df =

X	Y	Z
4	5	
6	8	
-2	-4	
.	.	
.	.	
.	.	

données numériques

X, Y

X = Temp

Y = Pression

X, Y

sont elles

corrélées ?

X, Y sont ≥ 0 corrélées si
les grandes valeurs de X vont avec
les grdes val de Y
ET les petites — X
les petites — Y

X	Y
100	235
-1000	-5000

X, Y sont ≤ 0 corrélées si
les grdes val de X vont avec les petites val de Y

X	Y
1	4
2	5
3	6

Produit Scalaire

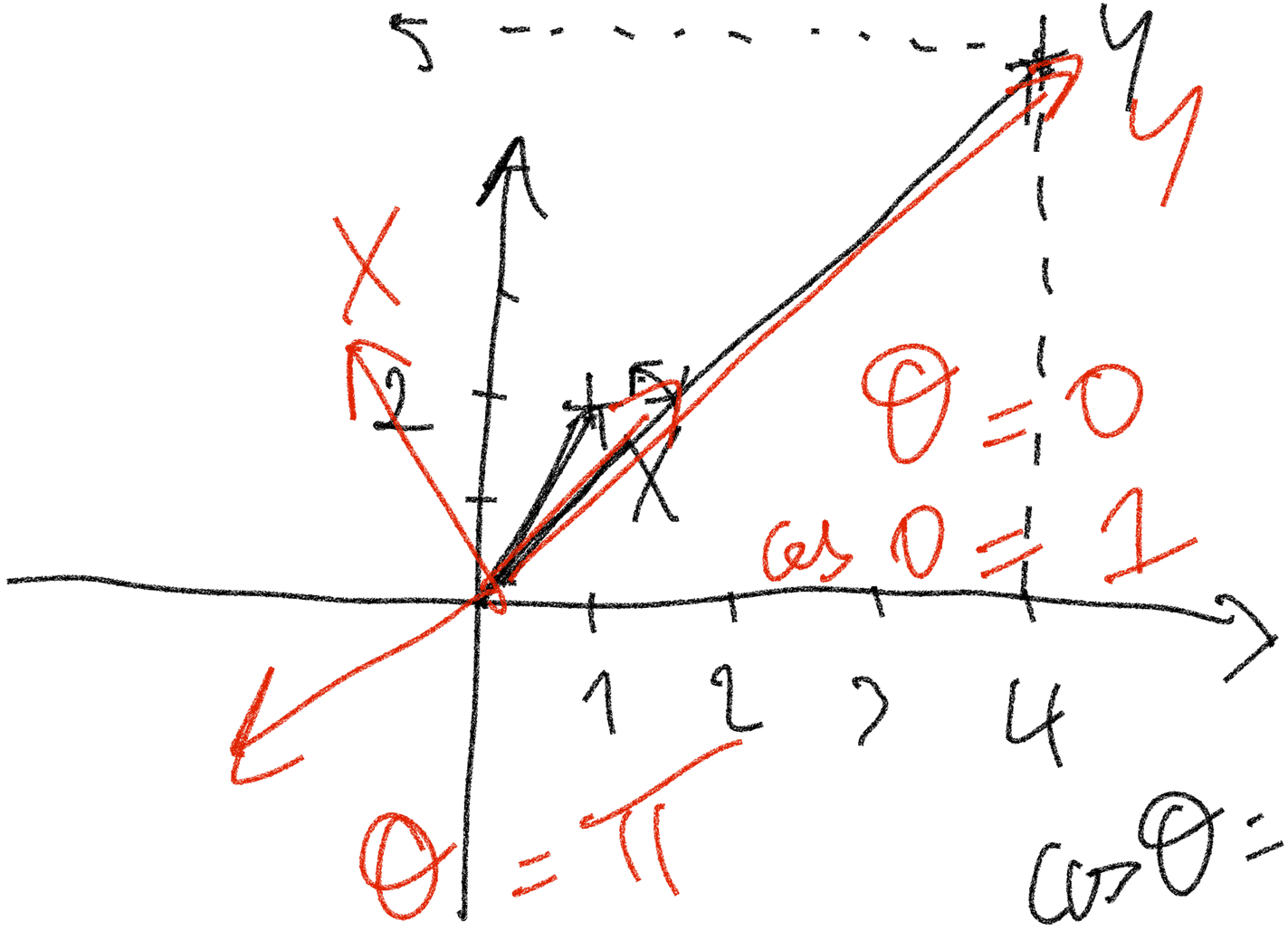
$$\langle x | y \rangle = 1 \times 4 + 2 \times 5 + 3 \times 6 = 32$$

