Modélisation statistique pour la Biologie (M1 bio-stat)

(Notes auxiliaires de cours)

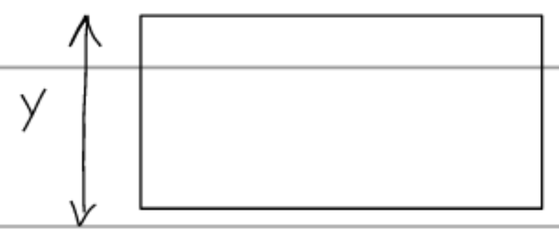
CHAPITRE I : Corrélation

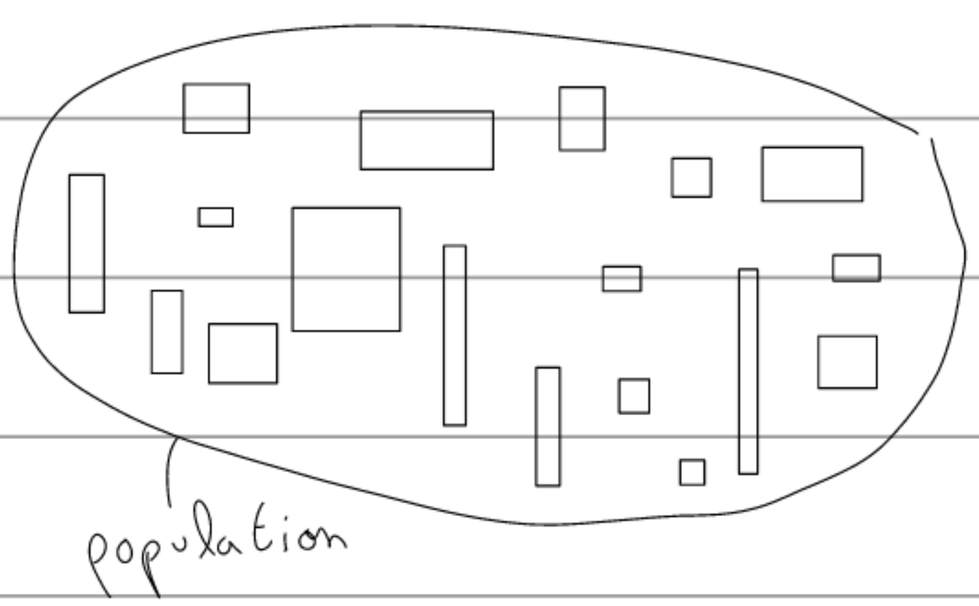
On a une POPULATION dont on considère 2 traits

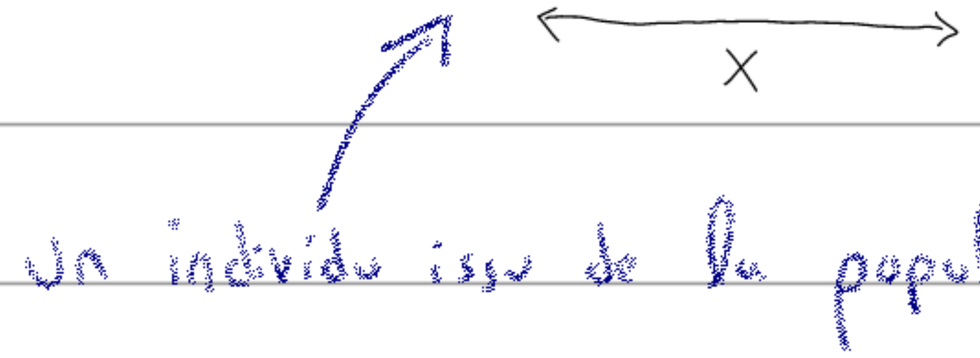
↳ eg : Population : une famille de Rectangles

