(x_i - \bar{X})^2}{N-1} \qquad \sigma^2(Y) = \frac{\sum_{j=1}^{N} n_{.j}(y_j - \bar{Y})^2}{N-1}$$

Distributions marginales

x_i	$n_{i.}$	$f_{i.}$	$\overline{X} = f_{i.}x_i$	$\sigma^2(X)$
1	25	0.42	0.42	0.14
2	18	0.3	0.6	0.6
3	17	0.28	0.84	1.34
$n_{.j}$	60		1.86	2,08 (∑)

Distributions marginales

y_j	$n_{.j}$	$f_{.j}$	$\overline{Y} = f_{.j}\, C_j$	$\sigma^2(Y)$
[2-6[	39	0.65	2.6	1.3
[6-10[	15	0.25	2	9.15
[10-14[	6	0.1	1.2	11. 86

Distributions conditionnelles

- **Fréquence conditionnelle de X sachant que $Y = y_j$:**

$$f_{i/j} = f_i^j = \frac{n_{ij}}{n_{.j}} \qquad \sum_{i=1}^{k} f_i^j = 1$$

f_i^j: proportion d'individus présentant la valeur x_i parmi l'ensemble des individus présentant la valeur y_j de Y.

- **Fréquence conditionnelle de Y sachant que $X = x_i$:**

$$f_{j/i} = f_i^j = \frac{n_{ij}}{n_{i.}} \qquad \sum_{j=1}^{l} f_i^j = 1$$

f_i^j: proportion d'individus présentant la valeur y_i parmi l'ensemble des individus présentant la valeur x_i de X.

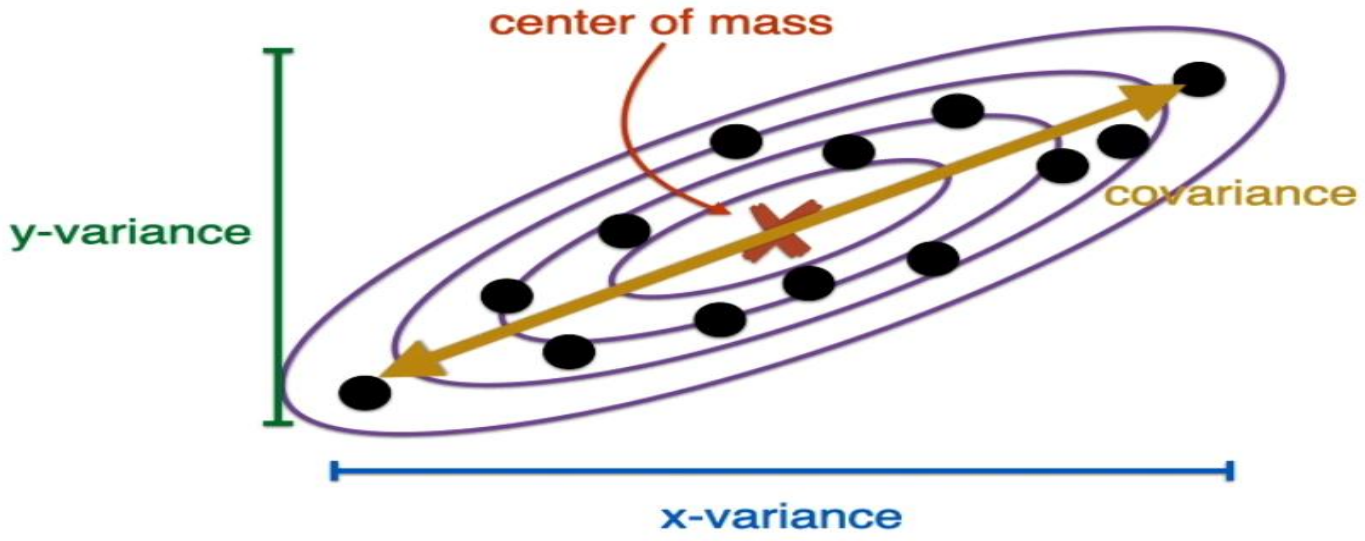
Distributions conditionnelles

Exemple: La distribution conditionnelle de X sachant que $Y=y_2$ ([6-10[)

