Chapitre 1 — Statistique Descriptive

I- Vocabulaire:

a- Observations: Ce sont les données relatives à un phénomène obtenues au cours d'une enquête ; ou à la suite d'expériences

b- Population: C'est l'ensemble des éléments sur lequel porte l'étude (Etudiants, bovidés, pièces mécaniques ...)

c- Individu: Un élément de la population est appelé individu.

d- Caractère: C'est le trait particulier, ou l'aspect des individus qui nous intéresse ; par exemple: poids, tailles, Couleurs.

Notation: On note par:

$\Omega \equiv$ la population

$\omega \equiv$ individu ($\omega \in \Omega$)

$C \equiv$ l'ensemble des valeurs du caractère auquel on s'intéresse

Exemple:

$\Omega \equiv$ l'ensemble des étudiants(es) de 2ème année Agronomie (année 2003-2004). (population).

$\omega \equiv$ étudiant (individu)

On s'intéresse à la note obtenue par les étudiants en 1ère année

$C = \{10, 10{,}25, \dots, 17\}$ (Caractère).

Variable Statistique:

Soient Ω (population) et C (le caractère) les ensembles sur lesquels porte notre étude.

L'application: X de Ω vers C qui à chaque individu ω associe sa valeur du caractère $X(\omega)$ est appelée variable Statistique notée V.S.

$$X : \Omega \longrightarrow C$$
$$\omega \longmapsto X(\omega).$$

Exemple:

$$X : \{\text{Promotion de } 2^{\text{ème}} \text{ année Agronomie}\} \longrightarrow \{10, \dots, 17\}$$
$$\omega \longmapsto X(\omega) = \text{note de } \omega.$$

Si Ω contient un nombre fini d'individus ($\text{card}\,\Omega = N$)

$$\Omega = \{\omega_1, \omega_2, \dots, \omega_N\}$$

On note $X(\omega_i) = x_i$.

$$X : \Omega \longrightarrow C$$
$$\{\omega_1, \dots, \omega_N\} \longmapsto \{x_1, x_2, \dots, x_N\}$$
$$x_1 < x_2 < \dots < x_N.$$

Si : $X : \Omega \longrightarrow C$ et $\forall \omega \in \Omega$; $X(\omega)$ est un réel alors on dira que X est une V.S quantitative.

par contre si $X(\omega)$ n'est pas un réel

$$X : \Omega = \{\text{ensemble des voitures des enseignants}\} \longrightarrow \{\text{Couleurs}\}$$

X est dite V.S. qualitative

Etude d'une variable statistique :

V.S discrète

Si $X : \Omega \longrightarrow C$ est une v.a.s et si C contient un nombre fini d'éléments on dira que X est une V.S. discrète.

Exemple : Dans un village habite 500 familles ; on s'intéresse au nombre d'enfants de chaque famille

Après rescensements nous avons les données

112 familles ont 0 enfant
170 " " " 1 enfant
128 " " " 2 enfants
60 " " " 3 enfants
30 " " " 4 enfants.

Ω = { familles du village } (population)

w = famille (individu)

$C = \{ \underset{x_1}{0}, \underset{x_2}{1}, \underset{x_3}{2}, \underset{x_4}{3}, \underset{x_5}{4} \}$ le caractère ≡ nombre d'enfants.

Définition : Le nombre d'éléments de Ω est appelé effectif total noté N.

Le nombre d'éléments de $\{ w \in \Omega ; X(w) = x_i \}$ s'appelle effectif partiel de la valeur x_i

il est noté n_i ($n_i = \text{Card}\{\omega \in \Omega ; X(\omega) = x_i\}$.

Exemple: effectif total $N = 500$.

effectif partiel de la valeur $x_4 = 3$ est $n_4 = 60$.

Proposition: La somme des effectifs partiels est égal à l'effectif total : $\sum_i n_i = N$

Définition: Soit $X : \Omega \longrightarrow \{x_1, x_2, \dots, x_n\}$.

$N = \text{Card}\,\Omega$ effectif total

$n_i =$ effectif partiel de x_i

Le nombre $f_i = \frac{n_i}{N}$ s'appelle fréquence partielle de la valeur x_i

Exemple:

$f_1 = \frac{112}{500} = 0{,}224$; $f_2 = \frac{170}{500} = 0{,}340$; $f_3 = \frac{128}{500} = 0{,}256$; $f_4 = \frac{60}{500} = 0{,}120$, $f_5 = \frac{30}{500} = 0{,}06$

Proposition: $\sum_i f_i = 1$.

$$\sum_i f_i = \sum_i \frac{n_i}{N} = \frac{\sum_i n_i}{N} = \frac{N}{N} = 1.$$

Définition: Soit $X : \Omega \longrightarrow \{x_1, x_2, \dots, x_n\}$ une V.S. discrète dont f_i la fréquence partielle de x_i ; alors la loi de la V.S X est donnée par le tableau :

x_i	x_2	x_3 ...	x_n
f_1	f_2	f_3 ...	f_n

Exemple: la loi de notre exemple est donnée par:

x_i	0	1	2	3	et
f_i	0,224	0,340	0,256	0,120	0,060

On doit lire: il y'a 22,4 % de familles qui ont 0 enfant.
il y'a 12 % de familles qui ont 3 enfants.

