S4 2019-2020

Statistiques

MINFO0401

Nathan TONNELLE

Table des matières

Statistique descriptive

Vocabulaire

Population

Une population est l'ensemble des individus ou objets sur lesquels portent une étude statistique, on le note P

Exemple

1) Logement d'une ville
2) Personnel d'une entreprise
3) Animaux d'un parc naturel

Echantillon

Un échantillon est une partie de la population e étudier sur laquelle porte l'étude statistique.
Une étude statistique portant sur un échantillon est appelée sondage.
On appelle étude associative ou recensement si elle porte sur l'ensemble de la population.

Caractère et modalité

Une étude statistique porte sur 1 ou plusieurs caractères communs à tous les individus de la population à étudier. Un caractère est aussi appelé variable.

Exemple

1) Surface de logement
2) Age, ancienneté, revenu
3) Régime alimentaire, espèces

Modalité

Les modalités d'un caractère sont les différentes valeurs que peut prendre se caractère sur les individu de la population étudiée

Exemple

1) R^+
2) [16,10)
3) Herbivore, carnivore, omnivore

On distingue 2 types de caractères :

- Caractère qualitatif :

Les modalités sont des attributs qualitatif (ex: régime alimentaire)

- Caractère quantitatif

Les modalités sont des quantités numériques (ex : age)
Les caractères quantitatifs sont de 2 types

- Variable discrète :

Les modalités de la variable appartiennent à un ensemble discret tel N, Z, N^2

- Variable continue :

Les modalités de la variable prennent des valeurs dans un ensemble continu tel que R

Remarque :

Pour étudier une variable continue on constitue des classes de valeurs possibles, ces classes sont des intervalles d'amplitude égale ou inégale est constitué alors de nouvelle modalité ou des caractères.

Attention :

Le découpage en classes peut influer sur les résultats et les interprétations que l'on peut faire.
S'il est trop important, il risque de faire apparaitre des irrégularités artificielles car les effectifs des classes seront trop faibles.
S'il est trop grossier, il conduira à une perte d'information.

Effectif et fréquence

L'effectif d'une modalité ou d'une classe de modalité est le nombre d'individu de la population correspondant à cette modalité ou à cette classe de modalités. On ne note n_i pour la $i^{ème}$ modalité.
La fréquence de la $i^{ème}$ modalité (ou classe) est donnée par le rapport de son effectif sur l'effectif total de la population noté n.
On a alors $f_i=n_i/n$

Remarque

Si la variable X possède K modalités $x_1,..,x_k$ alors

$\sum_{i=1}^{k} n_i=n$ $\sum_{i=1}^{k} f_i= 1$

Représentation graphique

Il existe différentes façons de représenter graphiquement des variables

Variables qualitatives

Diagramme circulaire

C'est un disque dans lequel chaque modalité est représenté par un secteur angulaire proportionnel à sa fréquence.

Pour 1<=i<=f $\alpha_i=360xf_i$

Diagramme en tuyaux d'orgue

C'est un diagramme formé de rectangles tous de même largeur et dont les hauteurs sont proportionnelles aux fréquences des modalités

Variable quantitative

Il y a 2 sortes de représentation graphique des variables quantitatives :

- Diagramme différentiel
- Diagramme intégrale

Variables discrètes

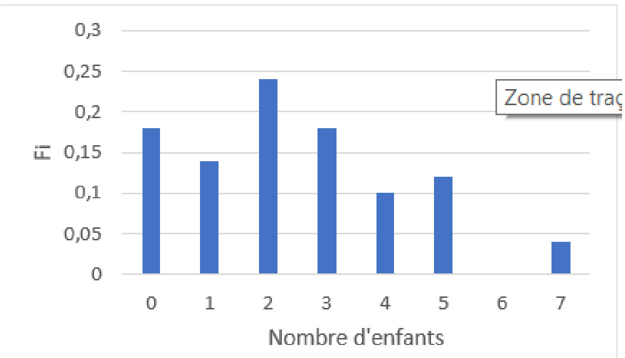
Le diagramme différentiel utilisé dans le cas d'une variable discrète est le diagramme à bâton.

Où chaque bâton est de longueur ou de hauteur proportionnel à la fréquence de la modalité correspondante

Exemple

Nombre d'enfants dans un échantillon de 50 familles.

Nombre d'enfant	n_i	f_i
0	9	0.18
1	7	0.14
2	12	0.24
3	9	0.18
4	5	0.10
5	6	0.12
6	0	0
7	2	0.04



La fonction cumulative

La fonction cumulative d'une variable X prise en un point x , noté F(x) est définie comme la proportion de la population pour laquelle la variable X prend des valeurs >= à x

Si les modalités de X sont $x_1, x_2, \ldots x_k$ alors pour $x_i <= x < x_{i+1}$, $F(x)=\sum_{j=1}^{i} f_j$

Exemple :

Si x<0 F(x)=0
0<=x<1F(x)=0.18
1<=x<2F(x)=0.32
2<=x<3F(x)=0.56
3<=x<4F(x)=0.74
4<=x<5 F(x)=0.84
5<=x<7 F(x)=0.86
x>=7 F(x)=1

