Suite des indicateurs statistiques dans le cadre d'une variable quantitative

- Mode :

Il existe une différence dans le calcul suivant que la variable est discrète ou continue :

a. Discret : valeur prise par la variable qui est la plus observée (associée à l'effectif le plus grand)
b. Continue : le mode est une classe -> classe modale
La classe associée à la hauteur la plus grande dans l'histogramme (hauteur = $\frac{fréquence}{amplitude\ de\ la\ classe}$)

Remarque : Le mode n'est pas forcement unique

- Variance : il existe 2 variances

- Brutes : $S_1^2\ (=\ S_n^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Individuelles : $S_1^2\ = \frac{1}{n}\sum_{i=1}^{p} n_i\ (x_i - \bar{x})^2$
- Regroupées : $S_1^2\ = \frac{1}{n}\sum_{i=1}^{k} n_i\ (x_i - \bar{x})^2$

- Seconde variance (la plus utilisée) :

- Brutes : $S^2\ (=\ S_{n-1}^2)\ = \frac{1}{n-1}\sum_{i=1}^{n}\ (x_i - \bar{x})^2$
- Individuelles : $S^2\ = \frac{1}{n-1}\sum_{i=1}^{p} n_i\ (x_i - \bar{x})^2$
- Regroupées : $S^2\ = \frac{1}{n-1}\sum_{i=1}^{k} n_i\ (x_i - \bar{x})^2$

Remarque : $S^2 = \frac{n}{n-1}\ S_1^2$

- La variance quantifie la dispersion des données autour de $\bar{x}$
- Problème : une grandeur avec unité : son unité est (unité des données)²

- L'écart-type : racine carrée de la variance :

- $S_1 = \sqrt{S_1^2}$
- $S = \sqrt{S^2}$ <- la préférée
- L'unit de l'écart-type = unité des données

- Coefficient de la variation : CV

CV = $\frac{S}{\bar{x}} * 100$: coefficient absolu (sans unité)

Interprétation :

- Si CV < 16% -> homogénéité des données
  ➔ La moyenne est représentative des données
- Si CV > 33% -> non homogénéité
  ➔ La moyenne n'est pas à elle seule représentative des données
- Si 16% < CV < 33% -> zone flou
  ➔ Cas par cas

- Etendue :

- Discret : (la plus grande valeur observée) – (la plus petite valeur observée)
- Continue : (la limite supérieure de la dernière classe) – (la limite inférieure de la première classe)
  C'est-à-dire : $[a_1, a_2[, [a_2, a_3[\dots [a_k, a_{k+1}[$
  $E = a_{k+1} - a_1$

Il existe 2 paramètres de forme :

- Paramètre de symétrie (skewness)
- Paramètre d'aplatissement (kurtosis)

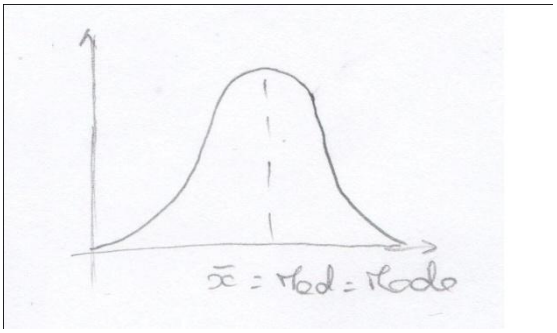
Objectif : comparaison à une loi normale

