Chapitre 3

Statistiques descriptives bivariées

Statistiques descriptives bivariées

= Décrire la relation entre deux variables dans un échantillon / une population :

- **Le score sur la variable 2 est dépendant du score sur la variable 1**
- **Le score sur la variable 1 permet de prédire le score sur la variable 2**
- **Variables liées versus indépendantes**
- **Différences sur une VD entre deux groupes est aussi une relation entre deux variables**

Statistiques descriptives bivariées

1. **Décrire la relation entre deux variables nominales**

- **Table de contingence**
- **Diagramme en barres superposées**
- **Technique du rapport de chances**
- **Outils statistiques liés à l'épidémiologie**

2. **Décrire la relation entre deux variables ordinales ou métriques**

- **Diagramme de dispersion**
- **Covariance et corrélation**
- **Droite de régression**
- **Coefficient de détermination**

Table de contingence

Représente la distribution des fréquences (ou %) d'une variable nominale à chaque modalité d'une autre variable nominale

Si les pourcentages sont différents, cela indique une relation entre les deux variables

Exemple : Prévalence des troubles anxieux selon le genre

Les tables de contingence

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	1881	6669	8550
Femmes	3817	7646	11463
Total	5698	14315	20013

Table de contingence 2 X 2

Les tables de contingence

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	9,4%	33,3%	42,7%
Femmes	19,1%	38,2%	57,3%
Total	28,5%	71,5%	100%

Les tables de contingence

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	22%	78%	100%
Femmes	33,3%	66,7%	100%

Les tables de contingence

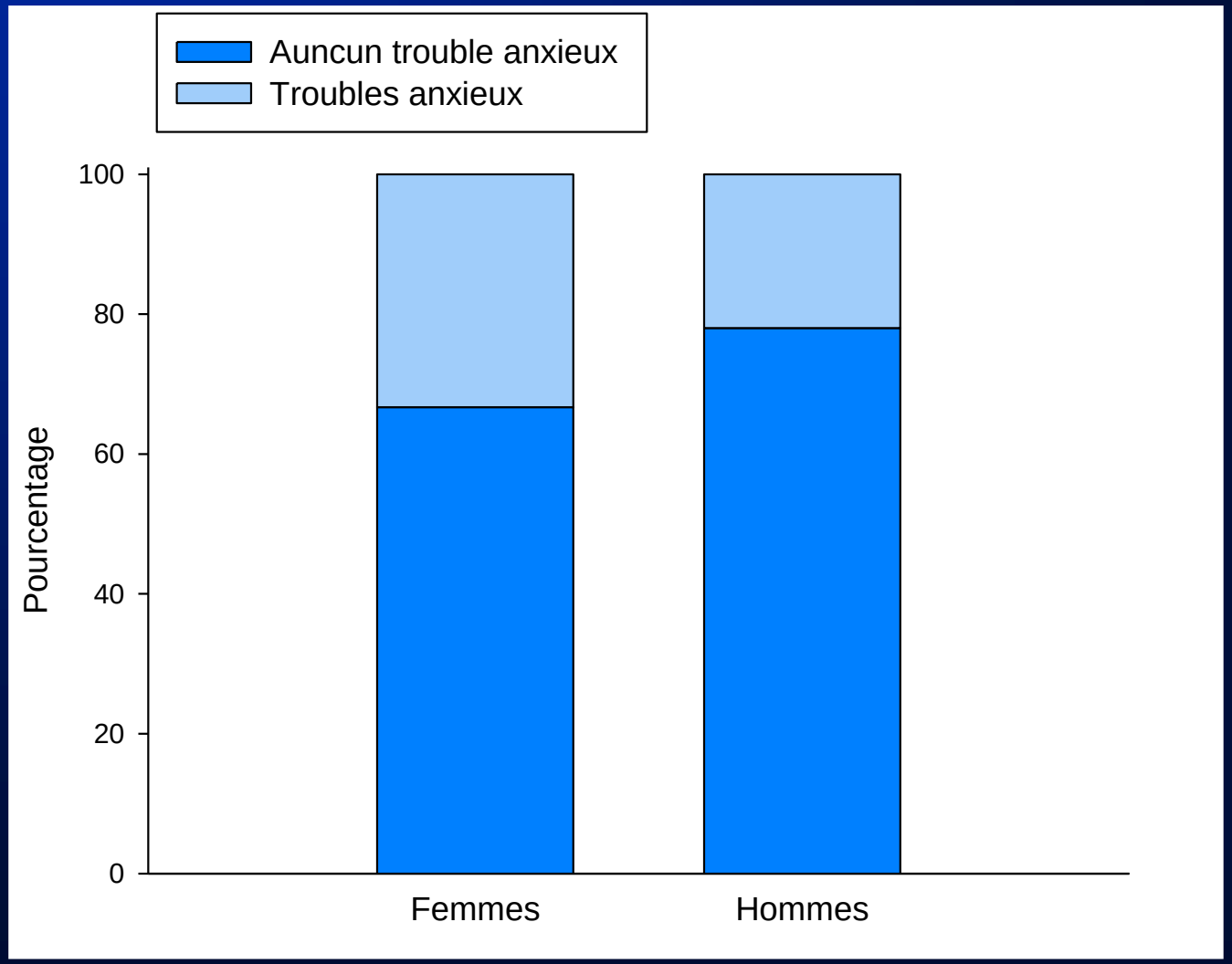
	Hommes	Femmes	Total
Trouble anxieux	33%	67%	100%
Aucun trouble anxieux	46,6%	53,4%	100%

Les tables de contingence

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	22%	78%	100%
Femmes	33,3%	66,7%	100%

Habituellement VI ici

Diagramme en barres superposées



Les outils mathématiques

- Coefficient Phi : chapitre 7
- Rapport des chances (odds ratio)
- Risque attribuable
- Risque relatif

Le rapport des chances

Pour les tables 2 X 2 (ou pour isoler deux modalités dans un table plus grande)

Calculer les chances dans deux modalités d'une variable, puis faire le rapport des chances

Exemple: Combien de fois plus de risques qu'une femme souffre de troubles anxieux par rapport à un homme ?

Le rapport des chances

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	1881	6669	8550
Femmes	3817	7646	11463
Total	5698	14315	20013

Chances chez les hommes : 1881 / 6669 = 0,282 **(ou 1 pour 3,5)**

Chances chez les femmes : 3817 / 7646 = 0,499 **(ou 1 pour 2)**

Le rapport des chances

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	1881	6669	8550
Femmes	3817	7646	11463
Total	5698	14315	20013

Rapport des chances : 0,499 / 0,282 = 1,77

Une femme a 1,77 fois plus de risques (chances) qu'un homme de souffrir de troubles anxieux

Le rapport des chances

	Trouble anxieux	Aucun trouble anxieux	Total
Hommes	1881	6669	8550
Femmes	3817	7646	11463
Total	5698	14315	20013

Rapport des chances : 0,499 / 0,282 = 1,77

Toujours les chances les plus élevées au numérateur et interpréter en conséquence

Statistiques descriptives en épidémiologie

- La prévalence
- Le risque absolu
- Le risque attribuable
- Le risque relatif
- Le rapport des chances

Statistiques descriptives en épidémiologie

	Malade	Non malade	Total
Exposé	A	B	TE
Non exposé	C	D	TNE
Total	TM	TNM	T

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Prévalence = proportion d'individus atteints

(TM x 100) / T = (165 x 100) / 1580 = 10,44 %

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Risque absolu = la prévalence dans l'un des deux groupes

Exemple: Risque absolu chez les fumeurs de cannabis =

A / TE = 63 / 368 = 0,171 (ou 17,1%)

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Risque attribuable = différence entre les risques absolus des deux groupes :

(A / TE) – (C / TNE) = 0,171 – 0,084 = 0,087 (ou 8,7%)

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Risque relatif = rapport entre les risques absolus des deux groupes :

(A / TE) / (C / TNE) = 0,171 / 0,084 = 2,03

Statistiques descriptives en épidémiologie

	Symptômes psychotiques	Pas de symptômes psychotiques	Total
Usage Cannabis	63	305	368
Pas d'usage de cannabis	102	1110	1212
Total	165	1415	1580

Rapport de chances = rapport entre les chances des deux groupes :

(A / B) / (C / D) = (63 / 305) / (102 / 1110) = 2,25

Statistiques descriptives en épidémiologie

Quelle mesure choisir ?

- **Par défaut : Rapport des chances**
- **Risque attribuable pour insister sur les faibles risques réel pour les maladies rares**
- **Risque relatif : facile à comprendre mais pas approprié pour les étude cas-témoins**

Risque attribuable pour les maladies rares

	Maladie rare	Pas de maladie rare	Total
Exposé	3	999 997	1 000 000
Non exposé	1	999 999	1 000 000
Total	4	1 999 996	2 000 000

Risque relatif = (3 / 1 000 000) / (1 / 1 000 000) = 3

Risque attribuable = (3 / 1 000 000) – (1 / 1 000 000) = 0,000002
soit deux personnes affectées de plus par million

Etudes cas-témoins

	Maladie rare	Pas de maladie rare	Total
Exposé	800	500	1300
Non exposé	200	500	700
Total	Cas : 1000	Témoins : 1000	2000

On ne connait pas la prévalence de la maladie (n'est pas 50%)
Risques absolus et relatifs ne sont pas pertinents.

Risque relatif : 2,15 (pas pertinent)
Rapport de chances : 4

Statistiques descriptives bivariées

1. **Décrire la relation entre deux variables nominales**

- **Table de contingence**
- **Diagramme en barres superposées**
- **Technique du rapport de chances**
- **Outils statistiques liés à l'épidémiologie**

