Chapitre 5
Statistiques descriptives bivariées

Exemple

Étude sur 5761 femmes de la survenue d'accouchement prématuré et de l'exposition à des événements stressants.

X : type d'accouchement

variable qualitative à 2 modalités

Y : score sur une échelle allant de 0 à 3.

variable quantitative discrète à 4 valeurs

X \ Y	0	1	2	3	totaux
à terme	4698	413	250	197	5558
prématuré	165	16	12	10	203
totaux	4863	429	262	207	5761

1 Organisation des données

1.1 Notations

▶ On notera $x_i, \ i = 1, \ldots, k$ les k modalités ou valeurs de la variable X

▶ On notera $y_i, \ i = 1, \ldots, \ell$ les ℓ modalités ou valeurs de la variable Y

▶ Les deux variables X et Y sont mesurées simultanément sur chacun des N individus de la population. On notera n_{ij} l'**effectif** correspondant au couple (x_i, y_j).

<u>Définition</u>

> On appellera **distribution jointe des effectifs de X et Y** l'ensemble des informations (x_i, y_j, n_{ij}) pour $i = 1, \ldots, k$ et $j = 1, \ldots, \ell$.

1.2 Tableau de contingence

Représentation de la distribution jointe du couple (X, Y) : on utilise un tableau à double entrée appelé

tableau de contingence

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_ℓ
x_1	n_{11}		n_{1j}		$n_{1\ell}$
$\cdots$					$\cdots$
x_i	n_{i1}		n_{ij}		$n_{i\ell}$
$\cdots$					$\cdots$
x_k	n_{k1}		n_{kj}		$n_{k\ell}$

Exemple :

12 : le nombre de femmes ayant accouché prématurément et ayant un score égal à 2.

Remarque : $$\sum_{i=1}^{k} \sum_{j=1}^{\ell} n_{ij} = N$$

2 Distributions Marginales

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

X \ Y	y_1	$\cdots$	y_j	$\cdots$	y_ℓ	totaux
x_1	n_{11}		n_{1j}		$n_{1\ell}$	$n_{1\bullet}$
$\cdots$					$\cdots$	$\cdots$
x_i	n_{i1}		n_{ij}		$n_{i\ell}$	$n_{i\bullet}$
$\cdots$					$\cdots$	$\cdots$
x_k	n_{k1}		n_{kj}		$n_{k\ell}$	$n_{k\bullet}$
totaux	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet \ell}$	$N = n_{\bullet\bullet}$

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

$$n_{i\bullet} = \sum_{j=1}^{\ell} n_{ij} = \text{total de la ligne i}$$

<u>Définition</u>

> Les k couples $(x_i, n_{i\bullet})$ définissent la **distribution marginale** de la variable X.

Remarque : $$\sum_{i=1}^{k} n_{i\bullet} = N$$

Exemple

X \ Y	0	1	2	3	totaux en ligne
à terme	4698	413	250	197	5558
prématuré	165	16	12	10	203

► Distribution marginale de X

X	à terme	prématuré	effectif total
effectifs	5558	203	5761

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

$$n_{\bullet j} = \sum_{i=1}^{k} n_{ij} = \text{total de la colonne j}$$

<u>Définition</u>

> Les ℓ couples $(y_j, n_{\bullet j})$ définissent la **distribution marginale** de la variable Y.

Remarque : $$\sum_{j=1}^{\ell} n_{\bullet j} = N$$

Exemple

X \ Y	0	1	2	3
à terme	4698	413	250	197
prématuré	165	16	12	10
totaux en colonne	4863	429	262	207

► Distribution marginale de Y

Y	0	1	2	3	effectif total
effectifs	4863	429	262	207	5761

3 Distributions conditionnelles

Exemple

Ligne 2 du tableau de contingence : distribution de la variable Y chez les femmes ayant eu un accouchement prématuré.

$Y\|_{X=\text{prématuré}}$	0	1	2	3	total
effectifs	165	16	12	10	203

Principe :
Comportement de l'une des deux variables quand l'autre a une valeur donnée.

Réponse :

▶ *À la ligne i du tableau de contingence*, on lit la distribution de la variable Y **sachant que** $X = x_i$, notée $Y|_{X=x_i}$.

Définition :

> 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** $X = x_i$.

▶ *À la colonne j du tableau de contingence*, on lit la distribution de la variable X **sachant que** $Y = y_j$, notée $X|_{Y=y_j}$.

Définition :

> 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** $Y = y_j$.

Exemple

Distribution conditionnelle de X sachant que la femme enceinte a subi un stress de niveau 2

$X\|_{Y=2}$	à terme	prématuré	total
effectifs	250	12	262

Obtention par la Colonne 3 du tableau de contingence.

