III Statistique Descriptive.
(En Analyse de donnée)

01 Echantillons et caractères

1. Echantillon (ou population).

C'est un ensemble de personnes ou d'objets sur lequel on fera une etude. Les elts de l'ensemble s'appels individus.

Remarque

En Analyse des donnée, le mot "echantillon" n'a pas le m sens en statistique où les individus sont supposés provenir d'une population plus vaste

Exemple :

Personnel d'entreprise.

Ensemble des etudiants en 1ère année de PC à l'institut de Mathematique

2. Les differents types de caracteres

Definition

Un caractere des individus d'une population est une application qui, a chaque individu de la population, fait correspondre un elt d'un ensemble appelé ensemble des modalités de caracteres

Exemple

Population : ensemble des etudiants en 1ère année de PC à USTHB

Caractere : couleur des cheveux

ensemble des modalités : Noire, Blond, Brun, Roux.

a) Caractères qualitatifs :

Un caractère qualitatif est un caractère dont les modalités échappent à la mesure, elles peuvent seulement être constatées.

Exemple :

Couleur des cheveux,

sexe d'une personne

b) Caractères quantitatifs.

On dit qu'un caractère est quantitatif, lorsqu'il est mesurable ou repérable. A chaque individu correspond alors un nbre qui est la mesure ou la valeur du caractère. A ce nbre on donne le nom de variable.

Exemples.

Taille, poids ou âge d'une personne.

Durée de vie d'une lampe

I Echantillon d'une variable.

I = ensemble de n individus

x = variable quantitative.

$$x : I \longrightarrow \mathbb{R}$$
$$i \longmapsto x(i) = x_i$$

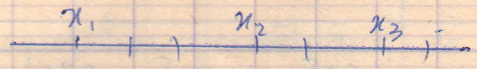
On appelle échantillon de la variable x, l'ensemble $E = \{x_i, i \in I\}$.

On suppose qu'à tout individu i est associé un poid $p_i > 0$ tq $\sum_{i=1}^{n} p_i = 1$.

1.1 Deux representations de $\mathcal{E}$.

On peut regarder $\mathcal{E}$ soit

- comme un nuage ensemble de $\mathbb{R}$.



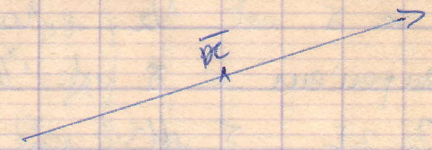
- comme un vecteur de $\mathbb{R}^n$.

$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ = vecteur de $\mathbb{R}^n$ associé à l'échantillon $\mathcal{E}$

2/(1) Trois) caractéristiques de valeurs centrale

a) La moyenne

$\bar{x} = \sum_{i=1}^{n} p_i x_i$



(Propre)

Son image sur la droite $\mathbb{R}$ est le centre de gravité des p^{ts} x_i munis des poids p_i

Propriétés

$\underset{1 \le i \le n}{\text{Min}}\, x_i \le \bar{x} \le \underset{1 \le i \le n}{\text{Max}}\, x_i$

b) La médiane

L'idée générale de la définition d'une médiane est de partager le domaine de la variable en deux parties, de manière à ce que (chacune) chacune des parties, contiennent, si possible, la moitié de l'effectif total, ou, si ce n'est pas possible de se rapprocher au mieux de (cette) cet objectif

Exemple.

