Chapitre 11

Tests statistiques sur les corrélations et régressions

Rappel

Relation entre X et Y :

	X	Y
Sujet 1	100	56
Sujet 2	80	87
…	…	…

$$\overline{X}, S_X$$

$$\overline{Y}, S_Y$$

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

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

$$S_{Y \cdot X}$$

Corrélation ou régression

Coefficient de corrélation

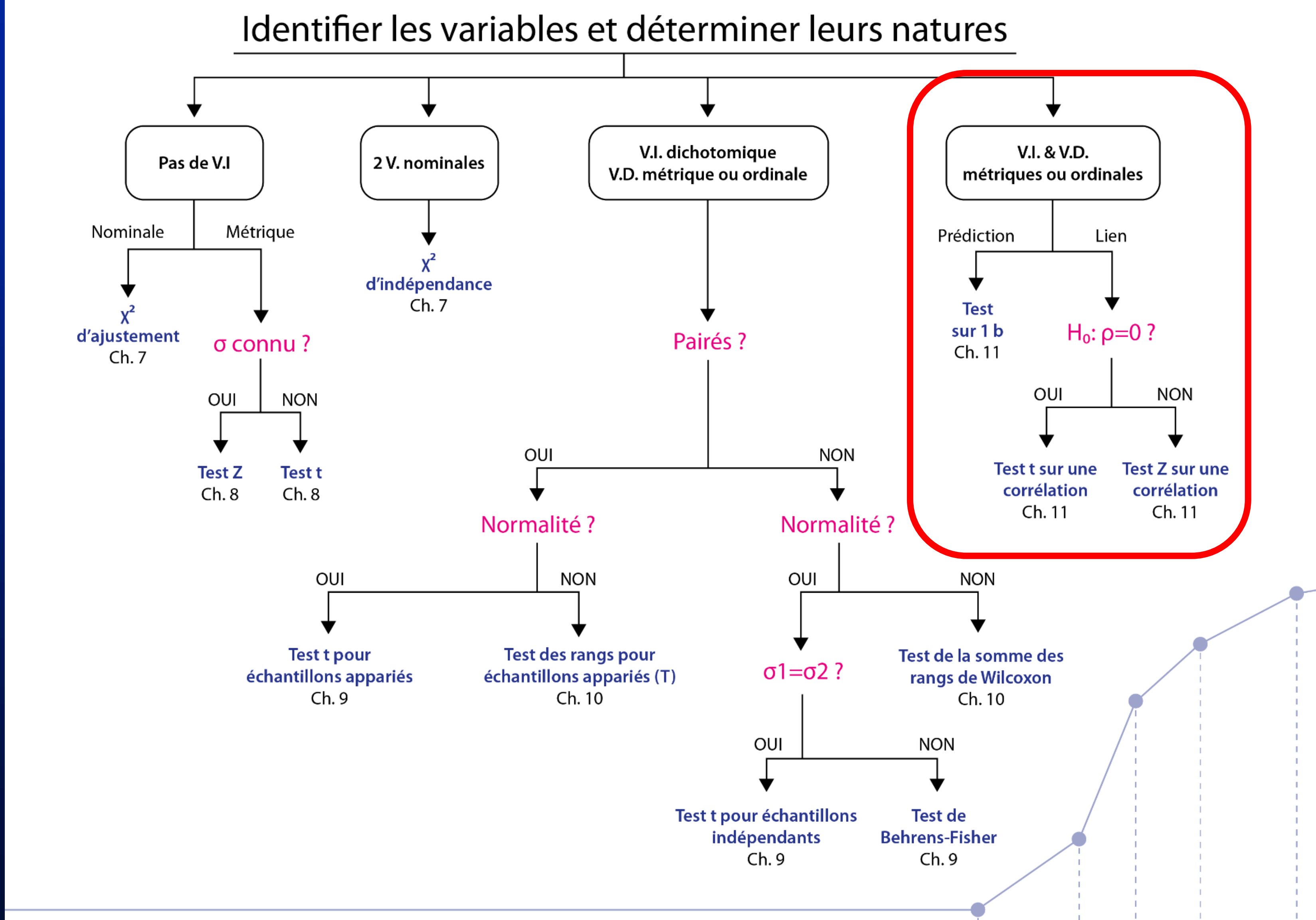
Droite de régression

= description

Quid de la population ?