Trait X : largeur

Trait Y : hauteur







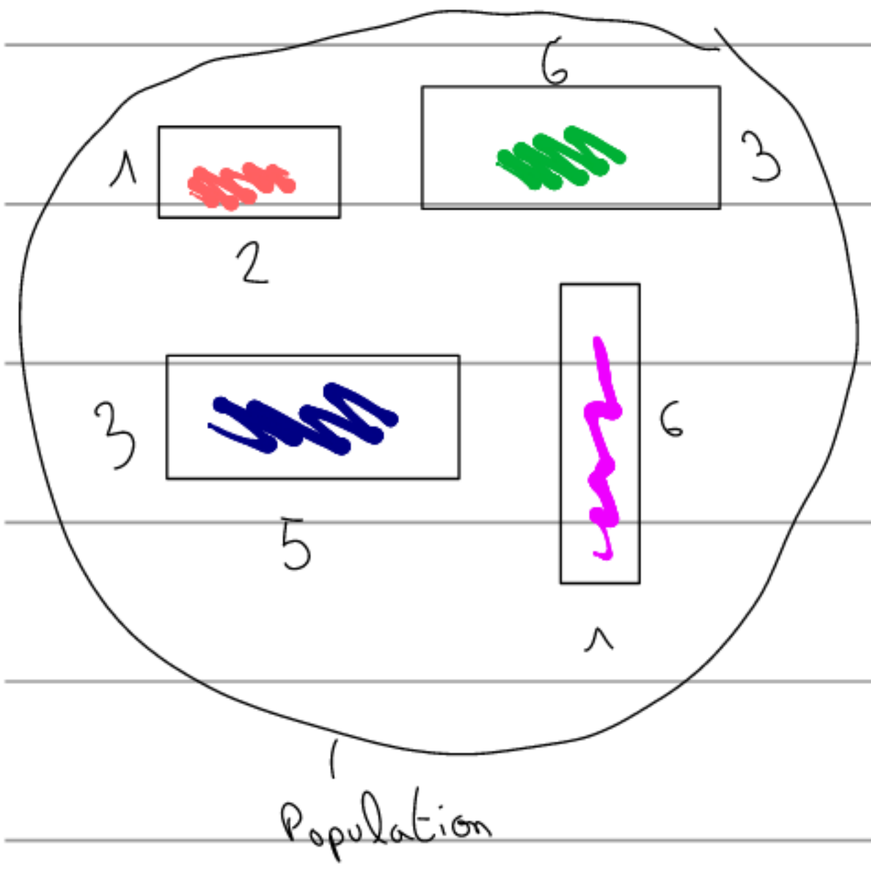
On regarde comment la valeur de X impacte la valeur de Y

↳ Est-ce que l'augmentation de X provoque celle de Y ?

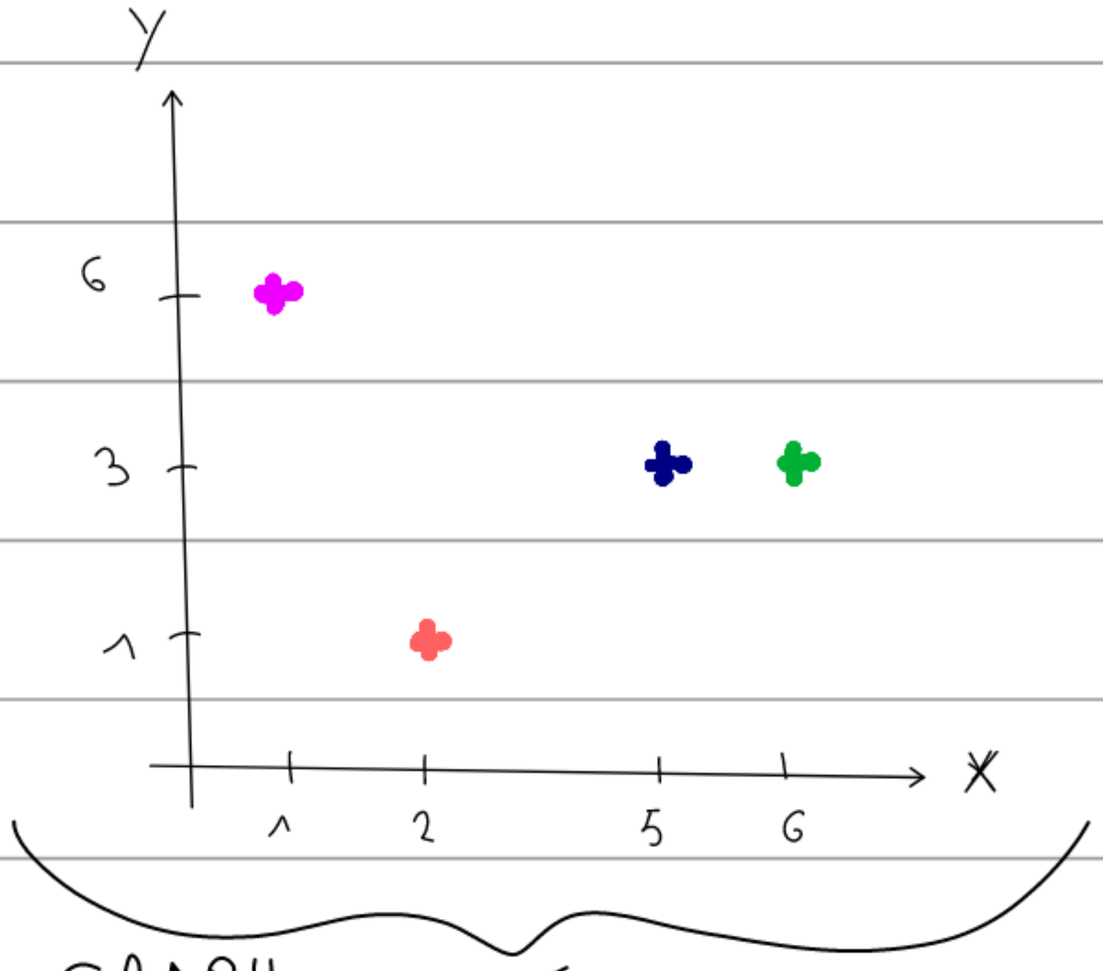
↳ Est-ce que l'augmentation de X provoque la diminution de Y ?