Variables continues

Si e_i et e_{i+1} sont les extrémités de la classe n°i, noté $[e_i, e_{i+1}]$, on notera c_i soit milieu et a_i son amplitude, 1<=i<=k

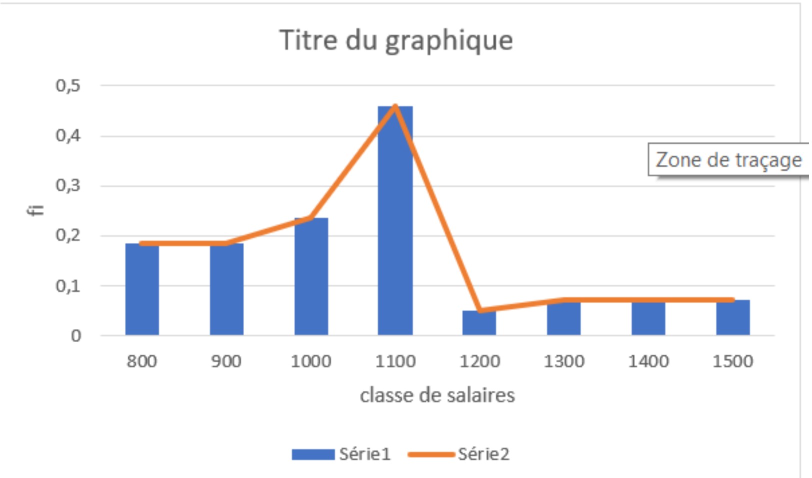
Définition (représentation graphique différentielle)

Un histogramme est une représentation graphique où chaque classe est représenté par un rectangle de base proportionnelle à son amplitude et de surface proportionnelle à sa fréquence. Ainsi la hauteur de la classe n°i est $h_i=f_i/a_i$. Un polygone statistique est un polygone reliant le milieu des bases supérieurs des rectangles de l'histogramme.

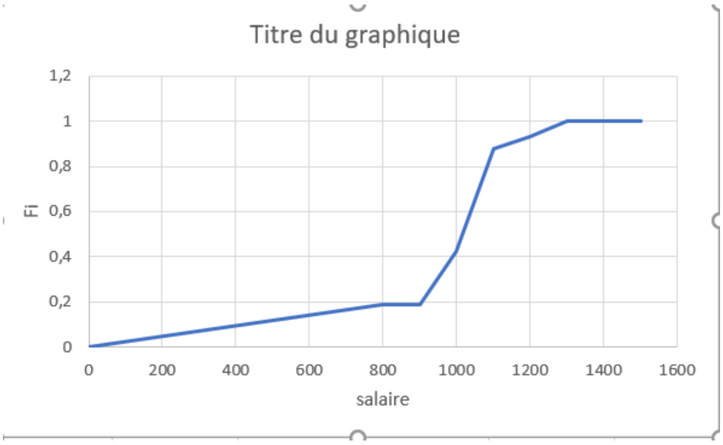
Exemple

Salaire mensuel net des ouvriers d'un établissement industriel.

Classes de salaire	n_i	f_i	a_i	$h_i=f_i/a_i$ (echelle x100)	F_i
[800,1000[	26	0.186	200	0.096	0.186
[1000,1100[	33	0.235	100	0.235	0.421
[1100,1200[	64	0.458	100	0.458	0.879
[1200,1300[	7	0.05	100	0.05	0.929
[1300,1500[	10	0.071	200	0.0355	1
total	140	1	700		



Courbe cumulative



Description numérique d'une variable

Paramètres de position

Médiane

Définition

La médiane mu(grec) (u) est la valeur de la variable x pour laquelle la moitié au moins des observations son supérieur ou égale et la moitié au moins des observations inférieures ou égales

Remarque

Dans le cas continue on détermine d'abords la classe médiane avant de calculer le point le médian par la méthode d'interpellation linéaire.

Exemple

1. (nombre d'enfants par famille) u=2 car (F_2=0.32 et F_3=0.56)
2. (salaires des ouvriers)

Classe médiane = [1100,1200[
Calcul de la médiane par la méthode d'interpellation linéaire

$$\leftrightarrow \frac{u-1100}{0.5-0.421}=\frac{1200-1100}{0.879-0.421}$$
$$\leftrightarrow \frac{u-1100}{0.079}=\frac{100}{0.458}$$
$$\leftrightarrow u-1100=0.079*\frac{100}{0.458}=17.249$$
$$\leftrightarrow u=1117.249$$

Le mode

Définition

Le mode est la valeur de la variable x ayant la plus grande fréquence

Remarque

Certaines séries statistiques peuvent avoir plusieurs modes. Dans le cas continu on parle de classe modale, on veillera cependant a tenir compte de l'amplitude des classe. La classe modale correspond à la classe ayant la plus grande hauteur h_i.