4 Proportions associées à un couple de variables

▶ trois notions de proportion (ou fréquence)

1. proportions du couple (x_i, y_j) ;
2. proportions marginales de X ou Y ;
3. proportions conditionnelles.

Exemple :
N=5761.
pour $(X, Y) = ($ à terme$, 0)$ la proportion est :

$$\frac{4698}{5761} = 0.815.$$

X \ Y	0	1	2	3
à terme	0.815	0.072	0.043	0.034
prématuré	0.029	0.003	0.002	0.002

La somme de toutes les proportions = 1

Définition 1.

La proportion du couple (x_i, y_j) est

$$p_{ij} = \frac{n_{ij}}{N}.$$

Exemple :

N=5761 ;

Proportions marginales pour X :

X	à terme	prématuré	total
effectifs	5558	203	5761
proportions	0.964	0.036	1

$$\frac{5558}{5761} = 0.964 \qquad \frac{203}{5761} = 0.036$$

Définition 2.

La proportion marginale de x_i est

$$p_{i\bullet} = \frac{n_{i\bullet}}{N}.$$

La proportion marginale de y_j est

$$p_{\bullet j} = \frac{n_{\bullet j}}{N}.$$

Exemple :

$X\|_{Y=2}$	à terme	prématuré	total
effectifs	250	12	262
proportions	0.954	0.046	1

$Y\|_{X=\text{prema.}}$	0	1	2	3	tot.
effectifs	165	16	12	10	203
proportions	0.813	0.079	0.059	0.049	1

Définition 3. :

La proportion conditionnelle de x_i sachant que $Y = y_j$ est

$$p_{i|Y=y_j} = \frac{n_{ij}}{n_{\bullet j}}$$

La proportion conditionnelle de y_j sachant que $X = x_i$ est

$$p_{j|X=x_i} = \frac{n_{ij}}{n_{i\bullet}}\ .$$

Remarque :

lien entre les différentes proportions

$$p_{ij} = p_{i|Y=y_j} \times p_{\bullet j} = p_{j|X=x_i} \times p_{i\bullet}$$

ou encore

$$p_{i|Y=y_j} = \frac{p_{ij}}{p_{\bullet j}} \quad \text{et} \quad p_{j|X=x_i} = \frac{p_{ij}}{p_{i\bullet}}$$

Remarque : lien entre les variables

On peut comparer les distributions conditionnelles de X sachant Y à la distribution marginale de X.

► Si ces distributions sont très proches, on peut conclure une certaine indépendance entre les deux variables.

► Si ces distributions sont très distinctes, cela signifie que les modalités de Y ont une influence sur la variable X et donc que les deux variables sont liées. (cf. exemple)

► De façon rigoureuse :

Les deux variables sont indépendantes si et seulement si $p_{ij} = p_{i\bullet} \times p_{\bullet j}$, ou $p_{i|Y=y_j} = p_{i\bullet}$.

Exemple :

X \ Y	0	1	2	3
à terme	0.966	0.963	0.954	0.952
préma.	0.034	0.037	0.046	0.048

X	dist marg. (prop.)
à terme	0.964
préma.	0.036

5 Étude de deux variables quantitatives

Notations

▶ si X et Y sont des variables quantitatives discrètes : x_i et y_j sont les valeurs prises.
▶ si X et Y sont des variables quantitatives continues : x_i et y_j désignent les centres des classes.

Exemple

Une entreprise employant 100 femmes relève pour chaque femme son âge, noté X, et le nombre de journées d'absence durant le mois de janvier, noté Y.

X \ Y	0	1	2	3
$[20, 30[$	0	0	5	15
$[30, 40[$	0	15	20	0
$[40, 50[$	15	10	5	0
$[50, 60[$	0	5	5	5

X \ Y	0	1	2	3	totaux
$[20, 30[$	0	0	5	15	20
$[30, 40[$	0	15	20	0	35
$[40, 50[$	15	10	5	0	30
$[50, 60[$	0	5	5	5	15
totaux	15	30	35	20	100

5.1 Principales caractéristiques

Moyennes des distributions marginales :

▶ Moyenne de X :

$$\mu(X) = \frac{1}{N}\sum_{i=1}^{k} n_{i\bullet} x_i = \sum_{i=1}^{k} p_{i\bullet} x_i$$

▶ Moyenne de Y :

$$\mu(Y) = \frac{1}{N}\sum_{j=1}^{\ell} n_{\bullet j} y_j = \sum_{j=1}^{\ell} p_{\bullet j} y_j$$