↳ ici X et Y sont quantitatives, ie. elles prennent des valeurs chiffrées.

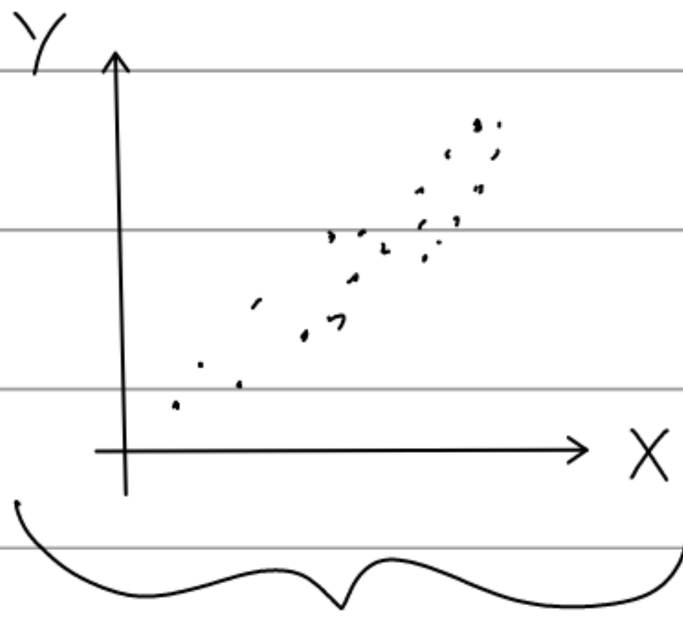
↳ On peut se rendre compte de la corréla° entre X et Y en traçant un graphique comme celui-ci :