Exemple

1. (salaire des ouvriers)

Classe modale =[1100,1200[
Equation (M M2)

$$a=\frac{0.458-0.235}{1200-1100}=0.223*10^{-2}$$
$$b=0.548-0.223*10^{-2}*1200=-2.218$$
$$y=0.223*10^{-2}x-2.218$$

Equation(M3 M4)

$$a = \frac{0.05 - 0.458}{1200 - 1100} = -0.408 * 10^{-2}$$
$$b = 0.458 + 0.408 * 10^{-2} * 1100 = 4.946$$
$$y = -0.408 * 10^{-2}x + 4.946$$

$$0.223 * 10^{-2}x - 2.218 = -0.408 * 10^{-2}x + 4.946$$
$$\leftrightarrow x = \frac{7.164}{0.631} * 10^{-2} = 135.34$$

la moyenne

Soit x une variable prenant les valeurs $x_1,\ldots,x_k$ avec les effectifs $n_1,\ldots,n_k$ respectivement (avec les fréquences $f_1,\ldots,f_k$). Alors la moyenne de la variable x est donnée par

$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i = \sum_{i=1}^{k} f_i x_i$

Où $n = \sum_{i=1}^{k} n_i$

Dans le cas continu, la moyenne d'une variable x est définie par

$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} n_i c_i = \sum_{i=1}^{k} f_i c_i$

Où les C_i sont les milieux des classes $[e_i,e_{i+1}[$

Propriétés

1) Linéarité

Si on considère la transformation Y=aX+b

Alors la moyenne de Y est $\bar{Y}$ = a $\bar{X}$ +b

2) $\sum_{i=1}^{k} n_i (x_i - \bar{x}) = 0$
3) Si on définit la fonction

$$L(c) = \sum_{i=1}^{k} n_i (x_i - c)^2$$

Alors L(c) prend son minimum pour c=$\bar{X}$

4) Si $P=P_1 \cup P_2$

Où la moyenne de X sur P_1 est $\bar{X}$, et l'effectif est n_1; la moyenne de X sur P_2 est $\bar{X}_2$ et l'effectif de P_2 est n_2, alors la moyenne de X sur P est :

$$\bar{X} = \frac{n_1 * \overline{X_1} + n_2 * \overline{X_2}}{n_1 + n_2}$$

Paramètre de dispersion

Etendu

L'étendu est la différence entre la plus grande et la plus petite valeur de la série statistique.

L'écart moyen absolu

On considère une série de n observations dont les modalités sont $x_1,\ldots,x_k$.

On note $\bar{X}$ sa moyenne empirique.

Alors l'écart-type moyen absolu est défini par

$$\boldsymbol{e_n = \frac{1}{n} \sum_{i=1}^{k} n_i |x_i - \bar{x}|}$$

Variance (écart quadratique moyen)

Si $x_1,\ldots,x_n$ sont les modalités d'une variance X observée n fois, alors la variance de X est définie par

$$\mathbf{Var(X)} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{n_i (x_i - \bar{x})}^2$$

Propriété

1) $\frac{1}{2} \sum_{i=1}^{k} n_i (x_i - \bar{x})^2 = \frac{1}{2} \sum_{i=1}^{k} n_i x^2 - \bar{x}^2$
2) Si $Y = aX + b$, alors $Var(Y) = a^2 Var(X)$

Preuve

$$Var(Y) = \frac{1}{n}\sum_{i=1}^{k} n_i(y_i - \bar{Y})^2$$
$$= 1/n \sum_{i=1}^{k} n_i(ax_i + b - (a\bar{X} + b))^2$$
$$= \frac{1}{n}\sum_{i=1}^{k} n_i(ax_i - a\bar{X})^2$$
$$= (a^2)\frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{X})^2$$
$$= a^2 Var(X)$$

3) Si P_1 et P_2 sont 2 sous populations d'une population P tel que les moyennes, les variences et les effectifs de X sont :

Pour P_1 : $\overline{X_1}, Var_1(X), n_1$
Pour P_2 : $\overline{X_2}, Var_2(X), n_2$
Alors la variance de X sur l'ensemble de P est

$$Var(X) = \frac{n_1 Var_1(X) + n_2 Var_2(X)}{n_1 + n_2} + \frac{n_1(\overline{X_1} - \bar{X})^2 + n_2(\overline{X_2} - \bar{X})^2}{n_1 + n_2}$$

Où $\bar{X} = \frac{n_1\overline{X_1} + n_2\overline{X_2}}{n_1 + n_2}$ est la moyenne de X sur P

$$Var(X) = moyenne\ des\ variances + variances\ des\ moyennes = variances\ intra + variance\ inter$$

Ecart type

L'écart type est défini par :

$$\sigma_x = \sqrt{Var(x)}$$

Intérêt :

X est en Km => en Km²
L'écart type exprime la dispersion dans la même unité de mesure que la variable X

Coefficient de variation

Le coefficient de variation est défini par

$$CV = \frac{\sigma_x}{\bar{X}}$$

Intérêt :

Ce coefficient est indépendant de l'unité de mesure. Il permet alors de comparer les dispersions de sens statistiques exprimés dans des unités de mesure différents.

