Chapitre 3 : Séries statistiques à deux variables :

L'objectif est d'analyser la distribution des valeurs des variables x et y et le lien éventuel entre elles.

1. Nuage des points :

Un graphique qui traduise les deux séries statistiques à l'aide de diagramme à 2 dimensions. Soit x et y deux variables statistiques numériques observées sur k individus. Dans un repère orthogonal$(o, \vec{\imath}, \vec{\jmath})$, l'ensemble des k points de coordonnées (x_i ; y_i) forme le nuage de points associé à cette série statistique.

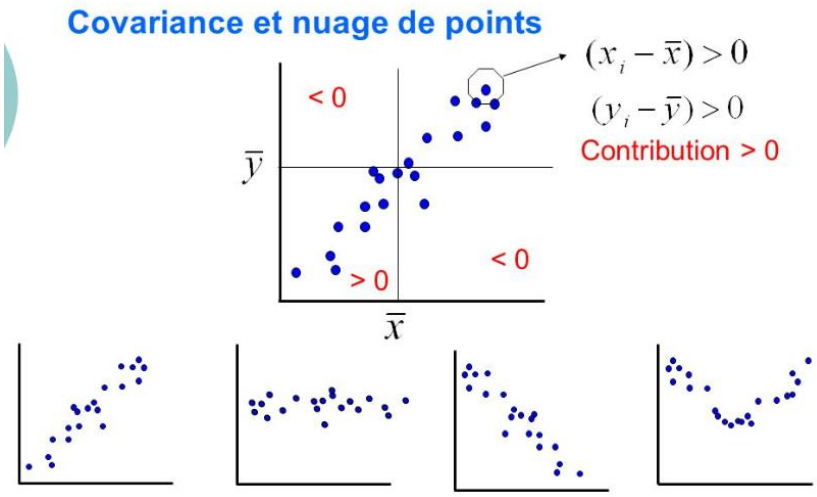

Figure 10 : Nuage de points

2. Tableaux de données (tableau de contingence)

- Tableau de contingence des effectifs :

- Variable discrètes : Paires de valeurs = (x_1, y_1) ; (x_2, y_2) ; (x_3, y_3) ; … ; (x_n, y_n)

Chaque effectif local n_{ij} correspond au nombre d'individus ayant l'abscisse x_i et l'ordonnée y_i.

Y / X	y_1	y_2	…	y_j	…	y_q
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2q}
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	…
.	.	.	…	.	…	.
.	.	.	…	.	…	.
.	.	.	…	.	…	.
x_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}

Tableau 6: Tableau de contingence « Variable discrètes »

- Variables continues : Ce sont les modalités centrales (x_1, x_2, x_3,…, x_p) et (y_1, y_2,

$y_3,\ldots, y_p$) des classes qui remplacent les modalités discrètes.

Chaque effectif local n_{ij} correspond au nombre d'individus dont leurs valeurs x appartiennent à la classe $[a_{i-1}, a_i[$ et leurs valeurs y appartiennent à la classe $[b_{j-1}, b_j[$.

Y / X		$[b_0,b_1[$	$[b_1,b_2[$	…	$[b_{j-1},bj[$	…	$[b_{q-1},b_q[$
		y_1	y_2	…	y_j	…	y_q
$[a_0,a_1[$	x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}
$[a_1,a_2[$	x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2q}
.	.	.	.	…	.	…	.
.	.	.	.	…	.	…	.
.	.	.	.	…	.	…	.
$[a_{i-1},a_i[$	x_i	n_{i1}	n_{i2}	…	n_{ij}	…	…
.	.	.	.	…	.	…	.
.	.	.	.	…	.	…	.
.	.	.	.	…	.	…	.
$[a_{p-1},a_p[$	x_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}

Tableau 7: Tableau de contingence « Variable continue »

- Tableau de contingence des fréquences : On garde les mêmes tableaux précédents et on divise tous les effectifs (locaux marginaux) par l'effectif total n.

3. Distribution marginales et conditionnelles, Covariance :

3.1. Distribution marginales :

On ajoute au tableau de contingence les totaux en ligne et en colonne.

