Plan

I. Corrélation et régression linéaire

II. Coefficient de corrélation

III. Régression linéaire simple

Annexes

I.1. Nature des variables

Le terme de corrélation est utilisé dans le langage courant pour désigner la liaison (relation / association) entre 2 variables quelconques.

En statistique, le terme de corrélation est réservé pour désigner la liaison entre 2 variables QUANTITATIVES (le plus souvent continues).

Corrélation / régression : liaison entre 2 variables quantitatives

I.2. Corrélation versus régression

Corrélation :

Liaison entre 2 variables quantitatives X et Y

Rôle symétrique (on peut permuter X et Y)

Rôle asymétrique

Régression :

Liaison entre 2 variables quantitatives X et Y

Rôle asymétrique uniquement :

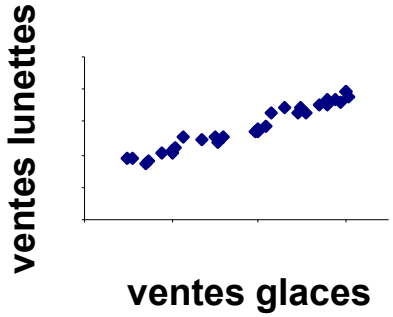
- **X = variable explicative / Y = variable expliquée**
- **X = variable indépendante / Y = variable dépendante**

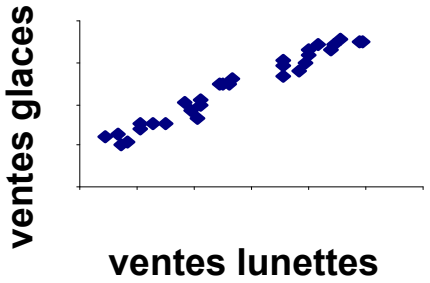
(on ne peut pas permuter X et Y)

I.2. Corrélation versus régression

1. Exemple : corrélation (positive)

- X = ventes de paires de lunettes de soleil en été
- Y = ventes de crèmes glacées en été
- Il existe une liaison entre X et Y :
  - Quand X augmente, Y augmente (météo estivale)
  - Quand X diminue, Y diminue (météo pluvieuse)
- La liaison est symétrique :
  - X est liée à Y, et Y est liée à X
  - mais X ne dépend pas de Y et Y ne dépend pas de X
  - on peut permuter X et Y en abscisses et en ordonnées
- Y ne peut pas être prédite par X

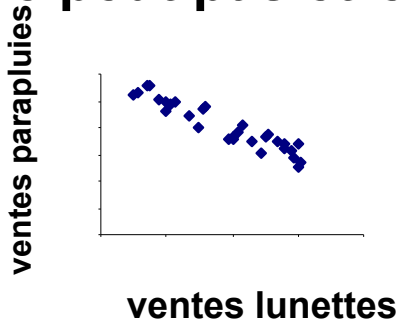


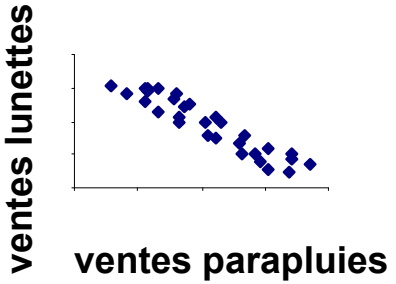


I.2. Corrélation versus régression

2. Exemple : corrélation (négative)

- X = ventes de paires de lunettes de soleil en été
- Y = ventes de parapluies en été
- Il existe une liaison entre X et Y :
  - Quand X augmente, Y diminue (météo estivale)
  - Quand X diminue, Y augmente (météo pluvieuse)
- La liaison est symétrique :
  - X est liée à Y, et Y est liée à X
  - mais X ne dépend pas de Y et Y ne dépend pas de X
  - on peut permuter X et Y en abscisses et en ordonnées
- Y ne peut pas être prédite par X

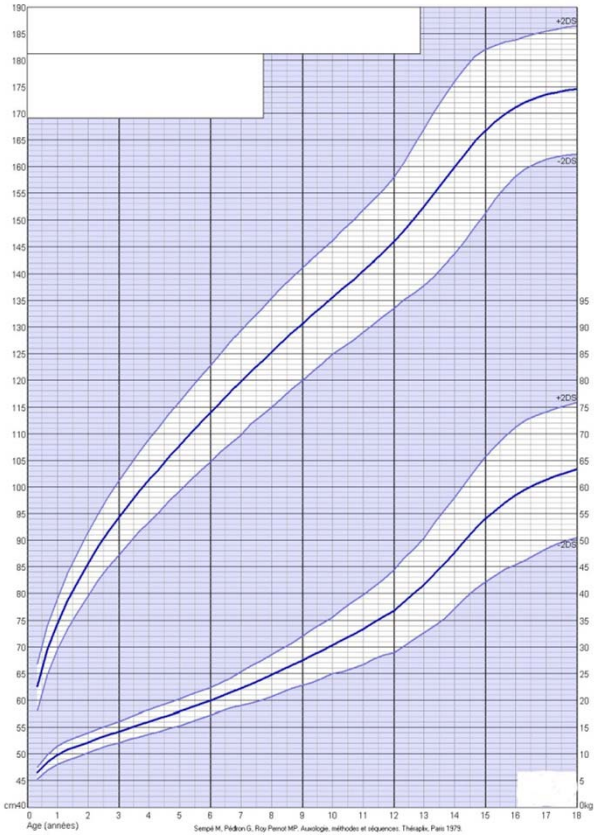




I.2. Corrélation versus régression

3. Exemple : régression

- X = âge (de 0 à 15 ans)
- Y = taille (cm)
- Il existe une liaison entre X et Y :
  - Quand l'âge augmente, la taille augmente
  - Quand l'âge diminue, la taille diminue
- La liaison est asymétrique :
  - la taille dépend de l'âge mais l'âge ne dépend pas de la taille
  - on ne peut pas permuter X et Y en abscisses et en ordonnées
- On peut prédire la taille par l'âge à l'aide d'une équation de droite ou de courbe de régression (cf carnet de santé)



I.2. Corrélation versus régression

	Corrélation	Régression
Variables	X = quantitative Y = quantitative	X = quantitative Y = quantitative
Symétrie de la liaison	Oui / Non Y liée à X X liée à Y	Non Y dépend de X -
Exemples	Y = conso. cannabis X = température moyenne annuelle	Y= taille X = âge
Prédiction	Non	Oui (équation)

I.3. Conditions d'application de la corrélation et de la régression linéaire simple