Problème de l'erreur d'échantillonnage

Choisir le bon test d'hypothèse



Notation

Echantillon	Population
r	ρ (rho)
b	β
a	α

Les tests statistiques

Différents tests sur une relation

$H_0 : \rho = 0$

$H_A : \rho \neq 0$

$H_0 : \rho = 0,5$

$H_A : \rho \neq 0,5$

$H_0 : \beta = 0$

$H_A : \beta \neq 0$

$H_0 : \beta = 0,5$

$H_A : \beta \neq 0,5$

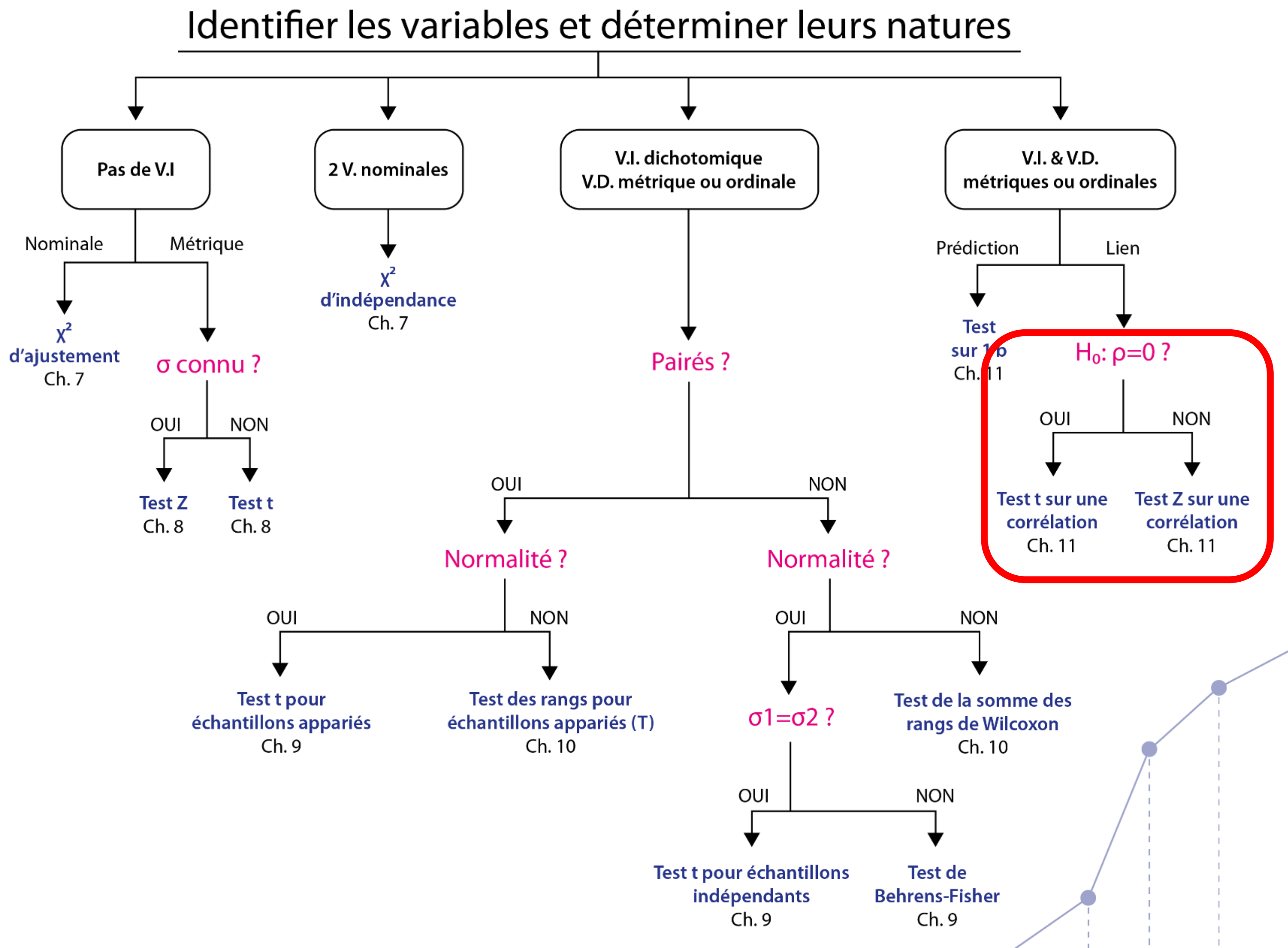
Tests sur un coefficient de corrélation

A distinguer:

$H_0 : \rho = 0$

$H_0 : \rho = 0,5$

Choisir le bon test d'hypothèse



Tester rho = 0

Pour ρ = 0, *r* est plus ou moins normalement distribué autour de zéro avec:

$$S_r = \sqrt{\frac{1 - r^2}{N - 2}}$$

Tester rho = 0

Dès lors:

$$t_{obs} = \frac{r}{S_r}$$

Avec N – 2 dl

Exemple

Y a-t-il une relation positive entre indice d’inégalité des revenus et prévalence des troubles de l’humeur ?

N = 9 **dl = 9 – 2 = 7**

r = 0,72

α = 0,05

Exemple

$H_0 : \rho = 0$

$H_A : \rho > 0$

$$S_r = \sqrt{\frac{1 - r^2}{N - 2}} = \sqrt{\frac{1 - 0{,}52}{9 - 2}} = 0{,}26$$

Exemple

$$t_{obs} = \frac{r}{S_r} = \frac{0{,}72}{0{,}26} = 2{,}77$$

$t_{0,05}$ = 1,895 (test unilatéral et 7 dl)

Puisque 2,77 > 1,895, rejeter H_0

Exemple

Il y a bien une corrélation positive significative entre indice d'inégalité des revenus et prévalence des troubles de l'humeur.

Tester rho = une valeur

$H_0 : \rho = 0,5$

