Statistiques :

Chapitre 1 : Série statistiques - Généralités

I) Définitions

Population : Ensemble d'éléments faisant l'objet d'une étude statistique

Exemples : Ensemble des élèves d'un amphi (âge, poids, taille…)
Ensemble des pièces fabriqué par une machine (taille, qualité…)
Ensemble des assurés d'une compagnie d'assurance (localisation, nombre de sinistres)

Echantillon : Partie de la population

Exemples : Groupe E des élèves
Pièces fabriqués le lundi
Une sélection au hasard de 100 pièces
Tous les assuré dans le 92

Unité statistique : Un élément de la population dont on veut étudier un ou plusieurs caractères. Il peut être qualitatif (couleur, lieu d'habitation…), ou quantitatif (taille, l'age…)

Variable statistique : Elle peut être discrète (des valeurs isolées) ou continue (valeurs sur un intervalle).

Série statistique : Ensemble des valeurs prises par une variable statistique sur l'ensemble de la population ou de l'échantillon

II) Série statistique d'un caractère quantitatif discret

Soit un échantillon de taille n, avec (x_1, x_2, x_3 … x_n), les valeurs possibles du caractère x.

Série statistique : les n valeurs prises par les n membres de l'échantillon.
Effectif total : n
Effectif partiel de x_i : C'est le nombre de fois n_i où la valeur x_i apparait. La somme de ces effectifs partiel est égal à $n : n_1 + n_2 + ... + n_p = n$

Fréquence relative de x_i : $f_i = \dfrac{n_i}{n}$ La somme de ces fréquences relatives est égal à n :

$f_1 + f_2 + ... + f_p = n$

Etendu de la série : Ecart entre la plus grande et la plus faible des valeurs de x_i

Exemple :

Population : 100 familles de 4 enfants
Caractère étudié : Nombre de garçon
Série statistique : 3, 1, 2, 2, 0, 3, 4, 2…

xi	0	1	2	3	4
ni	7	20	43	25	5

Effectif total : $n = 100$
Etendu : 4 - 0 = 4

Fréquence relative des familles de 2 garçon : $f_2 = \dfrac{43}{100} = 43\%$

III) Série statistique d'un caractère quantitatif continu

Le nombre p de valeurs possible de x_i est infini. On va alors constituer des classes en divisant l'étendue de la série en un certain nombre d'intervalle.

Exemple :
On pèse des nourrissons à la naissance (à 10g près). On a pesé 161 nourrissons qui pesaient entre 2,24kg et 4,48kg.
On fait 8 classes de 0,3kg :

Intervalle	[2,2 ; 2,5[	[2,5 ; 2,8[	[2,8 ; 3,1[	[3,1 ; 3,4[	[3,4 ; 3,7[	[3,7 ; 4,0[	[4,0 ; 4,3[	[4,3 ; 4,6[
Milieu de l'intervalle	2,35	2,65	2,95	3,25	3,55	3,85	4,15	4,45
n_i	5	11	24	40	42	20	13	6
Fréquence	3,1 %	6,8 %	14,9 %	24,8 %	26,1 %	12,4 %	8,1 %	3,7 %

IV) Série statistique d'un caractère qualitatif

On groupe les résultats en autant de classes qu'il existe de modalité du caractère
Exemple :

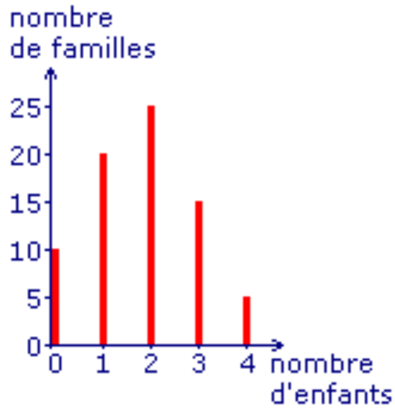
On étudie la couleurs des fleurs : 3 couleurs possibles (rouge, vert, bleu), réparti en 3 classes avec un effectifs par classe.