Indépendance des observations

Liaison linéaire entre X et Y

Distribution conditionnelle normale et de variance constante

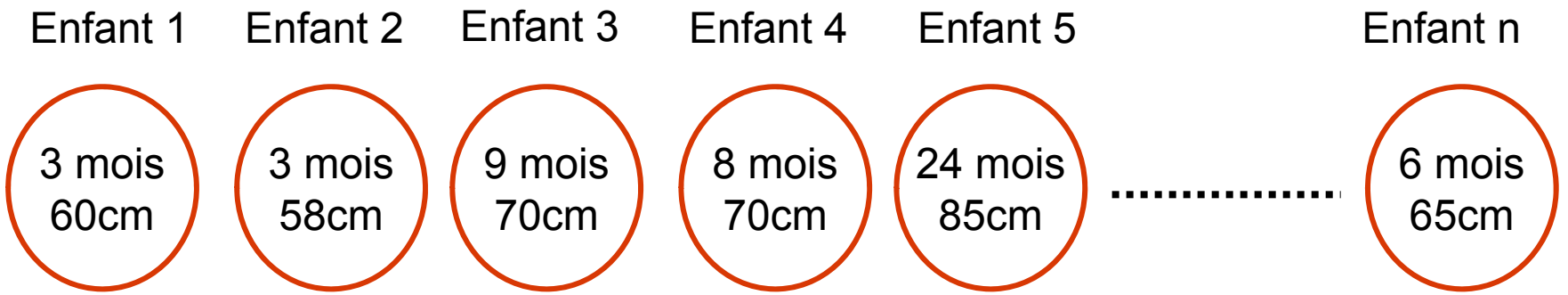
I.3. Conditions d'application de la corrélation et de la régression linéaire simple

1. Indépendance des observations

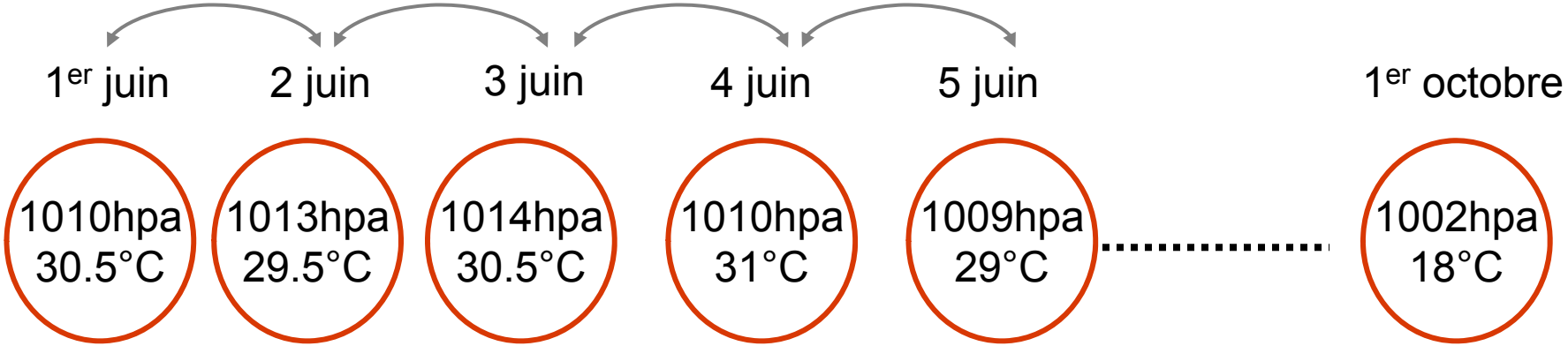
Ne pas confondre :

- **Indépendance des observations (condition d'application du test statistique)**
- **Indépendance des variables (hypothèse à tester)**

Observations indépendantes (et variables corrélées)



Observations corrélées (et variables corrélées)



I.3. Conditions d'application de la corrélation et de la régression linéaire simple

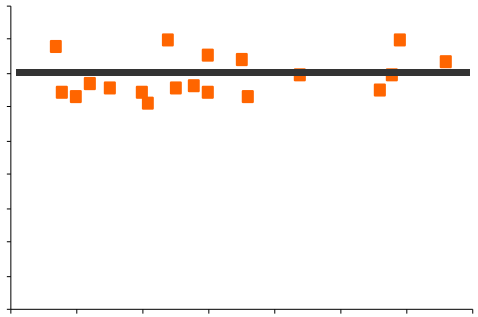
2. Liaison linéaire entre X et Y

Avant d'appliquer le test du coefficient de corrélation ou d'estimer la droite de régression, il faut vérifier - empiriquement (graphiquement) - que la liaison entre les 2 variables est de nature linéaire.

A défaut, l'interprétation du test du coefficient de corrélation ou du test de la pente de la droite de régression peut être erronée.

Coefficient de corrélation nul
Pente de la droite de régression nulle

Cas 1



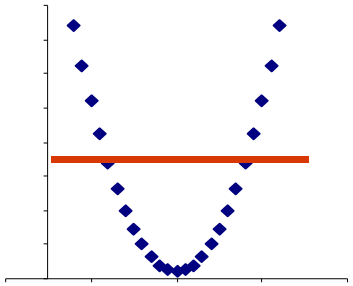
La nature de la liaison est linéaire (le nuage de points est résumé au mieux par une droite horizontale d'équation y = a)

La condition d'application est vérifiée

Il est possible d'utiliser le coefficient de corrélation et la régression linéaire simple pour quantifier la liaison entre les 2 variables (conclusion : X et Y sont indépendants [Y constant quelle que soit la valeur de X])

Coefficient de corrélation nul
Pente de la droite de régression nulle

Cas 2



Il existe une liaison entre X et Y mais cette liaison n'est pas linéaire : Y varie avec les valeurs de X.

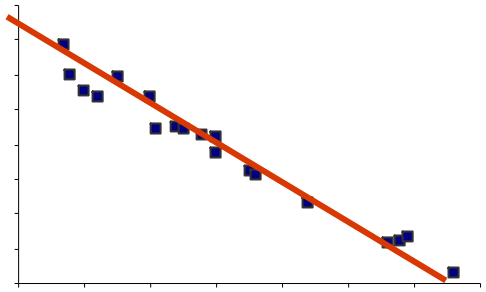
Le nuage de points n'est pas résumé au mieux par une droite mais plutôt par une fonction quadratique.