Y / X	y_1	y_2	…	y_j	…	y_q	Distribution marginale de Y n_{i*}	
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}	n_{1*}	Sommes des n_{ij} de la ligne
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2q}	n_{2*}	
. . .	. . .	. . .	… … …	. . .	… … …	. . .	. . .	
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	…	n_{i*}	
. . .	. . .	. . .	… … …	. . .	… … …	. . .	. . .	
x_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}	n_{p*}	
Distribution marginale de X n_{*j}	n_{*1}	n_{*2}	…	n_{*j}	…	n_{*q}	$n_{**} = n$	Effectif total
	Sommes des n_{1j} de la colonne						Effectif total	

Tableau 8: Distribution marginales « Variable discrète »

	Y / X	$[b_0,b_1[$	$[b_1,b_2[$	…	$[b_{j-1},b_j[$	…	$[b_{q-1},b_q[$	Distribution marginale de Y n_{i*}	
		y_1	y_2	…	y_j	…	y_q		
$[a_0,a_1[$	x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1q}	n_{1*}	Sommes des n_{ij} de la ligne
$[a_1,a_2[$	x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2q}	n_{2*}	
. . .	. . .	. . .	. . .	… … …	. . .	… … …	. . .	. . .	
$[a_{i-1},a_i[$	x_i	n_{i1}	n_{i2}	…	n_{ij}	…	…	n_{i*}	
. . .	. . .	. . .	. . .	… … …	. . .	… … …	. . .	. . .	
$[a_{p-1},a_p[$	x_p	n_{p1}	n_{p2}	…	n_{pj}	…	n_{pq}	n_{p*}	
distribution marginale de X n_{*j}		n_{*1}	n_{*2}	…	n_{*j}	…	n_{*q}	$n_{**} = n$	Effectif total
		Sommes des n_{1j} de la colonne						Effectif total	

Tableau 9 : Distribution marginales « Variable continue »

- En marge à droite (totaux en ligne) : la distribution de X : pour chaque indice i, l'effectif n_{i*} est le nombre total d'observations de la modalité x_i de X quelle que soit la modalité de Y. C'est-à-dire $n_{i*} = \sum_{j=1}^{q} n_{ij}$= Total de la ligne i.

Les p couples (x_i, n_{i*}) définissent la distribution marginale de la variable X.

- En marge en bas (totaux en colonne) : la distribution de Y : pour chaque indice j, l'effectif n_{*j} est le nombre total d'observations de la modalité y_j de Y quelle que soit la modalité de X. C'est-à-dire $n_{*j} = \sum_{i=1}^{p} n_{ij}$= Total de la colonne j.

Les q couples (y_j, n_{*j}) définissent la distribution marginale de la variable Y.

Remarque :

$$\sum_{i=1}^{p} n_{i*} = \sum_{j=1}^{q} n_{*j} = n$$

3.2. Distribution conditionnelles :

- La distribution des observations suivant les modalités de la variable Y sachant que la variable X prend la modalité x_i, est appelée distribution conditionnelle de Y pour $X = x_i$: A la ligne i du tableau de contingence, on lit la distribution de la variable Y sachant que $X = x_i$, notée $Y|_{X=xi}$.

- La distribution des observations suivant les modalités de la variable X sachant que la variable Y prend la modalité y_j, est appelée distribution conditionnelle de X pour $Y = y_j$: A la colonne j du tableau de contingence, on lit la distribution de la variable X sachant que $Y = y_j$, notée $X|_{Y=yj}$.

3.3. Covariance :

On appelle covariance de la série statistique double de variables x et y le nombre réel

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$$

Remarque :

$$\mathrm{Cov}(x, x) = V(x)$$

4. Coefficient de corrélation linéaire :

Le coefficient de corrélation linéaire est un nombre permettant de déterminer l'intensité d'un lien linéaire entre deux variables quantitatives.

Le coefficient de corrélation linéaire d'une série statistique de variables x et y est le nombre r défini par : $\rho = \frac{Cov\,(x\,,y)}{\sigma_x \sigma_y}$

Remarque:

- Le coefficient de corrélation est une valeur qui n'a pas d'unité et qui est toujours comprise entre -1 et +1
- Un coefficient de corrélation linéaire est positif indique un lien linéaire positif, alors que, si ρ est négatif, le lien linéaire entre les deux variables est négatif.
- Plus la valeur de ρ est près de -1 ou +1 plus le lien linéaire ente les deux variables est fort.

5. Droite de régression et droite de Mayer

5.1. Droite de régression

Une droite de régression est la droite qui s'ajuste le mieux à un nuage de points présentant une corrélation linéaire. La droite de régression sert à faire des prévisions. On parle de corrélation linéaire lorsque les points d'un nuage ont tendance à s'aligner. Plus la tendance est forte, plus la corrélation linéaire est forte.