X	n_{ij}	$f_{i/2}$
1	8	0.53
2	4	0.27
3	3	0.2
∑	15	1

Corrélation basée sur la covariance de données

- La forme du jeu de données (dataset) est déterminée par **la covariance**.
- L'élongation des points (individus) dans la direction diagonale.



$$\Sigma = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(x,y) & Var(y) \end{pmatrix}$$

Covariance de données

Considérons deux variables **numériques aléatoires** X, Y et un échantillon de n individus $\{(x_1, y_1),\ldots,(x_n, y_n)\}$.

$$E(X) = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{et} \quad E(Y) = \bar{Y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

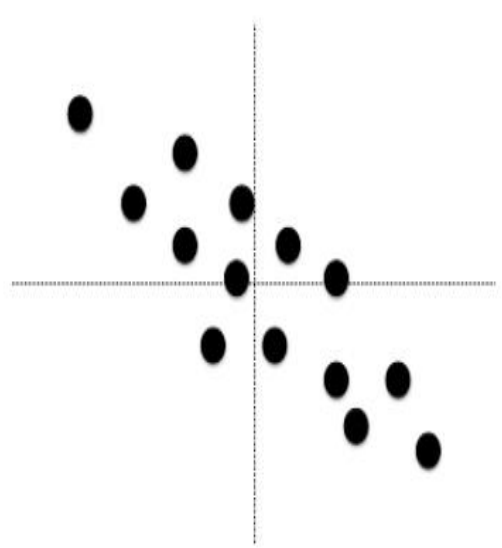
$$Cov(X,Y) = E\big((X - \bar{X}) * (Y - \bar{Y})\big) = \frac{\sum_{i=1}^{n}(x_i - \bar{X}) * (y_i - \bar{Y})}{n}$$

$$Cov(X,Y) = E(X.Y) - \bar{X}\bar{Y}$$

Quantifier la direction de la relation entre les variables X et Y.

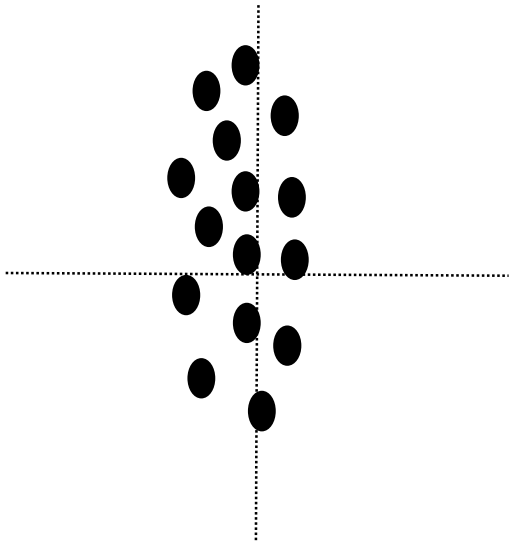
Si X > $\bar{X}$ et Y > $\bar{Y}$ $\Rightarrow Cov(X,Y)$ est positive.
Si X > $\bar{X}$ et Y < $\bar{Y}$ $\Rightarrow Cov(X,Y)$ est négative.
Si X et Y sont indépendants → $Cov(X,Y) = 0$ (ou presque null).
Variance est un cas spéciale de la covariance (i.e $cov(X,X) = \text{var (X)} = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n}$).