La condition d'application n'est pas vérifiée

→ Il ne faut pas utiliser le coefficient de corrélation ni la régression linéaire simple pour quantifier la liaison entre les 2 variables

Coefficient de corrélation non nul
Pente de la droite de régression non nulle

Cas 3



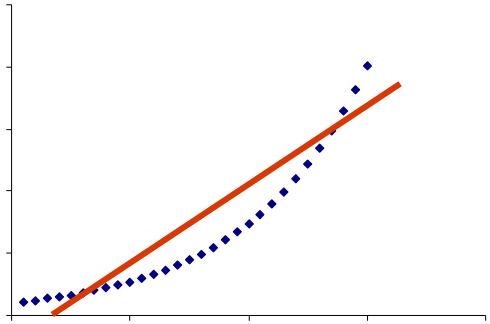
La nature de la liaison est linéaire (le nuage de points est résumé au mieux par une droite d'équation y = a+bx)

La condition d'application est vérifiée

Il est possible d'utiliser le coefficient de corrélation et la régression linéaire simple pour quantifier la liaison entre les 2 variables (conclusion : il existe une liaison linéaire entre X et Y)

Coefficient de corrélation non nul
Pente de la droite de régression non nulle

Cas 4



La nature de la liaison n'est pas linéaire (le nuage de points n'est pas résumé au mieux par une droite mais plutôt par une fonction exponentielle)

La condition d'application n'est pas vérifiée

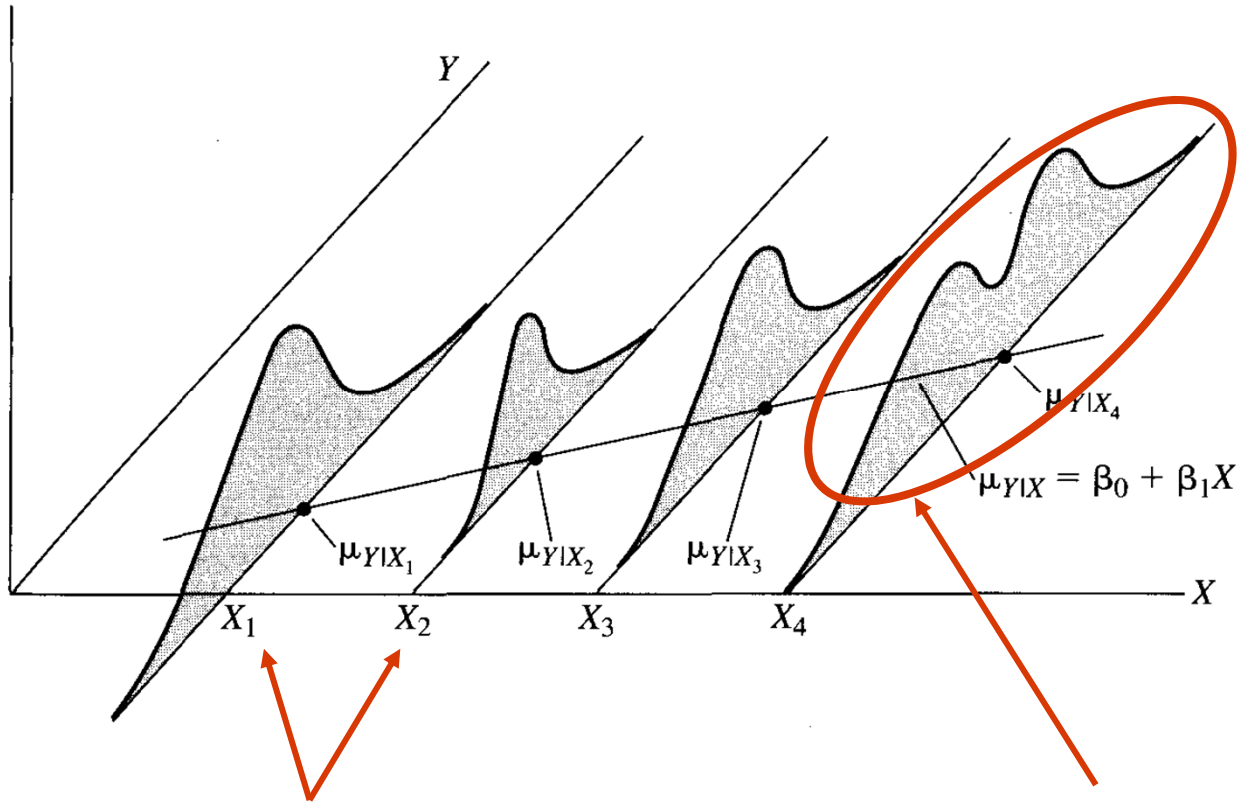
→ Il ne faut pas utiliser le coefficient de corrélation ni la régression linéaire simple pour quantifier la liaison entre les 2 variables

I.3. Conditions d'application de la corrélation et de la régression linéaire simple

3. Distribution conditionnelle normale et de variance constante

Distribution de Y normale et de variance constante pour chaque valeur de X

(difficilement vérifiable en pratique)



La variance de Y n'est pas constante pour les différentes valeurs de X

La distribution de Y n'est pas normale pour X = x4

La condition d'application n'est pas vérifiée

II.1. Covariance

- **Variance conjointe de 2 variables X et Y**

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

- **Cas particulier : X = Y → cov(X,Y) = cov(X,X) = var(X)**

$$\mathrm{cov}(X,X) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(X_i - \mu_X)}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N} = \mathrm{var}(X)$$

II.1. Covariance

- **X et Y indépendantes**

cas particulier Y constant quelle que soit la valeur de X

$$\mathrm{cov}(X, Y) = \frac{\sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)}{N} = 0$$

0 car Y_i = constante = μ_Y

II.1. Covariance

- **Equivalent de la formule de Huyghens pour la covariance**

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{n}$$

Rappel : $$\mathrm{var}(X)=\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}$$

II.2. Coefficient de corrélation

Le coefficient de corrélation entre 2 variables quantitatives X et Y est égal au rapport de la covariance de X et Y divisé par le produit des écart-types de X et Y.

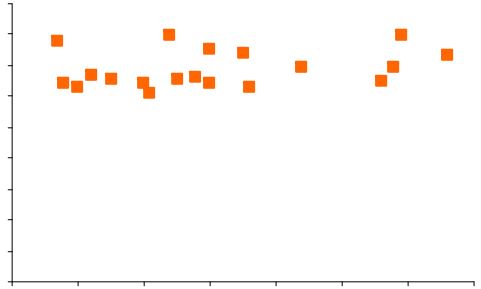
Le coefficient de corrélation est noté ρ dans la population.