Effectifs cumulés et fréquences cumulées:

Soit $X : \Omega \longrightarrow C = \{x_1, x_2, \dots x_m\}$.

n_i = effectif partiel de x_i $\quad i = 1, 2, \dots, m$.

f_i = fréquence partielle de x_i $\quad i = 1, \dots, m$.

alors le nombre $N_k = n_1 + n_2 + \dots + n_k$ s'appelle l'effectif cumulé de la valeur x_k.

et le nombre $F_k = f_1 + f_2 + \dots + f_k$ s'appelle fréquence cumulée de la valeur x_k. $\quad \left(F_k = \frac{N_k}{N} \right)$

Exemple:

x_i	0	1	2	3	4
n_i	112	170	128	60	30
N_i	112	282	410	470	500
f_i	0,224	0,340	0,256	0,120	0,06
F_i	0,1224	0,564	0,82	0,94	1

Interprétation: N_i = nombre d'individus dont la valeur du caractère est inférieure ou égale à x_i

$N_3 = 410$ = nombre de familles qui ont au plus 2 enfants

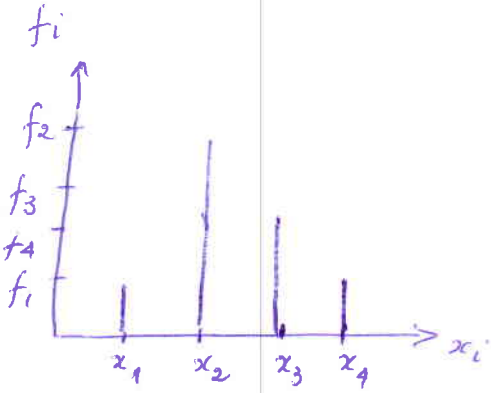
F_i = pourcentage d'individus dont la valeur du caractère est $\leq x_i$

$F_3 = 0,82 = 82\%$ de famille qui ont au plus 2 enfants.

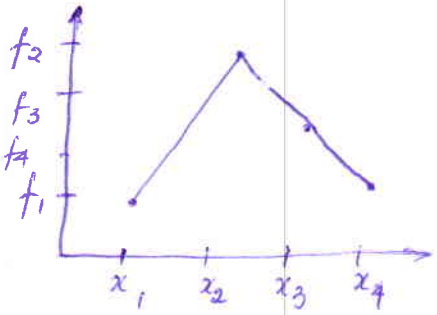
Représentation graphique:

Diagramme en batons:

On porte les fréquences partielles ou les effectifs partiels en ordonnée en fonction de x_i (ou les fréquences cumulées ou les effectifs cumulés).



Polygône des effectifs partiels.



Courbe cumulative des fréquences:

Soit $X : \Omega \longrightarrow C$ une v.s. discrète

Soit l'application: $F : \mathbb{R} \longrightarrow \mathbb{R}$, $x \longmapsto F(x)$

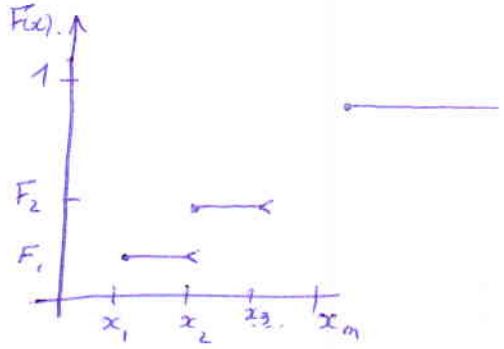
où $F(x)$ = pourcentage des individus dont la valeur du caractère est inférieure ou égale à x (F est appelée fonction de répartition).

La courbe représentative de cette fonction est appelée courbe cumulative des fréquences.

Propriétés:

1) $\forall i = 1, 2, \dots, m \quad F(x_i) = F_i$

2) $$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ F_i & \text{si } x \in [x_i, x_{i+1}[\\ 1 & \text{si } x \geq x_m. \end{cases}$$



3) On note $\lim_{\substack{x \to x_i \\ <}} F(x) = F(x_i - 0)$; $\lim_{\substack{x \to x_i \\ >}} F(x) = F(x_i + 0)$.

$F(x_i + 0) - F(x_i - 0) = f_i$

Variable statistique continue:

Soit : $X : \Omega \longrightarrow C$ ($\text{Card}(\Omega) = N$ = effectif total)

Si C contient un nombre "important" de valeurs alors on dira que X est une V. S. continue

Exemple:

Ω = Ensemble des nouveaux nés du mois de Février 2004 à l'hopital de Tlemcen. (1000)

ω = un nouveau né

$X(\omega)$ = poids de Ω (Kg).

$C = \{2, 2{,}01 \ldots, 4{,}5\}$.

dans ce cas on considère que $C = [2, 4{,}5]$ (un intervalle)

Lorsqu'il s'agira d'une V.S. continue C sera toujours considérés comme un intervalle. $X : \Omega \longrightarrow [x_{min}, x_{max}] = [\alpha, \beta]$

$\alpha = x_{min}$ = la plus petite valeur de x

$\beta = x_{max}$ = la plus grande valeur de x

Définition: le nombre noté e donné par $x_{max} - x_{min} = \beta - \alpha = e$ est appelé étendue de la variable statistique

$$e = x_{max} - x_{min} = \beta - \alpha.$$

—

Pour étudier une V.S. continue $X : \Omega \longrightarrow C = [\alpha, \beta]$ nous allons diviser l'intervalle C en m classes (tous intervalles) de même longueur h.

On prendra par convention $m \approx \sqrt{N}$

et on prendra h tel que $h \times m > e$.

en pratique h est la valeur la plus simple telle que $h > \frac{e}{m}$.

$C_1 = [\alpha, a_1[$, $C_2 = [a_1, a_2[$, $\ldots$, $C_m = [a_{m-1}, a_m[$ ($\beta \in [a_{m-1}, a_m[$)

Exemple:

Ω = Une récolte de pastèques (60 pastèques).

ω = une pastèque

$X(\omega)$ = diamètre de la pastèque (poids).

$C = [10, 16,1]$ Kgs

$N = 60$; étendue $e = 16,1 - 10 = 6,1$.

m nombre de classes $m \simeq \sqrt{N} = \sqrt{60}$

$m = 7$ classes.