$$\langle x | y \rangle < 32$$

-1 et 1

$$\|x\| \|y\| \quad -1 < S < 1$$

$$S = 0$$



x	y
1	4
2	5

$\theta = 0$

$\cos 0 = 1$

$\theta = \pi$

$\cos\theta = -1$

$\cos\theta = \dfrac{\langle x|y\rangle}{\|x\|\|y\|} \in [-1,1]$

$\cos\theta = 0$

x	y
1	4
2	5
3	6

$$\frac{\langle x|y \rangle}{\|x\| \cdot \|y\|}$$

$$\|x\| = \sqrt{1^2 + 2^2 + 3^2}$$

$$\|y\| = \sqrt{4^2 + 5^2 + 6^2}$$

$$\frac{1}{3}(1 + 2 + 3) = \frac{6}{3} = 2$$

4
6
7

Standardisation

$$X \longrightarrow \frac{X - \text{moyenne}(X)}{\text{écart-type}(X)}$$

X'	Y'
0	0

moyenne (X') / (Y') = 0

écart-type = 1

$$\left(\frac{\sum_{i=1}^{N} (X_i - \bar{X}) \times (Y_i - \bar{Y})}{\sqrt{\frac{1}{N-1}\sum X_i^2 - \bar{X}^2} \; \sqrt{\frac{1}{N-1}\sum Y_i^2 - \bar{Y}^2}} \right) \times \frac{1}{N-1}$$