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$$

$$-1 \leq \rho \leq +1$$

II.2. Interprétation du coefficient de corrélation

1. X et Y indépendantes : ρ = 0



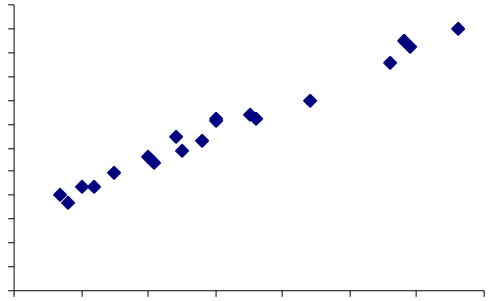
ρ = 0

- Y = fluctue autour d'une constante quelle que soit la valeur de X
- Nuage de points horizontal
- cov(X, Y) = 0

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} = 0$$

II.2. Interprétation du coefficient de corrélation

2. X et Y corrélées : ρ > 0



ρ > 0

- **Liaison linéaire croissante entre X et Y**

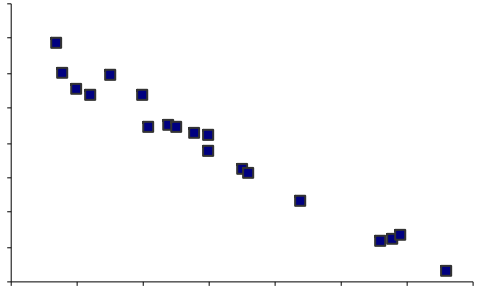
- **cov(X, Y) > 0**

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} > 0$$

NB : si Y = X → cov(X,Y) = var(X) et var(Y) = var(X) → ρ =1

II.2. Interprétation du coefficient de corrélation

2. X et Y corrélées : ρ < 0



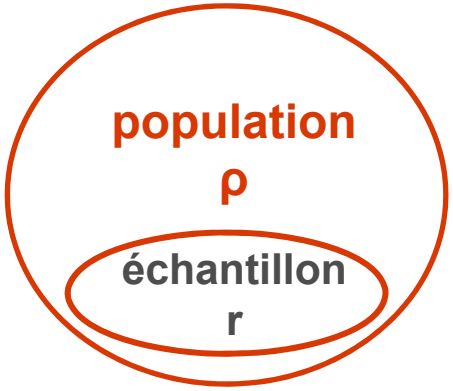
ρ < 0

- **Liaison linéaire décroissante entre X et Y**
- **cov(X, Y) <0**

$$\rho = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} < 0$$

NB : si Y = - X → cov(X,Y) = - var(X) et var(Y) = var(X) → ρ =-1

II.3. Estimation du coefficient de corrélation



Le coefficient de corrélation estimé sur un échantillon issu d'une population est noté r.

Il s'interprète comme le coefficient de corrélation ρ mesuré sur la population.

Il est calculé à partir des estimations de la covariance et des variances de X et de Y sur l'échantillon.

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{(n-1)} \qquad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - m_x)^2}{(n-1)} \qquad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - m_y)^2}{(n-1)}$$

II.3. Estimation du coefficient de corrélation

Par simplification des (n-1) au dénominateur de la covariance et de la variance de X et de la variance de Y, on obtient l'expression de l'estimateur du coefficient de corrélation r à partir d'un échantillon.

$$r = \frac{\sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n} (x_i - m_x)^2 \sum_{i=1}^{n} (y_i - m_y)^2}}$$

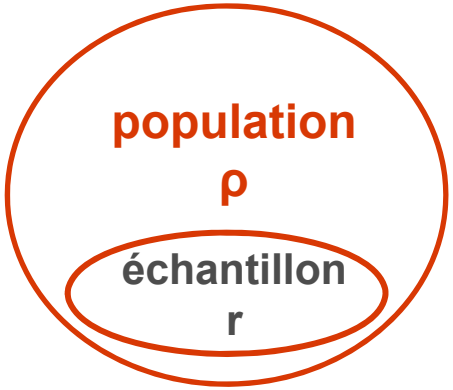
II.3. Estimation du coefficient de corrélation

Par simplification des (n-1) au dénominateur de la formule de Huyghens de la covariance et de la variance de X et de Y, on obtient une autre expression de l'estimateur du coefficient de corrélation r à partir d'un échantillon.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

II.4. Test du coefficient de corrélation

Après le calcul du coefficient de corrélation r estimé sur un échantillon, il faut déterminer si le coefficient de corrélation ρ est significativement différent de 0.



$r \approx \rho$

H0 : $\rho = 0$ (absence de liaison [linéaire] entre X et Y)

H1 bilatérale : $\rho \neq 0$ (existence d'une liaison entre X et Y)

II.4. Test du coefficient de corrélation

Sous l'hypothèse nulle (H0) :

Le rapport de l'estimateur du coefficient de corrélation r sur son écart-type suit une loi de Student à (n-2) degrés de liberté.

n est l'effectif de l'échantillon.

$$\frac{r}{s_r} \rightarrow t_{(n-2)ddl}$$

L'estimateur de l'écart-type du coefficient de corrélation est égal à :

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

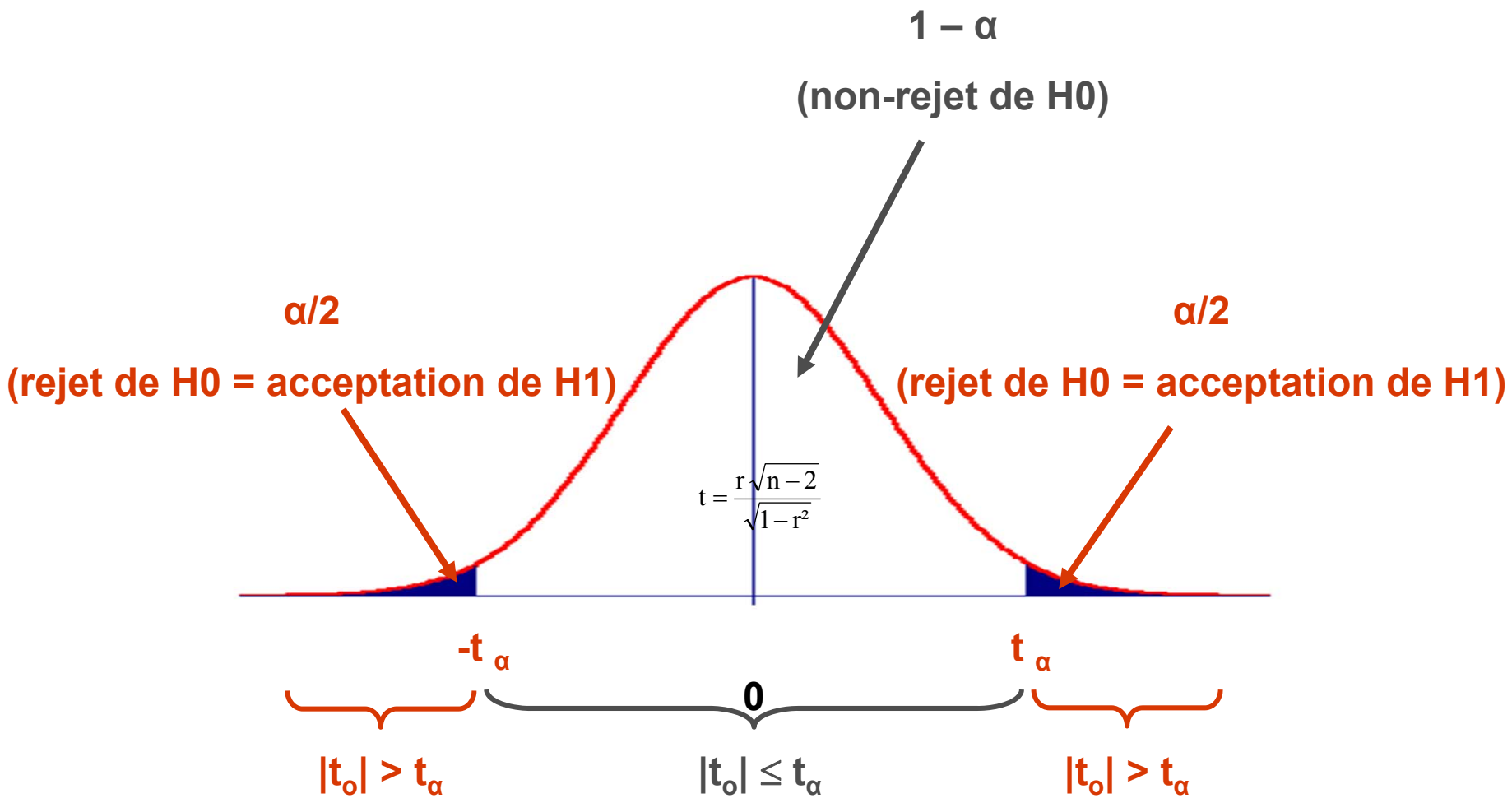
II.4. Test du coefficient de corrélation