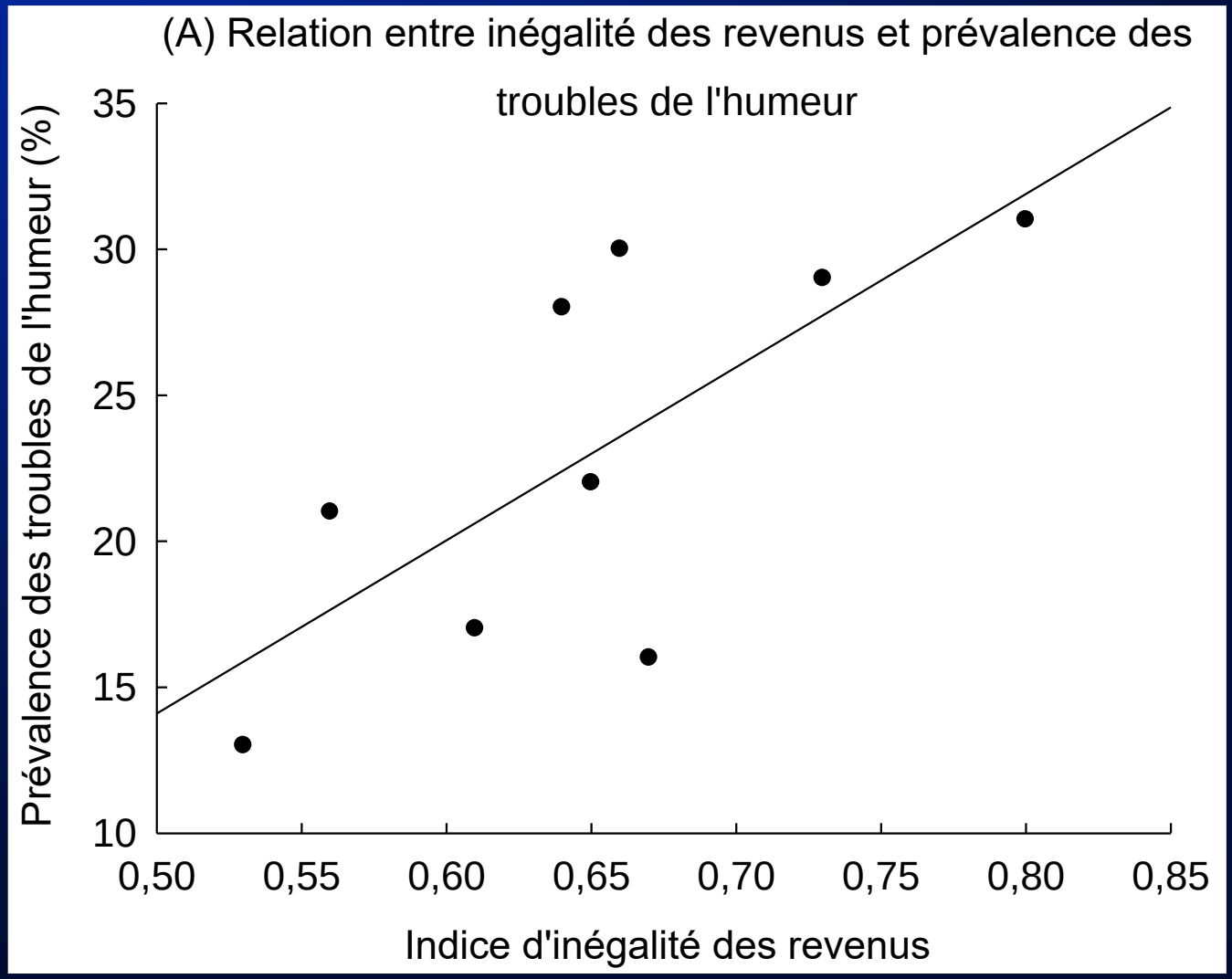
2. **Décrire la relation entre deux variables ordinales ou métriques**

- **Diagramme de dispersion**
- **Covariance et corrélation**
- **Droite de régression**
- **Coefficient de détermination**

Le diagramme de dispersion

Pays	Inégalité des revenus	Troubles de l'humeur (%)
Allemagne	0,67	16
Belgique	0,65	22
Espagne	0,56	21
Etats-Unis	0,80	31
France	0,73	29
Italie	0,61	17
Japon	0,53	13
Nouvelle-Zélande	0,66	30
Pays-Bas	0,64	28

Le diagramme de dispersion



Le diagramme de dispersion

Le prédicteur X (variable indépendante):

En abscisse = axe horizontal

Le critère Y (variable dépendante):

En ordonnée = axe vertical

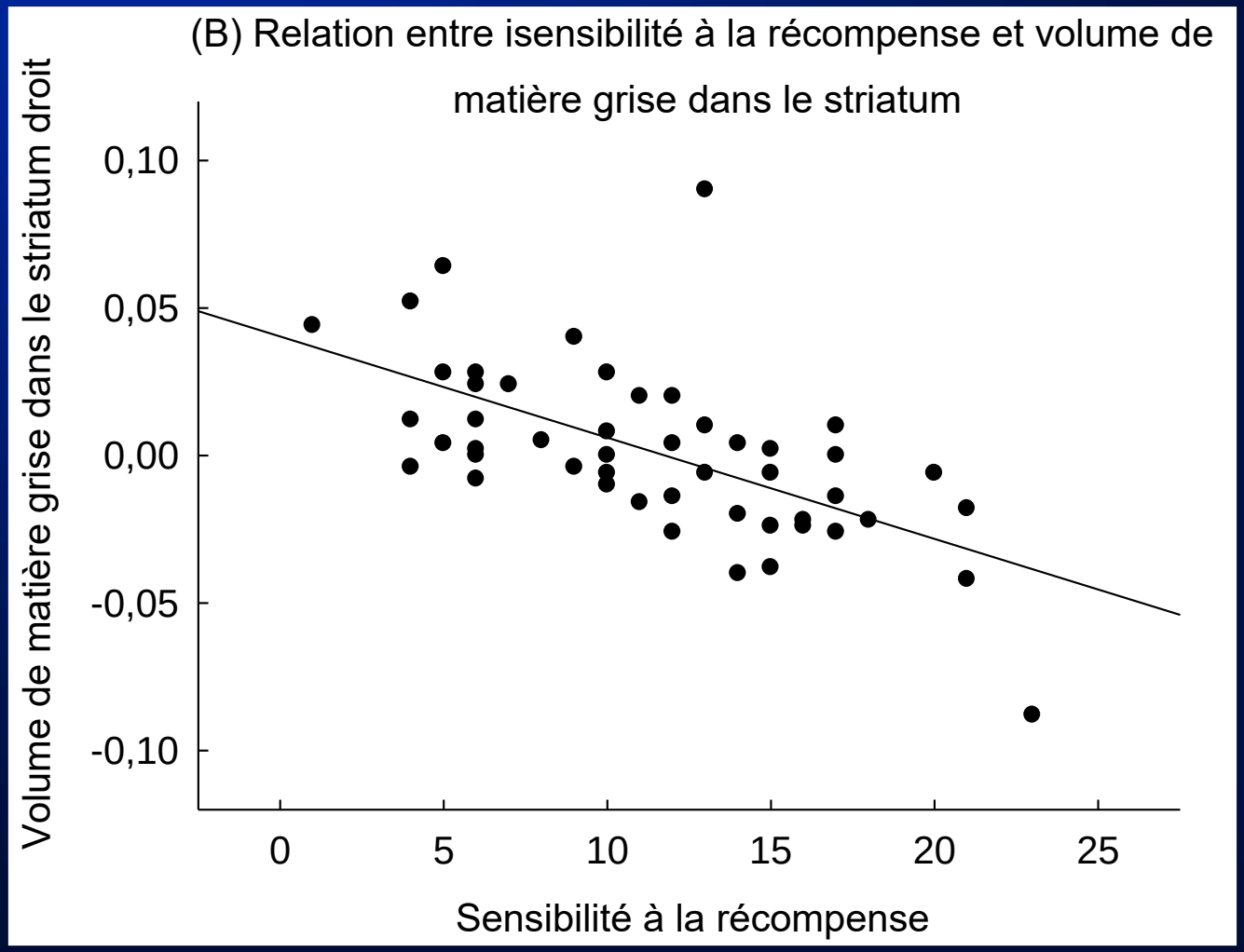
La droite de régression

La droite la mieux ajustée aux données

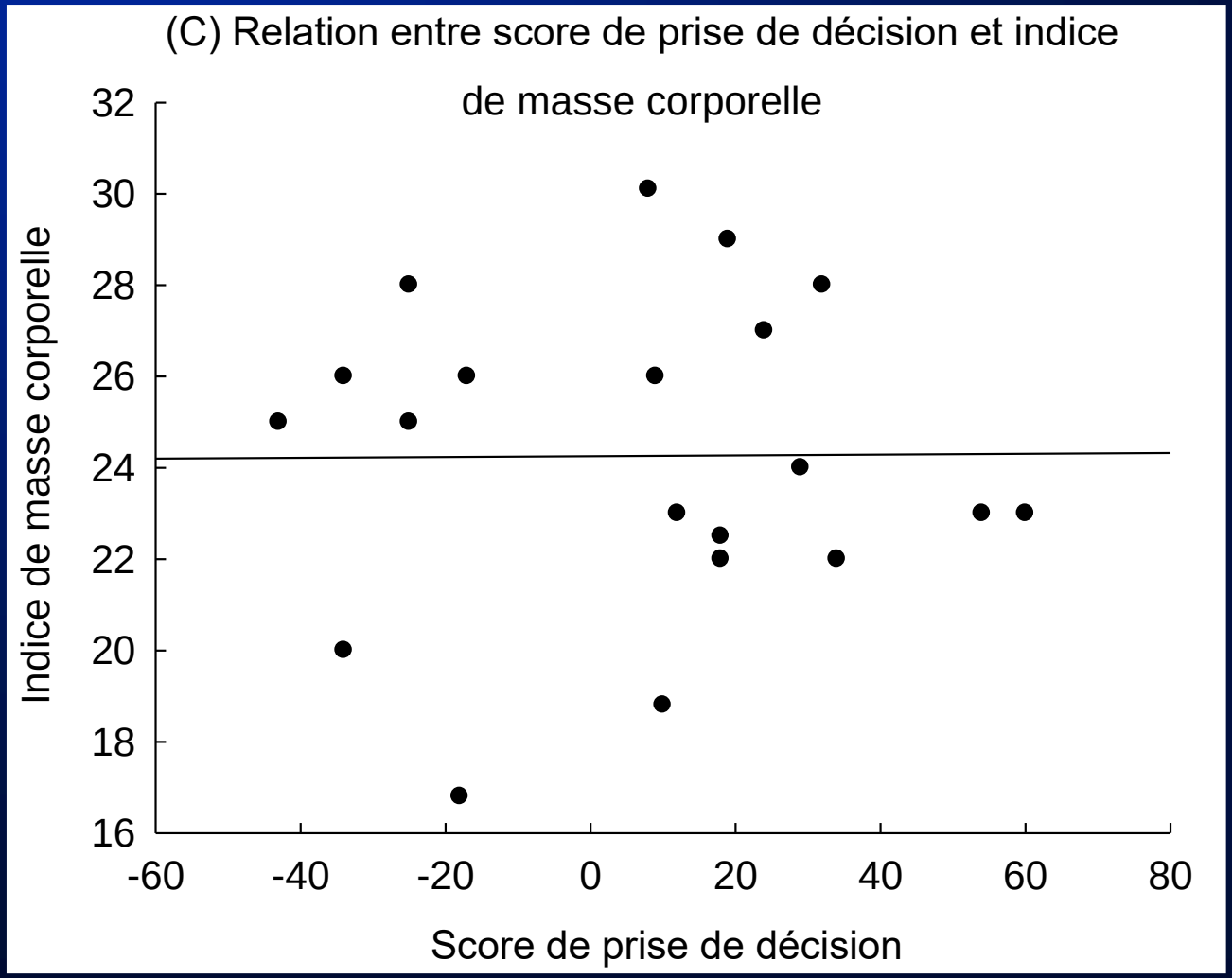
La droite de régression de Y prédit en fonction de X – « Y en X »