Population

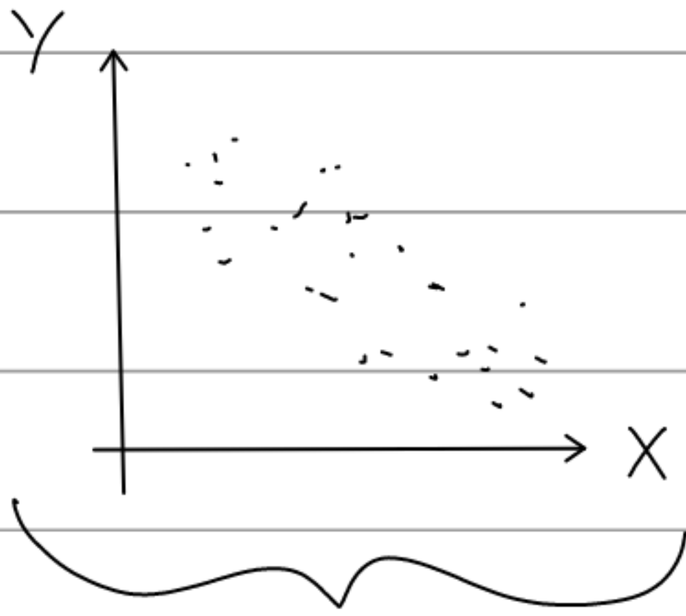


GRAPH associé.



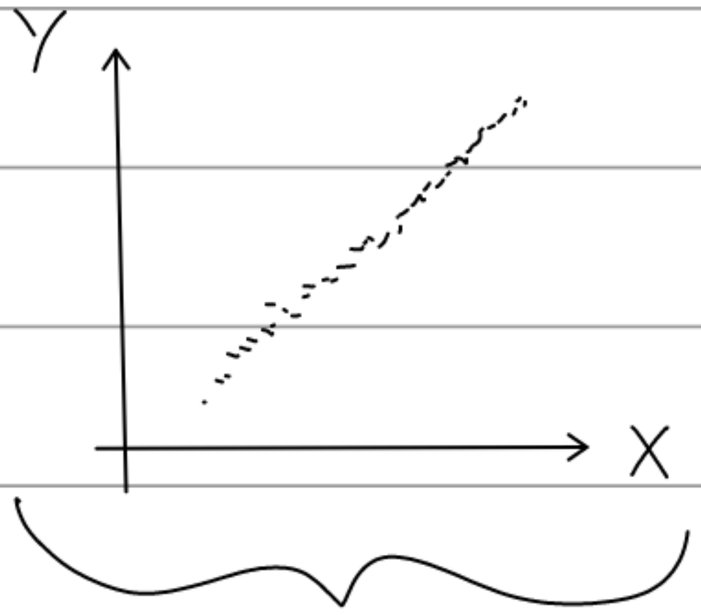
CORRélation POSITIVE

(ça monte)



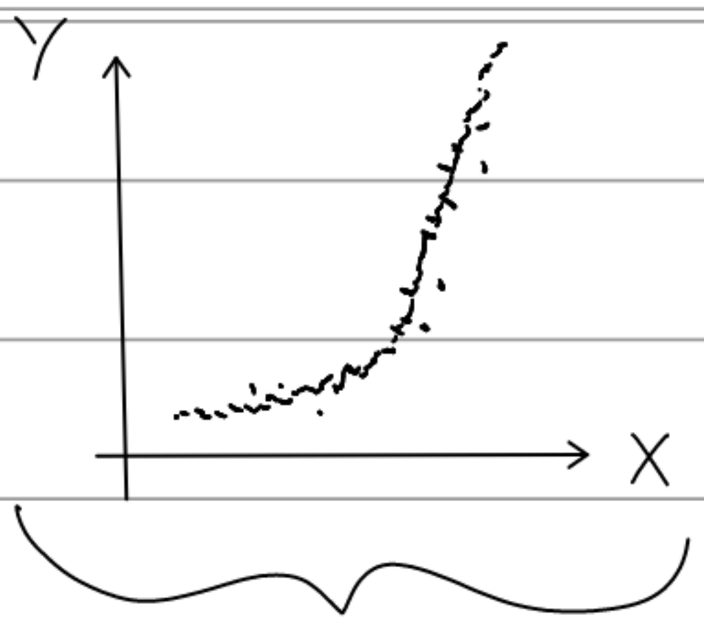
CORRélation NÉGATIVE

(ça descend)



Corrélation LINÉAIRE

(C'est droit)



Corrélation NON-LINÉAIRE

~~(C'est droit)~~

⚠ "X et Y sont corrélés" n'est pas la même chose que "il y a une relation cause-effet entre X et Y"

↳ des fois la corrélation apparaît par hasard ou est due à une variable sous-jacente inhérente à X et Y.