On calcule le coefficient de symétrie CB

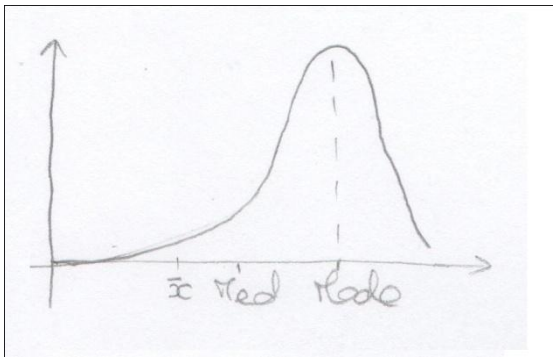
$$CB = \begin{cases} \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (\frac{x_i - \bar{x}}{5})^3 \ \ si\ brute \\ \frac{n}{(n-1)(n-2)} \sum_{i=1}^{p} n_i (\frac{x_i - \bar{x}}{5})^3 \ \ si\ individuelle \\ \frac{n}{(n-1)(n-2)} \sum_{i=1}^{\mathbb{R}} n_i (\frac{x_i - \bar{x}}{5})^3 \ \ si\ regroupée \end{cases}$$

Interprétation :

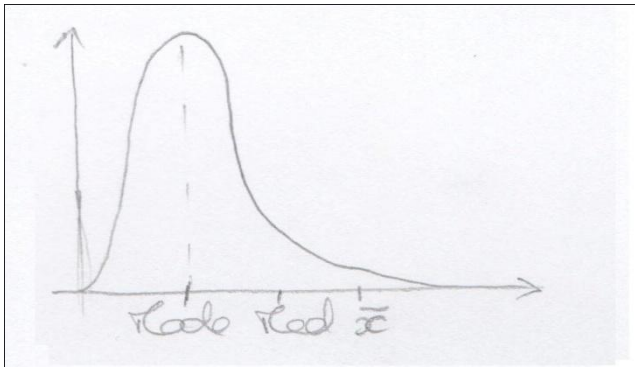
- Si « CB = 0 » -> la distribution des données est symétrique par rapport à $\bar{x}$



- Si « CB < 0 » -> distribution des données asymétrique à gauche



- Si « CB > 0 » -> asymétrie à droite

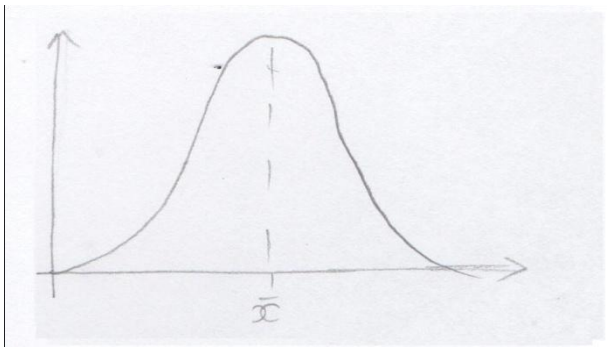


On calcule kurtosis : CA

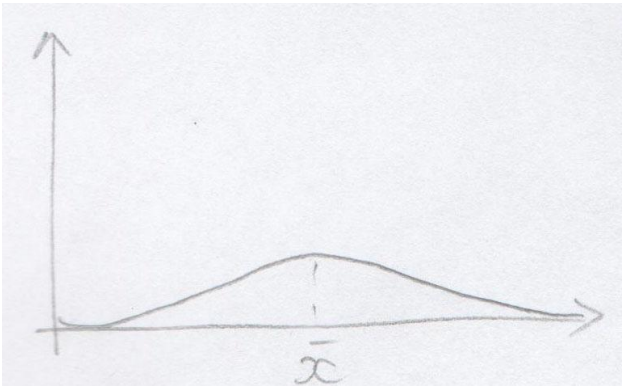
CA = $\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{5} - \frac{3(n-1)^2}{(n-2)(n-3)}$ *si brutes*

Interprétation :

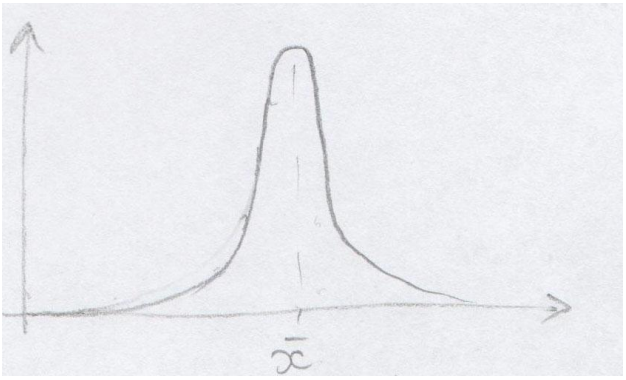
Si « CA =0 » -> la distribution est normalement étalée