Covariance de données



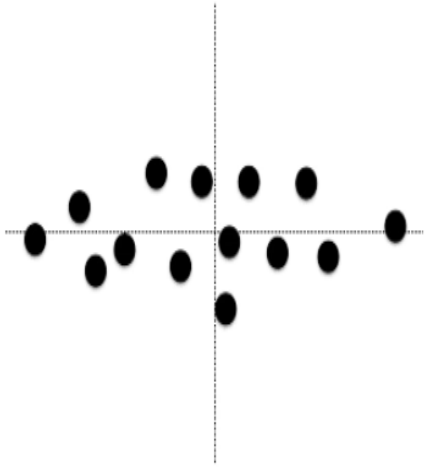
Covariance négative

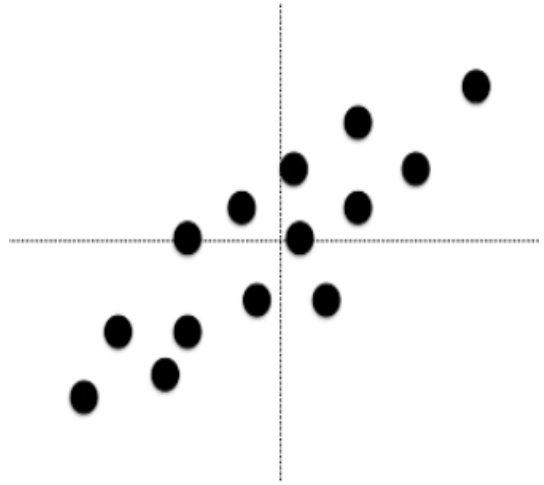
$Cov(X,Y) < 0$



Covariance nulle (ou presque nulle)

$Cov(X,Y) = 0$





Covariance positive

$Cov(X,Y) > 0$

Corrélation basée sur le coefficient de Pearson

Coefficient de corrélation linéaire de Pearson ($r_{X,Y}$): quantifie **l'intensité de la relation** entre deux variables **quantitatives** (numériques: discrètes).

- Considérons deux variables numériques X, Y et un échantillon de n individus $\{(x_1, y_1),\ldots,(x_n, y_n)\}$.

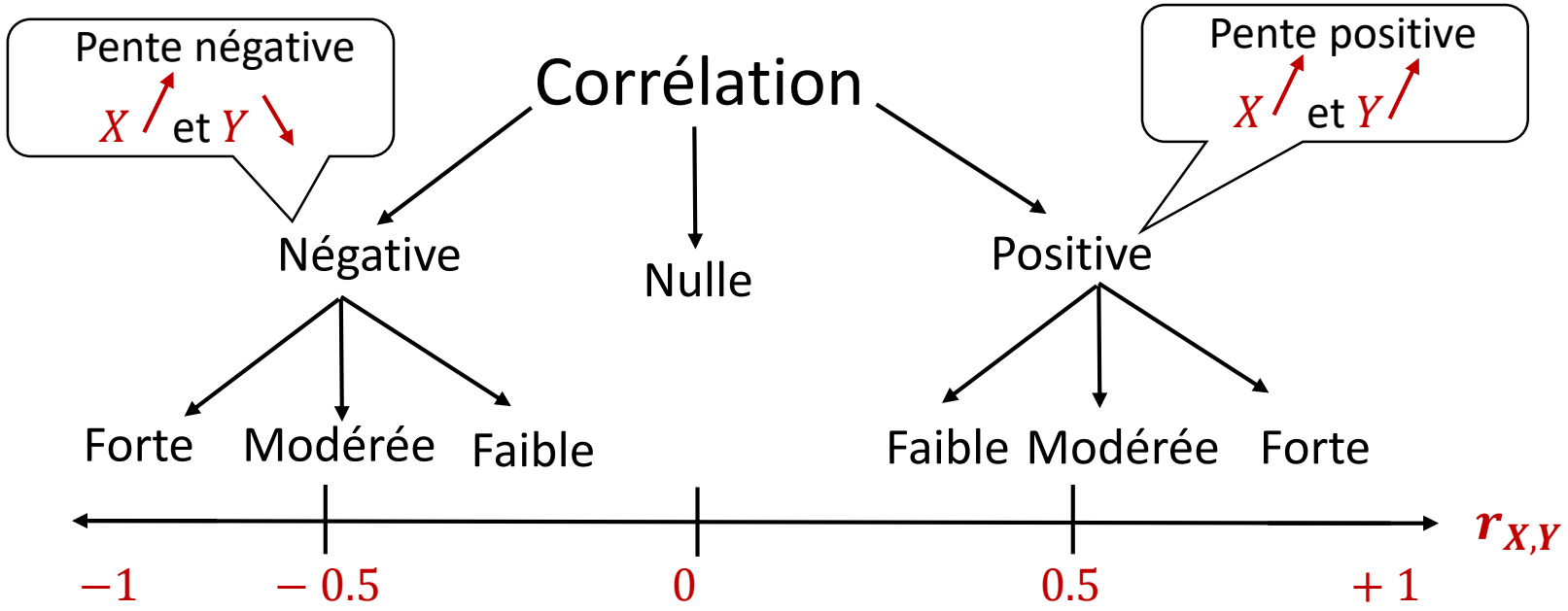
$$r_{X,Y} = \frac{\boldsymbol{Cov(X,Y)}}{\boldsymbol{\sigma_X \sigma_Y}} \qquad \sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2} \qquad \sigma_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2}$$