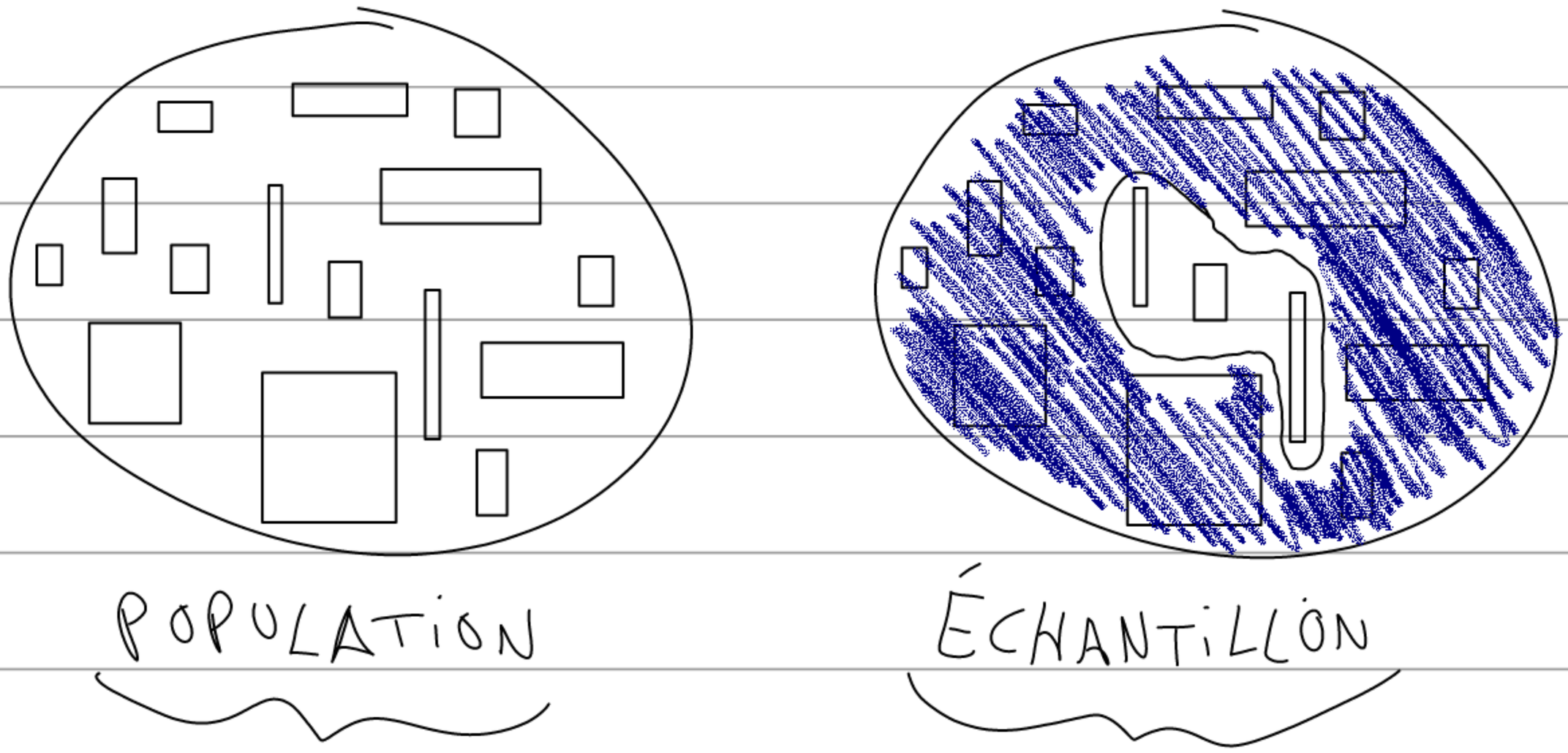
X = altitude / Y = température (Relation cause-effet)

X = Taille des enfants de 1989 / Y = Puissance de calcul des ordis (hasard)

X = nb de glaces vendues / Y = nb de Noyades

Variable sous-jacente : Z = Fréquentation de la plage

Malheureusement, en stat, on n'a pas toute la population à disposition mais seulement un échantillon.