Si « CA < 0 » -> distribution plus aplatie que la gaussienne



Si « CA > 0 » -> distribution plus pointue que la gaussienne



Remarque : f(x) = $\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i - E[x])2}{2\sigma^2}}$ -> la hauteur de la base $\frac{1}{\sqrt{2\pi\sigma}}$

Tout ce qui a été fait était pour une variable

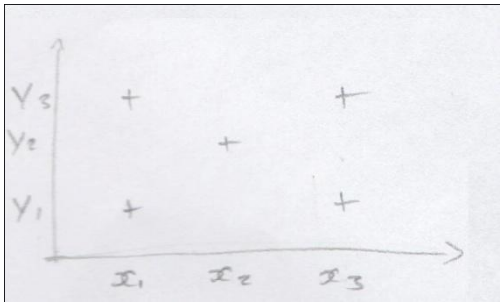
Que se passe-t-il si 2 variables ensembles ?

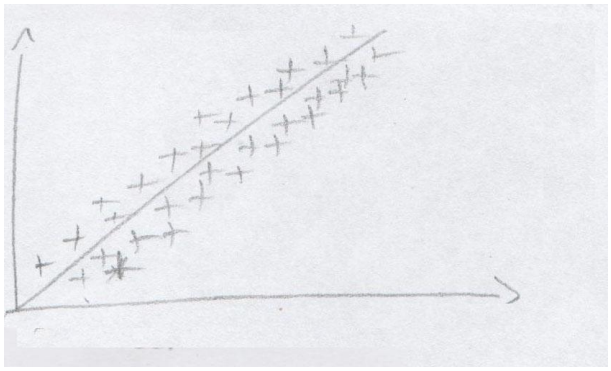
X	Y
x_1	y_1
…	…
x_n	y_n

1° étape : on étudie X et Y séparément

2° étape : recherche d'un éventuel lien entre X et Y

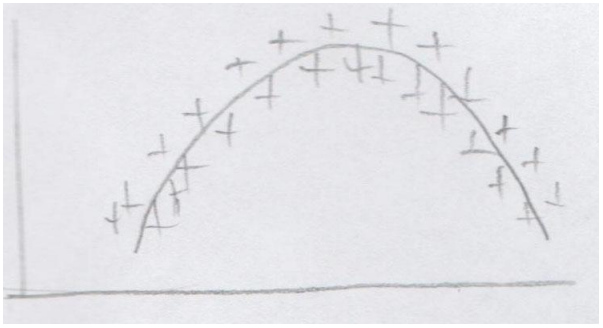
- <u>Nuage de points</u>



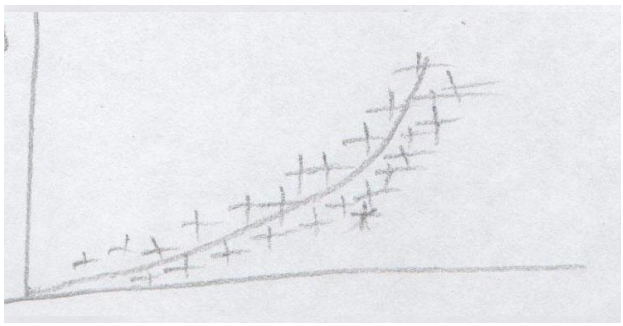


le nuage de point semble dirigé par une droite

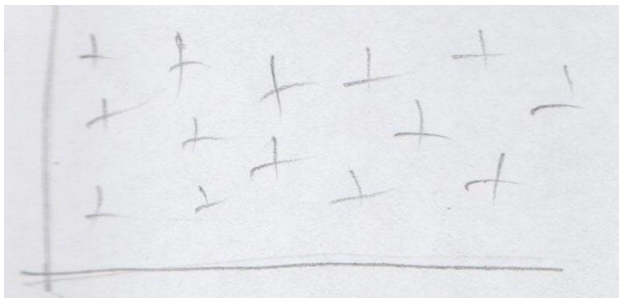
Relation linéaire : $\tilde{Y}$ = a + bX (Y = a+ b X = ε)