Permet de faire des prédictions

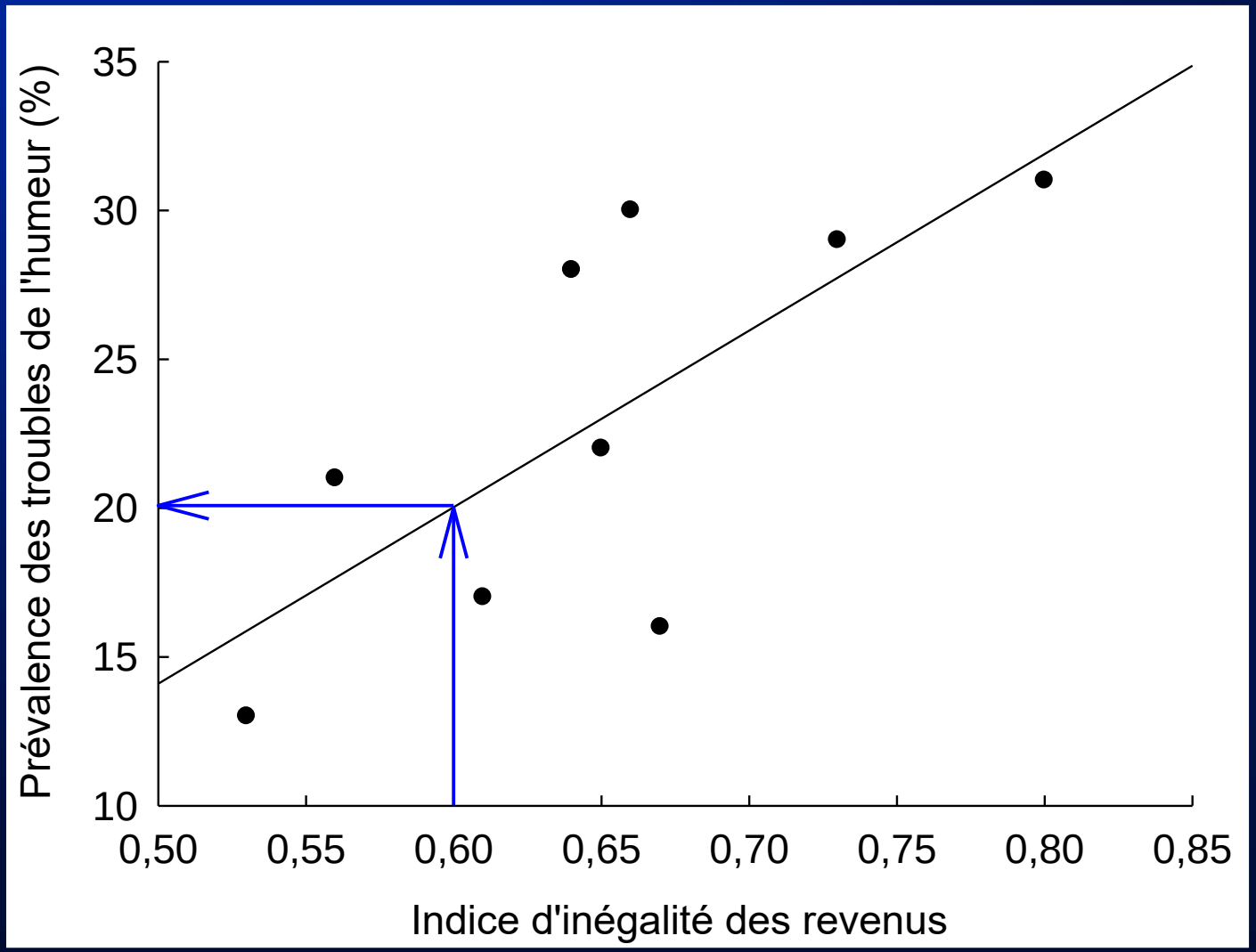
Le diagramme de dispersion



Le diagramme de dispersion



Prédiction $\hat{Y}$



$\hat{Y} = 20$

La covariance

Reflète le degré auquel deux variables varient ensemble

$$\mathrm{COV}_{XY} = \frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{N - 1}$$

Pays	X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)$
Allemagne	0,67	16	0,02	-7	-0,14
Belgique	0,65	22	0,00	-1	0,00
Espagne	0,56	21	-0,09	-2	0,18
Etats-Unis	0,80	31	0,15	8	1,20
France	0,73	29	0,08	6	0,48
Italie	0,61	17	-0,04	-6	0,24
Japon	0,53	13	-0,12	-10	1,20
N-Zélande	0,66	30	0,01	7	0,07
Pays-Bas	0,64	28	-0,01	5	-0,05
Somme	5,85	207	0,00	0	3,18
Moyenne	0,65	23			

La covariance

Exemple pour la covariance entre indice d'inégalité des revenus et prévalence des troubles de l'humeur:

$$cov_{XY} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{n - 1} = \frac{3,18}{8} = 0,3975$$

La covariance

$$\mathrm{cov}_{XY} = \frac{\sum XY - \dfrac{\sum X \sum Y}{N}}{N-1}$$

Le coefficient de corrélation de Pearson (*r*)

Problème de la covariance:

Sa valeur est fonction de S_X et S_Y et de leurs unités

Covariance de 0,3975 = ?

Soit faible relation entre variables
Soit petits écart-types

Le coefficient de corrélation de Pearson (*r*)

= Mesure du degré de relation entre deux variables

$$r = \frac{\mathrm{cov}_{XY}}{S_X S_Y}$$

La corrélation (*r*) entre X et Y

Indice de la relation entre X et Y

Valeur entre –1 et +1

Degré de regroupements des points autour de la droite

La corrélation (*r*) entre X et Y

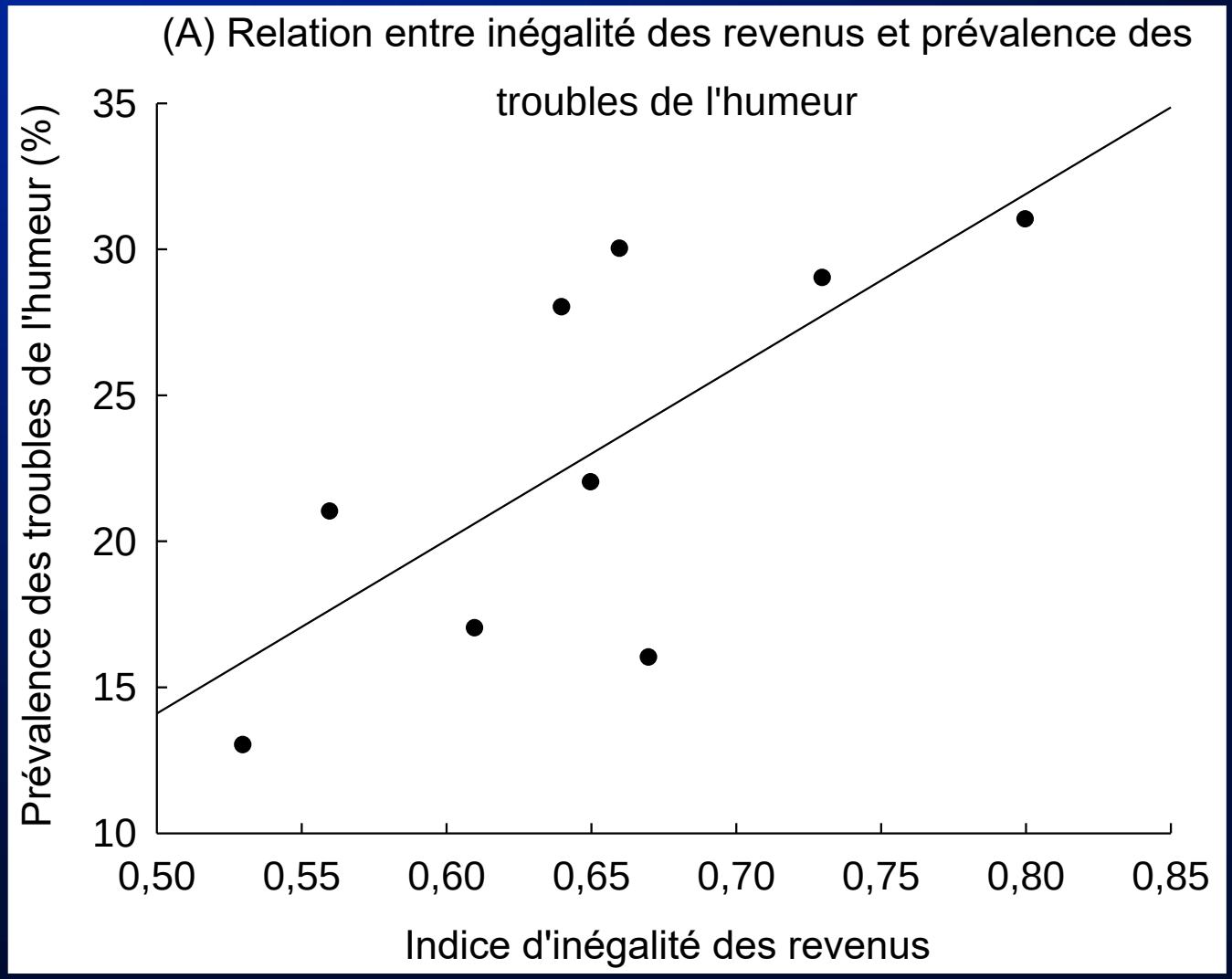
Le signe (+ ou –) indique le sens de la relation = corrélation positive / négative

