Séries statistiques à deux variables

F. Wlazinski

Licence d'économie

1 Tableau de contingence

Définition 1.1

Soient X et Y deux variables statistiques d'une population d'effectif total n.
Si la variable X possède p modalités et si la variable Y possède q modalités, on appelle *tableau de contingence* de X et Y le tableau où sont indiquées dans la première colonne les p modalités $(x_i)_{i=1,p}$ de la variable X et dans la première ligne les q modalités $(y_j)_{j=1,p}$ de la variable Y. A l'intersection de la ligne x_i et de la colonne y_i, on trouve l'effectif *partiel* $n_{i,j}$ correspondant à la conjonction des dites modalités pour la population.

Exemple 1.2

Pour une étude comportementale, on se propose d'étudier pour 200 lycéens une possible relation entre la catégorie socio-professionnelle de leur père et le choix de leur filière scolaire. Les résultats sont les suivants :

Catégories / Filières	Agriculteur ou dirigeant	Cadre supérieur	Cadre moyen	Ouvrier ou autre
ES	16	18	23	45
L	3	5	8	17
S	5	3	8	15
STT	7	6	9	12

Remarque 1.3

Définition 1.4

Les totaux par ligne et par colonne sont appelées les *distributions marginales*.

Exemple 1.5

Catégories / Filières	Agriculteur ou dirigeant	Cadre supérieur	Cadre moyen	Ouvrier ou autre	
ES	16	18	23	45	
L	3	5	8	17	
S	5	3	8	15	
STT	7	6	9	12	

Remarque 1.6

Remarque 1.7

Si X prend les valeurs $(x_1, x_2, \ldots, x_p)$ et si Y prend les valeurs $(y_1, y_2, \ldots, y_q)$, la somme de la ligne x_i est $n_{i,1} + n_{i,2} + \ldots + n_{i,q}$ c'est-à-dire $\sum_{j=1}^{q} n_{i,j}$ et la somme de la ligne y_j est $n_{1,j} + n_{2,j} + \ldots + n_{p,j}$ c'est-à-dire $\sum_{i=1}^{p} n_{i,j}$. L'effectif total n est $\sum_{i=1}^{p}\left(\sum_{j=1}^{q} n_{i,j}\right)$ ou $\sum_{j=1}^{q}\left(\sum_{i=1}^{p} n_{i,j}\right)$ suivant si l'on a fait les sommes par ligne ou par colonne en premier.

Définition 1.8

Si X prend les valeurs $(x_i)_{i=1,p}$, si Y prend les valeurs $(y_j)_{j=1,p}$ et si $n_{i,j}$ est l'effectif à la conjonction des modalités x_i et y_j alors la fréquence *empirique* est $f_{i,j} = \dfrac{n_{i,j}}{n}$.

Exemple 1.9

Catégories / Filières	Agriculteur ou dirigeant	Cadre supérieur	Cadre moyen	Ouvrier ou autre
ES				
L				
S				
STT				

Définition 1.10

On utilise les notations de la définition 1.8.
Les fréquences *en colonnes* sont les valeurs $\dfrac{n_{i,j}}{\sum_{i=1}^{p} n_{i,j}}$. et les fréquences *en lignes* sont les valeurs $\dfrac{n_{i,j}}{\sum_{j=1}^{q} n_{i,j}}$.

Exemple 1.11

Catégories / Filières	Agriculteur ou dirigeant	Cadre supérieur	Cadre moyen	Ouvrier ou autre	
ES					
L					
S					
STT					

Filières \ Catégories	Agriculteur ou dirigeant	Cadre supérieur	Cadre moyen	Ouvrier ou autre	
ES					
L					
S					
STT					

2 Moyenne, variance et covariance

Principe

Comme pour les séries à une variable, on pourra calculer les moyennes variances, et covariance uniquement pour les séries à deux variables quantitatives.

Exemple 2.1

En vue d'un éventuel remplacement, on s'intéresse aux nombres annuels (Y) d'interventions pour réparation sur les machines d'une entreprise en fonction de leur ancienneté (X) en années. Les données sont résumées par le tableau suivant.

Ancienneté \ Nombres de pannes	0	1	2
1	5	6	4
3	10	2	4
5	17	6	7
7	2	6	11

Définition 2.2

On utilise les notations de la définition 1.8.

La moyenne arithmétique (marginale) de X est $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p}\left(x_i \sum_{j=1}^{q} n_{i,j}\right) = \sum_{i=1}^{p} f_i x_i$.

La moyenne arithmétique (marginale) de Y est $\overline{Y} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{q}\left(x_i \sum_{i=1}^{p} n_{i,j}\right) = \sum_{j=1}^{q} f_j y_j$.

Remarque 2.3

On définit comme pour les séries d'une seule variable :

$\overline{X^2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p}\left(x_i^2 \sum_{j=1}^{q} n_{i,j}\right) = \sum_{i=1}^{p} f_i x_i^2$ et $\overline{Y^2} = \dfrac{1}{n}\displaystyle\sum_{j=1}^{q}\left(x_i^2 \sum_{i=1}^{p} n_{i,j}\right) = \sum_{j=1}^{q} f_j y_j^2$.

