# Chapitre 3

# Séries statistiques à deux variables

**Distribution statistique à deux dimensions**

* Soient deux variables X et Y définies sur une même population d’effectif total N (X et Y deux caractères quantitatifs ou qualitatifs ou une qualitative et l’autre quantitative).

* Définition : On appelle série statistique double associée aux caractères *X* et *Y*  l'ensemble des couples de donné par *{(x*1 ,*y*1),………., *(x*n ,*y*n)}
* *(x*i ,*y*j) nij
* nij l'effectif conjoint de *X*i et *Yj* : c'est le nombre d'individus pour lesquels X prend la valeur *x*p et Y la valeur yp,

**Présentation en tableau (tableau de contingence)**

Le [tableau de contingence](http://w3.mi.parisdescartes.fr/smel/lexique/table_contingence/table_contingence.html) est un moyen particulier de représenter simultanément deux caractères observés sur une même population, s'ils sont discrets ou bien continus et regroupés en classes. Les deux caractères sont x et y.



* ni. (total de la ligne i ) est le nombre total d’individus présentant la modalité xi indépendamment des modalités de Y.
* Les ni. sont appelés effectifs marginaux de X.

Telle que

* n.j (total de la colonne j ) est le nombre total d’individus présentant la modalité yj  indépendamment des modalités de X.
* Les n.j sont appelés effectifs marginaux de Y.

Telle que

* On appelle l’effectifs du couple de modalités (xi, yj), le nombre nij d’individus présentant simultanément les deux modalités.

* On appelle fréquence du couple de modalités (xi , yj), la proportion fij d’individus présentant simultanément les deux modalités.

* Les fi. sont appelés fréquence marginaux de X.

* Les f.j sont appelés fréquence marginaux de Y.

**Exemple**

Soit la répartition des salaires d’une entreprise selon le nombre d’enfant (X) et le salaire mensuel(Y) en 1000DA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  X y  | [2-6[ | [6-10[ | [10-14[ | ni. | fi. |
| 1  | 15  | 8  | 2  | 25  | 0.42  |
| 2  | 13  | 4  | 1  | 18  | 0.3  |
| 3  | 11  | 3  | 3  | 17  | 0.28  |
| n.j  | 39  | 15  | 6  | 60  |  |
| f.j | 0.65 | 0.25 | 0.1 |  | 1 |

**N=60** le nombre total des salariés de l’entreprise

**n22** salariés ont chacun 2 enfants et gagnent un salaire entre 6000 et 10000 DA

**n13** salariés ont chacun 1 enfant et gagnent un salaire entre 10000 et 16000 DA

**n2.** salariés ont chacun 2 enfants et gagnent un salaire entre 2000 et 14000 DA

**Distribution marginales**

Considérons la colonne de droite du tableau de contingence. Les effectifs ni. représentent les individus présentant la modalité xi indépendamment des modalités du second caractère étudié Y.

On dit qu’ils définissent la distribution marginale de X.

(cette série statistique est une série statistique à un seul caractère).

On définir alors la fréquence marginale de la modalité xi par:

**Moyennes marginales.**

Les moyennes marginales sont les moyennes des variables marginales X et Y :

**Variances marginales.**

Les variances marginales sont les variances des variables marginales X

**Exemple1**

Reprenons l’exemple précédent

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X (nb enfants)  | ni.  | fi. | fi..xi | fi..xi² |
| 1  | 25  | 0.42  | 0.42  | 0.42  |
| 2  | 18  | 0.3  | 0.6  | 1.2  |
| 3  | 17  | 0.28  | 0.84  | 2.52  |
| ∑  | 60  | 1  | 1.86  |  |

**Exemple2**

Reprenons l’exemple précédent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Y (salaires)  | n.j  | f.j  | cj  | f.j×yj  | f.j×yj ²  |
| [2-6[  | 39  | 0.65  | 4  | 2.6  | 10.4  |
| [6-10[  | 15  | 0.25  | 8  | 2  | 16  |
| [10-14[  | 6  | 0.1  | 12  | 1.2  | 14.4  |
| ∑  | 60  | 1  | -  | 5.8  | 40.8  |

**Distributions conditionnelles**

* C’est une distribution suivant l’une des deux caractères (X,Y) liée par une modalité de l’autre caractère.
* On dit qu’elle définit la **distributions conditionnelles de X sachant que Y=yj**

* on peut définir les fréquences conditionnelles associées par

* représente la proportion d’individus présentant la modalité xi parmi l’ensemble des individus présentant la modalité yj de Y.