Les moments

Définition

On appelle moment d'ordre t (t appartient à N) par rapport à une constante a d'une variable statistique x

$$m_t(a) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - a)^t$$

Les moments non centrés correspondent à

$$a = 0, ie : m_t = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^t$$

Les moments centrés correspondent à a=$\bar{X}$
$ie :\ \mu_t = m_t(\bar{X}) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{X})^t$

Remarque

$$m_1(0) = m_1 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \bar{X}$$

$$m_2(\bar{X}) = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \bar{X})^2 = Var(X)$$

Caractéristiques de forme

Coefficient d'asymétrie

Le coefficient d'asymétrie de Fisher est défini par

$$\gamma^1 = \frac{\mu_3}{\sigma_x^3}$$

Si A=>B alors $\bar{B}$ => $\bar{A}$

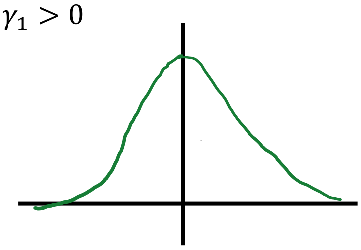
Remarque :

Si la distribution est symétrique, alors

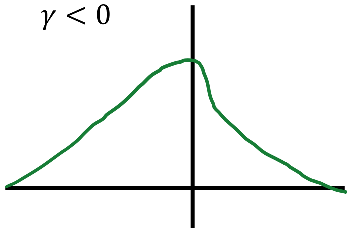
$$\gamma_1 = 0$$

Si $\gamma_1 \neq\ 0$ alors la distribution n'est as symétrique

$\gamma_1 > 0$



$\gamma < 0$

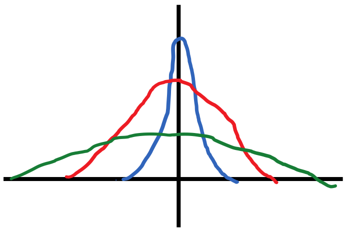


Distribution non symétrique, oblique à gauche et étroite à droite

Coefficient d'aplatissement (Kertosis)

$$\gamma_a = \frac{\mu_4}{\sigma_x^4} - 3$$

Il compare l'aplatissement de courbe statistique à la courbe de la loi N(0,1).



$\gamma_2 < 0$ $\gamma_2 < 0$ $\gamma_2 < 0$

Si la courbe est identique à celle de la loi N(0,1) alors $\gamma_2 = 0$.

Si $\gamma_2 \neq 0$, la courbe ne présente pas le même aplatissement que la courbe de la loi N(0,1).

Box-plot (boite à moustache)

Définition :

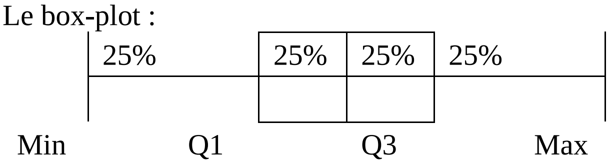
Premier quartile q1 est la valeur de la variable x pour laquelle au moins ¼ des observations lui sont inférieur ou égales et au moins ¾ des observations supérieures ou égales.

Le troisième quartile q3 est la valeur de la variable x pour laquelle au moins ¾ des observations lui sont inférieur ou égales et au moins ¼ des observations supérieures ou égales.

Remarque :

Dans le cas d'une variable continue, on détermine d'abord les classes contenant le premier quartile q1 et le 3eme quartile q3 avant de procéder par la méthode de l'interpellation linéaire au calcul de q1 et q3

Le box-plot :



Distributions à deux dimensions

On considère un couple de variables (X,Y) observé sur une population P de taille n. on suppose que la variable X possède les modalités $x_1,\dots,x_k$ et que les modalités de la variable y sont $y_1,\dots,y_l$. Ces modalités peuvent être des valeurs discrètes ou des classes de modalités (cas où les variables sont continues). L'observation du vecteur (X,Y) donne lieu à la table de contingence qui se présente sous la forme :

X\Y	y_1	y_2	…	Y_j	…	Y_l	
x_1	n_{11}	n_{12}		n_{1j}		n_{1l}	$n_{1}.$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2}.$
…							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{il}	$n_{i}.$
…							
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kl}	$n_{k}.$
	$n_{.1}$	$n_{.2}$		$n_{.j}$		$n_{.l}$	$n_{..} = n$

Où n_{ij} est l'effectif connu portant un nombre de fois où la modalité x_i de x et la modalité de y_j de Y sont observé, simultanément sur les individus de la population P.

Notations

$n_{i}. = \sum_{j=1}^{l} n_{ij}$, $n_{.j} = \sum_{i=1}^{k} n_{ij}$

$$n_{..} = \sum_{i=1}^{k} n_{i}. = \sum_{i=1}^{l} n_{.j} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n$$

La table de contingence donne la distribution des effectifs du couple (X,Y).

