

Statistiques

0. OBJECTIFS

Présentation de données à l'aide de schémas afin d'en faciliter l'observation.
Analyser les données à l'aide de valeurs caractéristiques.

1. STATISTIQUES A UNE VARIABLE.

1.1 Vocabulaire de base

Population : Tout ensemble P étudié en statistique s'appelle une population.

Individu : Les éléments de la population sont appelés individus.

Echantillon : Tout sous-ensemble de P s'appelle un échantillon.

Variable statistique : On appelle variable statistique quantitative toute application X de P dans \mathbb{R}

La variable statistique est discrète lorsqu'elle ne prend que des valeurs isolées, elle est continue si elle peut prendre toutes les valeurs d'un intervalle.

1.2 effectifs, fréquences.

1.2.1 Variable discrète

Soient $\{x_1; x_2; \dots; x_m\}$ les valeurs distinctes prises par la variable X . On les appelle aussi **modalités** du caractère X . (Dans la pratique on a $x_1 < x_2 < \dots < x_m$)

On appelle **effectif** de la modalité x_i le nombre n_i d'éléments de P ayant x_i pour image par X .

L'effectif total de la population est le nombre n tel que $n = \sum_{i=1}^{i=m} n_i$.

On appelle **fréquence** de la modalité x_i le réel $f_i = \frac{n_i}{n}$. (souvent donnée en pourcentage)

Les fréquences cumulées croissantes cumulent les fréquences associées aux valeurs du caractère inférieures ou égales à x_i .

On a donc $F_1 = f_1$ et $F_i = \sum_{j=1}^i f_j$ pour $i = 2, \dots, m$.

Les fréquences cumulées décroissantes cumulent les fréquences associées aux valeurs du caractère supérieures ou égales à x_i .

On a donc $F'_1 = f_1 + f_2 + \dots + f_m = 1$ et $F'_i = 1 - \sum_{j=1}^{i-1} f_j$ pour $i = 2, \dots, m$.

1.2.2 Variable continue

On partage l'intervalle I sur lequel X prend ses valeurs en intervalles disjoints appelés **classes** (en général de même amplitude).

$I = [x_0; x_1[\cup [x_1; x_2[\cup \dots \cup [x_{m-1}; x_m[$ avec $x_0 < x_1 < \dots < x_m$

L'effectif n_i associé à l'intervalle $[x_{i-1}; x_i[$ s'appelle l'effectif de la classe. On désigne souvent une classe par $(c_i; n_i)$ c_i étant le centre de la classe et n_i son effectif.

L'effectif cumulé croissant arrêté à la classe $[x_{i-1}; x_i[$ est le nombre $N_i = n_1 + n_2 + \dots + n_i$ des valeurs inférieures à x_i .

L'effectif cumulé décroissant arrêté à la classe $[x_{i-1}; x_i[$ est le nombre

$N'_i = n_i + n_{i+1} + \dots + n_m$ des valeurs supérieures à x_{i-1} .

On appelle fréquence de la classe le réel $f_i = \frac{n_i}{n}$.

1.3 Graphiques

Diagramme en bâton (discret) ou histogramme (continu)

Polygone des effectifs ; polygone des fréquences cumulées

Courbe cumulative des effectifs : C'est la représentation de la fonction de répartition définie sur \mathbb{R} par $x \mapsto$ (somme des effectifs $x_i < x$).

1.4 Caractéristiques de position

Mode : C'est la valeur du caractère correspondant à l'effectif le plus grand. Lorsqu'il s'agit de classe on parle de classe modale.

Médiane : C'est la valeur du caractère correspondant à un effectif cumulé égal à la moitié de l'effectif total.

Quartiles : Les 3 quartiles partagent la série en quatre séries de même taille.

25% des observations sont inférieures au 1^{er} quartile Q_{25} ;

50% des observations sont inférieures au 2^{ème} quartile Q_{50} ;

75% des observations sont inférieures au 3^{ème} quartile Q_{75} .

rem : on peut aussi définir les 9 déciles ou les 99 centiles qui partagent la série respectivement en 10 ou en 100 séries de même taille.

Moyenne : C'est le nombre noté \bar{x} ou parfois $E(x)$ défini par :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^m n_i x_i = \frac{n_1 x_1 + n_2 x_2 + \dots + n_m x_m}{n} .$$

Lorsqu'il s'agit de classe on remplace les valeurs x_i par les valeurs c_i correspondant aux centres des classes.

Prop : $E(X + b) = E(X) + b$ Changement par translation.
 $E(aX) = aE(X)$ Changement par homothétie.

rem : Ces changements peuvent être utilisés pour obtenir un calcul plus simple de $E(X)$

1.5 caractéristiques de dispersion

Etendue : C'est la différence entre la plus grande et la plus petite valeur de la série.

Ecart interquartile : C'est le nombre $Q_{75} - Q_{25}$.

Variance et écart type

La variance d'une série statistique est la moyenne de carrés des écarts à la moyenne.

$$V(X) = \frac{1}{n} \sum_{i=0}^m n_i (x_i - \bar{x})^2$$

Prop : $V(X + b) = V(X)$ Changement par translation.
 $V(aX) = a^2 V(X)$ Changement par homothétie.

Conséquence : $V(X) = \frac{1}{n} \sum_{i=1}^m n_i x_i^2 - \bar{x}^2$

L'écart type d'une série statistique est la racine carrée de la variance : $s(X) = \sqrt{V(X)}$

2. STATISTIQUES A DEUX VARIABLES

2.1. définition , présentation

Def : On appelle série statistique double pour les caractères X et Y l'application qui à chaque élément de Ω associe le couple $(x_i; y_i)$ où $\{x_1; x_2; \dots; x_m\}$ sont les valeurs du caractère X et $\{y_1; y_2; \dots; y_m\}$ sont celles du caractère Y .