Variables nominales : Pas de classement possible (couleur, profession…)
Variable ordinales : Si les modalités peuvent être ordonnées (taille vestimentaire…)
Variable dichotomiques : Si seulement 2 valeurs (sexe)

V) Représentation graphique des séries statistiques

1) Cas discret :

Diagramme en bâtons : n_i en fonction de x_i (ou avec les fréquences)



Polygone des effectifs/fréquence absolues :
On va rejoindre l'extrémité des bâtons par des segments

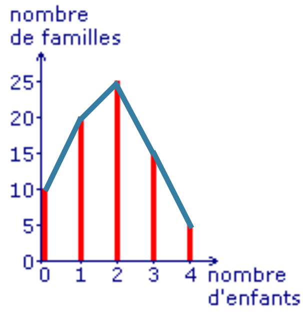
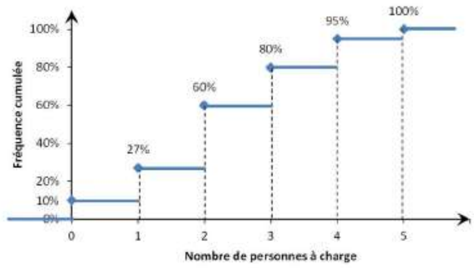


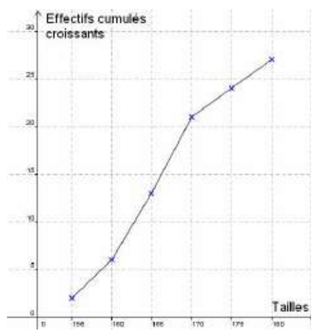
Diagramme cumulatif :



Effectif cumulé jusqu'à la i^{ème} valeur (n_1 +…+n_i)

x_i	0	1	2	3	4	5
n_i	10	17	33	20	15	5
$\sum ni$	10	27	60	80	95	100

Polygone des effectifs cumulés :

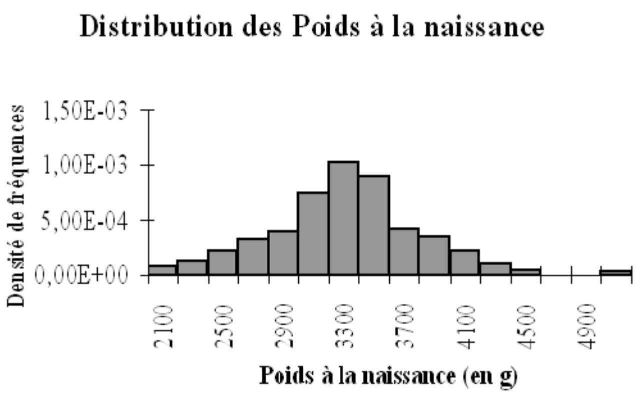


(On place les points a gauche de la classe)

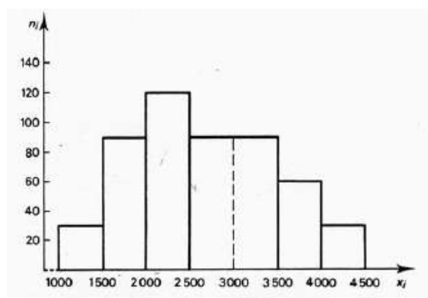
2) Cas continu :

Diagramme en bâton impossible car il y a trop de valeurs de x_i.

Histogramme :



La surface de est proportionnelle à l'effectif (on rend donc les classes égales quand elles ne le sont pas).



Chapitre 2 : Paramètres de position et de dispersion

Objectif : On veut obtenir quelques paramètres pour condenser l'information contenu dans une série statistique.

Paramètres de position : donnent un ordre de grandeur de ce qui est mesuré (x_i) et l'existence de valeurs centrales.

Paramètres de dispersion : donnent la dispersion des valeurs de la séries autours de paramètres de position.

I) Paramètres de position

On a une série statistique avec n valeurs $x_1, x_2, \cdots, x_n$.

1) Moyenne arithmétique :

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Série à caractères discret : On a p de valeurs $x_1, x_2, \cdots, x_n$.