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\boldsymbol{-1 \leq r_{X,Y} \leq 1}$$

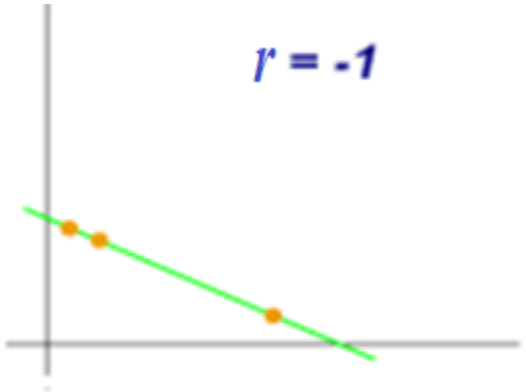
Coefficient de corrélation linéaire

- Si $r_{X,Y} \gg 0$ (valeur large proche de 1) $\Rightarrow$ Corrélation linéaire forte positive (variables redondantes) entre X et Y
- Si $r_{X,Y} = 0\ (ou\ proche\ de\ 0)$ $\Rightarrow$ Les variables X et Y ne sont pas corrélés linéairement (variables indépendantes)
- Si $r_{X,Y} < 0\ \ (proche\ de\ -1)$ $\Rightarrow$ Corrélation linéaire forte négative (variables redondantes) entre les variables X et Y.

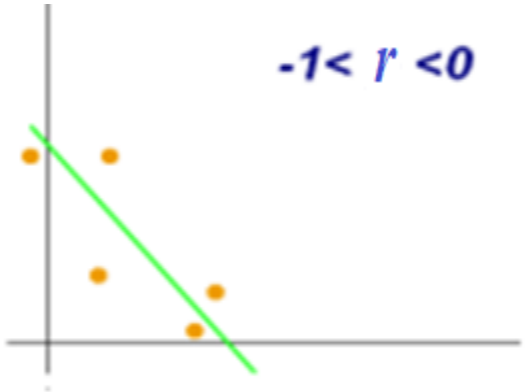


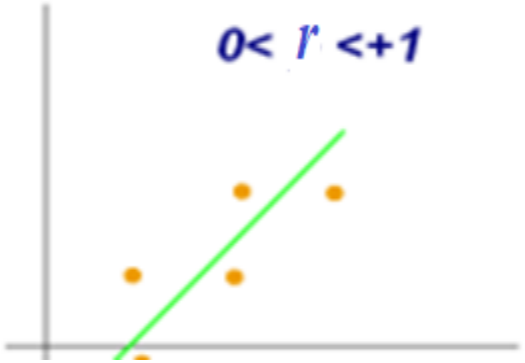
Coefficient de corrélation linéaire

Corrélation linéaire forte négative

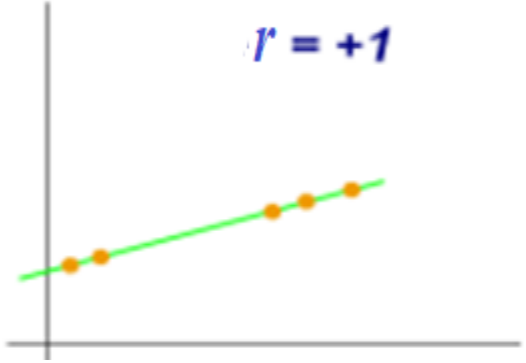


Corrélation linéaire modérée négative

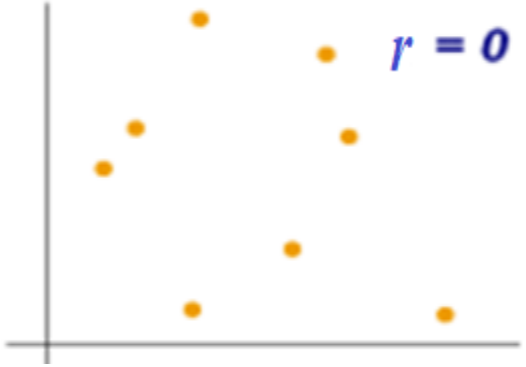




Corrélation linéaire modérée positive



Corrélation linéaire forte positive



Absence de corrélation

Corrélation basée sur le test de carré de Pearson

Test de carré de Pearson

✓ Test statistique où la statistique de test (variable aléatoire) suit une loi du x^2 sous l'hypothèse nulle.