Les résultats peuvent être présentés sous forme de données groupées ou sous forme de données non groupées.

Ex 1 : Données non groupées

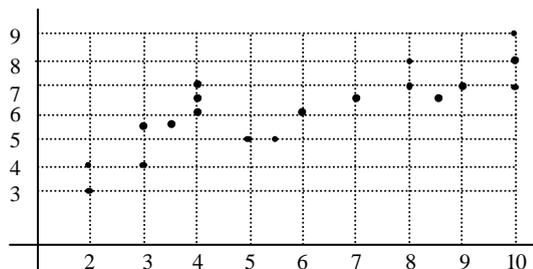
x_i	2	2	3	3	3,5	4	4	4	5	5,5	6	7	8	8	8,5	9	10	10	10	10
y_i	3	4	4	5,5	5,5	6	6,5	7	5	5	6	6,5	7	7,5	6,5	7	7	7,5	8	9

On peut choisir de regrouper les données en classes de largeur 2 et d'affecter à chaque classe la valeur centrale de l'intervalle.

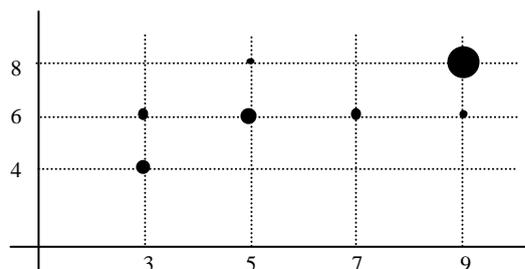
	intervalles en x	[2 ; 4 [[4 ; 6 [[6 ; 8 [[8 ; 10]
intervalles en y		3	5	7	9
[3 ; 5 [4	3	0	0	0
[5 ; 7 [6	2	4	2	1
[7 ; 9 [8	0	1	0	7

graphiques : nuages de points

Données non groupées



Données groupées



2.2 Ajustement linéaire : méthode de Mayer

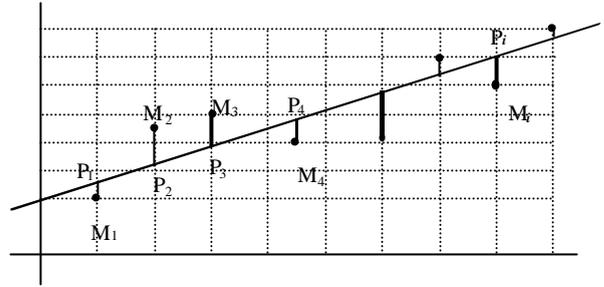
Dans certaines situation le nuage de point semble donner un ensemble de points presque alignés. On va alors dans ces cas faire un ajustement linéaire c'est à dire trouver une droite d'équation $y = a.x + b$ qui permet de représenter cette série.

Méthode de Mayer : On partage le nuage en deux nuages de même effectif (à une unité près). On cherche alors pour chacun leur point moyen notés respectivement G_1 et G_2 .

La droite (G_1G_2) est la droite d'ajustement selon la méthode de Mayer.

2.3 Droites de régression de y en x.

On appelle droite de régression de y en x la droite telle que la somme $S = \sum_{i=1}^n P_i M_i^2$ soit minimale.

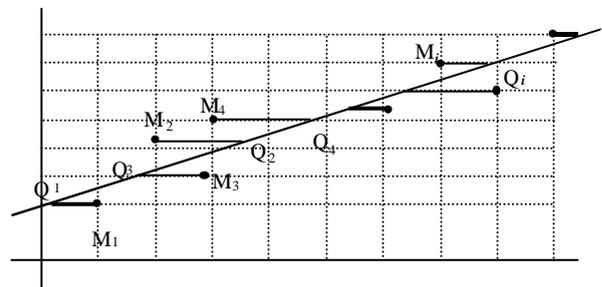


Cette droite a pour équation $y - \bar{y} = m(x - \bar{x})$ avec $m = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$

On peut aussi écrire $m = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ ou encore $m = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}{s_x^2}$

2.4 Droites de régression de x en y.

On appelle droite de régression de x en y la droite telle que la somme $S = \sum_{i=1}^n Q_i M_i^2$ soit minimale.



Cette droite a pour équation $y - \bar{y} = m'(x - \bar{x})$ avec $\frac{1}{m'} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$

On peut aussi écrire $m' = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}$ ou encore $m' = \frac{s_y^2}{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}}$

2.5 Corrélation linéaire

Def : On appelle covariance des variables X et Y le réel

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

Les coefficients directeurs des droites de régression s'écrivent alors $m = \frac{s_{xy}}{s_x^2}; m' = \frac{s_y^2}{s_{xy}}$

Pour comparer les directions des deux droites de régression on est amené à comparer leurs coefficients directeurs m et m' .

$$\text{or } \frac{m}{m'} = m \times \frac{1}{m'} = \frac{s_{xy}}{s_x^2} \times \frac{s_{xy}}{s_y^2} = \left(\frac{s_{xy}}{s_x s_y} \right)^2$$

Les droites seront alors d'autant plus proches l'une de l'autre que le rapport $\frac{m}{m'}$ sera proche de 1.

ou encore que la quantité $r = \frac{s_{xy}}{s_x s_y}$ sera proche de 1.

Def : La quantité $r = \frac{s_{xy}}{s_x s_y}$ s'appelle le coefficient de corrélation linéaire des variables X et Y de la série statistique double.