Le test du coefficient de corrélation consiste à calculer la grandeur to et à la comparer à la valeur seuil tα sur la table de la loi de Student à (n-2) degrés de libertés.

$$t_o = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Conditions d'application

- **indépendance des observations**
- **liaison linéaire entre X et Y**
- **distribution conditionnelle normale et de variance constante**



Abscisses : valeurs possibles de t sous H0 (ρ = 0)

t_o : valeur observée/calculée de t sur l'échantillon

Détermination du degré de signification associé à t_o (*P*-value)

Exemple :

- **to = 2.12**
- **n = 20**

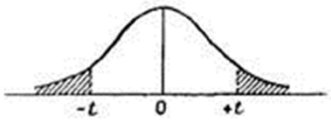
0.02 < *P* <0.05

P < α → rejet de H0

(n-2) = 18 ddl

Table de *t* (*).

La table donne la probabilité α pour que *t* égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

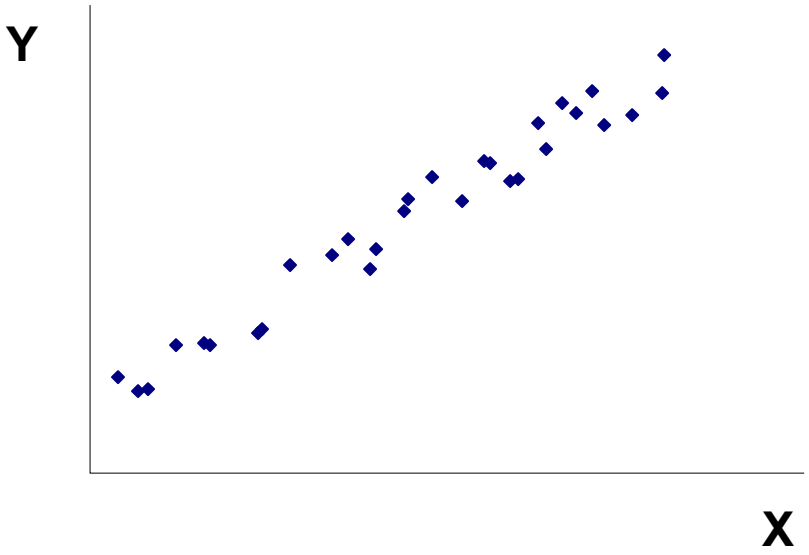
Rappel : P-value = probabilité d'observer une valeur plus grande que t_o sous l'hypothèse nulle H0

Plan

III.1. Régression linéaire simple

La régression s'adresse à un type de problème où les 2 variables quantitatives continues X et Y ont un rôle asymétrique : la variable Y dépend de la variable X.

La liaison entre la variable Y dépendante et la variable X indépendante peut être modélisée par une fonction de type $Y = \alpha + \beta X$, représentée graphiquement par une droite.



$Y = \alpha + \beta X$

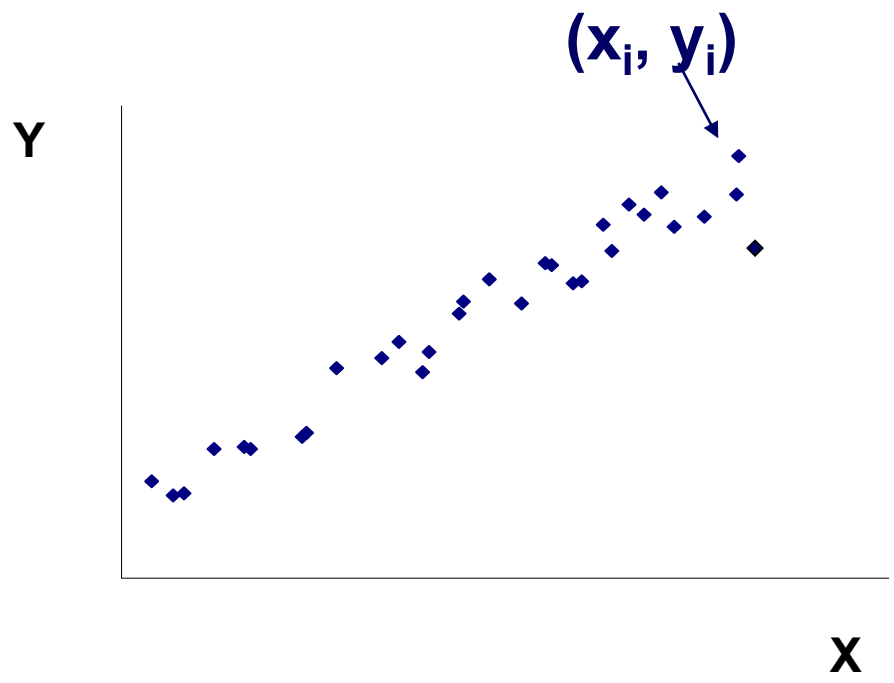
Y : variable dépendante (expliquée)

X : variable indépendante (explicative)

α : ordonnée à l'origine (valeur de Y pour x = 0)

β : pente (variation moyenne de la valeur de Y pour une augmentation d'une unité de X)