Exemple

$$\mu(Y) = \frac{1}{100}(15 \times 0 + 30 \times 1 + 35 \times 2 + 20 \times 3) = 1.6$$

$$\mu(X) = \frac{1}{100}(20\times 25+35\times 35+30\times 45+15\times 55) = 39$$

Variances des distributions marginales :

▶ Variance et écart-type de X :

$$V(X) = \left(\frac{1}{N} \sum_{i=1}^{k} n_{i\bullet} x_i^2 \right) - \mu(X)^2$$

$$\sigma(X) = \sqrt{V(X)}$$

▶ Variance et écart-type de Y :

$$V(Y) = \frac{1}{N} \sum_{j=1}^{\ell} n_{\bullet j} y_j^2 - \mu(Y)^2$$

$$\sigma(Y) = \sqrt{V(Y)}$$

Exemple

$$V(X) = 1615 - 39^2 = 94 \text{ donc } \sigma(X) = 9.67$$
$$V(Y) = 3.5 - 1.6^2 = 0.94 \text{ donc } \sigma(Y) = 0.97$$

Moyennes et variances des distributions conditionnelles :

► Moyenne de X sachant $Y = y_j$

$$\mu(X_{|Y=y_j}) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i = \sum_{i=1}^{k} p_{i|Y=y_j} x_i$$

► Variance de X sachant $Y = y_j$

$$V(X_{|Y=y_j}) = \frac{1}{n_{\bullet j}} \sum_{i=1}^{k} n_{ij} x_i^2 - (\mu(X_{|Y=y_j}))^2$$

► Celles de Y sachant X se déduisent de même.