La valeur absolue indique le degré de relation entre les variables

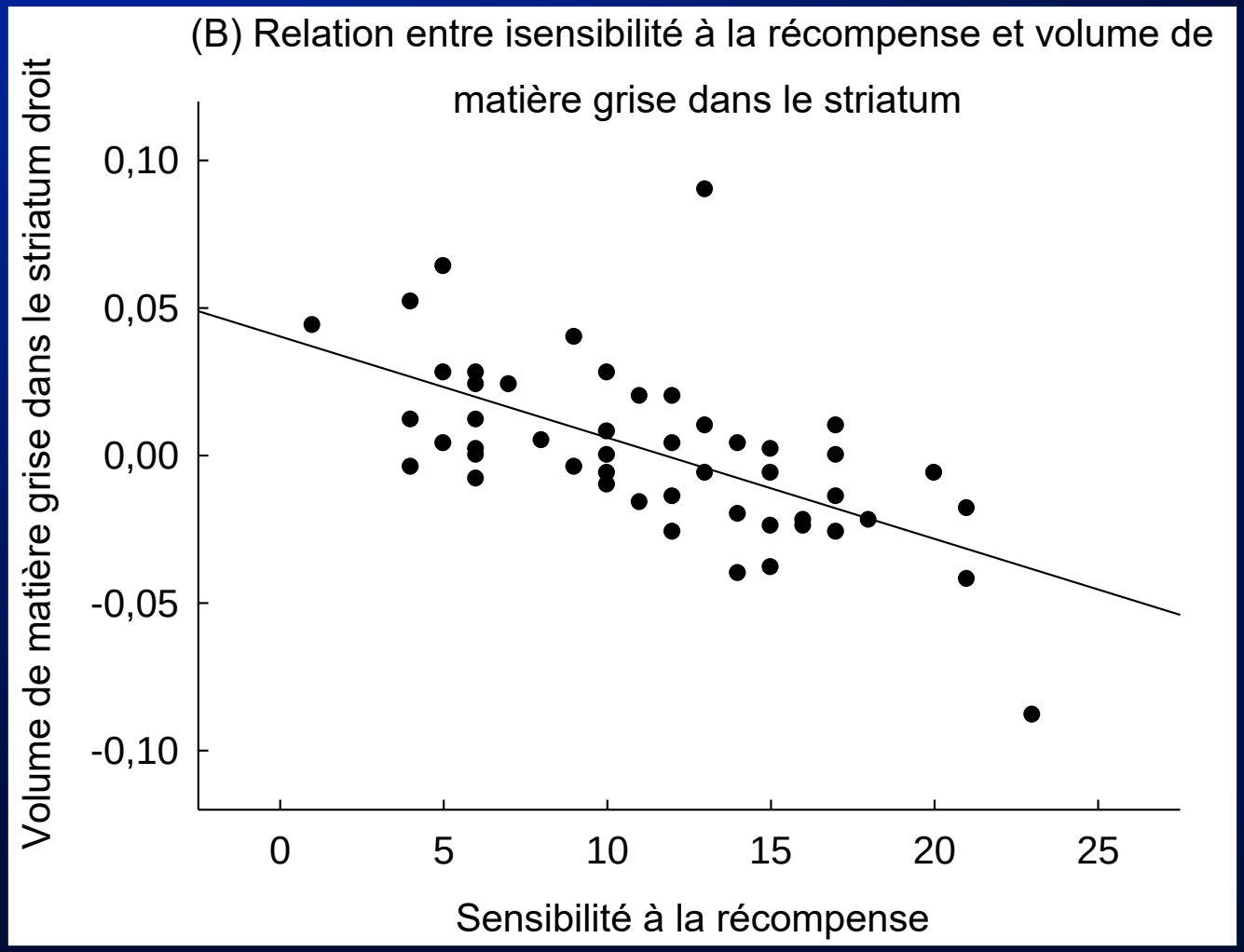
La corrélation (*r*) entre X et Y

Corrélation	Interprétation
0,80 – 1,00	Corrélation importante, de grande taille
0,50 – 0,80	Corrélation de taille moyenne
0,20 – 0,50	Corrélation de petite taille
0,00 – 0,20	Corrélation négligeable

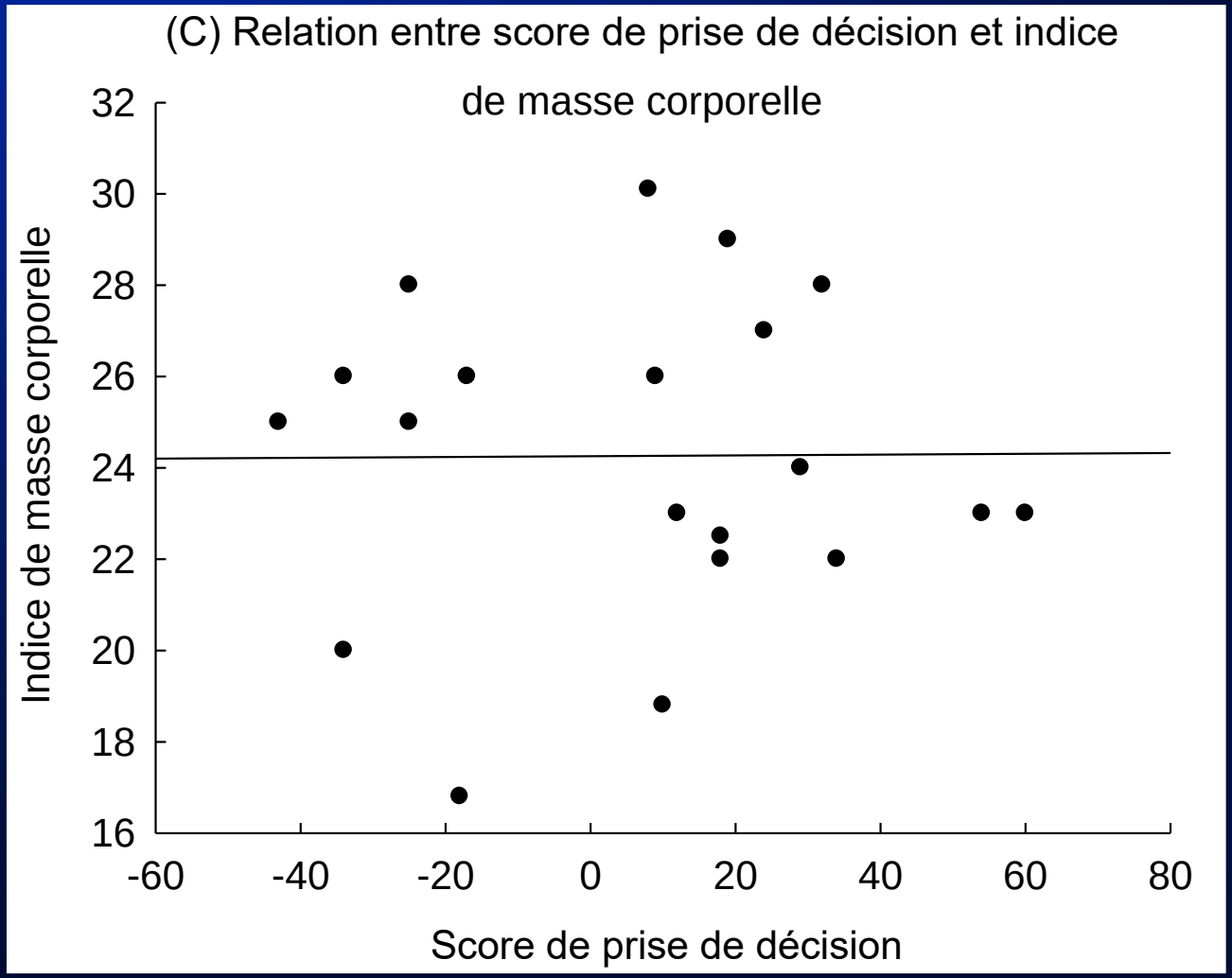
Corrélation positive : r = 0,72



Corrélation négative : r = –0,61



Corrélation négligeable : r = 0,007



Calcul du coefficient de corrélation

Revenus (X)	Troubles humeur (Y)	X^2	Y^2	XY
0,67	16	0,4489	256	10,72
0,65	22	0,4225	484	14,30
0,56	21	0,3136	441	11,76
0,80	31	0,6400	961	24,80
0,73	29	0,5329	841	21,17
0,61	17	0,3721	289	10,37
0,53	13	0,2809	169	6,89
0,66	30	0,4356	900	19,80
0,64	28	0,4096	784	17,92
$\Sigma X = 5,85$	$\Sigma Y = 207$	$\Sigma X^2 = 3,8561$	$\Sigma Y^2 = 5125$	$\Sigma XY = 137,73$

Calcul du coefficient de corrélation (*r*)

$\Sigma X = 5{,}85$	$\Sigma Y = 207$	$\Sigma X^2 = 3{,}8561$	$\Sigma Y^2 = 5125$	$\Sigma XY = 137{,}73$

$$cov_{XY} = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{N-1} = \frac{137,73 - \frac{5,85 \times 207}{9}}{8} = 0,3975$$

Calcul du coefficient de corrélation (*r*)

$\Sigma X = 5{,}85$ $\Sigma Y = 207$ $\Sigma X^2 = 3{,}8561$ $\Sigma Y^2 = 5125$ $\Sigma XY = 137{,}73$

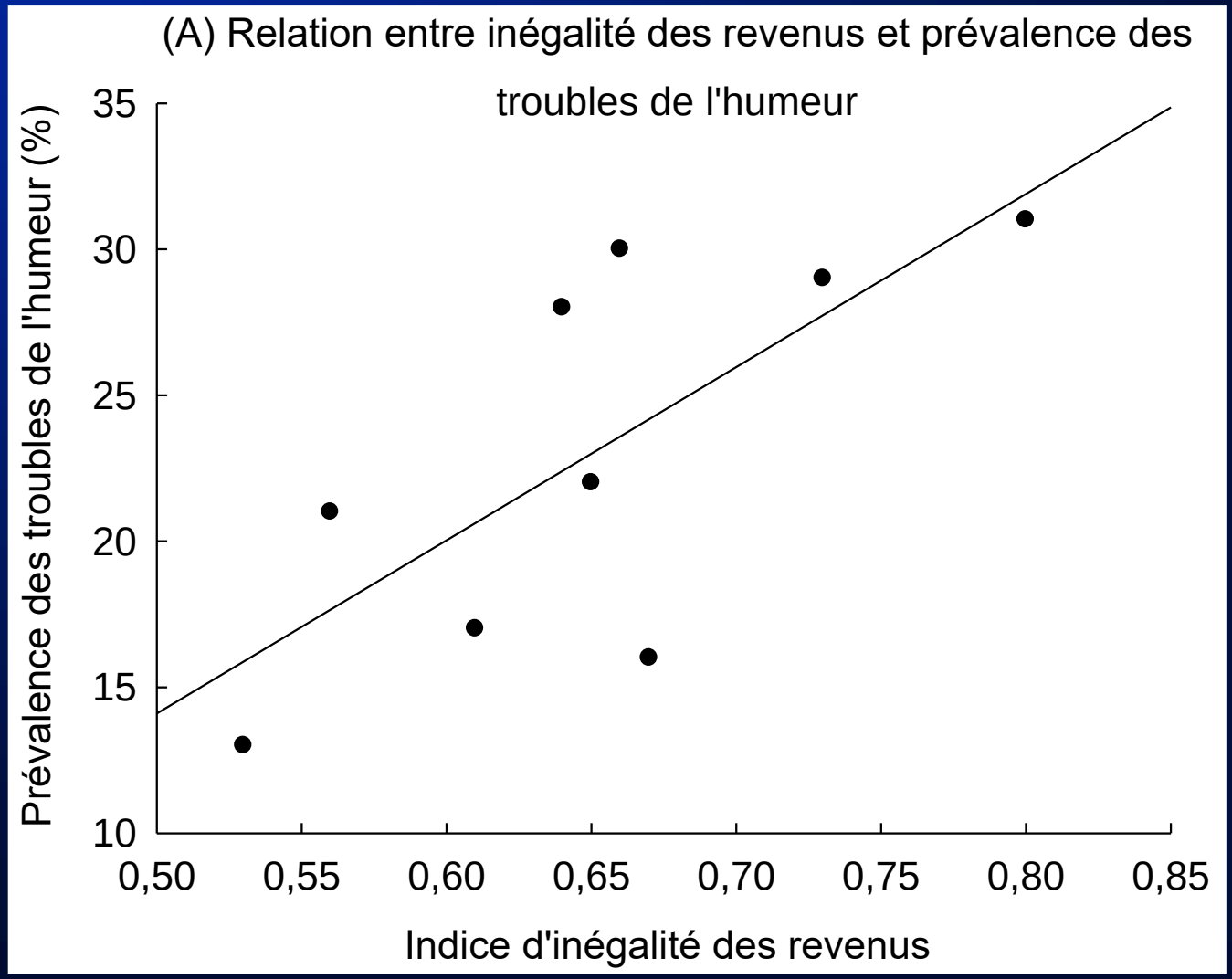
$$S_X = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}} = \sqrt{\frac{3,8561 - \frac{5,85^2}{9}}{8}} = 0,0819$$

$$S_Y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{N}}{N-1}} = \sqrt{\frac{5125 - \frac{207^2}{9}}{8}} = 6,7454$$