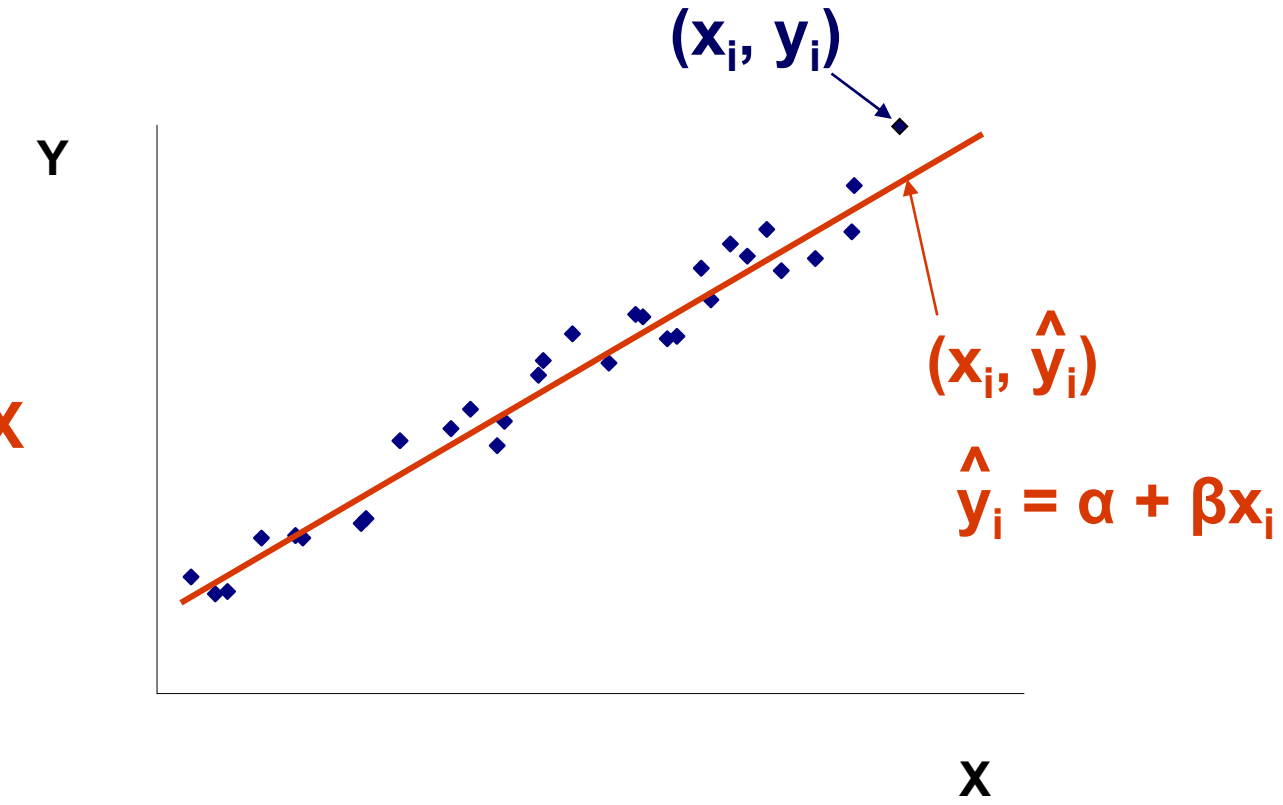
III.2. Estimation par la méthode des moindres carrés



Chaque individu i est caractérisé par un couple de coordonnées (xi, yi) et est représenté par un point sur le graphique.

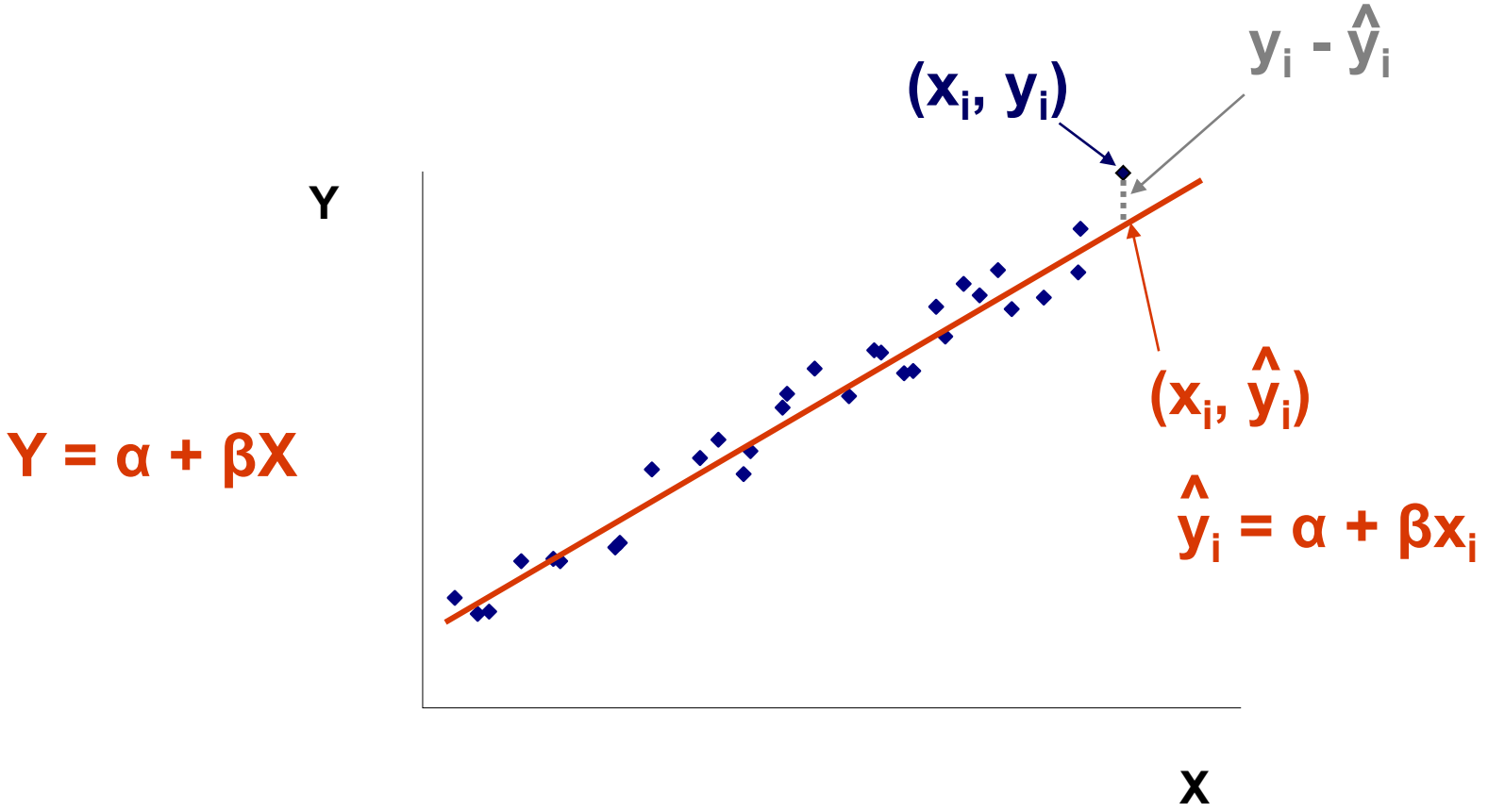
L'ensemble des individus forme un nuage de points.