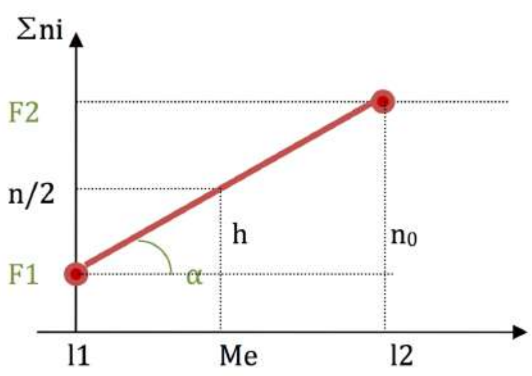
$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{n}$$

Série à caractères discret : On a p classes de centre $x_1, x_2, \cdots, x_n$.

$$\bar{x} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_p x_p}{n}$$

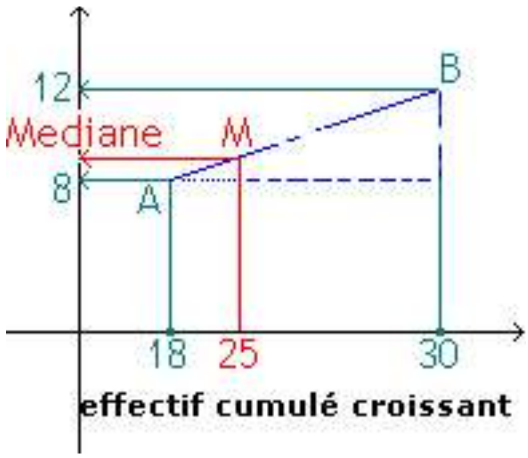
2) Médiane

Série à caractères discret : On va ordonner les valeurs de la série et on va prendre la valeur qui coupe la série en 2.
2 méthodes :
Graphique : On trace les effectifs cumulés, et on prend la valeur sur laquelle tombe l'effectif ½.
Analytique : La médiane tombera au milieu d'un intervalle $[l1, l2[$. On suppose les valeurs uniformément répartis dans l'intervalle.



$$Me = l1 + \left(\frac{n}{2} - F1\right)\frac{l2 - l1}{F2 - F1}$$

Autre méthode :



Grâce au théorème de Thalès, on peut trouver la médiane.

3) Quartiles

Série à caractères continu : On utilise la même méthode que la médiane, mais avec des valeurs de l_1, l_2, F_1, F_2 différentes

4) Mode ou valeur dominante

C'est la valeur la plus fréquente (on peut avoir plusieurs modes).

Série à caractères continu : On cherche la classe modale (où l'effectif est le plus élevé).

Le mode est : $Mod = l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2}(l_2 - l_1)$

Δ_1 est la différence entre la hauteur de la classe -1 et de la classe modale.

Δ_2 est la différence entre la hauteur de la classe +1 et de la classe modale.

II) Paramètres de dispersion

1) Etendue

Etendue : Ecart entre la plus grande valeur et la plus petite valeur

2) Ecart moyen

Soit une série statistique de valeurs x_i, de i allant à n, de moyenne $\overline{x}$.

Ecart par rapport à x_i : $e_i = \left|x_1 - \overline{x}\right|$

Ecart moyen : $\overline{e} = \frac{e_1 + e_2 + ... + e_n}{n}$

3) Variance et écart type

Variance : $\sigma x^2 = \frac{\left(x_1 - \overline{x}\right)^2 + \left(x_2 - \overline{x}\right)^2 + ... + \left(x_n - \overline{x}\right)^2}{n}$

Ecart type : $\sigma x = \sqrt{\sigma x^2}$

Propriété variance : $\sigma x = a \times \sigma z \Leftrightarrow \sigma x^2 = a^2 \times \sigma z^2$

4) Ecart interquartile

$Q_3 - Q_1$

5) Coefficient de variation :

Intérêt pour calculer valeurs sans unité (exemple : taille, poids) :

$$CdV = \frac{\sigma x}{\bar{x}}$$

Chapitre 3 : Ajustement de séries statistiques

On se donne une série statistique, continu ou discrète.

Cas 1 : Quand on trace n_i en fonction de x_i , les points sont à peu près aligné

Cas 2 : Quand on trace n_i en fonction de x_i , les points forment à peu près une parabole.

Problème : identifier et déterminer la courbe d'ajustement.
Cela permet de remplacer p couples de x_i, n_i par 2 ou 3 paramètres, de calculer des valeurs de n_i ou des x_i non relevé.