$H_A : \rho < 0,5$

Pour $\rho \neq 0$, *r* n'est pas distribué normalement

Tester rho = une valeur

Solution de Fisher:

Transformer r en r′ selon:

$$r' = 0{,}5 \times \ln\left(\frac{1+r}{1-r}\right)$$

Tester rho = une valeur

r′ est normalement distribué autour de ρ′

avec :

$$S_{r'} = \frac{1}{\sqrt{N-3}}$$

Tester rho = une valeur

Dès lors:

$$Z = \frac{r' - \rho'}{\sqrt{\dfrac{1}{N-3}}}$$

Exemple

La corrélation entre indice d'inégalité des revenus et prévalence des troubles de l'humeur est-elle inférieure à 0,9 ?

N = 9 **dl = 9 – 2 = 7**

r = 0,72

α = 0,05

Exemple

$H_0 : \rho = 0,9$

$H_A : \rho < 0,9$

$r = 0,72 \longrightarrow r' = 0,908$

$\rho = 0,9 \longrightarrow \rho' = 1,472$

Exemple

r	r′	r	r′
…	…	…	…
0.690	0.848	0.890	1.422
0.695	0.858	0.895	1.447
0.700	0.867	0.900	1.472
0.705	0.877	0.905	1.499
0.710	0.887	0.910	1.528
0.715	0.897	0.915	1.557
…	…	…	…

Exemple

$$Z = \frac{r' - \rho'}{\sqrt{\frac{1}{N-3}}} = \frac{0{,}908 - 1{,}472}{\sqrt{\frac{1}{6}}} = -1{,}38$$

Exemple

$H_0 : \rho = 0{,}9$

$H_A : \rho < 0{,}9$

$P(r < 0{,}72) = P(Z < -1{,}38) =$

$P(Z > 1{,}38) = 0{,}0838$

Exemple

Puisque 0,0838 > 0,05, ne pas rejeter H_0

Nous n’avons pas suffisamment de preuves pour affirmer que la corrélation entre indice d’inégalité des revenus et prévalence des troubles de l’humeur est inférieure à 0,9.