Distribution des fréquences du couple (X,Y)

La fréquence de la modalité (x_i,y_j) de (X,Y) est définie par

$f_{ij} = \frac{n_{ij}}{n}$, 1<= i<=k et 1<=j<=l

Notation

Pour 1<=i<=k et 1<=j<=l , on note $f_{i}. = \sum_{j=1}^{l} f_{ij} = \frac{n_{ij}}{n}$

$$f_{.j} = \sum_{i=1}^{k} f_{ij} = \frac{n_{ij}}{n}$$

Remarque

$$\sum_{i=1}^{k} f_{i}. = \sum_{j=1}^{l} f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1$$

Distribution marginale

La distribution marginale des effectifs de X est de définir par ($n_{1.}, \ldots, n_{k.}$). la distribution marginale des fréquences de X est ($f_{1.}, \ldots, f_{k.}$).

Les distributions marginales des effectifs et des fréquences de la variable Y sont respectivement ($n_{.1}, \ldots, n_{.l}$) et ($f_{.1}, \ldots, f_{.l}$).

Moyennes et variances marginales

La moyenne et la variance marginales de X sont définies par

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i}.x_i = \sum_{i=1}^{k} f_{i}.x_i$$

$$Var(X) = \frac{1}{n}\sum_{i=1}^{k} n_{i}.(x_i - \bar{X})^2 = \sum_{i=1}^{k} f_{i}.(x_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{k} n_{i}.x_i^2 - \bar{X}^2 = \sum_{i=1}^{k} f_{i}.x_i^2 - \bar{X}^2$$

La moyenne de la variance marginale de Y sont :

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\, y_j = \sum_{j=1}^{l} f_{.j}\, y_j$$

$$Var(Y) = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\left(y_j - \bar{Y}\right)^2 = \sum_{j=1}^{l} f_{.j}\left(y_j - \bar{Y}\right)^2 = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\, y_j^2 - \bar{Y}^2 = \sum_{j=1}^{l} f_{.j}\, y_j^2 - \bar{Y}^2$$

Distributions conditionnelles

Définition

La fréquence conditionnelle de X=x_i sachant Y=y_j est définie, pour $1 \le i \le k \; et \; 1 \le j \le l$, par

$$f_{\frac{i}{j}} = \frac{n_{ij}}{n_{.j}}$$

La fréquence conditionnelle de Y=y_j sachant X=x_i est définie pour tous $1 \le i \le k \; et \; 1 \le j \le l$ par

$$f_{\frac{j}{i}} = \frac{n_{ij}}{n_{i.}}$$

La distribution conditionnelle de X sachant Y=y_j est

$$\left(\frac{n_{1j}}{n_{.j}}, \frac{n_{2j}}{n_{.j}}, \ldots, \frac{n_{kj}}{n_{.j}}\right)$$

La distribution conditionnelle de Y sachant X=x_i est

$$\left(\frac{n_{i1}}{n_{i.}}, \frac{n_{i2}}{n_{i.}}, \ldots, \frac{n_{il}}{n_{i.}}\right)$$

Remarque

$$f_{\frac{i}{j}} = \frac{n_{ij}}{n_{.j}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{.j}}{n}} = \frac{f_{ij}}{f_{.j}}$$

Par conséquent

$$f_{ij} = f_{.j} * f_{\frac{i}{j}}$$

$$f_{\frac{j}{i}} = \frac{n_{ij}}{n_{i.}} = \frac{\frac{n_{ij}}{n}}{\frac{n_{i.}}{n}} = \frac{f_{ij}}{f_{i.}}$$

Conséquence

$$f_{ij} = f_{i.} * f_{\frac{j}{i}}$$

Indépendance des variables X et Y

Les variables X et y sont indépendantes lorsqu'on a

$$f_{\frac{i}{j}} = f_{i.} \; ou \; f_{\frac{j}{i}} = n_{.j} \qquad\qquad 1 \leq i \leq k \; et \; 1 \leq j \leq l$$

Conséquence

Si X et Y sont indépendants alors

$$f_{ij} = f_{i.} * f_{.j}$$

Cela peut s'exprimer aussi sous la forme

$$\frac{n_{ij}}{n} = \frac{n_{i.}}{n} * \frac{n_{.j}}{n} \leftrightarrow n_{ij} = \frac{n_{i.} * n_{.j}}{n}$$

Moyennes et variances conditionnelles

La moyenne conditionnelle de X sachant Y=y_j est définie par

$$\overline{X_j} = \sum_{i=1}^{k} f_{\frac{i}{j}} x_i = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} x_i$$

$1 \leq j \leq l$

La moyenne conditionnelle de Y sachant X=x_i est

$$\overline{Y_i} = \sum_{j=1}^{l} f_{\frac{j}{i}} y_j = \frac{1}{n_{i.}} \sum_{i=1}^{l} n_{ij} y_j$$