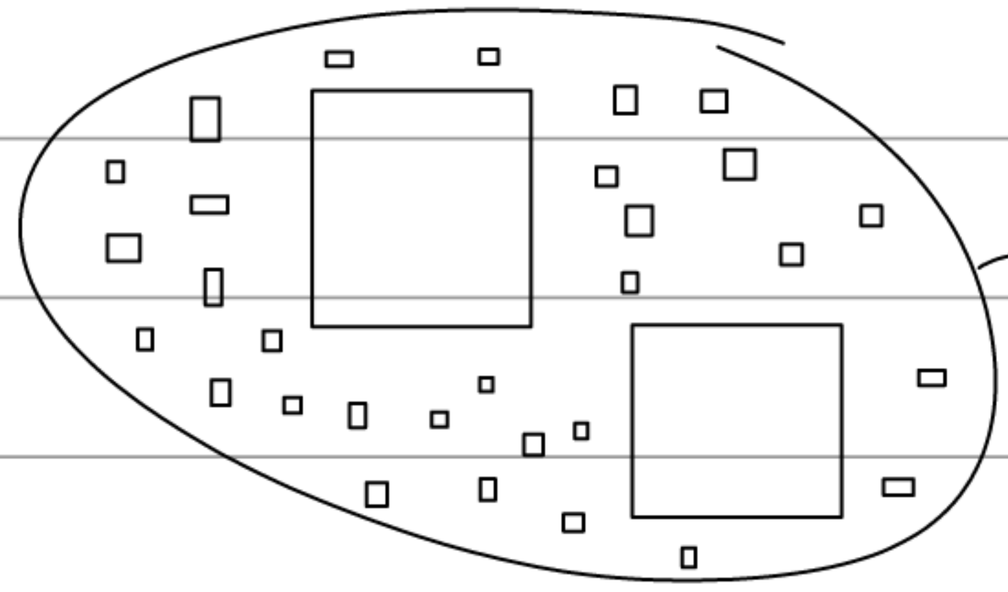


On cherche à décrire au mieux la popula° avec l'échantillon dont on dispose.

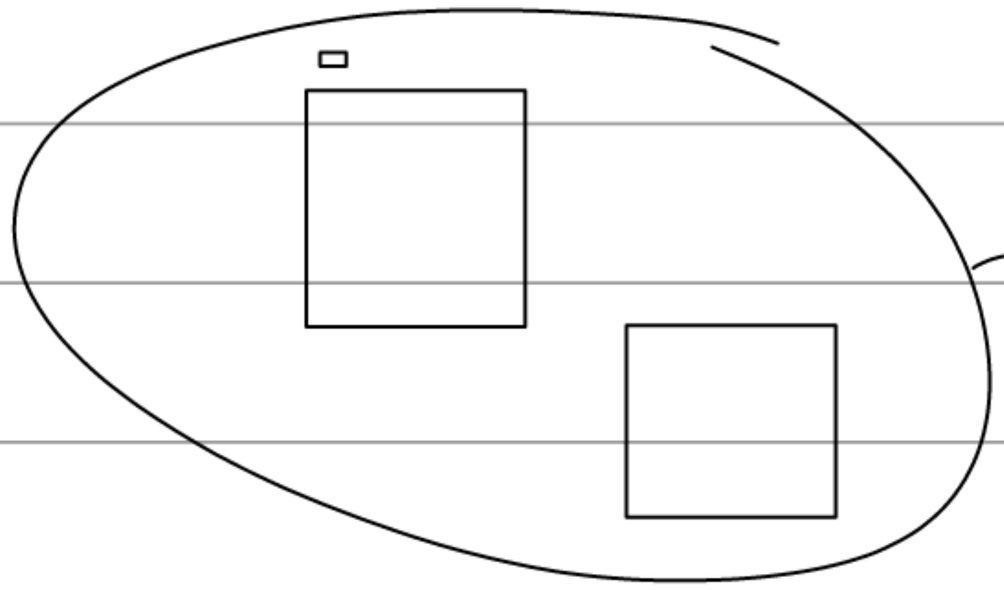
↳ Plus l'échantillon est grand, mieux on décrit la popula°.