III.2. Estimation par la méthode des moindres carrés



La droite de régression Y = α + βX est la droite qui résume le mieux le nuage de points. Intuitivement, il s'agit de la droite dont les points du nuage sont en moyenne les plus proches (c'est-à-dire la droite qui passe à la plus faible distance de chaque point du nuage, en moyenne).

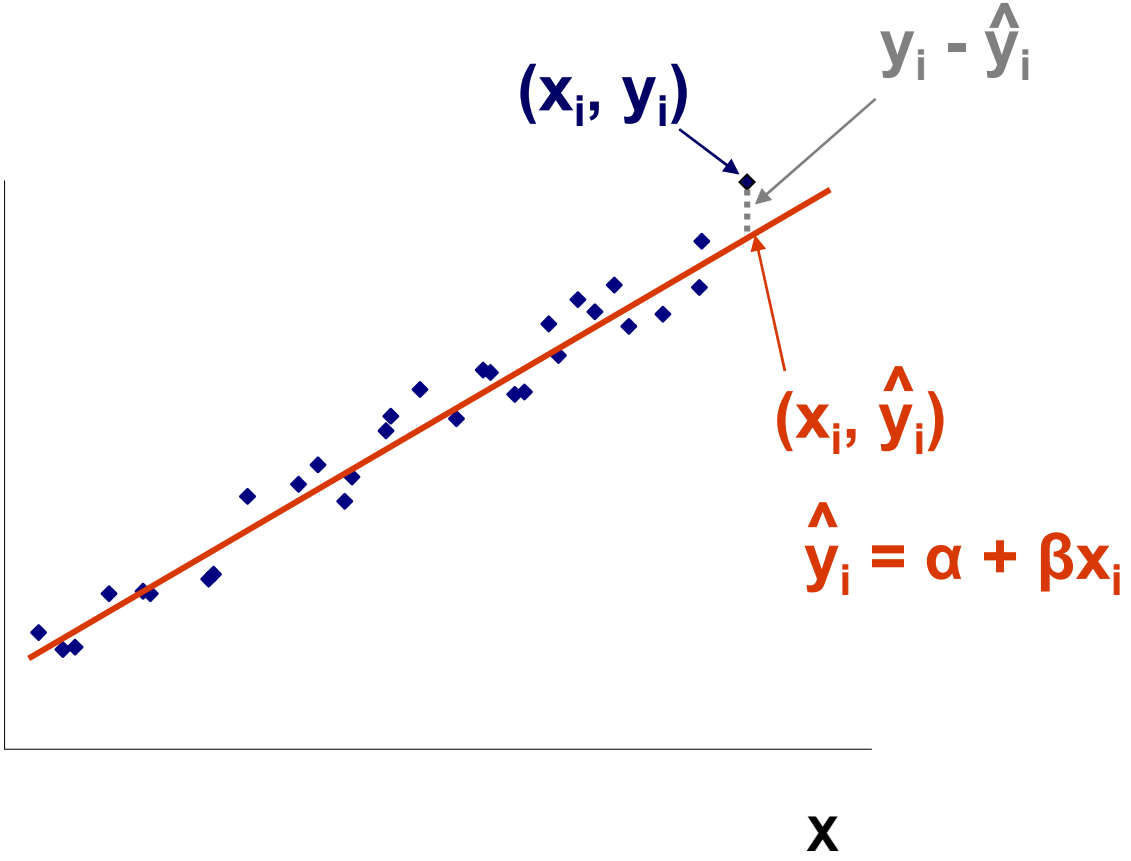
III.2. Estimation par la méthode des moindres carrés



La distance d'un point à la droite est la distance verticale entre l'ordonnée du point observé (xi, yi) et l'ordonnée du point correspondant sur la droite (xi, ^yi) .

Cette distance d'un point à la droite (yi - ^yi) peut être positive ou négative et la somme des distances à la droite s'annule.

III.2. Estimation par la méthode des moindres carrés



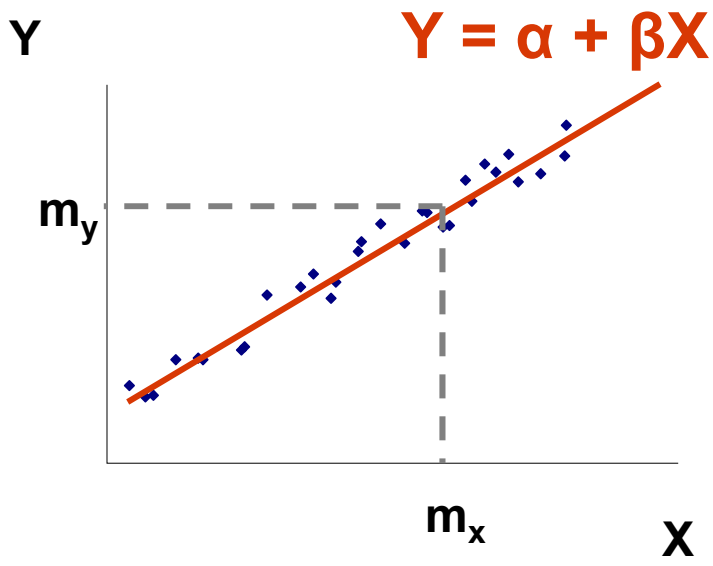
$Y = \alpha + \beta X$

$\hat{y}_i = \alpha + \beta x_i$

$$SCE = \Sigma_i (y_i - \hat{y}_i)^2$$

Pour s'affranchir du signe, on calcule la somme des carrés des distances de chaque point à la droite. La droite de régression est la droite qui minimise la somme des carrés des écarts. Elle est aussi appelée droite des moindres carrés.