$1 \leq i \leq l$

Résultat

Les moyens marginales et les moyennes conditionnelles sont liées par les relations suivantes :

$$\overline{X} = \sum_{j=1}^{l} f_{.i} \overline{X_j}$$

$$\overline{Y} = \sum_{i=1}^{k} f_{i.} \overline{Y_i}$$

Les variances conditionnelles de X sachant Y=y_j et sachant X=x_i sont définies respectivement par

$$Var_j(X) = \sum_{i=1}^{k} f_{\frac{i}{j}}(x_i - \bar{X}_j)^2$$

$1 \leq j \leq l$

Et

$$Var_i(Y) = \sum_{j=1}^{l} f_{\frac{j}{i}}(y_j - \bar{Y}_i)^2$$

$1 \leq i \leq k$

Les variances marginales de X et Y peuvent être décomposées dans les termes suivants :

$$Var(X) = \sum_{j=1}^{l} f_{.j} Var_j(X) + \sum_{j=1}^{l} f_{.j}\left(\bar{X}_j - \bar{X}\right)^2$$

= moyennes des variances conditionnelles et variances des moyennes conditionnelles

$$Var(Y) = \sum_{i=1}^{k} f_{i.} Var(Y) + \sum_{i=1}^{k} f_{i.}(\bar{Y}_i - \bar{Y})^2$$

Notion de corrélation

Coefficient de corrélation linéaire

Définition

Le coefficient de corrélation linéaire entre deux variables X et Y est défini par

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Où

$$cov = \frac{1}{n} \sum_{(i=1)}^{k} \sum_{j=1}^{l} n_{ij} x_i y_j - \bar{X}\bar{Y} = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij} x_i y_j - \bar{X}\bar{Y} = \sum_{i=1}^{k} \sum_{j=1}^{l} f_{ij}(x_i - \bar{X})(y_i - \bar{Y})$$

Propriétés

- $-1 \leq \mathrm{f}_{\mathrm{X,Y}} \leq 1$
- si X et Y sont indépendant alors $\mathrm{f}_{\mathrm{X,y}} = 0$

Remarque

Si X et Y ne sont pas corrélées, cela n'implique pas forcément que X et Y sont indépendants.
Il y a identité entre indépendance et non corrélation uniquement dans le cas de variable gaussiennes.

Propriétés (suite)

- Si X et Y présentaient une liaison fonctionnelle, i.e., il existe une fonction f tant que $f(X) = Y$ $, alors\ on\ a\ f_{X,Y}^2 = 1$

Rapports de corrélation

On considère deux variables X et Y prennent les valeurs $x_1,\ldots,x_k$ et $y_1,\ldots,y_l$ respectivement.
On note Xbarre et Ybarre les moyennes empiriques de X et Y respectivement.
On note var(X) et var(Y) les variances marginales de X et Y.
Si $\bar{X}_j$est la moyenne de X sachant $Y = y_j$, $Y_i$la moyenne de Y sachant X=x_i ,
On note $var_{inter}(X) = \frac{1}{n}\sum_{j=1}^{l} n_{.j}\left(\bar{X}_j - \bar{X}\right)^2 = \sum_{j=1}^{l} f_{.j}\left(\bar{X}_j - \bar{X}\right)^2$
Et $Var_{inter}(Y) = \frac{1}{n}\sum_{i=1}^{k} n_{i.}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{k} f_{i.}(\bar{Y}_i - \bar{Y})^2$

Définition

Le rapport de corrélation de X en Y est défini par

$$\eta_{\frac{X}{Y}}^2 = \frac{Var_{inter}(X)}{Var(X)} = \frac{\sum_{j=1}^{l} f_{.j}\left(\bar{X}_j - \bar{X}\right)^2}{\sum_{i=1}^{k} f_{i.}(x_i - \bar{X})^2}$$

Le rapport de corrélation de Y en X est défini par

$$\eta^2_{\frac{Y}{X}} = \frac{Var_{inter}(Y)}{Var(Y)} = \frac{\sum_{i=1}^{k} f_{i.}(\overline{Y_i} - \overline{Y})^2}{\sum_{j=1}^{l} f_{.j}\left(Y_j - \overline{Y}\right)^2}$$

Propriété

1) $0 \leq \eta^2_{\frac{X}{Y}} \leq 1 \;\; ET \;\; 0 \leq \eta^2_{\frac{Y}{X}} \leq 1$
2) Si $\eta^2_{\frac{X}{Y}} = 0$, cela veut dire que $\overline{X_j} = \overline{X}$ pourtant $1 \leq j \leq l$

   Le conditionnement de X par Y ne donne aucune liaison significative
3) Si $\eta^2_{\frac{X}{Y}} = 1$, cela veut dire que $\sum_{j=1}^{l} f_{.j}\left(\overline{X_j} - \overline{X}\right)^2 = Var(X)$

   X est lié de façon complète à Y
4) Plus $\eta^2_{\frac{X}{Y}}$ est proche de 1 plus la liaison de X à Y est importante

Courbes de régression

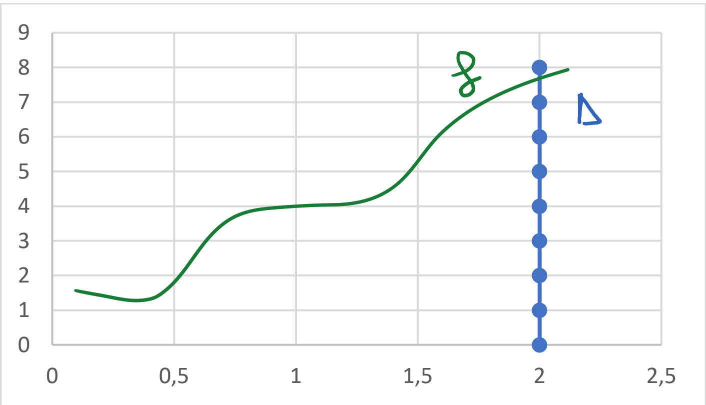
Les courbes de régression ont pour objet de donner une représentation graphique sur le plan de la distribution conjointe du vecteur (X,Y). Il y a deux courbes de régression. La courbe de régression de X en Y et la courbe régression de Y en X.