↳ Parfois le comportement de certains individus est radicalement différent de celui des autres. Cela peut entraîner une mauvaise analyse si l'échantillon est trop petit :

Exemple

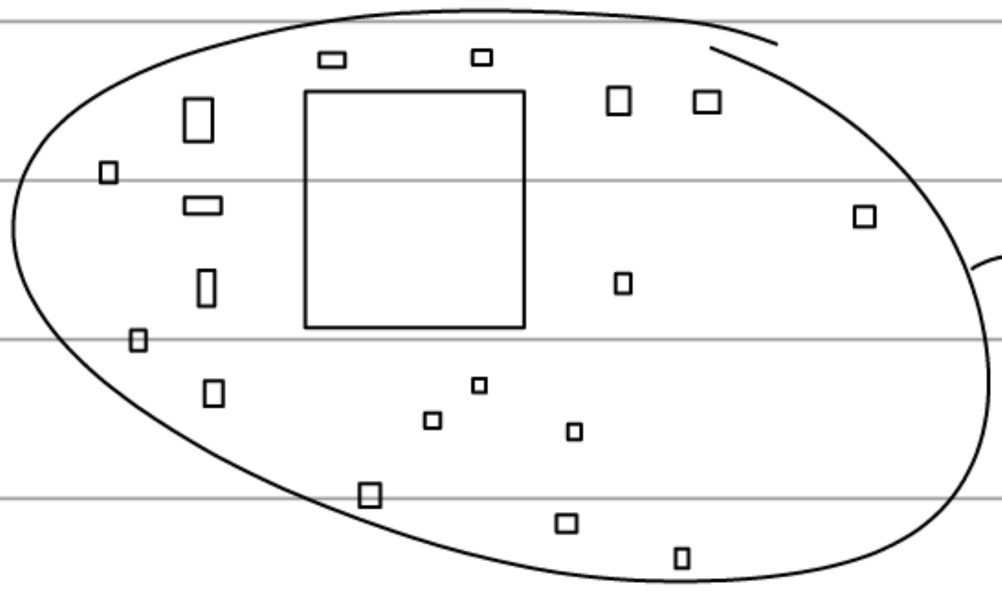


- Population



- Échantillon 1 (mauvais)

↳ conclusion : c'est une population de grands rectangles



- Échantillon 2 (meilleur)

↳ conclusion : c'est une population de petits rectangles

Note : Pour distinguer les valeurs relatives à l'échantillon, on ajoute un chapeau au dessus des notations :

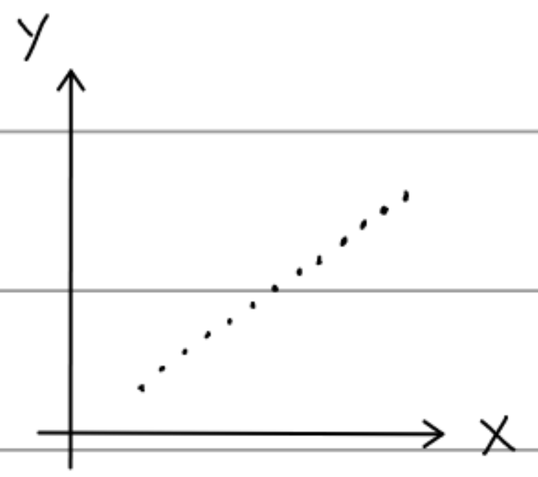
ρ RHO "population"

$\hat{\rho}$ RHO "échantillon"

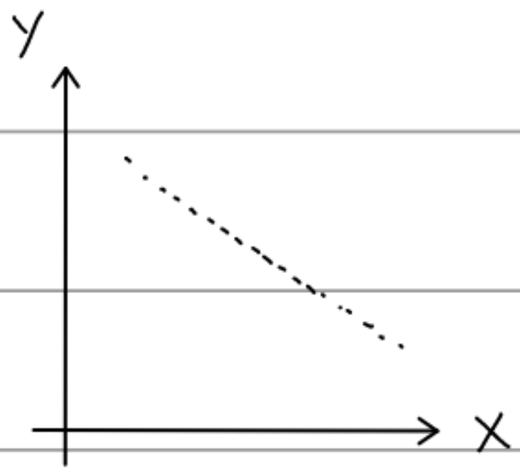
Pour mesurer la corrélation linéaire, on dispose du coefficient de corrélation linéaire :