Calcul du coefficient de corrélation (*r*)

$$r = \frac{cov_{XY}}{S_X S_Y} = \frac{0,3975}{0,0819 \times 6,7454} = 0,72$$

Calcul de la droite de régression



Calcul de la droite de régression

$$\hat{Y} = bX + a$$

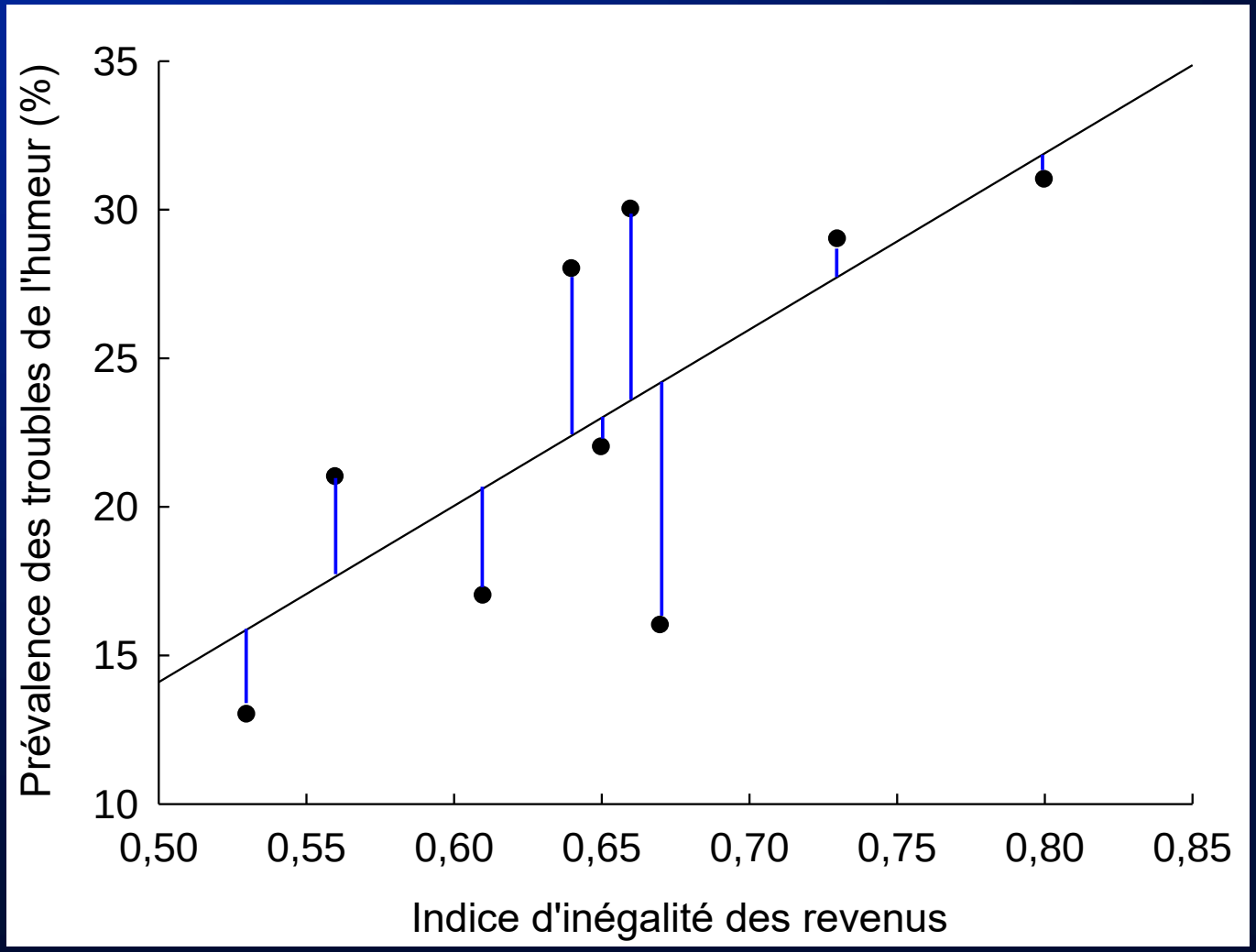
$\hat{Y}$ = la valeur prédite de Y
b = la pente de la droite de régression
a = l'ordonnée à l'origine
X = la valeur du prédicteur

Calcul de la droite de régression

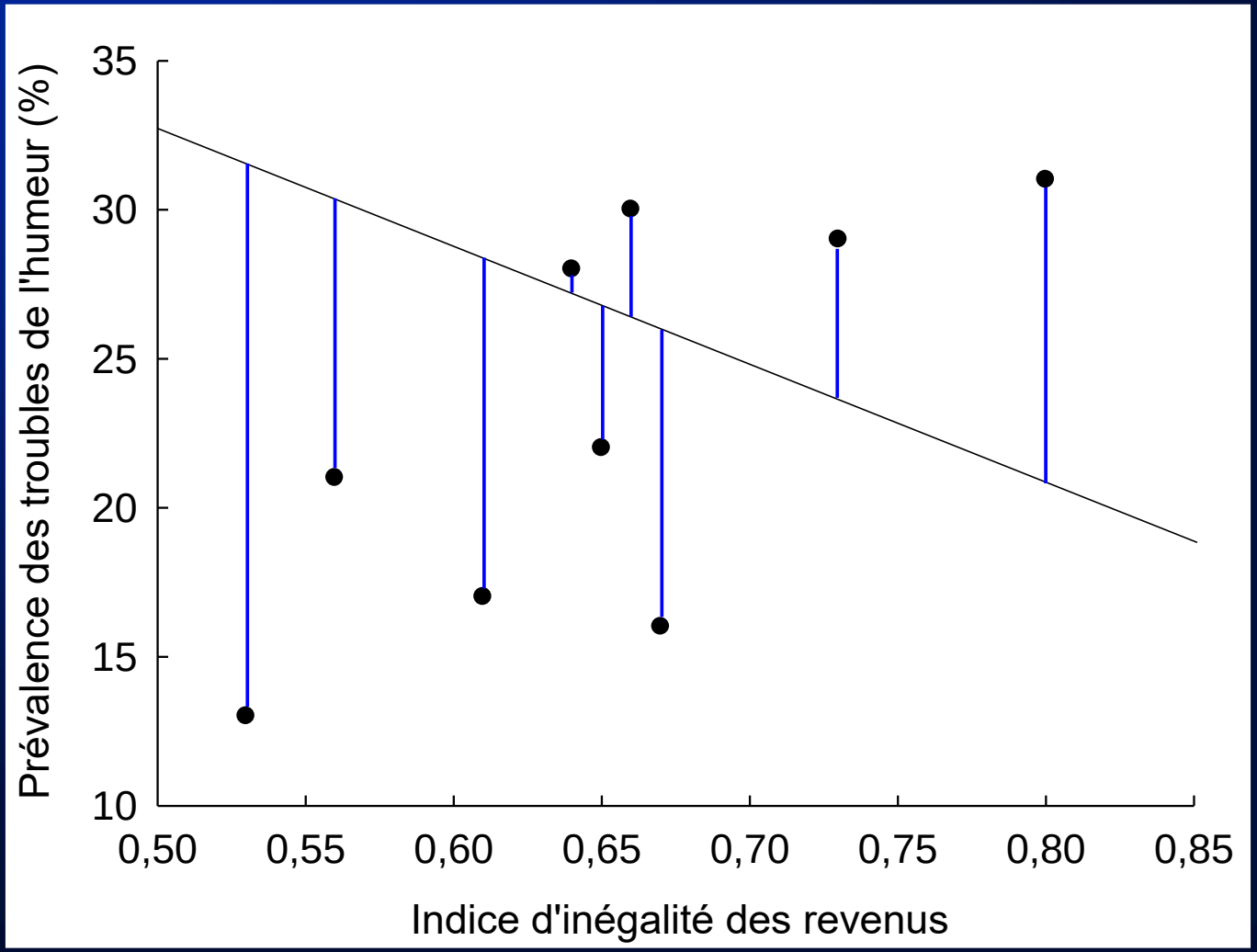
$$\hat{Y} = bX + a$$

Comment trouver les valeurs des coefficients a et b pour la droite la mieux ajustée aux données ?

Droite bien ajustée



Droite mal ajustée



Calcul de la droite de régression

- Droite qui minimise les écarts (ou résidus)
- Résidu = $Y - \hat{Y}$ = erreur de prédiction
- Droite qui minimise la somme des écarts au carré
  (technique des moindres carrés)

Pays	X	Y	Score prédit (Ŷ)	Résidu (Y-Ŷ)	Résidu carré
Allemagne	0,67	16	24,19	-8,19	67,02
Belgique	0,65	22	23,00	-1,00	1,00
Espagne	0,56	21	17,66	3,34	11,15
Etats-Unis	0,80	31	31,90	-0,90	0,81
France	0,73	29	27,75	1,25	1,57
Italie	0,61	17	20,63	-3,63	13,16
Japon	0,53	13	15,88	-2,88	8,30
N.-Zélande	0,66	30	23,59	6,41	41,04
Pays-Bas	0,64	28	22,41	5,59	31,28
Moyenne	0,65	23	23	0,00	19,48

Calcul de la droite de régression

On cherche la droite qui minimise les erreurs de prédiction = qui minimise $\Sigma(Y-\hat{Y})^2$

$$b = \frac{\mathrm{COV}_{XY}}{S_X^2} \qquad a = \overline{Y} - b\overline{X}$$

Exemple pour l'inégalité des revenus

$cov_{XY} = 0,3975$ $S_X^2 = 0,0067$ $\overline{X} = 0,65$

$\overline{Y} = 23$

$$b = \frac{cov_{XY}}{S_X^2} = \frac{0,3975}{0,0067} = 59,328$$

$$a = \overline{Y} - b\overline{X} = 23 - (59,328 \times 0,65) = -15,563$$

Exemple pour l'inégalité des revenus

L'équation de la droite de régression est donc:

$\hat{Y} = bX + a = 59{,}328\ X - 15{,}563$

Prédire un score particulier

Exemple: quelle serait la prévalence des troubles de l'humeur pour un pays ayant 0,80 comme indice d'inégalité des revenus ?

$$\hat{Y} = 59{,}328\ X - 15{,}563$$

$$\hat{Y} = (59{,}328 \times 0{,}80) - 15{,}563 = 31{,}90$$

Que sont a et b ?