I) Méthode de lissage

Pas réellement une méthode d'ajustement. Permet de réduire les variations de la courbe pour faciliter ensuite la détermination de la courbe d'ajustement.

Plusieurs méthodes :
- Méthode des points médians (cf formulaire)
  - On trace la courbe
  - On détermine les maxima et minima locaux
  - On construit une enveloppe supérieur et inférieur qui passe par ces extrema
  - Tracer un trait vertical pour chaque valeur de x_i et trouver le point médian.
  - Tracer la nouvelle courbe qui relie tous les points médian
- Méthode de la moyenne mobile
  - On remplace chaque y par $\dfrac{y_{i-1}+y_i+y_{i+1}}{3}$
  - On trace la nouvelle courbe

II) Méthode d'ajustement

1) Méthode graphique

- Essentiellement pour un ajustement linéaire
- Méthode : On trace une droite sur un calque et on positionne le calque pour avoir autant de points au dessus et au dessous de la droite
  Avec 2 points de la droite positionné, on déduit l'équation de la droite.

Remarque : Méthode subjective qui donne plusieurs résultats possibles.

2) La méthode des moindres carrés

Méthode :
- Sur la courbe d'ajustement, on note y'_i les valeurs ajustées de y .
- L'écart des y par rapport a cette courbe est donné par la quantité

$$(y_1 - y'_1)^2 + (y_2 - y'_2)^2 + ... + (y_p - y'_p)^2$$

- La forme de la courbe d'ajustement peut être une droite, une parabole, exponentielle…
- Une fois choisie, on cherche les paramètres de l'équation qui minimise l'écart.

Remarques : Pas de subjectivité dans la méthode, ainsi, tout le monde aura le même résultat.

Exemples :

1) Ajustement linéaire

La courbe ajustée a pour equation $y = ax + b$ donc $y'_i = ax_i + b$

On cherche alors le minimum de $\sum_{i=1}^{p}(y_i - y'_i)^2 \Rightarrow f(a,b) = \sum_{i=1}^{p}(y_i - ax_i - b)^2$

Pour trouver le minimum, on va calculer les dérivée partiel par rapport à a et à b .

Dérivée par rapport à a : $\dfrac{df(a,b)}{da} = \sum_{i=1}^{p}\left(-2x_iy_i + 2x_ib + 2ax_i^2\right) = -2\sum_{i=1}^{p}(y_i - ax_i - b)x_i$

Dérivée par rapport à b : $\dfrac{df(a,b)}{db} = \sum_{i=1}^{p}(-2y_i + 2ax_i + 2b) = -2\sum_{i=1}^{p}(y_i - ax_i - b)$

On attend le minimum quand la dérivée passe de négative à nulle.

$$\frac{df(a,b)}{db} = 0 \Rightarrow \sum_{i=1}^{p}(y_i - ax_i - b) = 0$$

$$\sum_{i=1}^{p} y_i = a\sum_{i=1}^{p} x_i + \sum_{i=1}^{p} b$$

$$\boxed{\sum_{i=1}^{p} y_i = a\sum_{i=1}^{p} x_i + pb} \text{ (1)}$$

$$\frac{1}{p}\sum_{i=1}^{p} y_i = \frac{a}{p}\sum_{i=1}^{p} x_i + b$$

$$\boxed{\overline{y} = a\overline{x} + b} \text{ (2)}$$

$$\frac{df(a,b)}{da} = 0 \quad \Rightarrow \boxed{\sum_{i=1}^{p} x_iy_i = a\sum_{i=1}^{p} x_i^2 + b\sum_{i=1}^{p} x_i} \text{ (3)}$$

On isole a et b par combinaison linéaire et on obtient :

$$\begin{cases} a = \dfrac{\overline{x_iy_i} - \overline{x_i}\,\overline{y_i}}{\overline{x_i^2} - \left(\overline{x_i}\right)^2} \\ b = \overline{y_i} - a\overline{x_i} \end{cases}$$