Une population A a subi un examen noté de 0 à 5 la distribution des notes est donnée par le tableau suivant

Valeur des notes	0	1	2	3	4	5
effectifs	5	15	30	25	20	5
Frequence	0,05	0,15	0,30	0,25	0,20	0,05

Soit δ un nbre compris entre 2 et 3 ($2 < \delta < 3$)
la proportion de la population A qui a obtenue une note inferieur a δ est $1/2$. , la proportion qui a obtenue une note superieur a δ est $1/2$.
$\forall \delta \in]2,3[$, δ est une mediane
En duo que l'intervalle $]2\ 3[$ est un intervalle mediane

c) Moyenne des valeurs extremes.

$$\frac{1}{2}\left(\min_{1\le i\le n} x_i + \max_{1\le i\le n} x_i\right) = ME(x)$$

3/ Trois caracteristiques de dispersion

a) Variance et ecart-type.

$$Var(x) = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$
$$= \sum_{i=1}^{n} p_i x_i^2 - \bar{x}^2$$

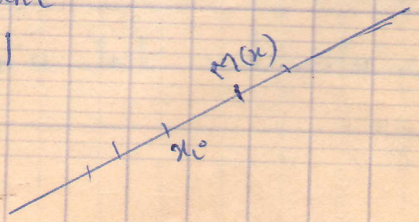
$$\sigma_x = \sqrt{Var(x)} = \sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2}$$

Remarques.
- si $Var(x) = 0$ E est forme de n valeurs egales
- L'ecart-type est "homogene a x" (c'est s'exprime avec la même unité que x)

b) Ecart moyen

$M(x)$ = mediane.

$$EC(x) = \sum_{i=1}^{n} p_i |x_i - M(x)|$$



c) Etendue

$$W(x) = \max_{1 \le i \le n} x_i - \min_{1 \le i \le n} x_i$$

4) Interpretation geometrique de la moyenne et de la variance dans $\mathbb{R}^n$

Exercice 7 de la serie

$$D = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \qquad \sum_{i=1}^{n} p_i = 1 \quad p_i \ge 0$$

a) $d(x,y) = N(x-y) = \sqrt{{}^t(x-y) D (x-y)}$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$D.(x-y) = \begin{pmatrix} p_1 (x_1 - y_1) \\ \vdots \\ p_n (x_n - y_n) \end{pmatrix}$$

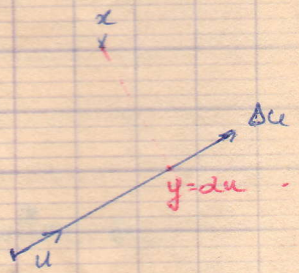
$${}^t(x-y) D (x-y) = (x_1 - y_1 ; \dots ; x_n - y_n) \begin{pmatrix} p_1 (x_1 - y_1) \\ \vdots \\ p_n (x_n - y_n) \end{pmatrix}$$

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

$$\Rightarrow \quad d(x,y) = \sqrt{\sum_{i=1}^{n} p_i (x_i - y_i)^2}$$

b) $u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \Delta u = \langle u \rangle$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$



$a.y = \alpha u$ projection D-orth de x sur Δu.

$$x - y \perp_D u \iff \langle x - y, u \rangle_D = 0$$

$$\iff \langle x, \alpha u \rangle_D = \alpha \langle u, u \rangle_D$$

$$\langle u, u \rangle_D = {}^t u D u = (1, \dots, 1) \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = 1$$

$$\alpha = \langle x, \alpha u \rangle_D = {}^t x D u = \sum_{i=1}^{n} p_i x_i = \bar{x}$$

Conclusion: l'abscisse de la projection D orthogonale de x sur Δu

est égale à $\bar{x}$, $y = \bar{x}u$ $\|u\|_2 = 1$

b- " $d(x, \Delta u) = \inf_{y' \in \Delta u} d(x, y') = d(x, y)$

$d(x, y) = d(x, \alpha u) = \sqrt{t(x - \alpha u) D (x - \alpha u)}$ $\alpha = \bar{x}$

$$\begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix} \begin{pmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} = {}^t\left(p_1 (x_1 - \bar{x}), \dots, p_n (x_n - \bar{x})\right)$$

$${}^t(x - \bar{x}u) D (x - \bar{x}u) = \sum_{i=1}^{n} p_i (x_i - \bar{x})^2$$

$$d(x, \Delta u) = \sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{x})^2} = \sigma_x$$

II - Echantillon d'un couple de variables.