III.2. Estimation par la méthode des moindres carrés

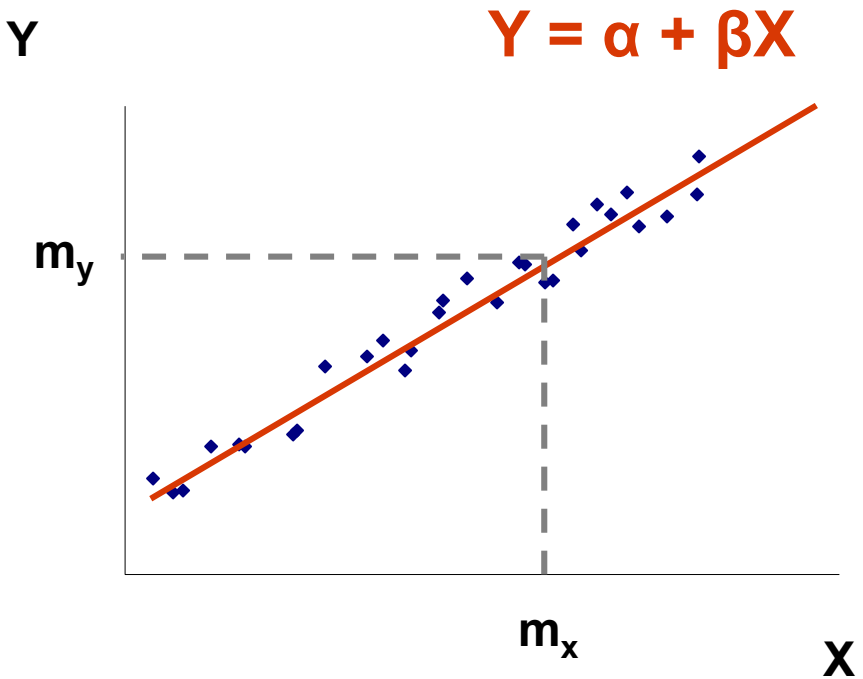


Une particularité de la droite de régression est de passer par le point moyen théorique de coordonnée (m_x, m_y).

III.2. Estimation par la méthode des moindres carrés

a et b sont les estimations de l'ordonnée à l'origine α et de la pente β de la droite de régression.

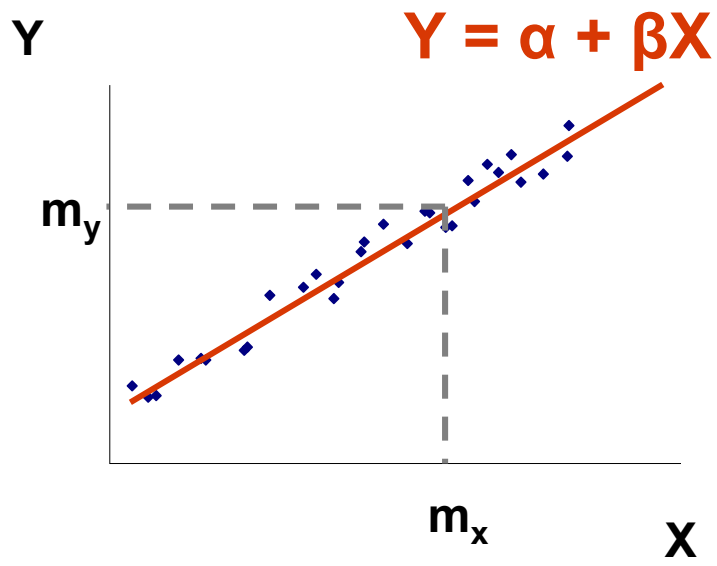
L'estimation de la pente de la droite de régression b est égale au rapport de la covariance de X et Y sur la variance de X.



$$b = \frac{\mathrm{cov}(X, Y)}{\mathrm{var}(X)}$$

$$b = \frac{\sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)}{\sum_{i=1}^{n} (x_i - m_x)^2}$$

III.2. Estimation par la méthode des moindres carrés

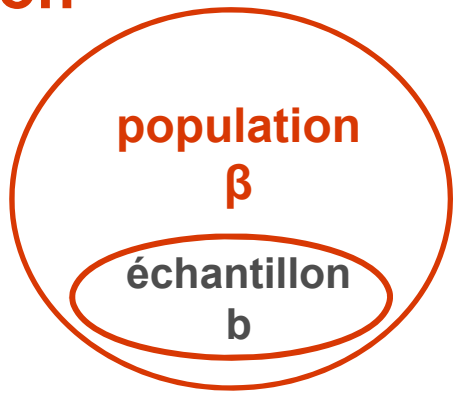


L'estimateur de l'ordonnée à l'origine a est déduit de la pente b et des coordonnées du point moyen (m_x, m_y) :

$$a = m_y - b\, m_x$$

III.3. Test de la pente de la droite de régression

$b \approx \beta$



La droite de régression d'équation $Y = \alpha + \beta X$ comporte 2 paramètres (α et β).

L'hypothèse nulle est que la pente β de la droite de régression de Y en X est égale à 0 (soit Y est égal à α, c'est-à-dire que la droite de régression est horizontale et qu'il n'y a pas de liaison entre X et Y).

H0 : $\beta = 0$ (droite de régression horizontale : $Y = \alpha$)

H1: $\beta \neq 0$

III.3. Test de la pente de la droite de régression

Sous l'hypothèse nulle (H0) :

Le rapport de l'estimateur de la pente b sur son écart-type suit une loi de Student à (n-2) degrés de liberté.

n est l'effectif de l'échantillon.

$$\frac{b}{s_b} \rightarrow t_{(n-2)ddl}$$

L'estimateur de l'écart-type de la pente est égal à : $s_b = \sqrt{\frac{\frac{s_y^2}{s_x^2} - b^2}{n-2}}$

III.3. Test de la pente de la droite de régression

Le test de la pente consiste à calculer la grandeur t_o et à la comparer à la valeur seuil t_α sur la table de la loi de Student à (n-2) degrés de libertés

$$t_o = \frac{b}{\sqrt{\frac{\frac{s_y^2}{s_x^2} - b^2}{n-2}}}$$

Conditions d'application

- **indépendance des observations**
- **liaison linéaire entre X et Y**
- **distribution conditionnelle normale et de variance constante**

Corrélation et régression

	Corrélation	Régression
Variables	**Quantitatives** **symétriques/asymétriques**	**Quantitatives** **asymétriques**
Test	**Coefficient de corrélation** $-1 \leq r \leq 1$	**Pente de la droite de régression**
Prédiction	**non**	**oui**
Conditions	**Indépendance des observations** **Liaison linéaire** **Distribution conditionnelle normale et de variance constante**	

Annexe : variance et covariance

- **Variance**
- $\mathrm{var}(X) = E(X^2) - [E(X)]^2$

$$\mathrm{var}(x) = \left(\frac{1}{n}\sum x^2\right) - \left(\frac{1}{n}\sum x\right)^2$$

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}$$

Annexe : variance et covariance

- **Covariance**
- cov(X,Y) = E(XY) – [E(X) x E(Y)]

$$\mathrm{cov}(x,y)=\left(\frac{1}{n}\sum xy\right)-\left[\left(\frac{1}{n}\sum x\right)\times\left(\frac{1}{n}\sum y\right)\right]$$

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}x_i y_i-\frac{\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{n}}{n}$$