Qualité de l'ajustement :

$$\underbrace{\sum_{i=1}^{p}\left(y_i - \overline{y}\right)^2}_{\substack{\text{Ecart total de } y_i \\ \text{par rapport à } \overline{y}}} = \underbrace{\sum_{i=1}^{p}(y_i - y'_i)^2}_{\substack{\text{Ecart résiduelle de} \\ y_i \text{ à la valeur estimé} \\ \text{de } y'_i}} + \underbrace{\sum_{i=1}^{p}\left(y'_i - \overline{y}^2\right)}_{\substack{\text{Ecart expliquée entre} \\ y_i \text{ et } \overline{y} \text{ ne dépendant} \\ \text{que de } x_i \text{ et pas de } y_i}}$$

On défini le coefficient de correlation par $R^2 = \dfrac{\text{écart expliqué}}{\text{écart total}}$ (valable même dans un cas non linéaire)

Si $R \approx 1$, l'ajustement est bon.

Si $R \approx 0$, l'ajustement est mauvais

2) Ajustement à l'aide d'une parabole

La courbe ajustée : $y = ax^2 + bx + c \Rightarrow y_i' = ax_i^2 + bx_i + c$

La méthode des moindres carré, on va chercher le minimum de la fonction $f(a,b,c) = \sum_{i=1}^{p}(y_i - y_i')^2$

Donc $f(a,b,c) = \sum_{i=1}^{p}(y_i - ax_i^2 - bx_i - c)^2$

Méthode : On va faire les dérivée partiels de a,b et c :

Dérivée par rapport à c : $\frac{df}{dc} = \sum_{i=1}^{p} -2(y_i - ax_i^2 - bx_i - c) = 0 \Leftrightarrow \sum_{i=1}^{p} y_i = a\sum_{i=1}^{p} x_i^2 + b\sum_{i=1}^{p} x_i + c \underbrace{p}_{\substack{nb\\de\\points}}$

Dérivée par rapport à b : $\frac{df}{db} = \sum_{i=1}^{p} -2x_i(y_i - ax_i^2 - bx_i - c) = 0 \Leftrightarrow \sum_{i=1}^{p} x_i y_y = a\sum_{i=1}^{p} x_i^3 + b\sum_{i=1}^{p} x_i^2 + c\sum_{i=1}^{p} x_i$

Dérivée par rapport à a : $\frac{df}{da} = \sum_{i=1}^{p} -2x_i^2(y_i - ax_i^2 - bx_i - c) = 0 \Leftrightarrow \sum_{i=1}^{p} x_i^2 y_y = a\sum_{i=1}^{p} x_i^4 + b\sum_{i=1}^{p} x_i^3 + c\sum_{i=1}^{p} x_i^2$

Exemple du formulaire n°4 :

	xi	yi	xi²	xi³	xi⁴	xi yi	xi² yi	yi²
	1	6	1	1	1	6	6	36
	2	5	4	8	16	10	20	25
	3	2	9	27	81	6	18	4
	5	3	25	125	625	15	75	9
	7	9	49	343	2401	63	441	81
Somme	18	25	88	504	3124	100	560	155

Donc d'après les equation obtenu grâce aux dérivées partiel :

$88a + 17b + 5c = 25$

$504a + 88b + 18c = 100$

$3124a + 504b + 88c = 560$

Par combinaison linéaire, on obtient :

$a = 0,61$

$b = -4,47$

$c = 10,41$

Quand on calcul le taux de corrélation, on obtient $R^2 = 0,97$

Chapitre 4 : Indices statistiques

Cela sert à suivre l'évolution d'une grandeur dans le temps :
- Indice des prix à la consommation (IPC)
- Indice de la production industrielle (IPI)

On distingue 2 notions d'indices :
- Indice simple (une seule grandeur étudié)
- Indice synthétique (plusieurs grandeurs)

I) Indice simple

Exemple : On considère un materiel qui coute 50€ en 2014 et 54€ en 2015

On peut regarder les prix de différentes façon :
- L'augmentation : 4€
- L'augmentation relative : 8%
- Le rapport de prix : $\frac{54}{50} = 1{,}08$
- L'indice simple : $\frac{54}{50} \times 100 = 108$

Définition : Soit P_0, la valeur à une date t_0 de référence d'une certaine grandeur.