I = ensemble de n individus.

x et y deux variables quantitatives

$$I \xrightarrow{x} \mathbb{R}$$
$$i^\circ \longmapsto x(i) = x_i$$

$$I \xrightarrow{y} \mathbb{R}$$
$$i^\circ \longmapsto y(i) = y_i$$

On appelle echantillon du couple (x, y), l'ensemble $\mathcal{E} = \{(x_i, y_i) \mid 1 \leq i \leq n\}$.

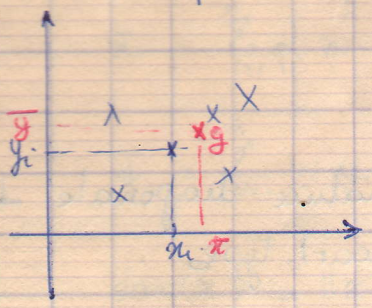
On suppose que chaque individu est muni d'un poids $p_i > 0$ $\left(\sum_{i=1}^{n} p_i = 1\right)$

A l'echantillon $\mathcal{E}$ correspondent les echantillons

$\mathcal{E}_x = \{x_i, 1 \leq i \leq n\}$ $\mathcal{E}_y = \{y_i, 1 \leq i \leq n\}$.

On peut decrire $\mathcal{E}_x$ et $\mathcal{E}_y$ separement, mais c'est insuffisant car il est bon de mettre en evidence les liens eventuels entre les variables x et y.

1- Representation de l'echantillon dans $\mathbb{R}^2$.



On represente $\mathcal{E}$ dans $\mathbb{R}^2$, par le nuage de n points (x_i, y_i)

$g = (\bar{x}, \bar{y})$ est le centre de gravité du nuage

$\bar{x} = \sum_{i=1}^{n} p_i x_i$

$\bar{y} = \sum_{i=1}^{n} p_i y_i$

2- Covariance et coefficient de corelation lineaire.

. Covariance de variable x et y.

$$Cov(x,y) = \sum_{i=1}^{n} p_i (x_i - \bar{x})(y_i - \bar{y})$$

. Coefficient de corelation lineaire des variables x et y

$$correl(x,y) = r(x,y) = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

. Matrice de variance-covariance associee à l'echantillon $\mathcal{E}$:

$$V = \begin{pmatrix} Var\,x & cov(x,y) \\ cov(x,y) & Var\,y \end{pmatrix}$$

. Matrice de corelation associee à l'echantillon $\mathcal{E}$

$$R = \begin{pmatrix} 1 & r(x,y) \\ r(x,y) & 1 \end{pmatrix}.$$

Remarques

* La covariance est sensible aux changements d'échelle sur x et y, le coefficient de corelation ne l'est pas

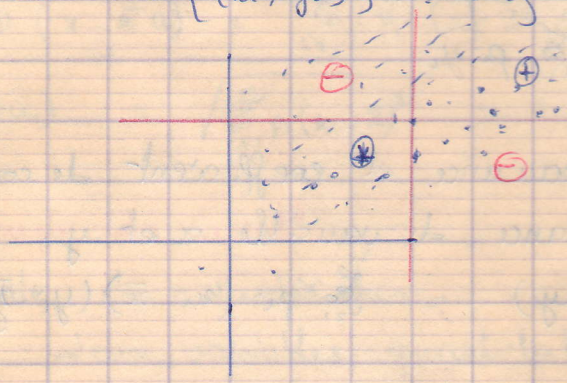
$r(ax, by) = r(x,y) \quad \forall a, b \in \mathbb{R}$.

* $R = D_{1/\sigma} \cdot V \cdot D_{1/\sigma}$

où $D_{1/\sigma} = \begin{pmatrix} 1/\sigma_x & 0 \\ 0 & 1/\sigma_y \end{pmatrix}$ matrice diagonale des inverses des écart-types.