On choisit h tel que $h > \frac{e}{m}$.

$\frac{e}{m} = \frac{6,1}{7} = 0,87$ $\quad h > 0,87$

On prendra $h = 1$ (par exemple).

On obtient ainsi les classes:

$C_1 = [10, 11[$; $C_2 = [11, 12[$, $C_3 = [12, 13[$, $C_4 = [13,14[$, $C_5 = [14, 15[$

$C_6 = [15, 16[$, $C_7 = [16, 17[$.

16,1

Effectif et fréquence d'une classe:

Soit $X : \Omega \longrightarrow C = [\alpha, \beta]$ une V.S. continue

Supposons qu'on divise C en n classes

$C_1 = [a_0, a_1[$, $C_2 = [a_1, a_2[$ ... $C_m = [a_{m-1}, a_m[$

$n_i = \text{Card}\{\omega \in \Omega, X(\omega) \in C_i\}$ s'appelle effectif partiel de la classe C_i

$f_i = \frac{n_i}{N}$ s'appelle fréquence partielle de la classe C_i

$N_k = n_1 + n_2 + \cdots + n_k$ s'appelle effectif cumulé de la classe C_k.

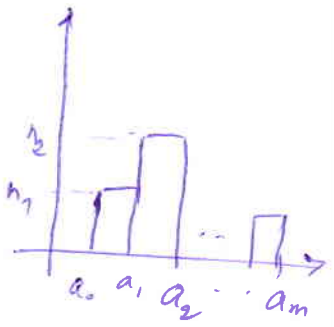
$F_k = f_1 + f_2 + \cdots + f_k$ s'appelle fréquence cumulée de la classe C_k.

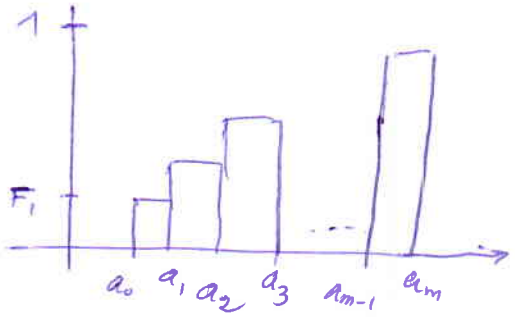
La loi de la variable statistique X est donnée par le tableau

c_i	c_1	c_2	...	c_m
f_i	f_1	f_2		f_m

Représentation graphique:

histogramme de effectifs (fréquences) partielles (cumulés)





Fonction de répartition:

$$X : \Omega \longrightarrow C = [a_0, a_1[\cup [a_1, a_2[\cup \cdots [a_{m-1}, a_m[$$

Soit $F : \mathbb{R} \longrightarrow \mathbb{R}$

$F(x)$ = pourcentage des individus dont la valeur du Caractère est inférieure ou égale à x. (fonction de répartition)

la courbe représentant F est appelée Courbe Cumulative des fréquences.

F est une fonction vérifiant:

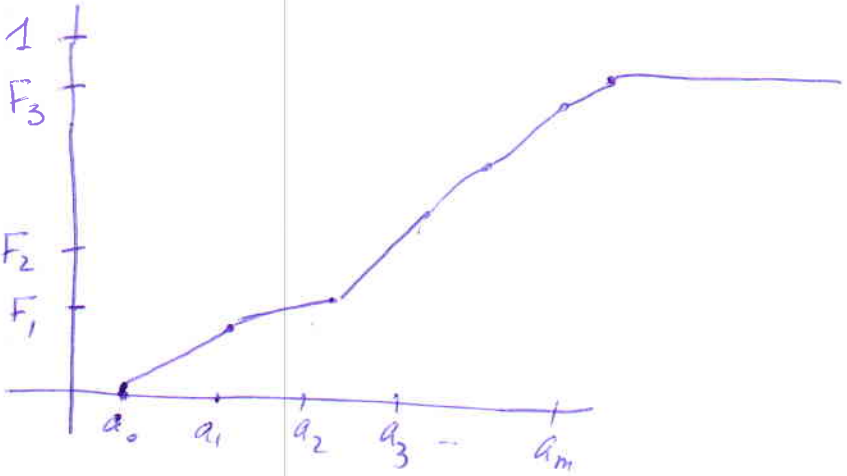
1- nulle sur l'intervalle $]-\infty, a_0[$

2- elle vaut 1 sur $[a_m, +\infty[$

3 elle passe par les points: $(a_1, 0)$, (a_1, F_1), (a_2, F_2), ..., $(a_m, F_n) = 1$

-11-

$$F(x) = \begin{cases} 0 & \text{si } x < a_0 \\ \frac{f_1}{h}(x - a_0) & \text{si } x \in [a_0, a_1[\\ F_i + \frac{f_{i+1}}{h}(x - a_i) & \text{si } x \in [a_i, a_{i+1}[\\ 1 & \text{si } x \geq a_m \end{cases}$$



• Indices de position et de dispersion :

• Indices de positions (ou de tendances centrales).

• Mode: noté M_0.

pour une V.S. discrète le mode noté M_0 est la valeur x_i du caractère qui a le plus grand effectif partiel (ou bien la plus grande fréquence partielle).