Remarque 2.4

Exemple 2.5

On reprend l'exemple 2.1.

Nombres de pannes / Ancienneté	0	1	2	
1	5	6	4	
3	10	2	4	
5	17	6	7	
7	2	6	11	

Définition 2.6

La *covariance* des variables X et Y est $\text{cov}(X;Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{p}\sum_{j=1}^{q}\big(n_{i,j} \times (\overline{X} - x_i)(\overline{Y} - y_i)\big)$.

Remarque 2.7

Exemple 2.8

Remarque 2.9

3 Liens entre deux variables sur une même population

Principe

On s'intéresse à l'étude, sur une population donnée, du ou des liens qui peuvent exister entre deux variables X et Y.

Individu	1	2	3	...	n
Valeurs de X	x_1	x_2	x_3	...	x_n
Valeurs de Y	y_1	y_2	y_3	...	y_n

Exemples 3.1

Ces exemples seront repris tout au long du cours.

• Exemple A : Le tableau suivant fournis les résultats scolaires de 6 étudiants à un examen ainsi que le temps qu'ils ont consacré à la révision pour cet examen.

Etudiants	1	2	3	4	5	6
X nombres d'heures de révision	6	16	10	3	1	12
Y notes à l'examen	11	20	8	8	1	9

• Exemple B : Le coût de production et les ventes de 5 produits fabriqués par une même entreprise sont donnés par le tableau suivant.

Produits	A	B	C	D	E
Coûts de production en euros	3	8	1	7	1
Ventes en milliers d'unités	11	20	7	8	1

• Exemple C : Les coûts en réparation et évolution d'un parc informatique d'une multinationnale sur 4 ans sont donnés par le tableau suivant.

Années	2014	2015	2016	2017
Coûts en million euros	6, 24	5, 75	6, 02	4, 79

• Exemple D : Le budget annuel d'une famille sur 6 ans est résumé par le tableau :

Années	2012	2013	2014	2015	2016	2017
X revenus en milliers d'euros	19	22	23	20	27	27
Y dépenses en milliers d'euros	18	18	21	22	20	21

Définition 3.2

Soient $X = (x_i)_{i=1,n}$ et $Y = (y_i)_{i=1,n}$ deux séries statistiques.
On appelle *nuage de points* l'ensemble des points $(x_i, y_i)_{i=1,n}$.

Exemple 3.3

Définition 3.4

Le *point moyen* G du nuage de points est le point de coordonnées $(\overline{X}; \overline{Y})$.

Exemple 3.5

Remarque 3.6

Exemple 3.7

On reprend l'exemple A :

Etudiants		A	B	C	D	E	F
Nombres d'heures de révision	x_i	6	16	10	3	1	12
Notes à l'examen	y_i	11	20	8	8	1	9

4 Ajustement affine

Remarque 4.1

Définition 4.2

Soient X et Y deux variables statistiques, de variances non nulles, sur une même population. On appelle *coefficient de corrélation linéaire* entre X et Y et on note $r_p(X, Y)$ ou $\rho(X, Y)$ le réel :
$r_p = r_p(X, Y) = \dfrac{\text{cov}(X; Y)}{\sigma_X \sigma_Y} = \dfrac{\text{cov}(X; Y)}{\sqrt{V(X)V(Y)}}$.

Remarque 4.3

Exemples 4.4

- Avec l'exemple A :

Etudiants		A	B	C	D	E	F		
Nombres d'heures	x_i	6	16	10	3	1	12		
Notes	y_i	11	20	8	8	1	9		

- Avec l'exemple D :

Années		2012	2013	2014	2015	2016	2017		
Revenus	x_i	19	22	23	20	27	27		
Dépenses	y_i	18	18	21	22	20	21		

Définition 4.5

On appelle droite de Mayer la droite qui passe par les deux points moyens des deux sous-séries obtenues en partageant en deux la série des couples (x_i, y_i) rangés selon l'ordre croissant des x_i.

Exemple 4.6

Remarque 4.7

Exemple 4.8

Définition 4.9

On appelle *droite de régression*, *droite d'ajustement* ou *droite des moindres carrés* de Y en X la droite d'équation $y = ax + b$ où $a = \dfrac{\text{cov}(X;Y)}{V(X)}$ et $b = E(Y) - a \times E(X)$.

Remarque 4.10

La droite de régression de Y en X passe par le point moyen $G\left(\overline{X}, \overline{Y}\right)$.

Remarque 4.11

Exemple 4.12

Remarque 4.13

Si le nuage de points semble suivre une fonction exponentielle ou une fonction puissance > 1 (Y croît plus vite que X), on tentera un ajustement affine non pas avec les séries (X, Y) mais avec les séries $(X, \ln Y)$. De plus, si $\ln Y = aX + b$, on aura alors $Y = e^b \times e^{aX}$.
Si le nuage de points semble suivre une fonction logarithme ou une fonction racine (Y croît moins vite que X), on tentera un ajustement affine non pas avec les séries (X, Y) mais avec les séries $(\ln X, Y)$.

Exemple 4.14

On considère les séries statistiques définies par le tableau suivant :

X	$0,1$	$0,5$	$1,2$	$1,4$	2	3
Y	3	$3,1$	$3,7$	4	5	$7,5$