Soit P_1, la valeur à une autre date t_1 de la même grandeur

L'indice simple de la grandeur à la date t_i, calculé sur la base 100 est : $I_{1/0} = \frac{P_1}{P_0} \times 100$

Propriété :

Un indice simple est réversible : $i_{1/0} \times i_{0/1} = 1$

Un indice simple est transitif : $i_{2/1} \times i_{1/0} = i_{2/0}$

Remarque :

On a 2 hausses de suite de 8%, on a donc une hausse de 16,64% et non 16%

II) Indice synthétiques

Plusieurs produits avec des prix L_i et des quantité Q_i
On peut définir 3 types d'indices :

- Indice des prix
- Indice des quantités
- Indice des valeurs

	t0 = 2014		t1 = 2015	
Produits	Prix	Quantités	Prix	Quantités
A	2	4	3	6
B	3	3	4	2
C	4	3	5	4

Exemple :
Indice des augmentation des prix :

Pondération de t_0 (indice de Laspeyres) :
Pondération : les quantités de 2014 ou de 2015

Moyenne des prix en 2014 (t_0): $2{,}9€$

Moyenne des prix en 2015 (t_1)(on conserve les quantité de 2014) : $3{,}9€$

D'où l'indice de la moyenne des prix : $I_{1/0} = \dfrac{3{,}9}{2{,}9} \times 100 = 134{,}5$

Sinon, indice de pondération de t_1 (indice de Paasche) :
Moyenne des prix en 2014 (t_0): $2{,}8€$

Moyenne des prix en 2015 (t_1)(on conserve les quantité de 2014) : $3{,}8€$

$$I'_{1/0} = \frac{3{,}8}{2{,}8} \times 100 = 136$$

Moyenne des indices des prix :

On calcul les indices simples de $A = \left(\dfrac{3}{2} \times 100\right)$, de $B = \left(\dfrac{4}{3} \times 100\right)$ et $C = \left(\dfrac{5}{4} \times 100\right)$

Pondération de t_0 (indice de Laspeyres) : $I_{1/0} = \dfrac{A \times 4 + B \times 3 + C \times 3}{4+3+3} = 138$

Pondération de t_1 (indice de Paasche) : $I'_{1/0} = \dfrac{A \times 6 + B \times 2 + C \times 4}{6+2+4} = 139$

Cas général :

Définitions :

Soit P_{i0} prix des produits à t_0

Soit P_{i1} prix des produits à t_1

Soit Q_{i0} quantités des produits à t_0

Soit Q_{i1} quantités des produits à t_1

Pondération de Laspeyres :
Indice moyenne des prix :

$$L_{1/0} = \frac{\sum P_{i1}Q_{i0}}{\sum P_{i0}Q_{i0}} \times 100$$

Moyenne des indices :

$$L'_{1/0} = \frac{\sum \frac{P_{i1}}{P_{i0}}Q_{i0}}{\sum Q_{i0}} \times 100$$

Pondération de Paasche :
Indice moyenne des prix :

$$P_{1/0} = \frac{\sum P_{i1}Q_{i1}}{\sum P_{i0}Q_{i1}} \times 100$$

Moyenne des indices :

$$P'_{1/0} = \frac{\sum \frac{P_{i1}}{P_{i0}}Q_{i1}}{\sum Q_{i1}} \times 100$$

Indice de Fischer : $F_{1/0} = \sqrt{L_{1/0} \times P_{1/0}}$

Chapitre V : Série Statistique à 2 caractères

I) Distribution marginales conditionnelle

a) Tableau de contingence

- Population de n unités statistiques
- Pour chaque unité, on va observer 2 caractères : poids + taille ou âge + département…
  - 1er caractère, on prend les valeurs $x_1, x_2 \dots x_p$
  - 2ème caractère, on prend les valeurs $y_1, y_2 \dots y_q$

Effectif partiel : n_{ij} = nombre d'observation du couple (x_i, y_i)
On fait un tableau à double entrée :

	y1	y2	yj	yq	Total
x1	n11	n12	n1j	n1q	n1o
x2	…	…	…	…	…
xi	ni1	ni2	nij	niq	nio
xp	np1	np2	npj	npq	npo
Total	no1	no2	noj	non	n