<u>a est l'ordonnée à l'origine</u>

La valeur de $\hat{Y}$ prédite pour X = 0

$$\hat{Y} = bX + a = (b \times 0) + a = a$$

Pour X = 0, $\hat{Y} = a$

Parfois ce n'est qu'une abstraction mathématique

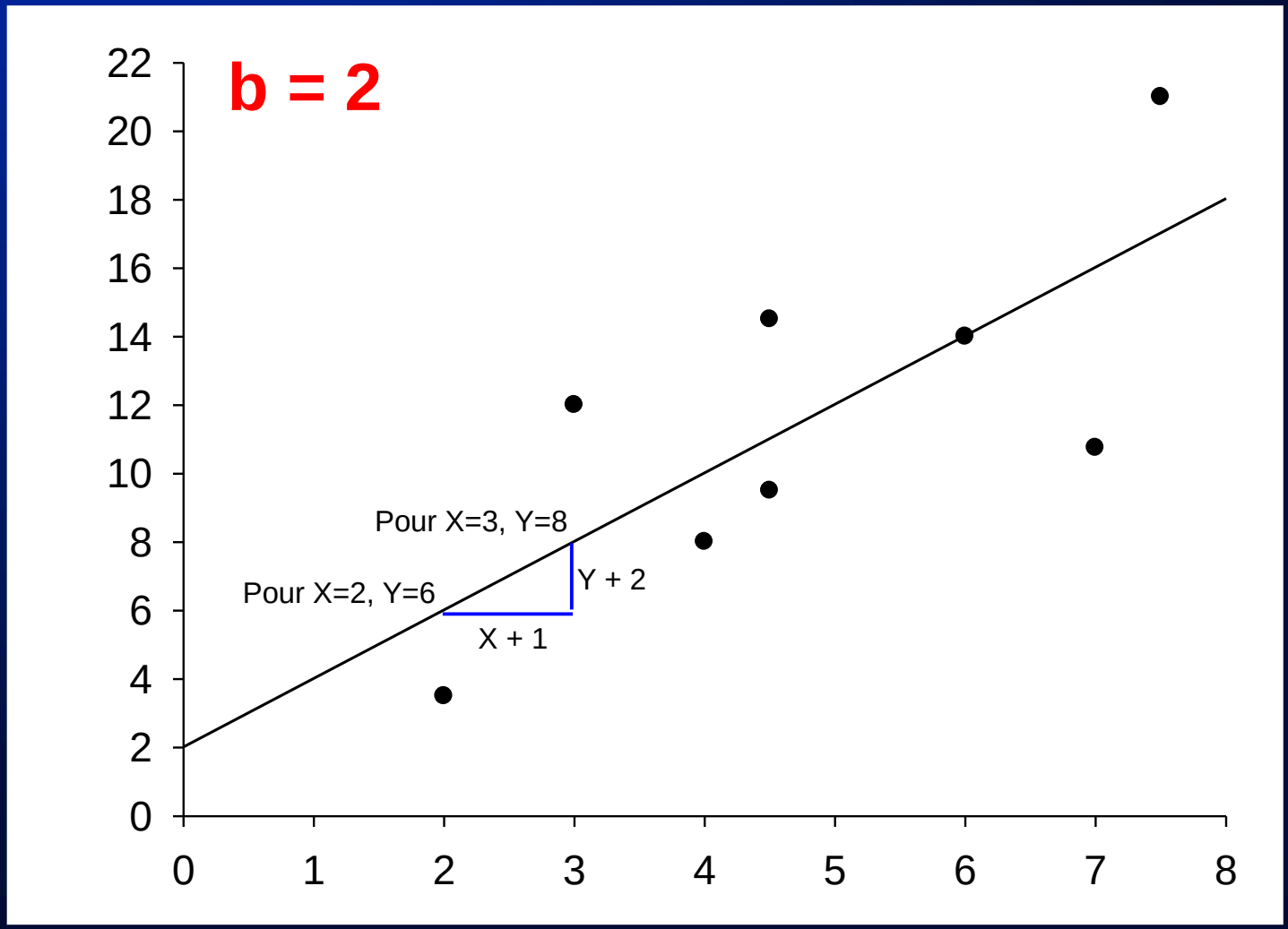
Que sont a et b ?

b définit la pente de la droite de régression

= changement de $\hat{Y}$ correspondant à un changement d'une unité de X

= mesure du taux de changement

Que sont a et b ?



Interprétation de b

Exemple pour l'inégalité des revenus

b = 59,328

Si le score d'inégalité des revenus augmente de 1 point, la prévalence augmente de 59%

Si le score d'inégalité des revenus augmente de 0,1 point, la prévalence augmente de 5,9%

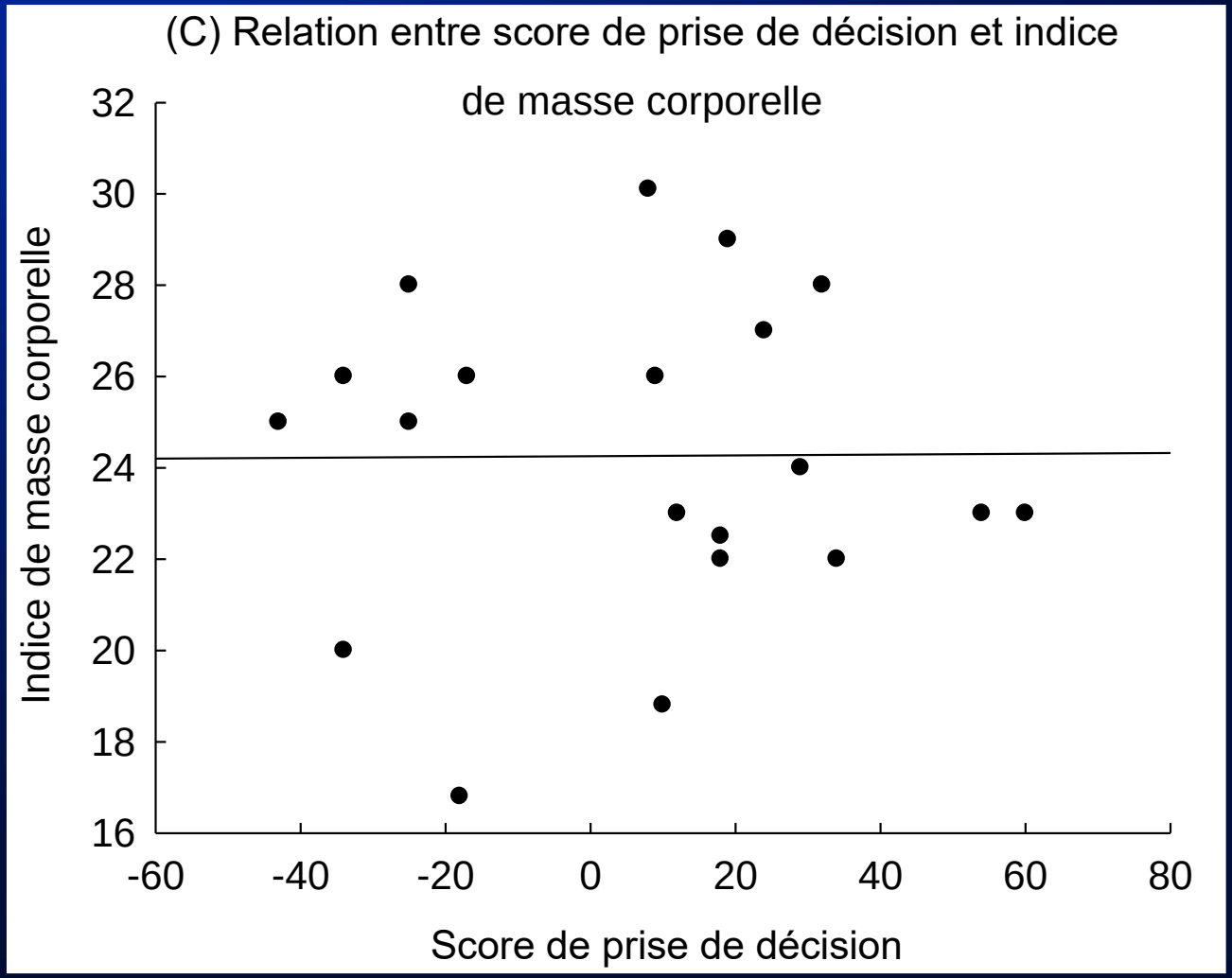
L’exactitude de la prédiction

La droite est-elle bien ajustée aux données ?

Les prédictions sont-elles bonnes ?

Indice mathématique de l’exactitude de la prédiction = erreur standard d’estimation

Droite mal ajustée



L’erreur standard d’estimation

- Basée sur les résidus

- Comme pour l’écart-type:
  - Mettre les résidus au carré
  - Faire la moyenne des résidus au carré = Variance résiduelle
  - Prendre la racine carrée de la variance résiduelle

Pays	X	Y	Score prédit (Ŷ)	Résidu (Y-Ŷ)	Résidu carré
Allemagne	0,67	16	24,19	-8,19	67,02
Belgique	0,65	22	23,00	-1,00	1,00
Espagne	0,56	21	17,66	3,34	11,15
Etats-Unis	0,80	31	31,90	-0,90	0,81
France	0,73	29	27,75	1,25	1,57
Italie	0,61	17	20,63	-3,63	13,16
Japon	0,53	13	15,88	-2,88	8,30
N.-Zélande	0,66	30	23,59	6,41	41,04
Pays-Bas	0,64	28	22,41	5,59	31,28
Moyenne	0,65	23	23	0,00	19,48

L’exactitude de la prédiction

Variance résiduelle ou variance de l’erreur:

$$S^2_{Y \cdot X} = \frac{\sum \left(Y - \hat{Y} \right)^2}{N - 2}$$

L’exactitude de la prédiction

Variance résiduelle ou variance de l’erreur:

$$S^2_{Y \cdot X} = \frac{SC_{résiduelle}}{dl}$$

Somme de carrés résiduelle

L'exactitude de la prédiction

Erreur standard d'estimation:

$$S_{Y \cdot X} = \sqrt{\frac{\sum (\mathbf{Y} - \hat{\mathbf{Y}})^2}{N - 2}}$$

Pays	X	Y	Score prédit (Ŷ)	Résidu (Y-Ŷ)	Résidu carré
Allemagne	0,67	16	24,19	-8,19	67,02
Belgique	0,65	22	23,00	-1,00	1,00
Espagne	0,56	21	17,66	3,34	11,15
Etats-Unis	0,80	31	31,90	-0,90	0,81
France	0,73	29	27,75	1,25	1,57
Italie	0,61	17	20,63	-3,63	13,16
Japon	0,53	13	15,88	-2,88	8,30
N.-Zélande	0,66	30	23,59	6,41	41,04
Pays-Bas	0,64	28	22,41	5,59	31,28
Somme	5,85	207	207	0,00	175,33

Calcul de l'erreur standard d'estimation pour l'inégalité des revenus

$$S_{Y \cdot X} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{N - 2}} = \sqrt{\frac{175,33}{7}} = 5$$

Les points s'écartent typiquement de 5 unités par rapport à la droite de régression

Manière simple de calculer $S_{Y \cdot X}$

Erreur standard d'estimation:

$$S_{Y \cdot X} = S_Y \sqrt{(1 - r^2)\frac{N-1}{N-2}}$$

Le coefficient r^2

r^2 = le pourcentage de la variabilité de Y qui est prédite (expliquée) par la variabilité de X

r^2 = coefficient de détermination

Le coefficient r^2

r^2 est compris entre 0 et 1

Coefficient simple à comprendre

Mais on perd l'information du sens de la relation

Exemple

Prédiction de la prévalence des troubles de l'humeur à partir des scores d'inégalité des revenus