X	Y

N

Score de corrélation de PEARSON

X	Y
1	4
2	5
3	6

$$\bar{x} = \frac{1+2+3}{3} = \frac{6}{3} = 2$$

$$\bar{y} = \frac{4+5+6}{3} = 5$$

$\frac{1}{N}$ // $\frac{1}{N-1}$

$$\sigma_x = \sqrt{Var_x} = \sqrt{\frac{(1-2)^2+(2-2)^2+(3-2)^2}{2}} = 1$$

$$\sigma_y = 1$$

x'	y'
-1	-1
0	0
1	1

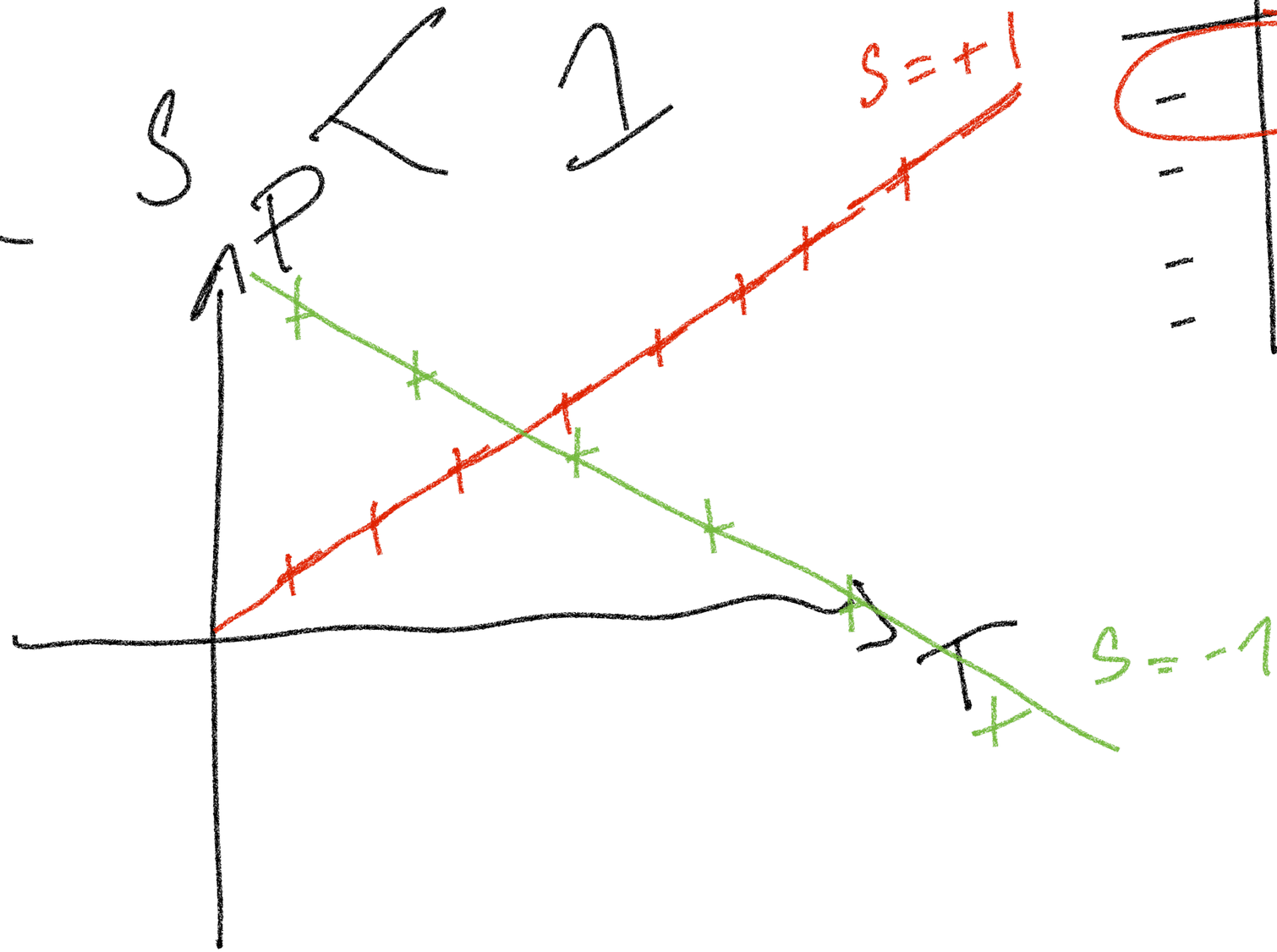
$$x \leftarrow \frac{x - \bar{x}}{\sigma_x}$$