Distribution statistique d'une seul caractère :
- x : valeurs $\left(x_i, n_{io} = effectif\ \ marginal\ \ de\ \ x_i\right)$ = nombre de valeur de x_i
- y : valeurs $\left(y_j, n_{oj} = effectif\ \ marginal\ \ de\ \ x_i\right)$ = nombre de valeur de x_i

$n_{io} = \sum_{j=1}^{q} n_{ij}$ et $n_{oj} = \sum_{i=1}^{p} n_{ij}$ avec $\sum_{i=1}^{p} n_{io} = \sum_{j=1}^{q} n_{oj} = n$

Terminologie :
n_{ij} = effectif partiel

$f_{ij} = \dfrac{n_{ij}}{n}$ = fréquence partiel

n_{io}, n_{oj} = effectif marginaux

$f_{io} = \dfrac{n_{io}}{n} \quad f_{oj} = \dfrac{n_{oj}}{n}$ = fréquence marginale

Exemple :
Distribution de $10\ 000$ jeunes salariés selon l'âge (x) et le salaire mensuel en milliers d'euros (y)

Tableau de contingence :

	[0,8-1,2[	[1,2-1,6[	[1,6-2,0[	Somme
19-21	700	300	100	1100
21-23	1400	2000	300	3700
23-25	1000	2900	1300	5200
Total	3100	5200	1700	10000

Fréquence partielle : $f_{32} = \frac{n_{32}}{n} = \frac{2900}{10000} = 29\%$

Fréquence marginale de x : $f_{1o} = \frac{n_{1o}}{n} = \frac{1100}{10000} = 11\%$

$y : f_{o1} = \frac{n_{o1}}{n} = \frac{3100}{10000} = 31\%$

c) Distribution conditionnelles :

Ce sont les distributions d'un caractère pour une modalité fixée de l'autre.

Fréquence conditionnelle :

$f_{i/j} =$ fréquence de $x = x_i$ sous la condition que $y = y_j$

$f_{j/i} = \frac{n_{ij}}{n_{io}}$ $\qquad$ $f_{i/j} = \frac{n_{ij}}{n_{oj}}$

$f_{i/j} =$ proportion d'unité ayant x_i parmi ceux ayant y_i

$=$ fréquence conditionnelle de x selon y

Exemple : $f_{i=1/j=2} = \frac{n_{12}}{n_{o2}} = \frac{300}{5200} = 5,8\%$

Donc $5,8\%$ des gens qui gagent entre $1,2$ et $1,6$ milliers d'euros ont entre 19 et 21 ans.

II) Caractéristiques des séries à 2 caractères

a) Caractéristiques des séries marginales

Moyenne marginale de x : $\sigma x^2 = \frac{1}{n}\sum_{i=1}^{p} n_{io}x_i = \frac{1}{n}\sum_{i=1}^{p}\sum_{i=1}^{q} n_{ij}x_i$

Moyenne marginale de y : $\sigma y^2 = \frac{1}{n}\sum_{j=1}^{p} n_{oj}y_j = \frac{1}{n}\sum_{j=1}^{q}\sum_{j=1}^{p} n_{ij}y_j$

Exemple :

$\bar{x} = \frac{1}{10000}(1100 \times 20 + 3700 \times 22 + 5200 \times 24) = 22,82$

$\bar{y} = \frac{1}{10000}(3100 \times 1 + 5200 \times 1,4 + 1700 \times 1,8) = 1,344$ milliers d'euros

Variance marginale de $x : \sigma x^2 = \frac{1}{n}\sum_{i=1}^{p} n_{io}\left(x_i - \overline{x}\right)^2 = \overline{x^2} - \left(\overline{x}\right)^2$

Variance marginale de $y : \sigma y^2 = \frac{1}{n}\sum_{i=1}^{q} n_{oj}\left(y_i - \overline{y}\right)^2 = \overline{y^2} - \left(\overline{y}\right)^2$