$r = 0,72$ $r^2 = 0,72^2 = 0,52$

52% de la variance des troubles de l'humeur peut être prédite à partir des scores d'inégalité des revenus

Corrélation ou régression

Les deux techniques sont liées :

$$b = r\frac{S_Y}{S_X} \qquad r = b\frac{S_X}{S_Y}$$

NB: Si b = 0 alors r = 0

Corrélation ou régression

Distinction théorique (officielle):

Corrélation : X et Y = variables aléatoires

Régression : X = variable fixe

Y = variable aléatoire

Corrélation ou régression

Distinction pratique:

Corrélation :

- Etudier simplement le degré de relation entre X et Y
- Identification d'une VI pas nécessaire

Régression :

- Etudier le taux de changement en unités de la VD
- Une VI est clairement identifiée

Quelques considérations supplémentaires

- Restriction de l'étendue
- Influence des scores extrêmes
- Linéarité de la relation
- Prédictions en dehors du champ des valeurs observées
- Interprétation en termes de causalité

Restriction de l'étendue

Limiter l'étendue des scores mesurés sur la variable X risque de sous-estimer la taille de la corrélation

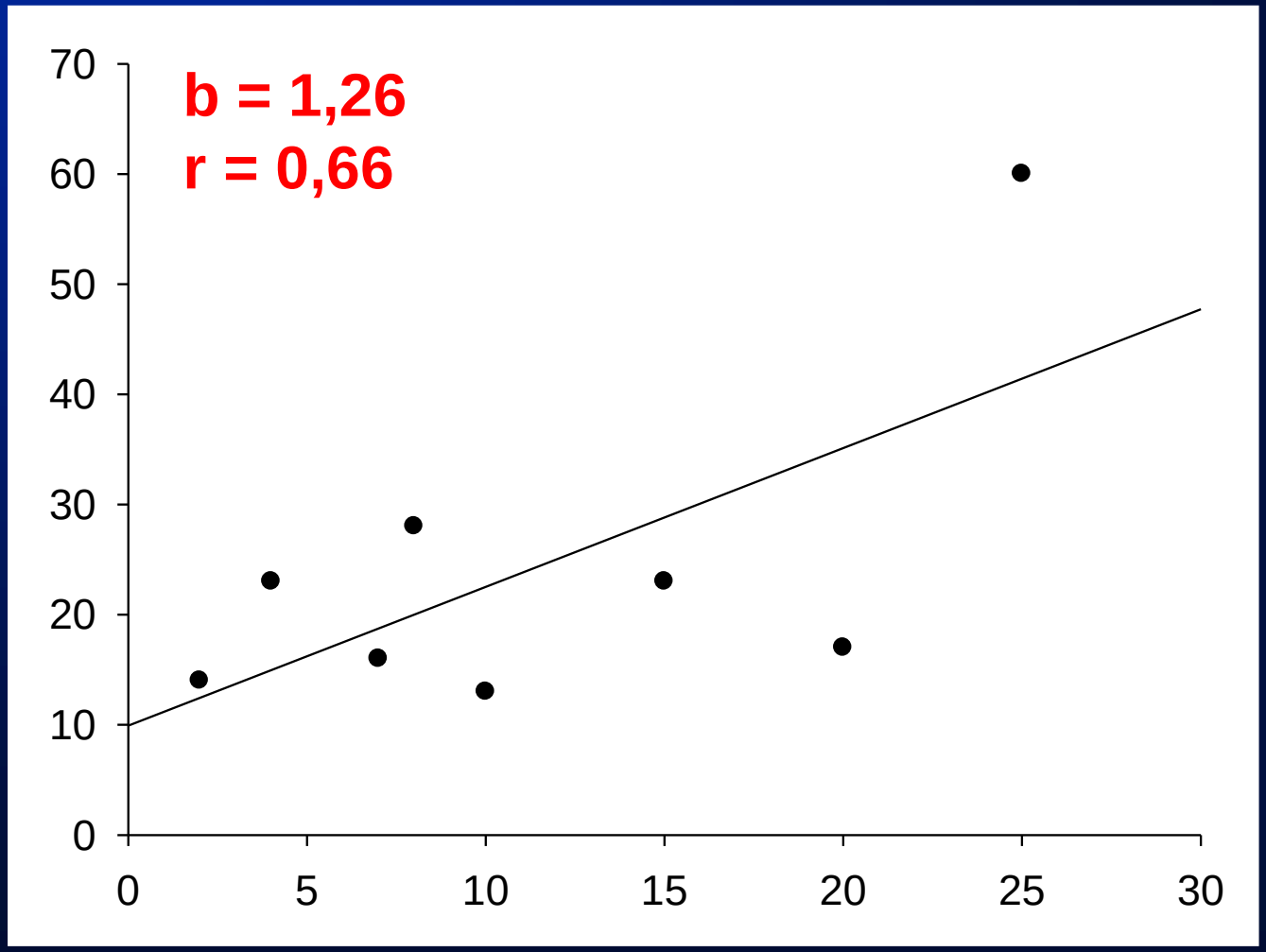
Exemple : r = 0,19 si on étudie seulement les pays avec des indices de 0,60-0,70.

Influence des scores extrêmes

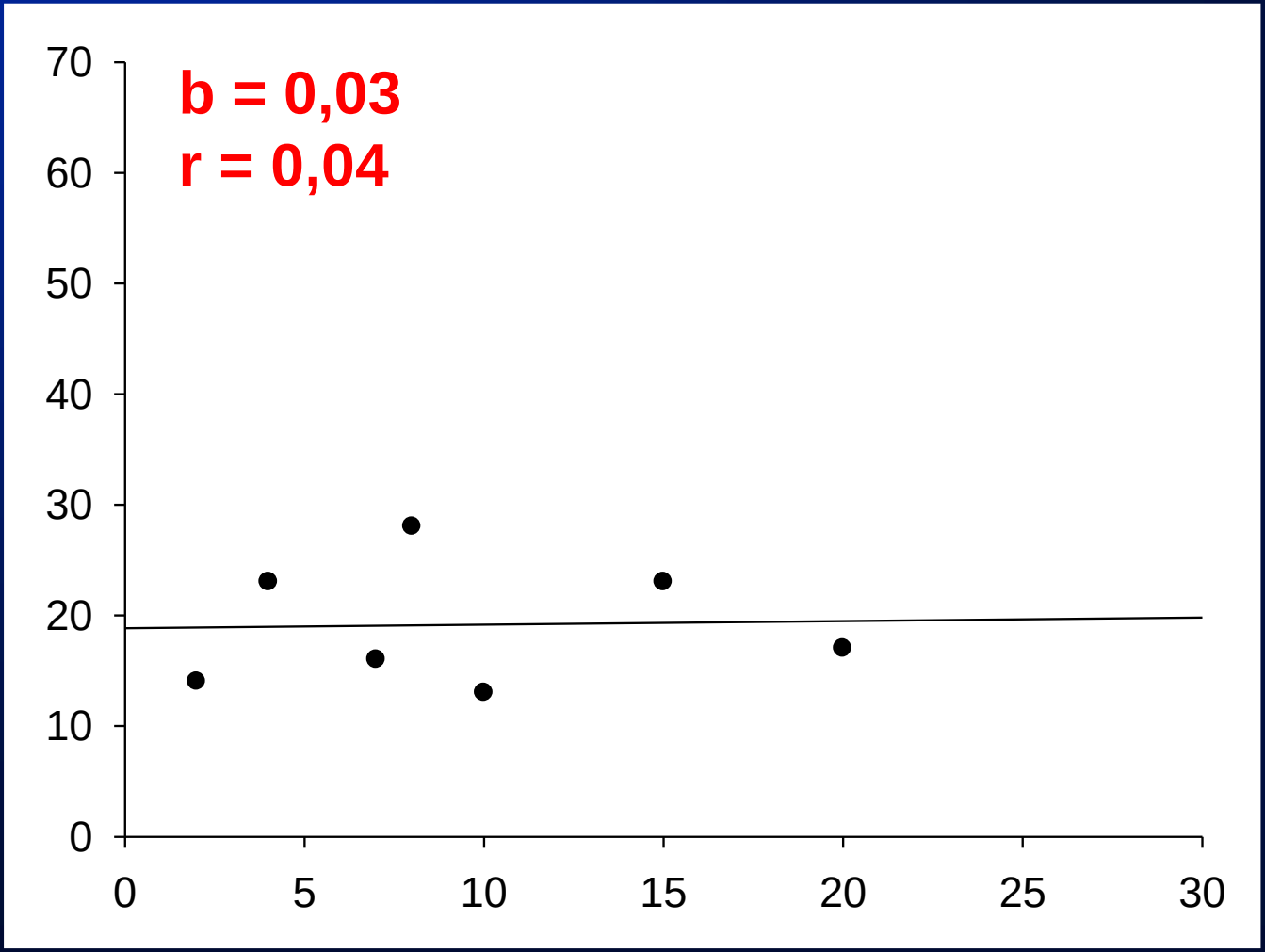
Attention avec l'interprétation d'une corrélation/régression qui repose essentiellement sur un minorité de points