La droite D d'équation y = ax + b est appelée droite de régression de y en x de la série statistique si la quantité suivante est minimale : $S = \sum_{i=1}^{n}[y_i - (ax_i + b)]^2$

Pour définir les coefficients a et b; on développe S et on considère successivement comme un trinôme en b; puis, b étant déterminé, comme un trinôme en a: On trouve :

$$a = \frac{Cov\,(x\,,y\,)}{\sigma_x^2}$$

$$b = \bar{y} - a\,\bar{x}$$

5.2. Droite de Mayer

Cet ajustement consiste à déterminer la droite passant par deux points moyens du nuage de point.

6. Courbes de régression, couloir de régression et rapport de corrélation

6.1. Courbes de régressions

Une courbe de régression permet d'analyser la relation entre deux variables (variable explicative et variable expliquée) et de mettre en avant la nature de cette relation sans faire aucune hypothèse préalable sur la forme de celle-ci. Elle fait correspondre les valeurs de la première variable avec les moyennes conditionnées de la seconde. On peut donc à partir de deux variables X et Y construire deux courbes de régression :

- La courbe de X en Y : elle met en relation les valeurs de Y (y_i) et les moyennes conditionnelles de X.
- La courbe de Y en X : elle met en relation les valeurs de X (x_i) et les moyennes conditionnelles de Y.

Mise à part le fait de donner une image, une représentation de la forme de la liaison existant entre deux variables, les courbes de régression présentent des propriétés importantes :

.elles résument le mieux la silhouette du nuage de points puisqu'elles sont en moyenne les plus proches des points de ce nuage (la somme et les moyennes des carrés des distances entre les points du nuage et les courbes de régression sont minimales)

.elles se coupent en un point qui représente le point de gravité du nuage de points (i.e. un point ayant pour coordonnées approximatives les moyennes marginales des deux variables).

6.2. Rapport de corrélation

Pour étudier la relation entre une variable qualitative et une variable quantitative, on décompose la variation totale en variation intergroupe(ou interclasse) et en variation intragroupe(ou intraclasse). Pour mesurer l'intensité de la relation, on peut calculer un paramètre appelé rapport de corrélation.

7. Ajustement fonctionnel

Lorsque l'on veut modéliser les données, la question qui se pose ensuite est de trouver L'équation des courbes de régression. Les courbes de régression correspondent généralement à des fonctions compliquées mais on peut chercher à les approcher par des fonctions plus simples. Le principe général est de partir d'une forme de fonction connue et de chercher les paramètres qui ajustent le mieux possible les courbes obtenues aux courbes de régression. Par exemple, si l'on part de l'idée que la courbe de régression, si l'on mettait de côté les erreurs de mesure, serait une droite, alors elle serait caractérisée par une équation de type

$$y = ax + b$$

Il faudrait alors identi.er les paramètres a et b pour connaître l'équation de la droite.
Si l'on imagine que la courbe de régression correspond à une parabole, l'équation cherchée serait de la forme :

$$y = ax^2 + bx + c$$

Et dans ce cas il faudrait trouver les valeurs des trois paramètres, a, b, et c.

7.1.Ajustement puissance

L'ajustement puissance est basé sur la courbe représentant par l'équation de type

$$y = ax^b$$

On remarque que

$$\ln y = a \ln x + \ln b$$

On pose V = ln y et U = ln x, on détermine l'équation de la droite de régression de v en u avec la méthode de Mayer ou la méthode des moindres carrés, l'équation obtenue est de la forme v = Au + B, on en déduit l'équation de la courbe de fonction puissance : $y = ax^b$ puisque A = b et B = ln a.

7.2.Ajustement exponentielle

L'ajustement exponentiel a une courbe de fonction exponentielle d'équation :

$$y = ab^x$$

On remarque que

$$y = x \ln b + \ln a$$

On pose z = ln y, on détermine l'équation de la droite de régression de z en x avec la méthode de Mayer ou la méthode des moindres carrés, l'équation obtenue est de la forme v = Au + B on en déduit l'équation de la courbe de fonction exponentielle

$$z = Ax + B$$

On en déduit l'équation de la courbe de fonction exponentielle :

$y = ab^x$ Puisque A = ln b et B = ln a