$$\widehat{\rho_{(X,Y)}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

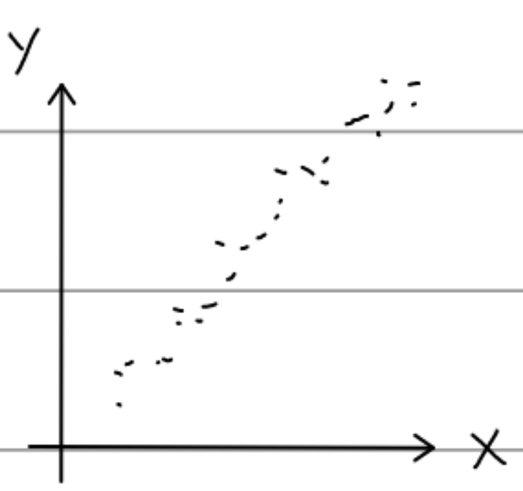
- $-1 \leq \rho(X,Y) \leq 1$
- $\rho > 0$ corrélation positive linéaire
- $\rho < 0$ corrélation négative linéaire
- $\rho = 1$ corrélation positive parfaite (points alignés)
- $\rho = -1$ corrélation négative parfaite (points alignés)
- $\rho = 0$ pas de corrélation linéaire
  ou bien $\approx$



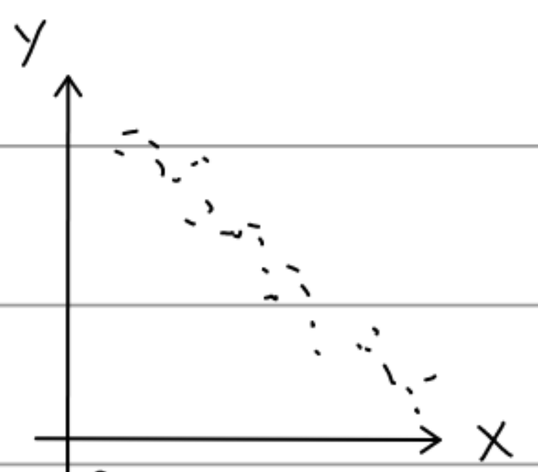
$\widehat{\rho} = 1$



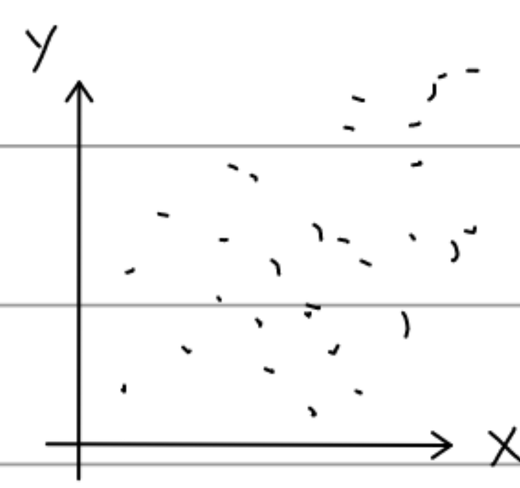
$\widehat{\rho} = -1$



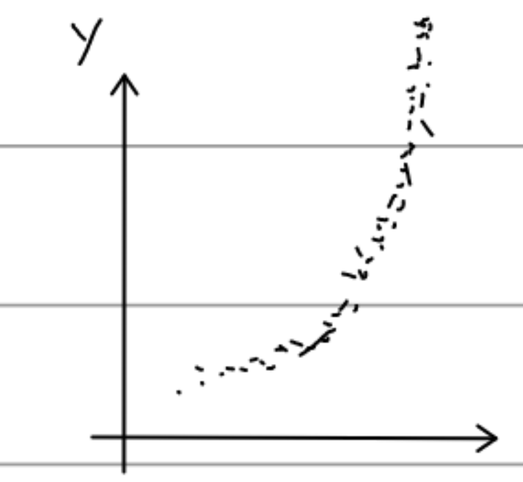
$\widehat{\rho} = 0,9$



$\widehat{\rho} = -0,9$



$\widehat{\rho} = 0,1$



$\widehat{\rho} = 0,1$

Mais corrélation non-linéaire positive