Ex(2) famille le mode est $x_2 = 1 = M_0$.

pour une V.S continue la classe modale est celle qui a le plus grand effectif partiel (ou la plus grande fréquence partielle).

le mode M_0 est le centre de la classe modale. C

Rq: le mode n'est pas nécessairement situé au centre

On peut avoir plusieurs modes

• Médiane: noté Me

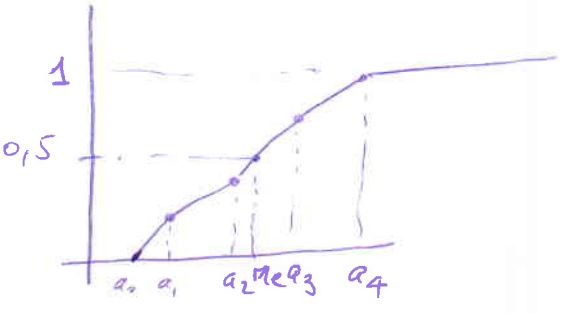
La médiane est la valeur de la variable statistique qui partage la population en deux

Ex: $Me = x_2 = 1$.

la médiane verifie $F(Me-0) < \frac{1}{2} \leq F(Me+0)$.

Si X est une V.S. Continue Me verifie $F(Me) = \frac{1}{2} = 0,5$

On retrouve Me graphiquement



ou alors par interpolation:

Soit $C_k = [a_k, a_{k+1}[$ la première classe telle que la fréquence cumulée est $\geq 0,5$ (50%). alors

$$\left(Me = a_k + \frac{0,5 - F_{k-1}}{F_k - F_{k-1}}\right)$$ F_k étant la fréquence cumulée.

$$Me = a_k + \frac{a_{k+1} - a_k}{f_k}(0,5 - F_{k-1})$$

b. Moyenne:

Soit $X : \Omega \longrightarrow C$ une V.S.

de loi : $\begin{array}{c|ccc} x_1 & x_2 & \dots & x_m \\ \hline f_1 & f_2 & & f_m \end{array}$ ou $\begin{array}{c|ccc} c_1 & c_2 & \dots & c_m \\ \hline f_1 & f_2 & & f_m \end{array}$.

la moyenne de la variable statistique est donnée par:

$$\bar{X} = \sum_{i=1}^{m} x_i f_i = x_1 f_1 + x_2 f_2 + \dots + x_m f_m$$

$$= \sum_{i=1}^{m} x_i \frac{n_i}{N} = \frac{1}{N}\left[\sum_{i=1}^{m} x_i \cdot n_i\right] \text{ si } X \text{ est discrète.}$$

et $\bar{X} = \sum_{i=1}^{m} f_i c_i = \frac{1}{N}\left[\sum_{i=1}^{m} n_i c_i\right]$ c_i est le centre de C_i

Ex: familles: $\bar{X} = \sum_{i=1}^{m} x_i f_i = 0 \times 0,224 + 1 \times 0,340 + 2 \times 0,256 + 3 \times 0,120 + 4 \cdot 0,06$

$= 1,45$

Propriétés:

1. $\sum_{i=1}^{m} f_i (x_i - \bar{x}) = 0$
2. Si $Y = aX + b$ $a, b \in \mathbb{R}$ alors $\bar{Y} = a\bar{X} + b$.

Preuve:

1. $\sum_{i=1}^{m} f_i (x_i - \bar{x}) = \sum_{i=1}^{m} f_i x_i - \sum_{i=1}^{m} f_i \bar{x} = \bar{x} - \bar{x} \sum_{i=1}^{m} f_i = \bar{x} - \bar{x} = 0$

2. $Y = aX + b$, $\bar{Y} = \sum_{i=1}^{m} f_i y_i = \sum_{i=1}^{m} f_i (a x_i + b) = a \sum_{i=1}^{m} f_i x_i + b \sum_{i=1}^{m} f_i$

$= a\bar{x} + b.$

Indices de dispersion:

• **Variance**: la variable Statistique $X : \Omega \longrightarrow \mathbb{C}$

ayant pour loi $\begin{array}{c|ccc} x_1 & x_2 & \dots & x_m \\ \hline f_1 & f_2 & & f_m \end{array}$ si X discrète

$\begin{array}{c|ccc} c_1 & c_2 & \dots & c_m \\ \hline f_1 & f_2 & & f_m \end{array}$

alors la variance de la variable X notée Var(X) est donnée par: $Var(X) = \sum_{i=1}^{m} f_i (x_i - \bar{x})^2$ ou $Var(X) = \sum_{i=1}^{m} f_i (c_i - \bar{x})^2$ c_i centre de C_i

$$= \sum_{i=1}^{m} f_i (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{m} f_i x_i^2 - 2\bar{x} \sum_{i=1}^{m} f_i x_i + \sum_{i=1}^{m} f_i \bar{x}^2$$

$$= \sum_{i=1}^{m} f_i x_i^2 - 2\bar{x}^2 + \bar{x}^2$$

$$= \sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2 \qquad \left(Var(X) = \sum_{i=1}^{m} f_i c_i^2 - \bar{x}^2 \right).$$

• **Ecart type**:

l'ecart type d'une V.S est donnée par:

$$\sigma_X = \sqrt{Var(X)} = \sqrt{\sum_{i=1}^{m} f_i (x_i - \bar{x})^2} = \sqrt{\left(\sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2 \right)}$$

Remarque:

σ_X donne une idée de la variation des valeurs de la V.S autour de $\bar{x}$.

si σ_X est petit alors les valeurs de la V.S sont regroupées autour de $\bar{x}$.

si σ_X est grand alors les valeurs de la V.S sont dispersées autour de $\bar{x}$.

