Objectifs d'apprentissage - A la fin de ce chapitre, je sais :
- étudier des séries statistiques à deux variables ☐
- lire et interpréter un nuage de points ☐
- discuter de la corrélation entre deux variables à l'aide du coefficient de corrélation linéaire et de la droite de régression ☐

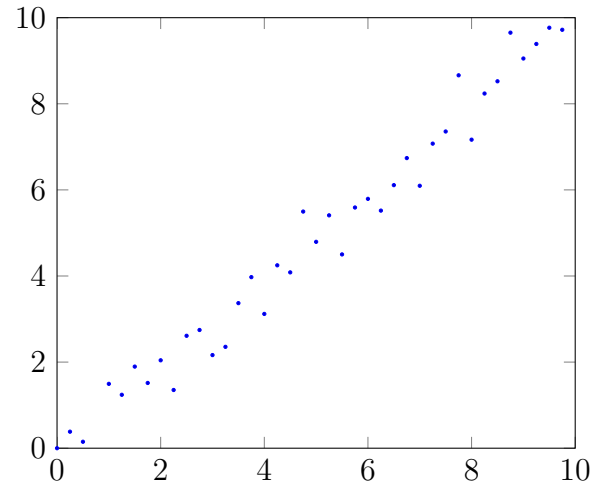
On s'intéresse dans ce chapitre à l'étude conjointe de deux séries statistiques que l'on note $x = (x_1, \dots, x_n)$ et $y = (y_1, \dots, y_n)$. On étudie alors les couples $[(x_1, y_1), \dots, (xn, yn)]$ que l'on appelle observations dans le cas de **statistiques bivariées**. Comme en première année, les statistiques sont le pendant « empirique », des notions vues en probabilité. D'où le qualificatif « empirique » addossé parfois à certains indicateurs : moyenne, variance, covariance...

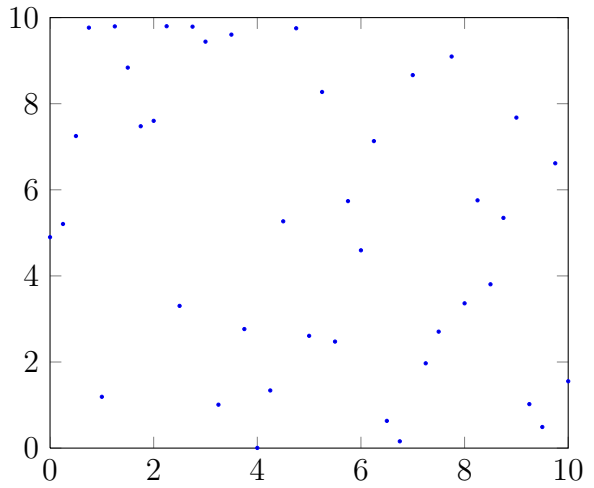
1 Nuage de points

Définition : on appelle **nuage de points** associé à la série statistique (x, y) l'ensemble des points M_k de coordonnées (x_k, y_k) (pour $k \in [\![1, n]\!]$) tracés dans un repère du plan.
Le **point moyen** du nuage est le point de coordonnées $(\overline{x}, \overline{y})$, où $\overline{x}$ désigne la moyenne (empirique) de la série x et $\overline{y}$ celle de la série y

Rappel : avec Python, si `x` et `y` contiennent deux listes de valeurs de même taille (par exemple deux séries statistiques), la commande : `plt.plot(x,y,'.')` (puis plt.show()) renvoie le graphique des points (x_i, y_i), autrement dit le **nuage de points.**

Exemples de nuages de points :





L'objectif principal du chapitre est de déterminer si une corrélation existe entre les deux variables x et y. Sur le premier nuage, cela semble être le cas car les variations de y semblent suivre celle de x avec une certaine régularité. Sur le second nuage, cela ne semble pas être le cas.

2 Covariance, coefficient de corrélation linéaire

Définition : on appelle **covariance** (empirique) de x et y et l'on note s_{xy} :

$$s_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) = \overline{(x - \overline{x})(y - \overline{y})}$$

Propriété - formule de Kœnig Huygens : $\boxed{s_{xy} = \overline{xy} - \overline{x} \times \overline{y}}$

Définition : on appelle **coefficient de corrélation linéaire** (empirique) de x et y et l'on note r_{xy} la valeur $\boxed{r_{xy} = \dfrac{s_{xy}}{s_x s_y}}$ (où s_x et s_y sont les écarts-type de x et y)

Propriétés : $\boxed{-1 \leqslant r_{xy} \leqslant 1}$ et $\boxed{|r_{xy}| = 1 \Leftrightarrow \exists (a,b) \in \mathbb{R}^2, y = ax + b}$

Remarque : la corrélation illustrée par ce coefficient est la possibilité d'établir une relation affine entre x et y donc plus ce coefficient est proche de 1 en valeur absolue, plus la corrélation (i.e. la relation affine) est forte.

3 Régression linéaire

L'objectif ici est de poursuivre l'étude de la corrélation affine entre x et y et de trouver la « meilleure droite » $y = ax + b$
Une méthode dite « **des moindres carrés** » est de chercher à minimiser le carré des écarts entre les y_i et les $ax_i + b$, soit trouver a et b qui minimise $d(a,b) = \sum_{k=1}^{n} (y_i - (ax_i + b))^2$

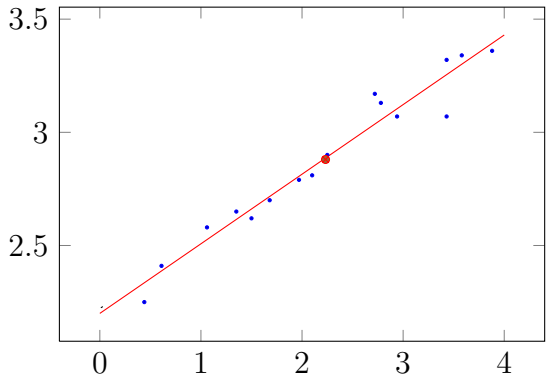
Définition et propriétés : on appelle **droite de régression linéaire** la droite

$$\boxed{y = ax + b \text{ où } a = \frac{s_{xy}}{s_x^2} \text{ et } b = \overline{y} - a \times \overline{x}}$$

cette droite, i.e. les valeurs de a et b, est celle qui minimise $d(a,b)$ et le point moyen du nuage appartient à cette droite.

Remarque : une régression linéaire fournit toujours une droite, cela ne veut pas dire qu'elle représente réellement un lien entre x et y. On considèrera que la corrélation est bonne si $|r_{xy}| \geqslant 0,85$
On cherche parfois un lien de causalité entre deux variables et on parle alors de variable à expliquer et de variable explicative.
Nous verrons cela principalement en informatique.

Un exemple de droite de régression linéaire (ci-contre) avec le point moyen. Dans ce cas, $r_{xy} \geqslant 0,97$



Transformation pour se ramener à un cas linéaire

On voit sur cet exemple que la modélisation affine :
$y = ax + b$ n'est pas toujours pertinente.
Par contre dans certains cas, un autre lien peut être mis en évidence, ici on cherchera une relation du type $y = \ln(x)$ et pour cela on pourra chercher une corrélation classique entre y et $z = \ln(x)$ (voir en informatique).