Intervalle de confiance

Calculer un intervalle de confiance pour rho à partir de r

Exemple: Estimer la corrélation entre indice d'inégalité des revenus et prévalence des troubles de l'humeur pour l'ensemble des pays.

Intervalle de confiance

$$IC(\rho') = r' \pm Z_{\alpha/2} \frac{1}{\sqrt{N-3}}$$

$$IC(\rho') = 0{,}908 \pm 1{,}96 \frac{1}{\sqrt{6}} = 0{,}908 \pm 0{,}8$$

Intervalle de confiance

$$0{,}108 < \rho' < 1{,}708$$

Devient

$$0{,}11 < \rho < 0{,}935$$

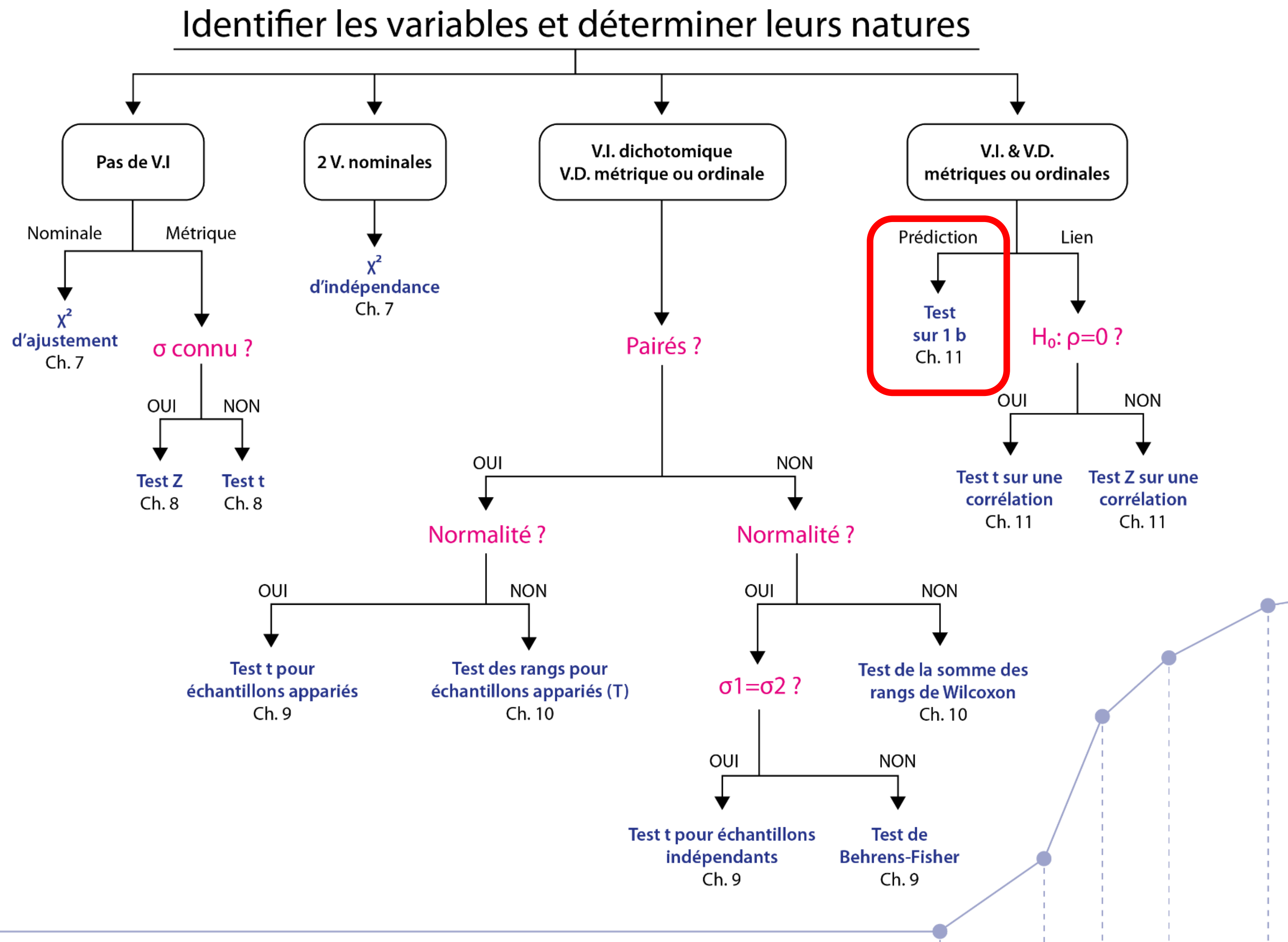
Exemple

r	r′	r	r′
…	…	…	…
0.725	0.918	0.925	1.623
0.730	0.929	0.930	1.658
0.735	0.940	0.935	1.697
0.740	0.950	0.940	1.738
0.745	0.962	0.945	1.783
0.750	0.973	0.950	1.832
…	…	…	…

Intervalle de confiance

On peut donc affirmer avec une certitude de 95% que la corrélation entre indice d'inégalité des revenus et prévalence des troubles de l'humeur dans la population générale se trouve entre 0,110 et 0,935.

Choisir le bon test d'hypothèse



Tester b

b est normalement distribué avec une erreur standard égale à :

$$S_b = \frac{S_{Y \cdot X}}{S_X \sqrt{N-1}}$$

Tester b

Dès lors:

$$t_{obs} = \frac{b - \beta}{S_b}$$

Avec N – 2 dl

Exemple

Peut-on affirmer que le taux de prévalence des troubles de l'humeur augmente de plus de 2% pour une augmentation de 0,1 unité sur l'indice d'inégalité des revenus ?