$$\langle x' | y' \rangle = (-1) \times (-1) + 0 \times 0 + 1 \times 1 = 2$$

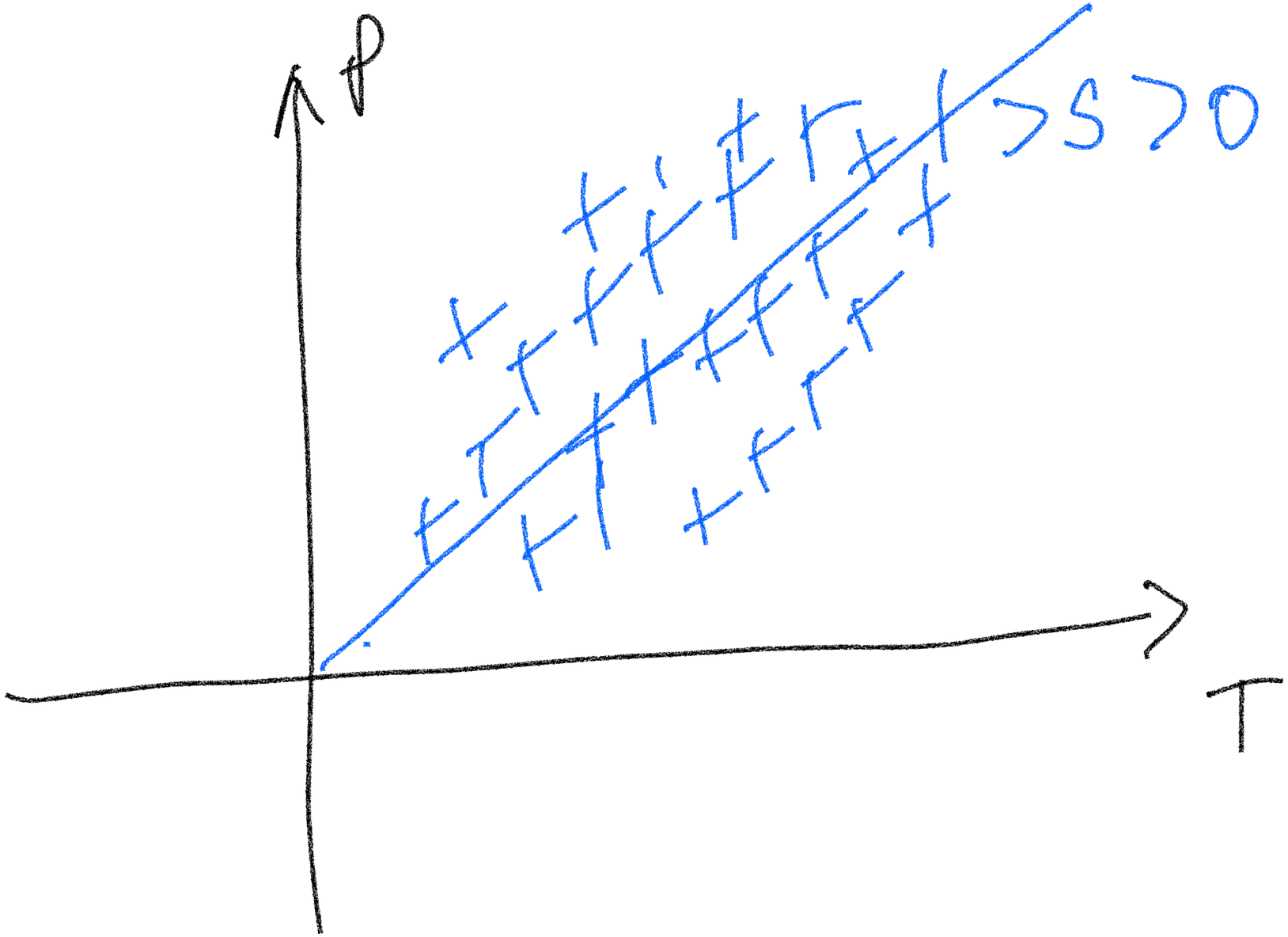
$$s_{Pearson} = \frac{2}{2} = 1$$

