QUELQUES FORMULES UTILES

MOYENNE

Un DS donne les notes suivantes: 10, 12, 4, 19, 7, 15, 2, 0.

La moyenne est notée $\overline{x}$ et on la calcule de la manière suivante:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Ce qui, dans notre exemple, donne une moyenne de 69/8 = 8,25

MÉDIANE

Parfois la moyenne n'a pas de sens/ne donne pas la bonne information, par exemple lorsque l'on parle du salaire en France.

Le salaire moyen serait aux alentours de 2,200 euros nets par mois. Mais les personnes qui gagnent beaucoup gagnent parfois VRAIMENT beaucoup et tirent ce chiffre vers le haut.

Il peut donc être plus logique de regarder la médiane: 50% des français gagnent moins que le salaire médian et naturellement 50% des français gagnent plus que le salaire médian (1772 euros).

La médiane est donc la valeur qui sépare l'échantillon équitablement en deux parties.

QUANTILES

Pour aller plus loin que la médiane, on peut considérer les quantiles. Reprenons l'exemple du salaire. Selon l'INSEE en 2013:

10% des français gagnent moins de 1,200 euros par mois: 1,200 est le 10ème centile ou le 1er décile.

30% des français gagnent moins de 1,471 euros par mois: 1,471 est le 30ème centile ou le 3ème décile.

50% des français gagnent moins de 1,772 euros par mois: 1,772 est le 50ème centile, le 5ème décile ou la médiane.

1% des français gagnent plus de 8,061 euros par mois: 8,061 est le 99ème centile.

VARIANCE

La variance représente la variabilité au sein d'un échantillon. On la calcule comme la moyenne des écarts à la moyenne.

Si les notes d'un DS sont: 10, 10, 10, 10, 10, il n'y a aucune variance. Elle vaut 0.

Si les notes sont: 10, 12, 8, 14 la moyenne est 11 et la variance vaut:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

C'est-à-dire dans notre cas: 1/4 (1 + 1 + 9 + 9) = 20/4 = 5.

ECART-TYPE

C’est la racine de la variance. Il est plus facile à interpréter.

Dans notre example, les notes sont 10, 12, 8, 14, la moyenne est 11, la variance vaut 5 et l’écart-type vaut 2,24.

En moyenne, les notes se trouvent donc à un écart de 2,24 de la moyenne 11.

LA COVARIANCE

A quel point deux variables varient dans le même sens.

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

La covariance est élevée si les variations de x et de y par rapport à leurs moyennes respectives vont dans le même sens. Elle est au contraire très négative si les directions sont opposées. Par exemple si x est toujours au dessus de sa moyenne quand y est en dessous et inversement.

LA CORRÉLATION

C'est la covariance "corrigée" par les écarts-types:

$$\rho(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

La corrélation vaut toujours entre -1 et 1. Elle vaut 1 si x et y ont exactement les mêmes variations, -1 s'ils vont exactement en sens inverse, et 0 s'ils ne font rien en commun.

Example: x = (-1, 2, -3, 2) et y = (-2, 4, -6, 4)

Variance de x: 18/4, variance de y: 72/4, covariance: 36/4, corrélation: 1