Définition

Si la variable X prend ses valeurs dans l'ensemble $\{x_1,\ldots,x_k\}$ et Y dans l'ensemble $\{y_1,\ldots, y_l\}$. On note $\overline{X_j}$ la moyenne conditionnelle de X sachant $Y=y_i$ et $\overline{Y_i}$ la moyenne conditionnelle de Y sachant $X=x_i$. Alors la courbe de régression de X en Y notée $C_{courbe\frac{X}{Y}}$ est la courbe passant par les points de coordonnées $\left(\overline{X_j}, y_j\right)$ $1 \leq j \leq l$. La courbe passait par les points de coordonnées$(x_i, \overline{Y_i})$ $1 \leq i \leq k$.

Commentaire

Lorsque $X=x_i$, la valeur qui synthétise le plus la variable Y est la moyenne conditionnelle de Y sachant $X = x_i$ $\overline{Y_i}$



$\mathcal{C}$

$y=ax+b$

Pour ajuster une droite à une courbe on utilise le critère des moindres carrées donné dans l'exemple par

$$C \cap C\;(C_{courbe}, D) = \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \sum_{i=1}^{n} (f(x_i) - ax_i - b)^2$$

Il faudra alors minimiser CMC(C,D) par rapport à a et b pour trouver la meilleure droite d'ajustement.

Liaison entre deux variables

Liaison nulle ente variables X et Y

Cela signifie qu'il n'y a pas d'influence d'une variable sur l'autre

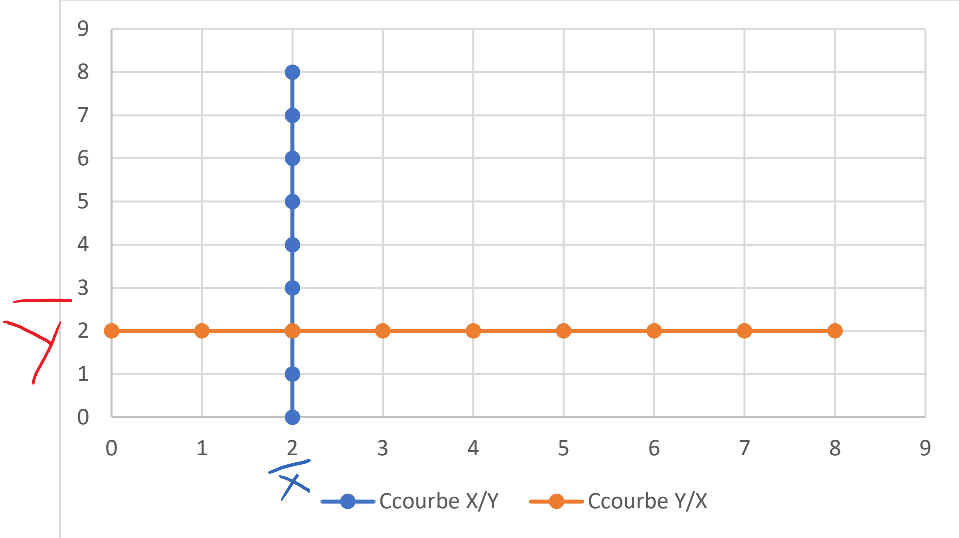
Les variables X et Y sont indépendantes, signifie que la variation de l'une des variables n'entraine pas d'effets sur l'autre variable.

Autrement dit, $f_{\frac{i}{j}} = f_{i.}$ $avec$ $1 \leq i \leq k$ et $f_{\frac{j}{i}} = f_{.j}$ $avec$ $1 \leq j \leq l$

Ainsi on a $\overline{X_j} = \overline{X}, 1 \leq j \leq l$ et $\overline{Y_i} = \overline{Y}\,, 1 \leq i \leq k$

Les courbes de régression ont pour équations alors :

$C_{courbe\frac{X}{Y}} = x = \bar{X}\ et\ C_{courbe\frac{Y}{X}} = y = \bar{Y}$



Liaison fonctionnelle

La liaison fonctionnelle est donnée par : $f(X) = Y$

Remarque

Lorsque X et Y présentent une liaison fonctionnelle, les courbes de régression $C_{courbe\frac{X}{Y}}\ et\ C_{courbe\frac{Y}{X}}$ sont confondues.