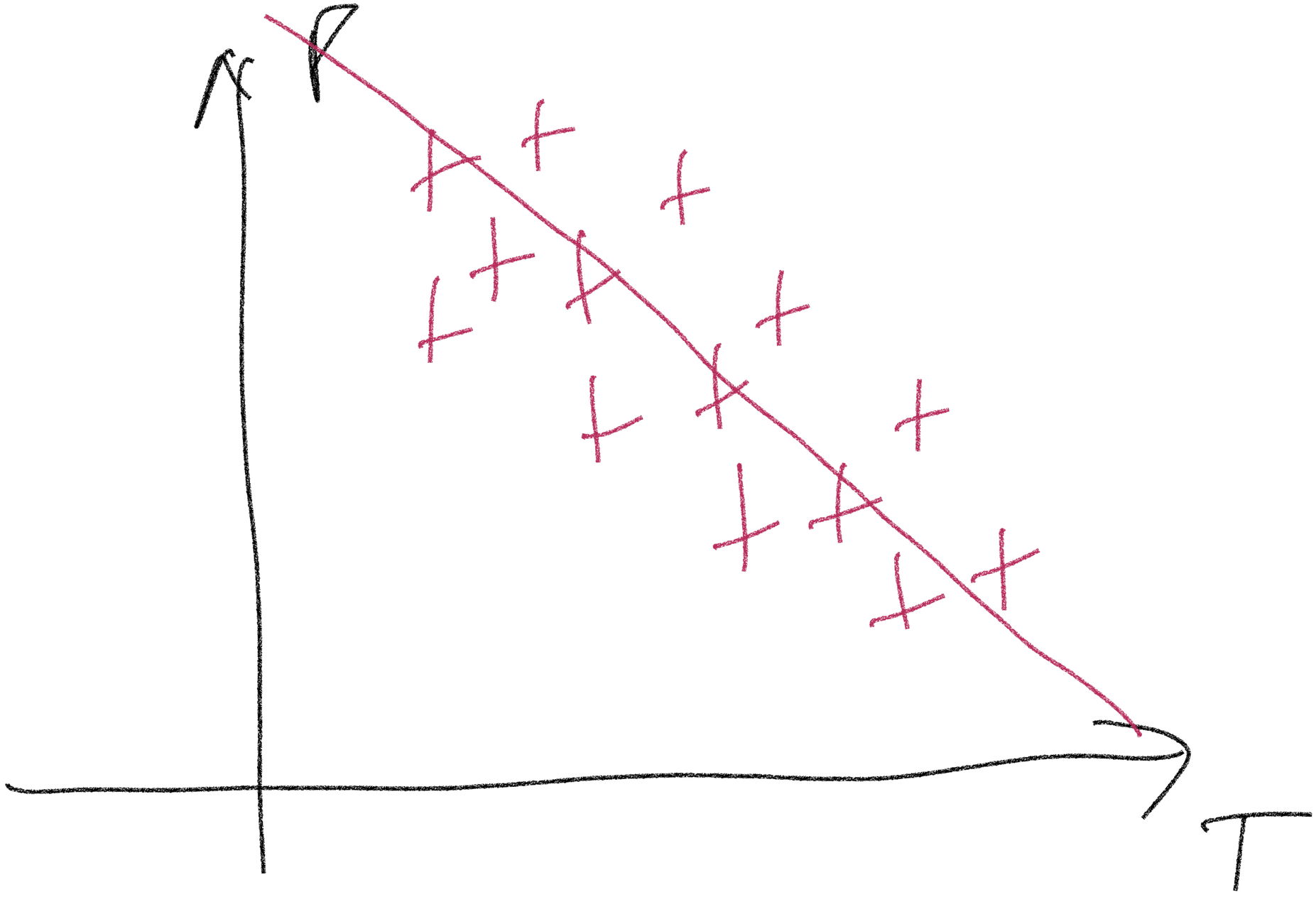
x, y sont parfaitement corrélés

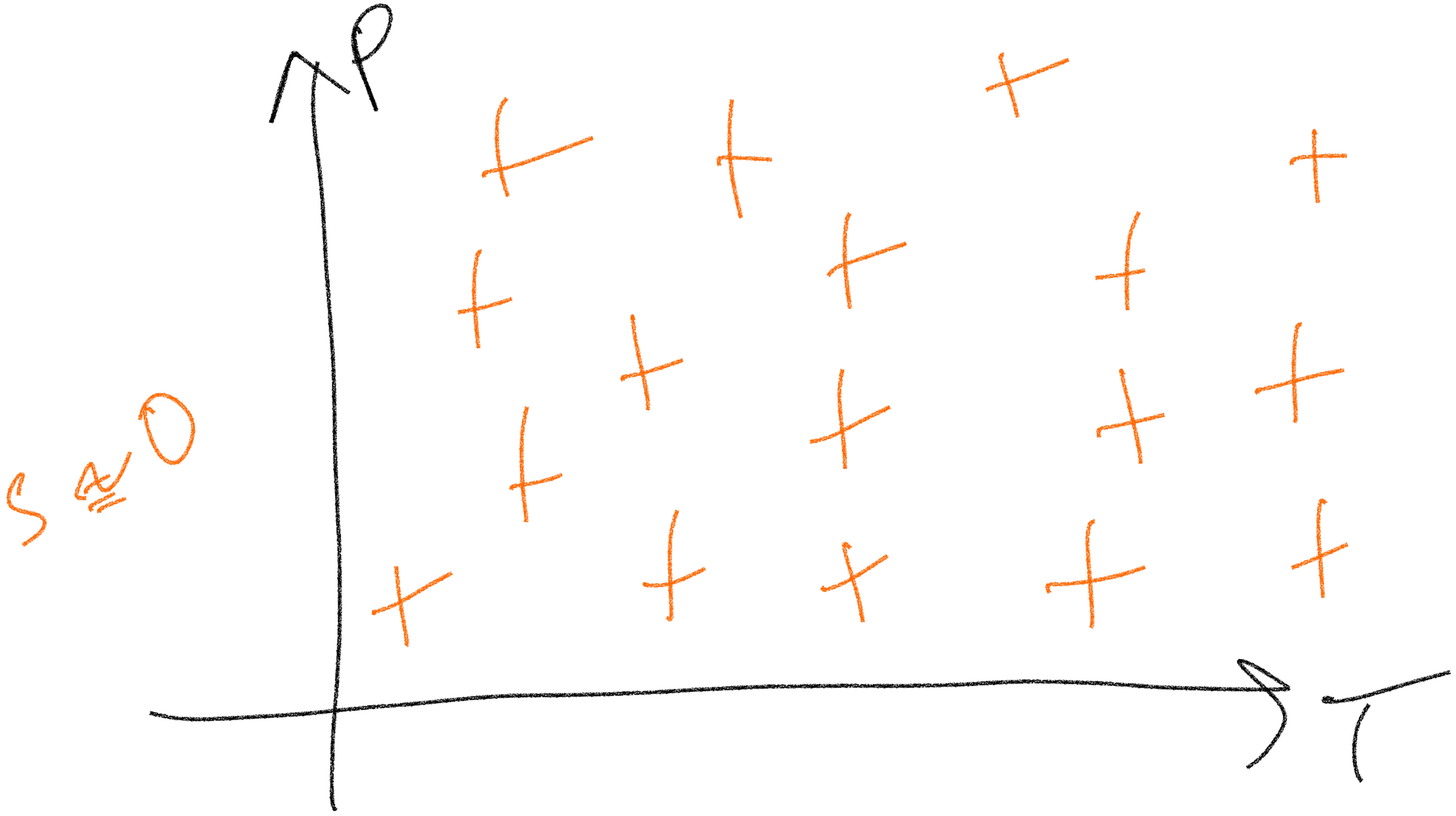
$-1 < S < 1$



T	P
-	-
-	-
-	-
-	-
-	-


p
T
S > 0


P
T


P
τ
S ≤ 0