✓ Il permet d 'évaluer :

(1) La qualité de l'ajustement ,

(2) L' homogénéité et

(3) L' indépendance des variables (les individus de deux variables, exprimées dans un tableau de contingence)

→ Accepter ou rejet d'une hypothèse

Corrélation basée sur le test de carré de Pearson

Test d'indépendance de carré de Pearson

- Supposant deux variables **qualitatives aléatoires** X ={x_i,…,x_c} et Y ={y_j,…,y_r} et un échantillon de n individus.
- Vérifier l'indépendance de ces deux variables ? (hypothèse à tester)

→ Créer la table de contingence avec les distributions jointes (X_i, Y_j)

$$x^2 = \sum_{i=1}^{c}\sum_{j=1}^{r} \frac{(\ o_{ij} - e_{ij})^2}{e_{ij}} \qquad e_{ij} = \frac{fréquence_{marginale}(X = x_i) * fréquence_{marginale}(\ Y = y_j)}{n}$$

o_{ij}: valeur Obervée (empirique) de *la distribution* jointe (X_i, Y_j)

e_{ij} : valeur Espérée (théorique) de *la distribution jointe*(X_i, Y_j)

$fréquence_{marginale}(X = x_i)$: nombre des individus ayant la valeur x_i

$fréquence_{marginale}(Y = y_j)$: nombre des individus ayant la valeur y_i

Corrélation basée sur le test de carré de Pearson

Test d'indépendance de carré de Pearson

Principe:

- Définir **les hypothèses nulle et alternative.**
- Sélectionner un **niveau de confiance** souhaité (niveau de signification , valeur p ou niveau alpha correspondant) pour le résultat du test.
- Calculer la statistique du test du chi carré x^2 .
- Déterminer les **degrés de liberté** (rows-1)*(cols-1) , de cette statistique.
- Comparer avec **la valeur critique** de la distribution du chi carré avec les degrés de liberté et le niveau de confiance sélectionné.
- Accepter ou rejeter l'hypothèse nulle.

Corrélation basée sur le test de carré de Pearson

Exemple :

Table des valeurs observées:

Diplôme

Etat civil

	Collège	Lycée	Bac	Master	Doctorat	Total
Non marié	18	36	21	9	6	90
Marié	12	36	45	36	21	150
Divorcé	6	9	9	3	3	30
Veuf	3	9	9	6	3	30
Total	39	90	84	54	33	300

Objectif: vérifier la relation de corrélation entre les deux variables aléatoires « diplôme » et « état civil ».

Corrélation basée sur le test de carré de Pearson

Test d'indépendance de carré de Pearson

Hypothèse nulle: Pas de relation entre les variable diplôme et état civil.

Hypothèse alternative: relation significative entre les variable diplôme et état civil.

niveau de signification=0.05

Table des valeurs espérées:

	Collège	Lycée	Bac	Master	Doctorat
Non marié	$\frac{90*39}{300} = 11.7$	$\frac{90*90}{300} = 27$	25.2	16.2	9.9
Marié	19.5	45	42	27	16.5
Divorcé	3.9	9	8.4	5.4	3.3
Veuf	3.9	9	8.4	5.4	3.3

Corrélation basée sur le test de carré de Pearson

Calcul x^2

Valeurs observées (o)	Valeurs espérées (e)	$(o_{ij} - e_{ij})$	$(o_{ij} - e_{ij})^2$	$\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
18	11.7	6.3	39.69	3.39
36	27	9	81	3
….	…..	……	…..	…
3	3.3	-0.3	0.09	0.02

$$x^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 3.39 + 3 + \cdots + 0.02 = 23.57$$

Degré de liberté = (rows-1)*(cols-1) =(4-1)*(5-1)=12

Corrélation basée sur le test de carré de Pearson

Test d'indépendance de carré de Pearson

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14

https://www.youtube.com/watch?v=f53nXHoMXx4&list=RDCMUCFUTIAx5h703U3gEcM1157Q&index=1