Exemple

$H_0 : \beta = 20$

$H_A : \beta > 20$

$N = 9$ $dl = 9 – 2 = 7$ $\alpha = 0,05$

$S_X = 0,0819$ $S_{Y.X} = 5,00$

$b = 59,328$

Exemple

$$S_b = \frac{S_{Y.X}}{S_X\sqrt{N-1}} = \frac{5}{0{,}0819\sqrt{8}} = 21{,}58$$

$$t_{obs} = \frac{b-\beta}{S_b} = \frac{59{,}328 - \mathbf{20}}{21{,}58} = \mathbf{1,82}$$

Exemple

$t_{obs} = 1{,}82$

$t_{0{,}05} = 1{,}895$ (test unilatéral et 7 dl)

Puisque 1,82 < 1,895, ne pas rejeter H_0

Exemple

Nous n’avons pas de preuves suffisantes pour affirmer que le taux de prévalence des troubles de l’humeur augmente de plus de 2% pour une augmentation de 0,1 unité sur l’indice d’inégalité des revenus.

Intervalle de confiance

Calculer un intervalle de confiance pour β à partir de b

Exemple: Estimer la pente de la régression entre indice d'inégalité des revenus et prévalence des troubles de l'humeur.

Intervalle de confiance

$$IC_{0,95} = b \pm t_{\alpha/2} S_b$$

$$IC_{0,95} = 59,328 \pm (2,365 \times 21,58)$$

$$= 59,328 \pm 51,037$$

Intervalle de confiance

$$8{,}291 < \beta < 110{,}365$$

Pour une augmentation de 0,1 point sur l'échelle d'inégalité des revenus, le taux de prévalence des troubles de l'humeur augmente entre 8 et 11%.