Propriétés :

Si $y = ax + b$. $\quad a, b \in \mathbb{R}$.

$$Var(y) = a^2 Var(x).$$

Preuve :

$$Var(y) = \sum_{i=1}^{m} f_i y_i^2 - \bar{y}^2 = \sum_{i=1}^{m} f_i (ax_i + b)^2 - (a\bar{x} + b)^2$$

$$= \sum_{i=1}^{m} f_i (a^2 x_i^2 + b^2 + 2ab x_i) - (a^2 \bar{x}^2 + b^2 + 2ab\bar{x}^2).$$

$$= a^2 \sum_{i=1}^{m} f_i x_i^2 + b^2 \sum_{i=1}^{m} f_i + 2ab \sum_{i=1}^{m} f_i x_i - a^2 \bar{x}^2 - b^2 - 2ab\bar{x}$$

$$= a^2 \sum_{i=1}^{m} f_i x_i^2 + b^2 + 2ab\bar{x} - a^2 \bar{x}^2 - b^2 - 2ab\bar{x}$$

$$= a^2 \left(\sum_{i=1}^{m} f_i x_i^2 - \bar{x}^2 \right).$$

Chapitre II

Variable Statistique à deux dimensions :

Définition : Soient X, Y deux Variables Statistiques définies sur la même population Ω ; le couple $Z = (X, Y)$ est appelé variable statistique à deux dimensions. ($Card(\Omega) = N$ effectif total).

l'ensemble des valeurs prises par la V.S (X, Y) est donné par $\{(X(\omega_i), Y(\omega_i)) ; i = 1, 2, \dots N\}$.

ou bien par le tableau :

ω_i	ω_1	ω_2	ω_N
$X(\omega_i)$	$X(\omega_1)$	$X(\omega_2)$	$X(\omega_N)$
$Y(\omega_i)$	$Y(\omega_1)$	$Y(\omega_2)$	$Y(\omega_N)$

et on note $X(\omega_i) = x_i$
$Y(\omega_i) = y_i$

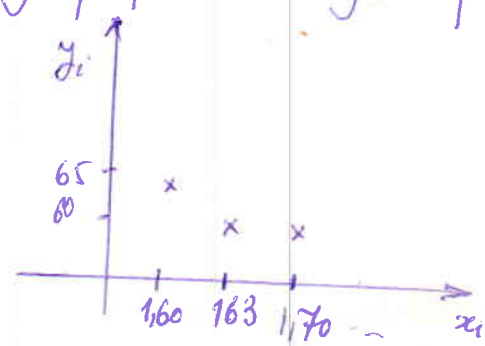
Exemple : Ω = ensemble de 10 personnes.

$X(\omega)$ = taille de ω (mètre). $Y(\omega)$ = poids(ω) (Kgs).

l'ensemble des valeurs de (X, Y) est donné par :

ω_i	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}
$X(\omega_i)$	1,60	1,63	1,70	1,72	1,75	1,77	1,80	1,81	1,87	1,90
$Y(\omega_i)$	65	60	63	63	68	69	68	75	80	81

Représentation graphique en nuage de points :



cette représentation s'appelle nuage de points (x_i, y_i).

Rq: Cette représentation se limite aux cas où $\text{car}(\Omega)$ fini.

2. Tableau de contingence des effectifs de la V.S (X,Y):

Soient $x_1, x_2, \dots, x_K$ les valeurs prises par la V.S X (ordonnées suivant les valeurs croissantes).

idem pour $y_1, y_2, \dots, y_M$ les valeurs prises par la V.S. Y.

Posons $n_{ij} = \text{Card}\{\omega \in \Omega ; (X(\omega), Y(\omega)) = (x_i, y_j)\}$

$=$ effectif de la valeur (x_i, y_j).

Le tableau de contingence des effectifs de la V.S (X,Y) est donné par le tableau suivant:

x_i \ y_j	y_1	y_2	$\dots$	y_j	$\dots$	y_M
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1M}
x_2	n_{21}	n_{22}	$\dots$	n_{2j}	$\dots$	n_{2M}
$\vdots$	$\vdots$					
x_i	$\vdots$			n_{ij}		
$\vdots$						
x_K	n_{K1}					n_{KM}

Exemple:

Ω = ensemble des familles d'une cité.

ω = une famille.

$X(\omega)$ = nbre d'enfants filles.

$Y(\omega)$ = nbre d'enfants garçons

Le tableau de contingence des effectifs de la variable (X,Y) est donnée par:

x_i \ y_j	y_1 = 0	y_2 = 1	y_3 = 2	y_4 = 3	y_5 = 4	
$x_1 = 1$	4	4	2	0	0	$\longrightarrow n_{1.} = 10$
$x_2 = 2$	9	16	4	0	0	$\longrightarrow n_{2.} = 29$
$x_3 = 3$	4	12	9	2	0	$\longrightarrow n_{3.} = 27$
$x_4 = 4$	1	6	1	1	2	$\longrightarrow n_{4.} = 11$
$x_5 = 5$	0	1	0	1	1	$\longrightarrow n_{5.} = 3$
	$n_{.1} = 18$	$n_{.2} = 39$	$n_{.3} = 16$	$n_{.4} = 4$	$n_{.5} = 3$	

Lecture du tableau:

par exemple: il y'a 12 familles qui ont: 3 filles (x_3) et 1 garçons (y_2)

$$n_{32} = 12.$$

On note $n_{i.} = \sum_{j=1}^{M} n_{ij}$; $n_{i.}$ = effectif de la valeur x_i

$n_{1.} = n_{11} + n_{12} + n_{13} + n_{14} + n_{15} = 4+4+2+0+0 = 10$

$n_{2.} = 9+16+4+0+0 = 29$; $n_{3.} = 4+12+9+2+0 = 27$

$n_{4.} = 1+6+1+1+2 = 11$; $n_{5.} = 0+1+0+1+1 = 3$

$n_{3.} = 27$ = effectif de la valeur x_3

= 27 familles ont 3 filles

de la même manière: $n_{.j} = \sum_{i=1}^{K} n_{ij} = n_{.j}$ = effectif de la valeur y_j.