+ Signe de la covariance et du coefficient de corrélation

Reprenons le nuage des n pts (x_i, y_i) qui représente dans $\mathbb{R}^2$ l'échantillon $\mathcal{E} = \{(x_i, y_i) ; 1 \le i \le n\}$



Si le nuage a une allure ↑, $cov(x,y) > 0$ $r(x,y) > 0$

Si le nuage a une allure ↘, $cov(x,y) < 0$

3- Représentation de l'échantillon dans $\mathbb{R}^n$.

Signification du coefficient de corrélation.

Dans $\mathbb{R}^n$, l'échantillon $\mathcal{E} = \{(x_i, y_i), i \le i \le n\}$ a pour image le couple de pts $(\underline{x}, \underline{y})$ où

$$\underline{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{et} \quad \underline{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$\mathbb{R}^n$ est muni de la métrique D_p.

$$D_p = \begin{pmatrix} p_1 & & 0 \\ & p_2 & \\ 0 & & \ddots \; p_n \end{pmatrix}$$

Nous avons vu que

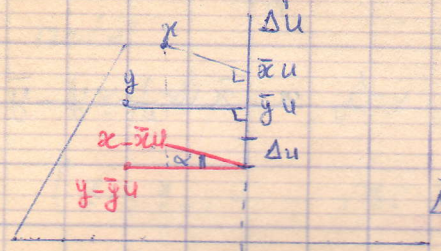
$$\bar{x} = \sum_{i=1}^{n} p_i x_i = {}^t x \, D_p \, u \qquad u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \; u \in \mathbb{R}^n.$$

$$\bar{y} = \sum_{i=1}^{n} p_i y_i = {}^t y \, D_p \, u$$

$$\mathrm{Var}(x) = \| x - \bar{x} u \|^2_{D_p}.$$

$$\mathrm{Var}(y) = \| y - \bar{y} u \|^2_{D_p}$$

Si on considère la décomposition en somme directe $\mathbb{R}^n = \Delta u \oplus \overline{\Delta}^{\perp} u$



où $\overline{\Delta}^{\perp} u$ désigne le sous espace vectoriel D_p-orthogonal à Δu. On a :

$x - \bar{x} u \in \overline{\Delta}^{\perp} u$ et $y - \bar{y} u \in \overline{\Delta}^{\perp} u$

Comme tout vecteur de $\mathbb{R}^n$, $x - \bar{x} u$ et $y - \bar{y} u$ symbolisent les caractères, on dit que ce sont les caractères x et y centrés, en effet leur moyenne

$$\sum_{i=1}^{n} p_i (x_i - \bar{x}) = 0 \qquad \sum_{i=1}^{n} p_i (x_i - \bar{x}) = 0$$

Les caractères centrés, $x - \bar{x} u$ et $y - \bar{y} u$ sont les projection D_p orthogonales sur $\overline{\Delta}^{\perp} u$ des vecteurs x et y.

Remarque.

Dans $\mathbb{R}^2$, centrer revient à mettre l'origine au centre de gravité g du nuage de n points $\quad g = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$

Interprétation du coefficient de corrélation

Soit α, angle que forment les vecteurs x et y centrés

$$\cos\alpha = \frac{\langle x - \bar{x}u, y - \bar{y}u \rangle}{\|x - \bar{x}u\|_{D_p} \cdot \|y - \bar{y}u\|_{D_p}} = r(x,y)$$

$\langle x, \bar{x}u ; -\bar{y}u, y \rangle_{D_p} = {}^t(x - \bar{x}u) D_p (y - \bar{y}u) = \sum p_i (x_i - \bar{x})(y_i - \bar{y}) = \text{cov}(x,y)$

Le coefficient de corrélation linéaire entre x et y est égal au $\cos\alpha$, où α est le D_p angle formé de $\overset{\perp}{\Delta} u$ par les caractères centrés

Remarque.

$-1 \leq r(x,y) \leq 1$

$r(x,y) = 0 \iff x - \bar{x}u \perp_{D_p} y - \bar{y}u$.