Niveau de signification=0.05

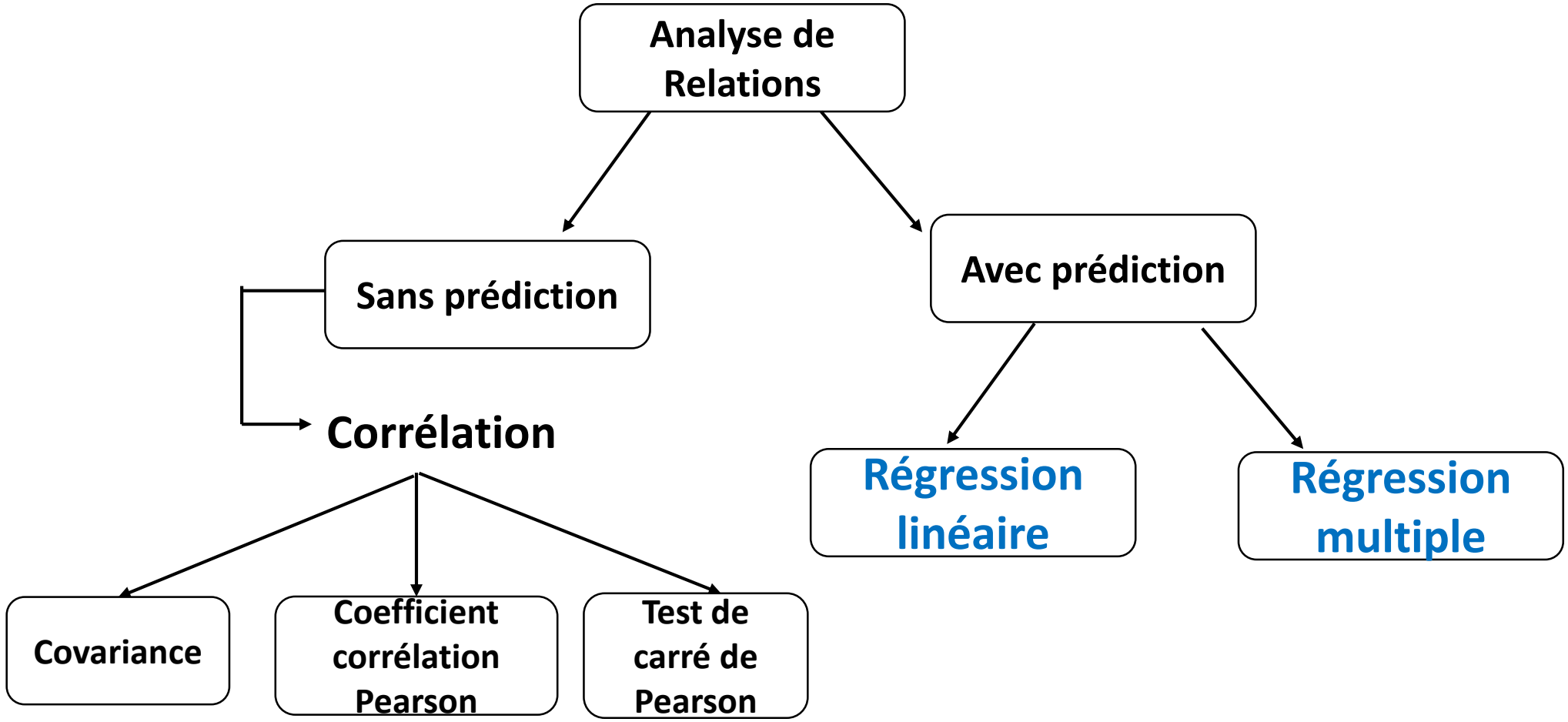
$$x^2_{Critique} = 21.03$$

x^2 =23.57

$$x^2 > x^2_{Critique}$$

→ Rejeter l'hypothèse nulle et
→ Accepter l'hypothèse alternative

Introduction aux statistiques bi-variées



Conclusion

- Etude statistique d'une série bi-variée.
- Analyse des relations entre deux variables aléatoires :

▪ Corrélation

1) Variables quantitatives:

- Covariance (**quantifier la direction de la relation**)
- Coefficient de corrélation linéaire (**quantifier l'intensité de la relation**)

2) Variables qualitatives:

- Test de carré de Pearson (**tester** si la liaison est **statistiquement significative).**

▪ Régression