Conditions d’application

Condition 1 (rappel)

Linéarité de la régression

La relation entre X et Y est linéaire

Conditions d'application

<u>Autres conditions</u>

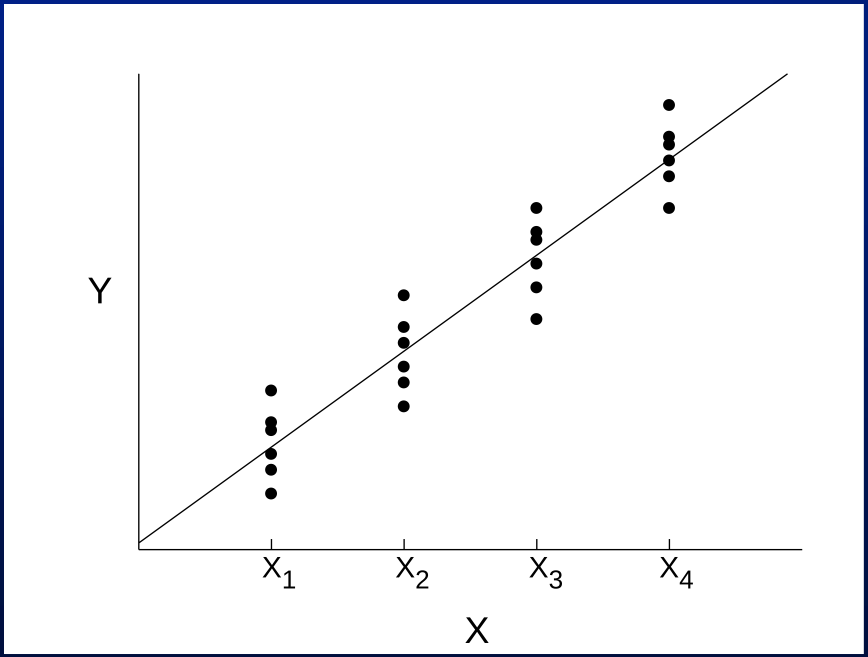
Uniquement pour les tests d'hypothèse

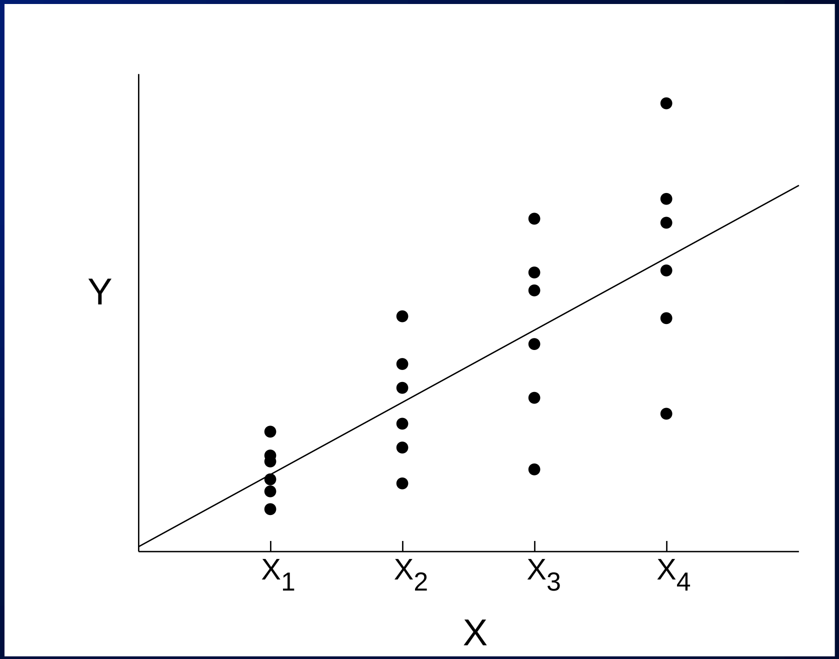
Conditions d'application de la régression

1. Homogénéité des variances

2. Normalité dans les vecteurs

Conditions d’application de la régression





Conditions d'application de la corrélation

<u>Distribution normale bivariée</u>:

- X et Y sont normalement distribués
- Pour tout X, les Y sont normalement distribués
- Pour tout Y, les X sont normalement distribués

En cas de violation des conditions

1. Tests très robustes

2. Tests non paramétriques sur les rangs