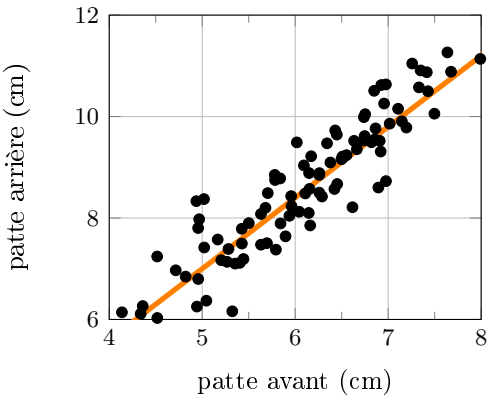
Statistiques à deux variables

2 variables par individu – Nuage de points

Exemple : On étudie une population de grenouilles. Pour chaque individu, on relève la taille d'une patte avant et la taille d'une patte arrière.

Chaque individu est associé à **deux valeurs**.

La paire de valeurs peut être interprétées comme des coordonnées $(x\,;\,y)$. Alors chaque individu est représenté par un point.
La population forme alors un **nuage de points**.



Point moyen

Si les données sont x_i et y_i alors G, centre du nuage, a les coordonnées $(\overline{x}\,;\,\overline{y})$.

Ajustement – modèle

Les points sont répartis selon un nuage de points de coordonnées $(x\,;\,y)$. On se demande s'il y a un lien entre x et y.

- On peut le supposer pour des raisons théoriques. On a produit un raisonnement qui nous amène à poser ce lien.
- On peut simplement constater que, d'après les expériences, une formule semble relier x et y sans qu'on sache le justifier théoriquement.

Ce lien peut prendre la forme $y = f(x)$. C'est un **modèle**. La fonction f peut être ajustée pour s'adapter aux résultats de mesure. Par exemple :

$$y = a \cdot x + b \quad \text{ou} \quad y = b \cdot \exp(-a \cdot x)$$

Ces deux exemples ont des paramètres a et b dont il faut choisir la valeur pour coller au mieux aux résultats observés.

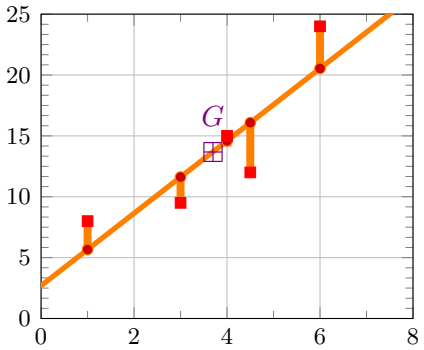
Un modèle ne doit pas avoir trop de réglages. Fermi disait : « *Avec quatre paramètres, je peux dessiner un éléphant, et avec cinq je lui fais agiter sa trompe* ». Cela veut dire que si un modèle a trop de paramètres de réglage, on arrivera toujours à l'ajuster à n'importe quel nuage de points et donc ce modèle, trop souple, n'a aucune signification.

Ajustement affine

Si les points sont à peu près alignés, on peut proposer un modèle affine $y = ax + b$.

Les points ne sont certainement pas exactement alignés. On cherche donc une droite passant au plus près des points. Il nous faut un critère pour ce *au plus près*.

On considère les écarts verticaux $\varepsilon_i = y_i - \hat{y}_i$ où $\hat{y}_i = a \cdot x_i + b$ est la valeur prédite par le modèle pour $x = x_i$. On cherche à minimiser le total $\sum_i \varepsilon_i^2$.



Pour un nuage $(x_i\,;\,y_i)$, le calcul de a et b se fait avec un machine !

On sait que G est sur la droite d'ajustement et : $a = \dfrac{cov(x,y)}{cov(x,x)}$ $\quad b = \overline{y} - a \cdot \overline{x}$

- $cov(x,y)$ est la covariance entre les x_i et y_i : $cov(x,y) = \dfrac{1}{n}\left(\sum_i x_i \cdot y_i\right) - \overline{x} \cdot \overline{y}$
- $cov(x,x) = \dfrac{1}{n}\left(\sum_i x_i^2\right) - \overline{x}^2 = V(x) = \sigma_x^2$

Remarque : *la calculatrice ne donne pas $cov(x,y)$. Si on veut faire le calcul en détail (encore une fois, c'est inutile...) il faut utiliser les résultats des $\sum_i$ visibles dans le panneau de résultats des statistiques à 2 variables.*

r^2 : coefficient de détermination linéaire de Pearson

$r = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$ est le coefficient de corrélation linéaire. On est certain que $-1 \leq r \leq 1$.

On utilise r^2 pour juger de la qualité de l'ajustement. r^2 est sans unité et on peut l'exprimer comme un pourcentage : $0\,\% \leq r^2 \leq 100\,\%$.

r^2 représente la part de la variance totale $V(y)$ qui est expliqué par le modèle $y = a \cdot x + b$.

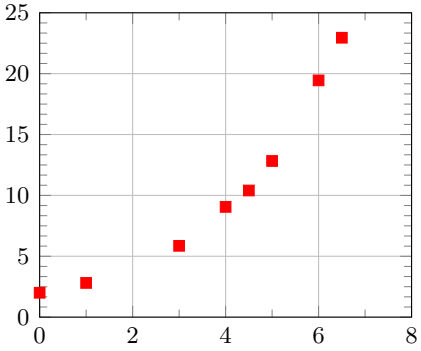
En général, on veut r^2 d'au moins 99 % mais cela dépend grandement de la situation : Dans certains cas on peut se conter d'un r^2 beaucoup plus faible, dans d'autres cas, des points mal alignés donnent pourtant un r^2 élevé.

Autre ajustement

Dans l'exemple ci-contre, les points ne sont pas assez bien alignés pour envisager un ajustement affine.

On peut imaginer que des considérations théoriques nous permettent de supposer qu'un ajustement de forme $y = \exp(a \cdot x + b)$ pourrait convenir.

Pour trouver a et b on essaie de se ramener à un ajustement affine : on pose $Y = \ln(y)$ de sorte que $Y = a \cdot x + b$.



Alors on calcule tous les $Y_i = \ln(x_i)$ puis on cherche l'ajustement affine du nuage $(x_i\,;\,Y_i)$. Une fois a et b trouvés, on a $Y = a \cdot x + b$ et donc $y = \exp(a \cdot x + b)$.