$n_{.1} = 4+9+4+1+0 = 18$, $n_{.2} = 4+16+12+6+1 = 39$, $n_{.3} = 2+4+9+1+0 = 16$

$n_{.4} = 0+0+2+1+1 = 4$, $n_{.5} = 0+0+0+2+1 = 3$.

$n_{.2} = 39$ = effectif de la valeur y_2

= nombre de familles qui ont 1 garçon

Propriétés:

$$\sum_{i=1}^{K} n_{i.} = \sum_{j=1}^{M} n_{.j} = N \quad \text{effectif total.}$$

Exemple:

$\sum n_{i.} = 80$ 80 familles dans la cité.

$\sum n_{.j} = 80$

- Loi de la variable Statistique (X, Y):

le nombre $f_{ij} = \frac{n_{ij}}{N}$ est appelé fréquence de la valeur (x_i, y_j)

la loi de la variable statistique (X, Y) notée $\mathcal{L}(X, Y)$ est donnée par le tableau:

x_i \ y_j	y_1	y_2	...
x_1	f_{11}	f_{12}	
⋮			

Remarque:

si on note $f_{i.} = \sum_{j=1}^{M} f_{ij}$ et $f_{.j} = \sum_{i=1}^{K} f_{ij}$

alors $\left(f_{i.} = \frac{n_{i.}}{N} \text{ et } f_{.j} = \frac{n_{.j}}{N} \right)$

$f_{i.}$ = la fréquence de la valeur x_i

$f_{.j}$ = la fréquence de la valeur y_j

Propriétés

$$\sum_{i=1}^{K} f_{i\cdot} = \sum_{j=1}^{M} f_{\cdot j} = 1.$$

en effet :

$$\sum_{i=1}^{K} f_{i\cdot} = \sum_{i=1}^{K}\sum_{j=1}^{M} f_{ij} = \sum_{i=1}^{K}\sum_{j=1}^{M} \frac{n_{ij}}{N} = \sum_{i=1}^{K} \frac{n_{i\cdot}}{N} = \frac{N}{N} = 1.$$

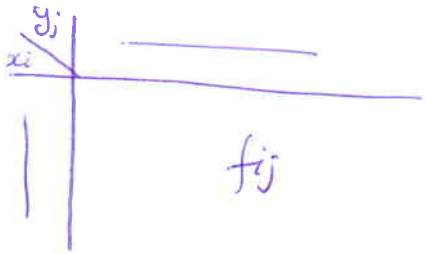
- Séries Marginales :

Soit (X,Y) une V.S à 2 dimensions.

La variable statistique X est appelée (série) V.S marginale

——— u ——— " Y ——— " ——— u ——— u ———.

Par la suite on supposera que la loi du couple (X,Y) est donnée



alors :

- la loi marginale de la V.S X notée $\mathcal{L}(X)$ est donnée par :

x_i	x_1	x_2	...	x_K
$f_{i\cdot}$	$f_{1\cdot}$	$f_{2\cdot}$		$f_{K\cdot}$

- la loi marginale de la V.S. Y notée $\mathcal{L}(Y)$ est donnée par :

y_j	y_1	y_2	...	y_M
$f_{\cdot j}$	$f_{\cdot 1}$	$f_{\cdot 2}$		$f_{\cdot M}$

Exemple: si l'on reprend l'exemple des familles.

La loi marginale de la V.S. X est donnée par:

x_i	1	2	3	4	5
$f_{i\cdot} = \frac{n_{i\cdot}}{N}$	0,125	0,3625	0,3375	0,1375	0,0375

y_j	0	1	2	3	4
$f_{\cdot j}$	0,225	0,4875	0,2	0,05	0,0375

- Moyennes marginales.

1/ $\bar{x} = \sum_{i=1}^{K} f_{i\cdot} x_i = \frac{1}{N} \sum_{i=1}^{K} x_i n_{i\cdot}$ s'appelle moyenne marginale de la V.S. X. ($Ex = 2,6 = \bar{x}$)

2/ $\bar{y} = \sum_{j=1}^{M} f_{\cdot j} y_j = \frac{1}{N} \sum_{j=1}^{M} n_{\cdot j} y_j$ s'appelle la moyenne marginale de la V.S. Y. ($N = \sum n_{i\cdot} = \sum n_{\cdot j}$).

(Ex ($\bar{y} = 1,1875$).

- Covariance du couple (X,Y)

La covariance du couple (X,Y) notée $Cov(x,y)$ est donnée par:

$$Cov(x,y) = \sum_{i=1}^{K} \sum_{j=1}^{m} f_{ij} (x_i - \bar{x})(y_j - \bar{y})$$

où $\bar{x}$ et $\bar{y}$ sont respectivement les moyennes de x et y.

Remarque:

- $Cov(x,y) = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{M} n_{ij} (x_i - \bar{x})(y_j - \bar{y})$ ($N = \sum n_{i\cdot} = \sum n_{\cdot j} = Card(\Omega)$.
- $Cov(x,y) = Cov(y,x)$
- $Cov(x,y) = \sum_{i=1}^{k} \sum_{j=1}^{m} f_{ij} (x_i y_j) - (\bar{x}\bar{y}) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{m} n_{ij} x_i y_j - (\bar{x}\bar{y})$

Remarques:

1) Si $x=y$ $\quad Cov(x,x) = \sum_{i=1}^{k} f_i x_i^2 - \bar{x}^2 = Var(x)$.

la notion de covariance généralise la notion de variance

2) la Covariance de (x,y) mesure le degré de dépendance des V.S x et y.