Exemple :

$$\overline{x^2} = \frac{1}{10000}\left(1100\times 20^2 + 3700\times 22^2 + 52000\times 24^2\right) = 522{,}6$$

$$\Rightarrow \sigma x^2 = 522{,}6 - (22{,}82)^2 = 1{,}85$$

$$\Rightarrow \sigma = 1{,}36$$

$$\overline{y^2} = \frac{1}{10000}\left(3100\times 1^2 + 5200\times 1{,}4^2 + 1700\times 1{,}8^2\right) = 1{,}88$$

$$\Rightarrow \sigma y^2 = 1{,}88 - (1{,}344)^2 = 0{,}0736$$

$$\Rightarrow \sigma = 0{,}271$$

b) Caractéristiques des distributions conditionnelles

Moyenne conditionnelle de x si y_j : $\overline{x_j} = \frac{1}{n_{oj}}\sum_{i=1}^{p} n_{ij}x_i$

Moyenne conditionnelle de y si x_i : $\overline{y_i} = \frac{1}{n_{io}}\sum_{j=1}^{q} n_{ij}y_j$

Variance conditionnelle de x si y_j est fixé : $\sigma_j^{\,2}(x) = \frac{1}{n_{oj}}\sum_{i=1}^{p} n_{ij}\left(x_i - \overline{x}\right)^2 = \overline{x_j^{\,2}} - \left(\overline{x_j}\right)^2$

Variance conditionnelle de y si x_i est fixé : $\sigma_i^{\,2}(y) = \frac{1}{n_{io}}\sum_{j=1}^{q} n_{ij}\left(y_j - \overline{y}\right)^2 = \overline{y_i^{\,2}} - \left(\overline{y_i}\right)^2$

Exemple :

$$\underbrace{\overline{x_1}}_{j=1} = \frac{1}{3100}\left(700\times 20 + 1400\times 22 + 1000\times 24\right) = \frac{68800}{3100} = 22{,}194$$

= moyenne d'age de ceux qui gagnent entre 800 et 1200 euros

$$\underbrace{\sigma_1^{\,2}(x)}_{j=1} = \frac{1}{3100}\left(700\times 20^2 + 1400\times 22^2 + 1000\times 24^2\right) - (22{,}194)^2 = 2{,}15$$

$$\Rightarrow \sigma_1(x) = 1{,}47$$

Dispersion d'age de ceux qui gagnent entre 800 et 1200 euros.

$$\underbrace{\overline{y_2}}_{i=2} = \frac{1}{3700}\left(1400\times 1 + 2000\times 1{,}4 + 300\times 1{,}8\right) = 1{,}281$$

Salaire moyen de ceux qui ont entre 21 et 23 ans

$$\sigma_2(y)^2 = \frac{1}{3700}\left(1400\times 1^2 + 2000\times 1{,}4^2 + 300\times 1{,}8^2\right) - (1{,}281)^2 = 0{,}0596$$

$\sigma_2(y) = 0{,}24$ = dispersion au salaire pour les jeunes entre 21 et 23 ans.

c) Relation entre les caractéristiques marginales et conditionnelles

- Sur les moyennes :
  - On a $\overline{x} = \frac{1}{n}\sum_{j=1}^{q} n_{oj}\overline{x_j}$
  - $\overline{y} = \frac{1}{n}\sum_{i=1}^{p} n_{io}\overline{y_i}$

d) Covariance de x et y

Définition : Cela sert à quantifier le degré d'indépendance de x et y .

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left(x_i - \overline{x}\right)\left(y_j - \overline{y}\right) = \overline{xy} - \overline{x}\,\overline{y}$$

III) Représentation graphique des séries à 2 caractères

Série $\left(x_y, y_j\right)$ d'effectif n_{ij}

Courbe de régression :

On va tracer les moyennes conditionnelles en fonction de x_i ou y_j

Les moyennes conditionnels sont en y, et les valeurs de x_i ou y_i seront en x.

III) Séries à caractères indépendant

Définition : Les caractères sont indépendants si une variation de l'un n'entraine pas une variation de l'autre, ou encore, les fréquences conditionnelles $f_{i/j}$ ne dependent pas de j .

Conséquences : $f_{i/j} = f_{io}$ et $f_{j/i} = f_{oj}$

Alors : $f_{ij} = f_{io} \times f_{oj}$

$\overline{x} = \overline{x_j}$ et $\overline{y} = \overline{y_i}$

Si x et y sont indépendants alors $\operatorname{cov}(x,y) = 0$ ou $\overline{xy} = \overline{x}\,\overline{y}$

$\operatorname{cov}(x,y) = 0$ n'entraine pas l'indépendance.