Techniques statistiques pour repérer les points excessivement influents

Influence des scores extrêmes



Influence des scores extrêmes



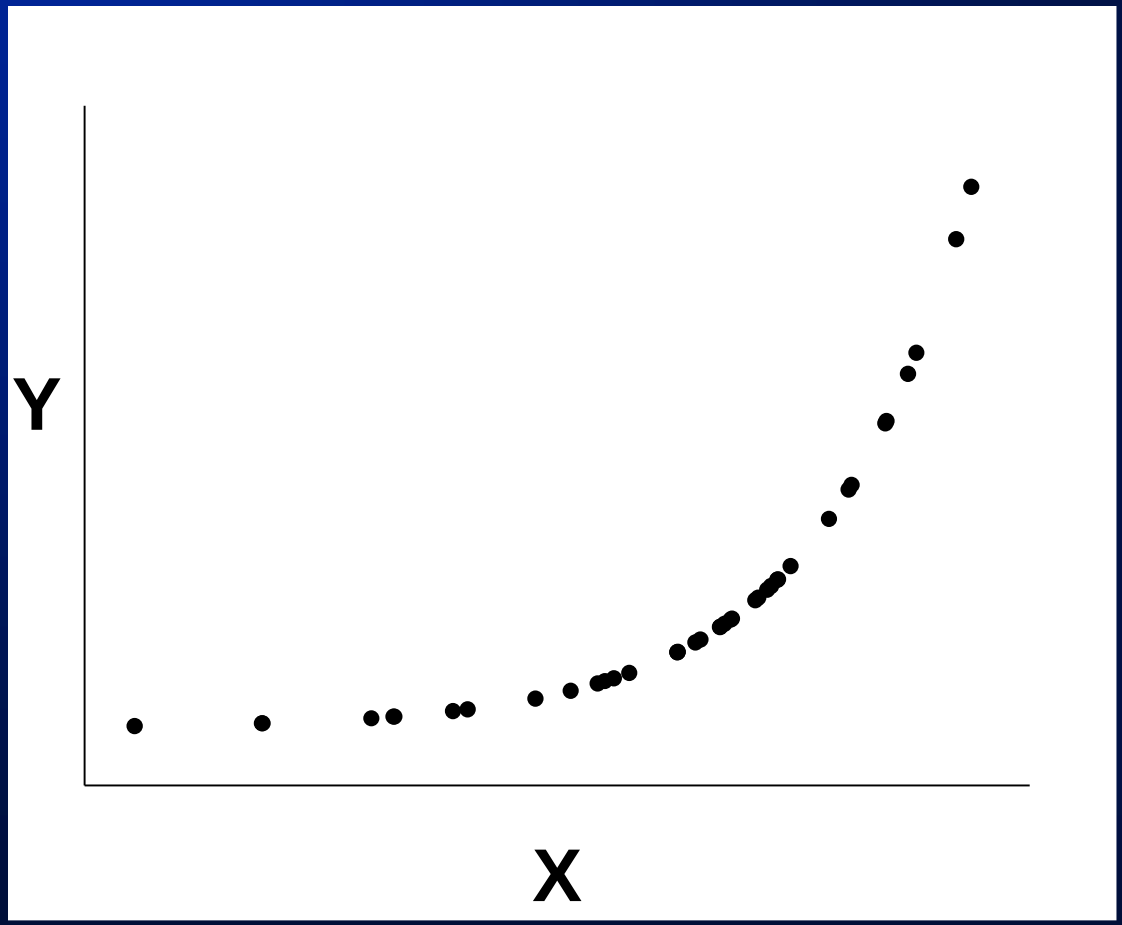
Linéarité de la relation

Les techniques étudiées permettent de mesurer une relation linéaire

D'autres types de relation impliquent d'autres traitements statistiques

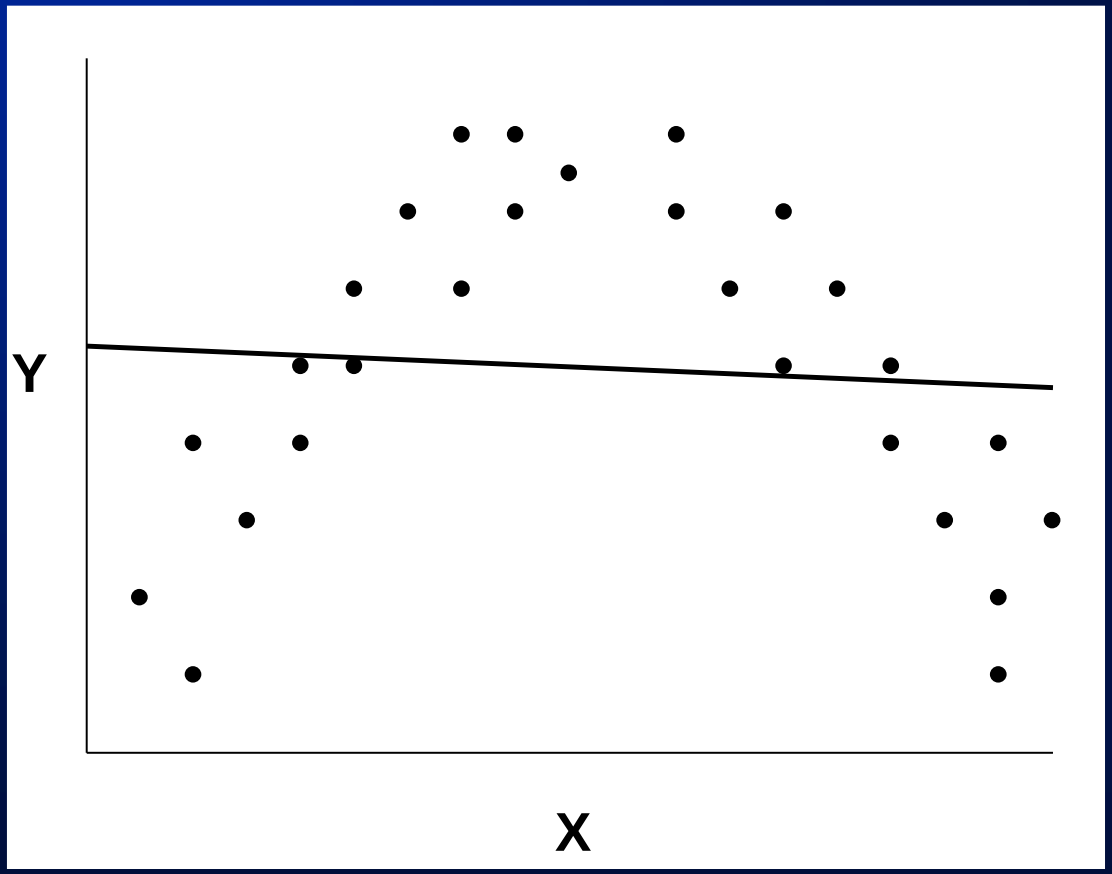
Absence de corrélation linéaire ≠ absence de relation

Exemple de relation curvilinéaire



Pour ces données: r = 0,81

Exemple de relation curvilinéaire



Pour ces données: r = –0,07

Prédictions en dehors du champ des valeurs observées

Attention aux prédictions de scores pour des valeurs en dehors de l'échantillon

Risque important qu'ils soient absurdes ou erronés

La forme de relation peut être différente en dehors des valeurs X testées

Prédictions en dehors du champ des valeurs observées

Exemple avec l'inégalité des revenus

(indices entre 0,5 et 0,8 dans l'échantillon)

Quelle prédiction pour un score de 0,1 ?

$\hat{Y}$**= (59,328 x 0,1) – 15,663 = – 9,63**

Prévalence négative = absurde !!

Interprétations en termes de causalité

Une corrélation ne signifie pas lien de cause à effet !!!

Exemple: corrélation significative entre nombre de machines à lessiver et nombre d'obèses

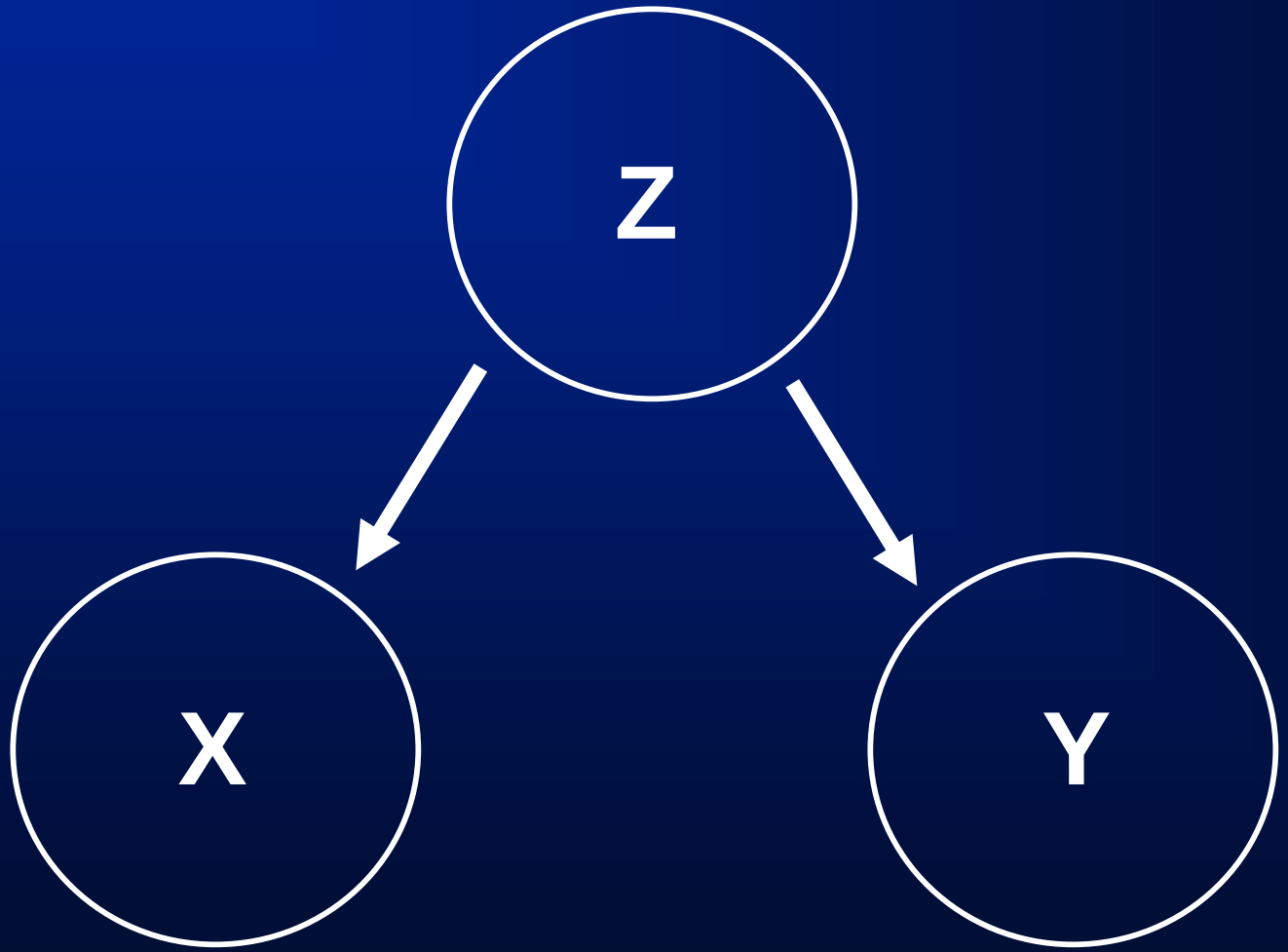
Interprétations en termes de causalité

Exemple: X = consommation de tabac
Y = capacité pulmonaire

Interprétations en termes de causalité

Pas toujours clair si X cause Y ou Y cause X
Exemple: Revenus et Santé

Interprétations en termes de causalité



Interprétations en termes de causalité

Si la corrélation est due à une variable exogène:

Fausse corrélation ou Corrélation absurde

Rapporter les statistiques descriptives bivariées

Variables nominales

Texte, tableau de contingence avec % ou diagramme en barres

Faire simple (ex. réduire les modalités montrées)

Variables ordinales ou métriques

Texte, diagrammes de dispersion ou matrice de corrélations

Rapporter les statistiques descriptives bivariées

« Un échantillon total de 1580 personnes a été recruté pour l'étude. Dans cet échantillon, 368 personnes consommaient du cannabis (23% de l'échantillon). Chez les consommateurs de cannabis, 17% étaient atteints de troubles psychotiques pour seulement 8% des personnes ne consommant pas de cannabis. La technique du rapport des chances indique que les consommateurs de cannabis ont 2,25 fois plus de risques de souffrir de symptômes psychotiques ».

Rapporter les statistiques descriptives bivariées

« Il y a une corrélation positive de taille moyenne entre l'indice d'inégalité des revenus et le taux de prévalence des troubles de l'humeur (r = 0,72) ».

Rapporter les statistiques descriptives bivariées

Variables	A	B	C	D	E
A	1,00				
B	0,42	1,00			
C	-0,33	-0,12	1,00		
D	0,87	0,53	-0,53	1,00	
E	0,24	0,16	-0,21	0,35	1,00