<u>Exemple</u>

$\mu(X_{|Y=1}) = \frac{25\times0+35\times15+45\times10+55\times5}{30} = 41.67$

$V(X_{|Y=1}) = \frac{35^2\times15+45^2\times10+55^2\times5}{30} - 41.67^2$

$V(X_{|Y=1}) = 1791.67 - 41.67^2 = 55.28$

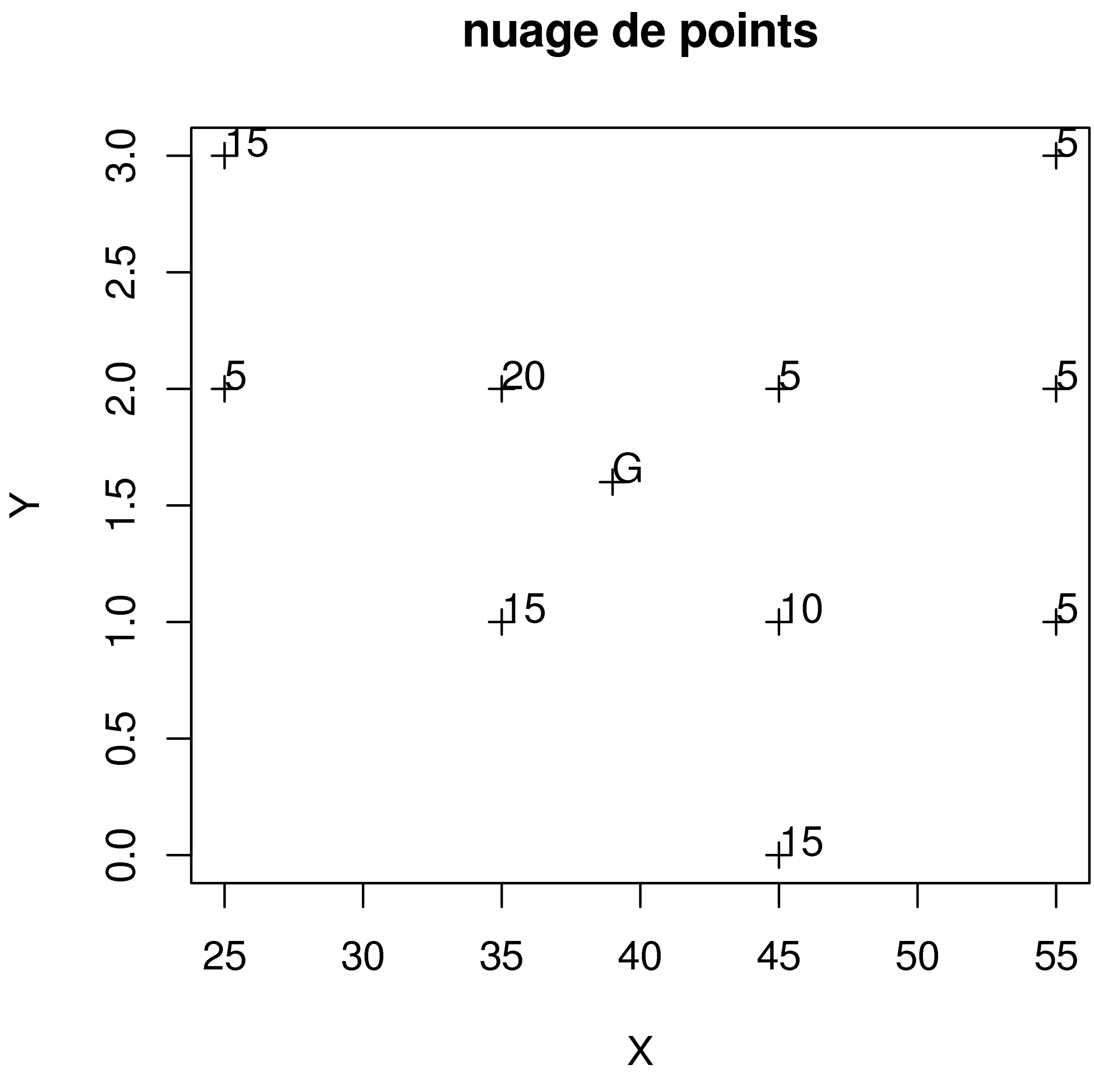
$\sigma(X_{|Y=1}) = 7.44$

5.2 Représentation graphique

On peut représenterer la distribution du couple (X, Y) par un **nuage de points** de coordonnées (x_i, y_j), chaque point étant affecté du "poids" n_{ij}.

Le **centre de gravité** du nuage est alors le point (non observé) de coordonnées $(\mu(X); \mu(Y))$.

Exemple



nuage de points
Y
3.0
2.5
2.0
1.5
1.0
0.5
0.0
15
5
5
20
5
5
G
15
10
5
15
25
30
35
40
45
50
55
X

5.3 Covariance, Correlation

► Outils pour mesurer la dépendance linéaire entre deux caractères quantitatifs X et Y.

<u>Définition</u>

La **covariance** de X et Y est le nombre réel défini par

$$\operatorname{cov}(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{\ell} n_{ij}(x_i - \mu(X))(y_j - \mu(Y))$$

Formule pratique de calcul

$$\operatorname{cov}(X,Y) = \left(\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{\ell} n_{ij} x_i y_j\right) - \mu(X)\mu(Y)$$

<u>Exemple</u>

$\operatorname{cov}(X,Y) = 58.5 - 39 \times 1.6 = -3.9$

Propriétés

$$\mathrm{cov}(X, Y) = \mathrm{cov}(Y, X) \quad \text{et} \quad \mathrm{cov}(X, X) = V(X).$$

Remarques :

- ▶ dépendance aux unités utilisées
- ▶ prend n'importe quelle valeur réelle.

D'où définition du **coefficient de corrélation linéaire** :

Définition

Le **coefficient de corrélation linéaire de** X et Y est défini par

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

Propriétés

$\mathrm{corr}(X,Y) \in [-1,1]$

$\mathrm{corr}(X,Y) = \mathrm{corr}(Y,X)$ et $\mathrm{corr}(X,X) = 1.$

Exemple

$\mathrm{corr}(X,Y) = \frac{-3.9}{9.7*0.97} = -0.414$

Le coefficient de corrélation est un coefficient sans dimension. Il mesure la présence et l'intensité de la liaison linéaire entre X et Y.

1. $\text{corr}(X, Y) = 1$: **liaison linéaire exacte** $Y = aX + b$ avec $a > 0$;
2. $\text{corr}(X, Y) = -1$: **liaison linéaire exacte** $Y = aX + b$ avec $a < 0$;
3. $\text{corr}(X, Y) = 0$: **non corrélation** : on a indépendance possible, mais non certaine ;
4. $\text{corr}(X, Y) > 0$: **liaison relative**, X et Y ont tendance à varier dans le **même sens** ;
5. $\text{corr}(X, Y) > 0$: **liaison relative**, X et Y ont tendance à varier dans le **sens contraire** ;
6. $|\text{corr}(X, Y)| > 0.9$ la liaison linéaire est considérée comme **forte**.

Remarque :

il faut bien se garder au vu de la seule valeur du coefficient de corrélation, d'émettre des interprétations abusives.

Ex des chaussures et de la culture générale tous deux liés à l'âge ! !

Par contre il existe des outils permettant d'étudier plus en détail les relations linéaires entre deux caractères et permettant (dans une certaine mesure) d'extrapoler à partir de données existantes et de faire de la prévision !