X, Y relation non linéaire

$$X = \alpha Y + \beta$$

$$X = Y^2$$

$$X = Y^3$$

$$\vdots$$

$S \geq 0$

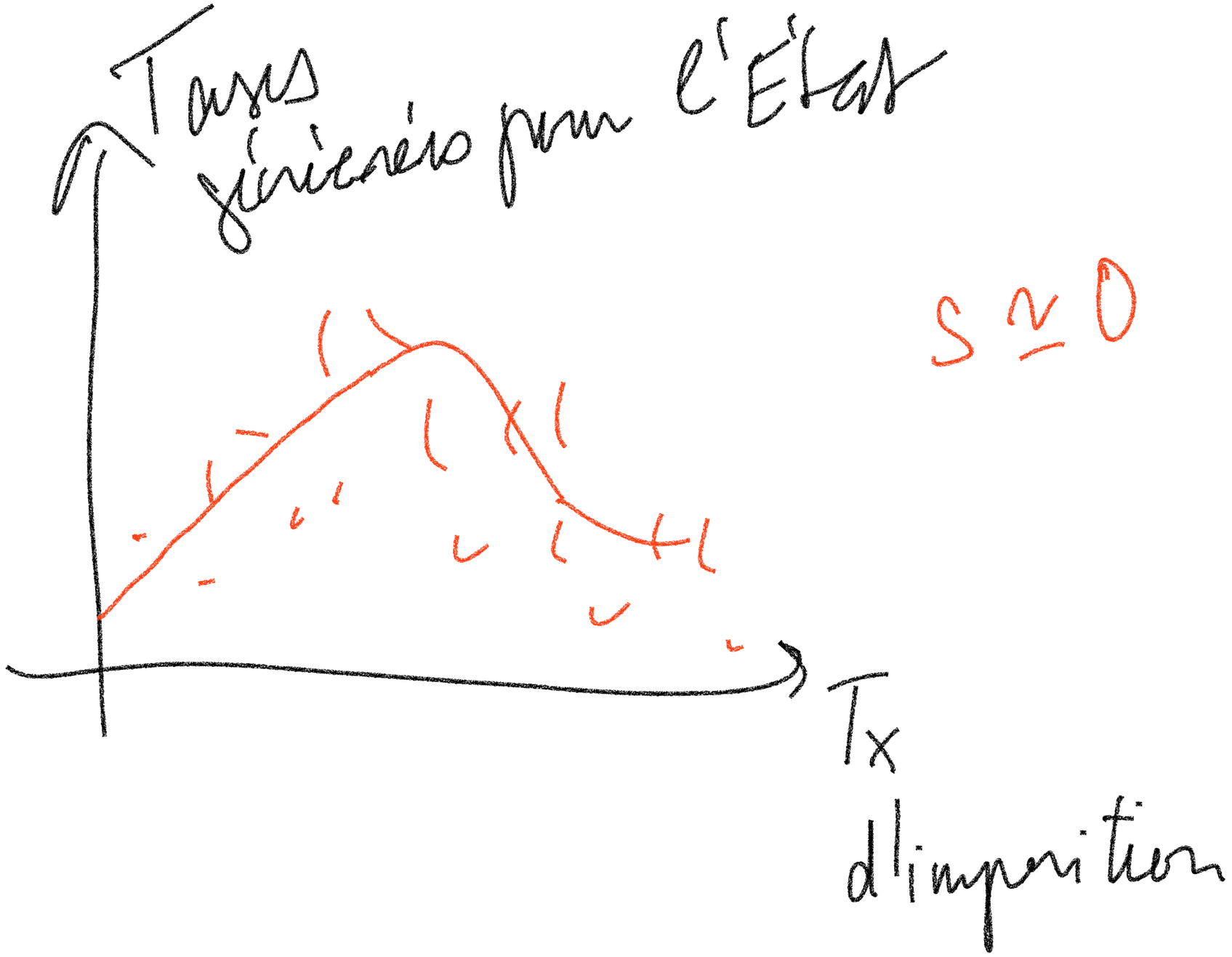

Taux
prélèvements pour l'État
S ≥ 0
Tx
d'imposition

Table de corrélation

df = | A | B | C |

moyenne
écart-type, etc.

df.corr()

	A	B	C
A	1,0	0,3	-0,7
B	0,3	1,0	0,1
C	-0,7	0,1	1,0

Spearman

Kendall