• Lois conditionnelles:

La distribution ou la loi conditionnelle de x sachant que $y=y_j$ notée $\mathcal{L}(x / y=y_j)$. (loi de x sachant que $y=y_j$) ; c'est la loi de la variable statistique qui prend les valeurs $x_1, x_2, \dots, x_k$ avec les fréquences $\frac{f_{1j}}{f_{\cdot j}}, \frac{f_{2j}}{f_{\cdot j}}, \dots, \frac{f_{kj}}{f_{\cdot j}}$

c'est donc la loi :

x_i	x_1	x_2	...	x_k
$\frac{f_{ij}}{f_{\cdot j}}$	$\frac{f_{1j}}{f_{\cdot j}}$	$\frac{f_{2j}}{f_{\cdot j}}$		$\frac{f_{kj}}{f_{\cdot j}}$

et la valeur: $\bar{x}_j = \sum_{i=1}^{k} \frac{f_{ij}}{f_{\cdot j}} x_i$ = moyenne conditionnelle sachant que $y=y_j$.

de la même manière la distribution ou la loi de y sachant que $x=x_i$ notée $\mathcal{L}(y / x=x_i)$ est la loi de la V.S qui prend les valeurs $y_1, y_2, \dots, y_m$ avec les fréquences $\frac{f_{i1}}{f_{i\cdot}}, \frac{f_{i2}}{f_{i\cdot}}, \dots \frac{f_{im}}{f_{i\cdot}}$

c'est donc la loi:

y_i	y_1	y_2	...	y_m
$\frac{f_{ij}}{f_{i\cdot}}$	$\frac{f_{i1}}{f_{i\cdot}}$	$\frac{f_{i2}}{f_{i\cdot}}$		$\frac{f_{im}}{f_{i\cdot}}$

et de la même manière la valeur :

$$\bar{y}_i = \sum_{j=1}^{m} \frac{f_{ij}}{f_{i.}} y_j$$ est la moyenne conditionnelle

sachant que $x = x_i$

Exemple récapitulatif :

Le tableau de contingence d'une V.S à deux dimensions est donné par :

x_i \ y_j	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$	
$x_1 = 1$	18	4	18	$\rightarrow n_{1.} = 40$
$x_2 = 2$	20	60	20	$\rightarrow n_{2.} = 100$
$x_3 = 3$	2	16	42	$\rightarrow n_{3.} = 60$
	$n_{.1} = 40$	$n_{.2} = 80$	$n_{.3} = 80$	$\Rightarrow N = 200$

1- Donner la loi de (X, Y), $\mathcal{L}(X, Y)$, (les lois marginales de x et y, les moyennes — " — " —)
les lois conditionnelles $\mathcal{L}(X / Y = y_j)$ les moyennes conditionnelles $\bar{x}_1, \bar{x}_2, \bar{x}_3$

x_i \ y_j	$y_1 = -1$	$y_2 = 0$	$y_3 = 1$	
$x_1 = 1$	0,09	0,02	0,09	$f_{1.} = 0,2$
$x_2 = 2$	0,1	0,3	0,1	$f_{2.} = 0,5$
$x_3 = 3$	0,01	0,08	0,21	$f_{3.} = 0,3$
	$f_{.1} = 0,2$	$f_{.2} = 0,4$	$f_{.3} = 0,4$	$= 1$

2- Donner la loi marginale de x.

x_i	$x_1 = 1$	$x_2 = 2$	$x_3 = 3$
$f_{i.}$	$0,2 = f_{1.}$	$f_{2.} = 0,5$	$f_{3.} = 0,3$

• moyenne marginale de x.

$$\bar{x} = \sum_{i=1}^{3} f_{i\cdot}\, x_i = 0{,}2 \times 1 + 0{,}5 * 2 + 0{,}3 \times 3 = 2{,}1$$

$$\bar{x} = 2{,}1.$$

• Loi marginale de Y

y_j	$y_1 = -1$	$y_2 = 0$	$y_3 = +1$
$f_{\cdot j}$	$f_{\cdot 1} = 0{,}2$	$f_{\cdot 2} = 0{,}4$	$f_{\cdot 3} = 0{,}4$

• moyenne marginale de Y

$$\bar{y} = \sum_{j=1}^{3} f_{\cdot j}\, y_j = (-1)\, 0{,}2 + 0 \cdot 0{,}4 + 0{,}4 = 0{,}2$$

$$\bar{y} = 0{,}2.$$

• $Cov(x,y) = \sum_{i=1}^{k} \sum_{j=1}^{km} f_{ij}\,(x_i y_j) - (\bar{x}\bar{y})$

$$= \sum_{i=1}^{3} \sum_{j=1}^{3} f_{ij}\,(x_i y_j) - (\bar{x}\bar{y}).$$

$$= f_{11}(x_1 y_1) + f_{12}(x_1 y_2) + f_{13}(x_1 y_3) + f_{21}(x_2 y_1) + f_{22}(x_2 y_2)$$

$$+ f_{23}(x_2 y_3) + f_{31}(x_3 y_1) + f_{32}(x_3 y_2) + f_{33}(x_3 y_3).$$

$$= 0{,}09\,(-1)\,1 + 0{,}02 \times 0 \times 1 + 0{,}09\; 1 \times 1 + 0{,}1\;\; 2 \times -1 + 0{,}3 \cdot 0 \times 2$$

$$+ 0{,}1 \times 1 \times 2 + 0{,}01 \times 3 \times (-1) + 0{,}08\; 3 \cdot 0 + 0{,}21 \times 3 \times 1 - 2{,}1 \times 0{,}2$$

$$= 0{,}18 - 0{,}42.$$

$$= -0{,}24$$

• Donner les lois conditionnelles : $\mathcal{L}(X/Y=1=y_1)$, $\mathcal{L}(X/Y=0=y_2)$ $\mathcal{L}(X/Y=1=y_2)$. et les moyennes conditionnelles correspondantes :