Liaison relative

Dans ce cas le nuage de points du vecteur (X,Y) est résumé par les courbes $C_{courbe\frac{X}{y}}\ et\ C_{courbe\frac{Y}{X}}$ qui se coupent au point centre de gravité $G(\bar{X}, \bar{Y})$.

Ces courbes donnent des informations sur la nature de la liaison entre les variables X et Y.

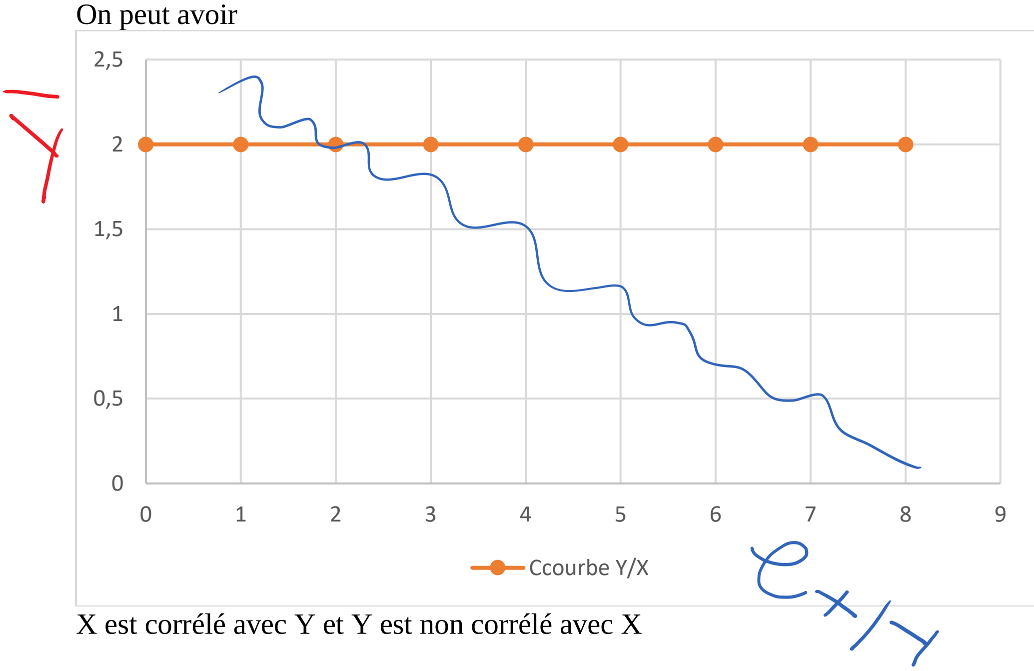
- On dira que la corrélation est positive si les deux variables varient dans le même sens.
- On dira que la corrélation est négative si les varient dans les sens opposés.
- On dira que la corrélation entre X et Y est linéaire si les courbes de régression sont des droites parallèles aux axes.

La corrélation est d'autant plus grande que l'angle formé par les courbes $C_{courbe\frac{X}{Y}}\ et\ C_{courbe\frac{Y}{x}}$ est petit

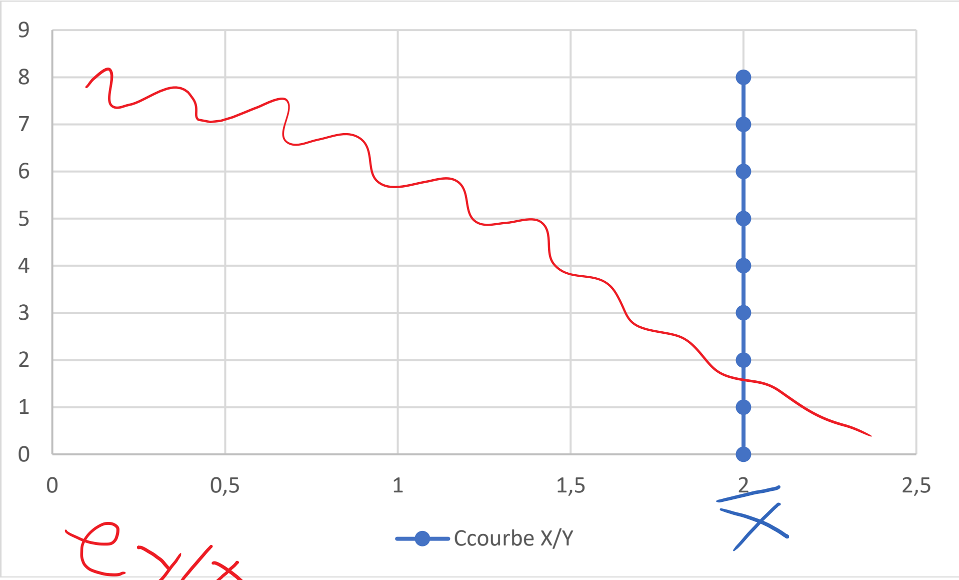
Remarque

A la différence avec l'indépendance la corrélation n'est pas propriété réciproque.

On peut avoir



X est corrélé avec Y et Y est non corrélé avec X



Y dépend de X et X ne dépend pas de Y.