* On dit qu’elle définit la **distributions conditionnelles de Y sachant que X=xi**

* on peut définir les fréquences conditionnelles associées par

* représente la proportion d’individus présentant la modalité yj parmi l’ensemble des individus présentant la modalité xi de X.

**Exemple**

la distribution conditionnelle de X sachant que Y=y2

|  |  |  |
| --- | --- | --- |
| X (nb enfants)  | ni/2 {y2 [6-10[}  | fi/2  |
| 1  | 8  | 0.53  |
| 2  | 4  | 0.27  |
| 3  | 3  | 0.2  |
| ∑  | 15  | 1  |

* Les fréquences conditionnelles associées sont définies par:

* fi/2 représente la proportion d’individus présentant la modalité y2 parmi l’ensemble des individus présentant la modalité xi de X.

**Covariance**

X et Y deux variables quantitatives définies sur une même population composée de n individus.

On appelle covariance du couple de variable (X,Y), le nombre noté COV(X,Y), défini par:

La covariance est un paramètre qui peut servir à établir une dépendance linéaire entre deux variables quantitatives.

Lorsque les données organisées dans un tableau à double, la formule de la covariance s’écrire

 **Exemple**

Reprenons l’exemple précédent, calculer la covariance

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  Y(Cj)X  | 4  | 8  | 12  |  |  |  |
| 1  |  |  |  |  |  |  |
| 2  |  |  |  |  |  |  |
| 3  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

**Coefficient de corrélation linéaire**

La coefficient de corrélation définir par: 

* -1 < r < 1
* Si r = 1 ou r = -1 alors points parfaitement alignés



**Exemple**

 Reprenons l’exemple précédent, calculer la coefficient de corrélation

**Nuage de points**

* l’ensemble des points *Mij* de coordonnées (*xi*;*yj*), est appelé *le nuage de points* associé à cette série statistiques à deux variables.
* Le point *Mij* correspond son axe abscisse est la valeur *xi* , et son ordonnée la valeur *yj* .



**Point moyen du nuage**

* Notons la moyenne des valeurs *xi* et la moyenne des valeurs *yi*. Avec les notations précédentes, on a :

 et

* Le point *G* de coordonnées est appelé *le point moyen* du nuage de points associé à cette série statistique à deux variables.

**Droite de régression**
L’idée est de transformer un nuage de point en une droite. Celle-ci doit être la plus proche possible de chacun des points. On cherchera donc à minimiser les écarts entre les points et la droite



Pour cela, on utilise la méthode des moindres carrées. Cette méthode vise à expliquer un nuage de points par une droite qui lie *Y* à *X*, c’est à dire, 
L’erreur est la différence entre le point réellement observé et le point prédit par la droite. La méthode des moindres carrés consiste à chercher la valeur des paramètres *a* et *b* qui minimise la somme des erreurs élevées au carré.





**Méthode d’ajustement linéaire de Mayer**

Elle consiste à partager un nuage de points rangés dans l’ordre croissant de leurs abscisses en deux sous groupes de même effectif. Chacun des deux sous-groupes est alors remplacé par le point dont les coordonnées sont respectivement :

en abscisse, la moyenne arithmétique des abscisses des points du sous-groupe.

en ordonnée, la moyenne arithmétique des ordonnées des points du sous-groupe.

Par ces points que l’on nomme G1 et G2 passe une seule droite qui sera la droite d’ajustement. On détermine l‘équation de la droite à partir des coordonnées de ces deux points.

 ***G1 : y1= a x1 + b***

 ***G2 : y2= a x2 + b***

a et b sont les inconnues que l'on cherche.

x1 et y1 sont les coordonnées du premier point (G1)

x2 et y2 sont les coordonnées du deuxième point (G2)



Exemple

 