• $\mathcal{L}(X/Y=y_1=-1)$.

x_i	1	2	3
f_{i1} / $f_{\cdot 1}$	$\frac{0,09}{0,2}=0,45$	$\frac{0,1}{0,2}=0,5$	$\frac{0,01}{0,2}=0,05$

$\bar{x}_1 = 1\times 0,45 + 2\times 0,5 + 3\times 0,05 = 1,6$ (moyenne conditionnelle sachant que $Y=y_1$)

• $\mathcal{L}(X/Y=y_2=0)$.

x_i	1	2	3
f_{i2} / $f_{\cdot 2}$	$\frac{0,02}{0,4}=0,05$	$\frac{0,3}{0,4}=0,75$	$\frac{0,08}{0,4}=0,2$

$\bar{x}_2 = 1\times 0,05 + 2\times 0,75 + 3\times 0,2 = 2,15$. (moyenne conditionnelle sachant que $Y=y_2$)

• $\mathcal{L}(X/Y=y_3=+1)$.

x_i	1	2	3
f_{i3} / $f_{\cdot 3}$	$\frac{0,09}{0,4}=0,225$	$\frac{0,1}{0,4}=0,25$	$\frac{0,21}{0,4}=0,525$

$\bar{x}_3 = 1\times 0,225 + 2\times 0,25 + 3\times 0,525 = 2,3$. (moyenne conditionnelle sachant que $Y=y_3$).

• Calculer : $\bar{x}_1 \cdot f_{\cdot 1} + \bar{x}_2 f_{\cdot 2} + \bar{x}_3 f_{\cdot 3} = 1,6 \cdot 0,2 + 2,15 \cdot 0,4 + 2,3 \times 0,4$

$= 2,1 = \bar{x}$.

Proposition :

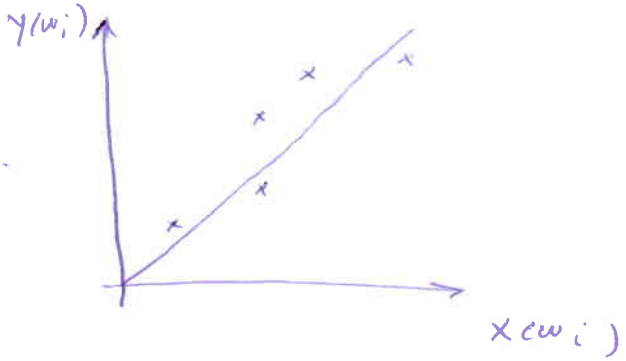
$$\bar{x} = \sum_{j=1}^{m} \bar{x}_j f_{\cdot j} \quad \text{et} \quad \bar{y} = \sum_{i=1}^{k} \bar{y}_i f_{i\cdot}$$

$\bar{y}_i$, $\bar{x}_j$ étant les moyennes conditionnelles et $\bar{x}$, $\bar{y}$ les moyennes marginales.

• Corrélation et droite de régression:

Il arrive que dans une étude statistique qu'on mesure sur chaque individu w 2 caractères (2 V.S); et par la suite on examine s'il existe une dépendance (ou corrélation) entre les 2 V.S

(cela peut être remarqué après représentation en nuage de points).



Rappel:

$$Cov(x,y) = \sum_{i=1}^{k}\sum_{j=1}^{m} f_{ij}\, x_i y_j - \bar{x}\bar{y}$$

$$\bar{x} = \sum_{i=1}^{k} f_{i\cdot}\, x_i, \quad \bar{y} = \sum_{j=1}^{m} f_{\cdot j}\, y_j$$

$$Var(x) = \sum_{i=1}^{k} f_{i\cdot}\, x_i^2 - \bar{x}^2, \quad Var(y) = \sum_{j=1}^{m} f_{\cdot j}\, y_j^2 - (\bar{y})^2$$

$$\sigma_x = \sqrt{Var(x)}\ ; \quad \sigma_y = \sqrt{Var(y)}$$

Définition: On appelle coefficient de corrélation entre les V.s x et y le nombre ρ défini par : $\rho = \dfrac{Cov(x,y)}{\sigma_x\, \sigma_y}$

Propriétés: • le coefficient ρ mesure le degré de liaison linéaire entre x et y.

• $|\rho| \leq 1$ (ie $-1 \leq \rho \leq 1$).

• Si $\rho = 1$ ou $\rho = -1$ x et y sont liées par une relation linéaire
• Si $\rho \approx +1$ ou $\rho \approx -1$ tous les points x_i, y_j se trouvent près d'1 même droite
• Si $\rho = 0$ il y'a indépendance entre x et y.

Droite de regression:

Nous cherchons à déterminer une droite d'équation:

$y = ax + b$ qui puisse approcher le nuage de points dans le sens où $Q(\alpha, \beta) = \frac{1}{N} \sum_{i=1}^{N} (Y(w_i) - \alpha X(w_i) - \beta)^2$ est minimale

$Q(\alpha, \beta)$ représente la moyenne de la somme des carrées de la distance des points $(X(w_i), Y(w_i))$ et $(X(w_i), \alpha X(w_i) + \beta)$.

On cherche donc à minimiser la quantité $Q(\alpha, \beta)$

et on aura $Q(a, b) = \min Q(\alpha, \beta)$.

_Définition: a et b sont tels que:

$$a = \frac{Cov(X,Y)}{Var(X)} \quad \text{et} \quad b = \bar{y} - a\bar{x} = \bar{y} - \frac{Cov(X,Y)}{Var(X)}\bar{x}$$

la droite de régression de y en x est donnée par

$$y = ax + b = \frac{Cov(X,Y)}{Var(X)} \cdot x + \left(\bar{y} - \frac{Cov(X,Y)}{Var(X)}\bar{x}\right)$$