III/ Echantillon de p variables.

I est un ensemble de n individus sur chacun desquels on a mesuré p caractères x^j, $1 \leq j \leq p$

$$I \xrightarrow{x^j} \mathbb{R}$$
$$i \longmapsto x^j(i) = x^j_i$$

On appelle échantillon des p variables $\{x^1, \dots, x^p\}$, l'ensemble $\mathcal{E} = \{(x^1_i, x^2_i, \dots, x^p_i),\ i = \overline{1,n}\}$.

Soit le tableau suivant.

$$X = \begin{array}{c} \\ 1 \\ i \\ \\ n \end{array} \begin{pmatrix} x^1_1 & x^2_1 & \dots & x^j_1 & \dots & x^p_1 \\ \vdots & & & & & \\ x^1_i & x^2_i & \dots & x^j_i & \dots & x^p_i \\ \vdots & & & & & \\ x^1_n & x^2_n & \dots & x^j_n & \dots & x^p_n \end{pmatrix} \Bigg\} \text{n lignes}$$

p (Lignes) colonnes

La i^{eme} ligne de X symbolise l'individu "i"

La j^{eme} colonne de X symbolise le caractère x^j

Le vecteur $\underline{x}_i = \begin{pmatrix} x_i^1 \\ \vdots \\ x_i^j \\ \vdots \\ x_i^p \end{pmatrix} \in \mathbb{R}^p$ symbolise l'individu "i"

Le vecteur $\underline{x}^j = \begin{pmatrix} x_1^j \\ \vdots \\ x_n^j \end{pmatrix} \in \mathbb{R}^n$ symbolise le caractère x^j

$\mathbb{R}^p$ est l'espace "les individus", l'échantillon E, y est représenté par un nuage de n pts

$\mathbb{R}^n$ est l'espace "les caractères", l'échantillon E, y est représenté par un nuage de p pts

On pourra noter

$$X = (\underline{x}^1, \underline{x}^2, \dots, \underline{x}^j, \dots, x^p).$$

$$^tX = (\underline{x}_1, \underline{x}_2, \dots, \underline{x}_i, \dots, \underline{x}_n)$$

Matrice de variance, de matrice de corrélation, tableau de description

Les elts de l'ensemble I sont munis des poids, p_i $(p_i > 0)$

$\sum_{i=1}^{n} p_i = 1$

Soit D_p la matrice diagonale

$$D_p = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$$

D_p est symétrique définie positive.

D_p est matrice sur $\mathbb{R}^n$ = espace des caractères

DS $\mathbb{R}^p$, le pt moyen qui uniformise E est :

$$g = \begin{pmatrix} \bar{x}^1 \\ \bar{x}^j \\ \bar{x}^p \end{pmatrix} = \begin{pmatrix} \langle \underline{x}^1, u \rangle_{D_p} \\ \langle \underline{x}^j, u \rangle_{D_p} \\ \langle \underline{x}^p, u \rangle_{D_p} \end{pmatrix} \qquad u \in \mathbb{R}^n \qquad u = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

mettre l'origine de $\mathbb{R}^p$ en g, c'est centrer les caractères.

Le tableau X est alors dit centré.

Les matrices de covariance et de corrélation associées à l'échantillon $\mathcal{E}$ sont les matrices symétriques suivantes

$$V = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1j} & \cdots & V_{1p} \\ V_{21} & V_{22} & & & & \\ \vdots & & & & & \\ V_{i1} & & & V_{ij} & & \\ \vdots & & & & & \\ V_{p1} & & & & & V_{pp} \end{pmatrix} \quad \text{où } V_{ij} = \text{cov}(x^i, x^j)$$

$$R = \begin{pmatrix} 1 & r_{12} & & r_{1p} \\ r_{21} & & & r_{2p} \\ \vdots & & & \\ r_{p1} & & & 1 \end{pmatrix} \quad \text{où } r_{ij} = r(x^i, x^j)$$