lien quadratique : Y ~= P(X) avec P un polynôme



lien exponentiel : Y ~= c * e^{aX}



pas de lien

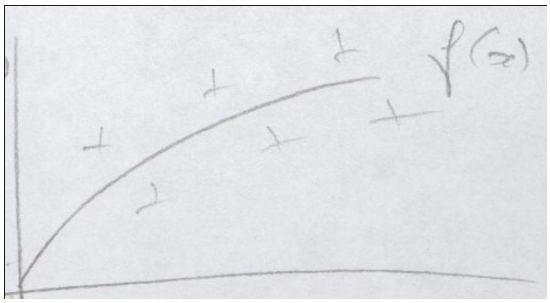
Question : existe-t-il une fonction f telle que Y ~= f(X)

Comment trouver f ?

On va mettre un postulat sur la forme de f

- ➔ Linéaire f(x) = a + bx
- ➔ Exponentielle : f(x) = ce^{ax}
- ➔ Quadratique : f(x) = $a_0 + a_1x + a_2x^2 + \dots + a_nx^n$

Problème : trouver les différentes constantes



principe : numériser la somme des distances élevées au carré

Min : $\sum_{i=1}^{n}(y_i - f(x_i))^2$

Dans le cadre linéaire :

On cherche a et b réel tq $\sum_{i=1}^{n}\left(y_i - (a + bx_i)\right)^2$ soit minimale

Dans le cas quadratique

On cherche a_0, a_1, …, a_n tel que $\sum_{i=1}^{n}(y_i - (a_0 + a_1 x_i + \ldots + a_n x_i^n))^2$ minimale

➔ Méthode des moindres carrées

Principe de l'optimisation

On calcule les différentes dérivées partielles

On résout le système annulant simultanément toutes les dérivés partielle -> obtention de point critique

On cherche le minimum parmi ces points critique

Proposition

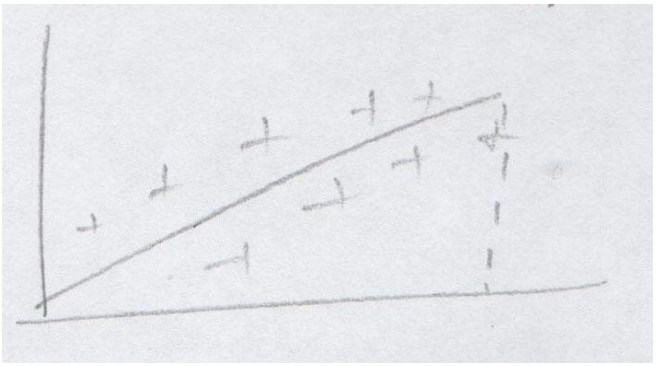
Si on suppose que f(x) = a + bx

La solution de la minimisation est données par

$$\left| \begin{array}{c} a = \bar{y} - b\bar{x} \\ b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})} = \dfrac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{\sum {x_i}^2 - n\,(\bar{x})^2} \end{array} \right.$$

$$notation$$

On note $\hat{y} = f(x)$: prédiction



Quantification du lien

Cadre général : coefficient de détermination : r²

$r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ avec $\hat{y}_i = f(x_i)$

Propriétés

- $0 \leq r^2 \leq 1$ car on peut montrer que $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Interprétation :

Plus r² est proche de I meilleur est le modèle f

Inconvénient : Pour calculer r² il faut avoir entièrement déterminé f

Cas particulier des cadres linéaires

On peut montrer que

Coefficient de détermination = (coefficient de corrélation)

Coefficient de corrélation = r

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}} \quad avec\ S_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$