Rappel :

Statistique unidimensionnelle :

Un fabriquant de chaussures soute avoir une idée des pointures des habitants de la cité afin d'optimiser ça production.

Pour ceci, il suffit de conduire une étude en relevant les pointures des habitants. A la fin de l'étude, l'information est un ensemble X contenant les pointures enregistrées.

Cet ensemble de valeurs n'est pas vraiment informatif (surtout si le nombre de valeurs récoltées est trop large). Plutôt, il est plus utile de représenter les valeurs sous forme d'un histogramme.

Une façon de résumer un ensemble de valeurs d'une manière concise est de prendre la **moyenne** (en statistique : **L'espérance**). La moyenne donne une idée sur le **centre** des valeurs d'une distribution.

$$\text{Esperance (X) = E(X)} = \frac{\sum_{i=1}^{n} X_i}{n}$$

avec n le nombre des valeurs de X

Le calcule de l'espérance consiste simplement à sommer sur toutes les valeurs et diviser sur le nombre de valeurs.

Cependant, il est possible d'avoir plusieurs distributions trop différentes avec le même centre. Par exemple, soit X et Y deux ensembles des notes pour deux classes différentes avec le même nombre d'étudiants:

X={8,9,10,11,12}

Y={2,6,10,14,18}

X et Y sont largement différents alors que E(X) = E(Y)= 10 (le même centre)

Pour exprimer la **dispersion des valeurs**, on utilise un autre paramètre. Il s'agit de la **variance**.

$$\text{Var(X)} = \frac{\sum_{i=1}^{n} (X_i - E(X))^2}{n} = E(\ ((X - E(X))^2\ )$$

$$E(\ ((X - E(X))^2\ ) = E(\ X^2 + E(X)^2 - 2XE(X)\ )$$

$$= E(\ X^2) + E(\ E(X)^2\ ) - E(\ 2XE(X)\ )\ )$$

Rappel : $pour\ C\ une\ constante\ E(C\ ) = C$

$E(X)^2\ est\ une\ constante\ et\ donc\ E(\ E(X)^2) =\ E(X)^2$

De même $E\big(\ 2XE(X)\big) = 2E(X)E(X) = 2E(X)^2$

$$= \ E(X^2) + \ E(X)^2 \ - 2E(X)^2$$

$$\mathrm{Var(X)} = \ E(X^2) - \ E(X)^2$$

La variance de X est la moyenne des distances au carré entre les valeurs de X (X1, X2…, Xn) et le centre de X (E(X)).

L'écart type de X, σ_X est utilisé pour représenter la dispersion des valeurs dans X. Ce paramètre représente la distance typique entre les valeurs de X (X_1, X_2,…, X_n} et l'espérance de X.

$$\sigma_X = \sqrt{\mathrm{Var(X)}}$$

La différence entre σ_X et V(X) est que σ_X possède la même métrique que X alors que Var (X) aura la métrique de X au carré. Par exemple : si X représente le poids en Kilogramme, Var (X) serai en Kilogramme^2 et σ_X en Kilogramme.

Par exemple pour X={8,9,10,11,12}

E(X)=10, Var (X)= 2 et $\sigma_X = \sqrt{2}$

Par exemple pour Y={2,6,10,14,18}

E(Y)=10, Var (Y)= 32 et $\sigma_Y = \sqrt{32}$

Statistique bidimensionnelle :

Dans l'exemple précédant, chaque individu de la population été décrit par une seule valeur. Par fois, les individus sont représentés par 2 ou plusieurs valeurs. Par exemple, soit X les valeurs suivantes associant les surfaces des maisons en **m^2** avec leur prix en milliard (**mil**).

Surface (m^2)	120	100	79	85	95	70	89
Prix (mil)	1.5	1.2	0.8	0.75	0.93	0.65	0.85

Dans la statistique bidimensionnelle, on s'intéresse à la relation entre les deux valeurs décrivant un individu d'une population. La relation la plus simple entre deux valeurs est l'équation de la ligne droite (ligne de régression).

A partir de la ligne droite, il est possible de faire des prédictions. Par exemple si on peut définir une équation de la forme ***« Prix = a x Surface +b »,*** il sera possible d'estimer le prix d'une maison à partir de sa surface.

La méthode des moindres carrés nous donne les paramètres de cette équation :

$$\hat{Y}=a\,X+b$$

$$a=\frac{\mathrm{Cov(X,Y)}}{\mathrm{Var(X)}}$$

$$b=E(Y)-\frac{\mathrm{Cov(X,Y)}}{\mathrm{Var(X)}}E(X)$$

(car la ligne passera toujours par le point E(X), E(Y))

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}=E(\big(X-E(X)\big)\big(Y-E(Y)\big))$$

La covariance quantifie comment les valeurs de X et Y varient ensemble (quand X augmente par rapport à la moyenne E(X), comment Y change par rapport à la moyenne E(Y) ?)

La variance de X est la même que la Covariance de X avec X :

$$Cov(X,X)=E\Big(\big(X-E(X)\big)\big(X-E(X)\big)\Big)=Var(X)$$

La corrélation, notée r_{XY} , est un paramètre qui décrit à quelle point la ligne de régression décrit la relation entre Y et X.

$$r_{XY}=\frac{\mathrm{Cov(X,Y)}}{\sigma_X\sigma_Y}$$

r_{XY} est toujours comprise entre -1 et 1.

Une valeur de r_{XY} qui tend vers 1 indique que la relation entre Y et X **est forte et positive (quand X augmente, Y augment aussi).**

Une valeur de r_{XY} qui tend vers -1 indique que la relation entre Y et X **est forte et négative (quand X augmente, Y diminue**).

Une valeur de r_{XY} qui tend vers 0 indique que la relation entre Y et X est **faible**.