# STATISTIQUE DESCRIPTIVE BIVARIEE

Dans beaucoup de recherches statistiques, on ne s'intéresse pas qu'à un seul caractère mais à plusieurs en même temps. Quand on étudie deux caractères X et Y sur une population donnée, c'est en général parce qu'on cherche à savoir s'il existe un lien entre eux et qu'elle est l'intensité du lien.

Exemple de relations possibles entre les variables suivantes : taille et âge ; diabète et poids, taux de cholestérol et régime alimentaire, niche écologique et population, ensoleillement et croissance végétale, toxine et réaction métabolique, survie et pollution, effets et doses ...Les caractères étudiés peuvent être aussi bien qualitatifs que quantitatifs[1].

## 1.1 Distributions statistiques à deux variables

On considère une population de N individus mesurés simultaniment par les deux caractères X et Y qui peuvent être qualitatives ou quantitatives, et qui peuvent ne pas être de même nature. Les k modalités[2] de X sont désignées par $x_1, \ldots, x_j, \ldots, x_k$ ; les l modalités de Y

[1] Les types de variables ont été définis dans le chapitre précédent.

[2] Dans le cas d'une variable quantitative continue la $i^{ème}$ modalité d'une variable désigne le centre de la $i^{ème}$ classe.

sont désignées par $y_1, \ldots, y_j, \ldots, y_l$.

## 1.2 Tableau statistique

La répartition des N observations, ou distribution conjointe, suivant les modalités de X et Y se présente sous forme d'un tableau à double entrée, appelée **tableau de contingence** ou tableau à double entrée ou tableau croisé ou parfois tableau de corrélation(tableau de k lignes et de l colonnes).

| X\Y | $y_1$ | $y_2$ | ... | $y_j$ | ... | $y_l$ | TOTAL |
|---|---|---|---|---|---|---|---|
| $x_1$ | $n_{11}$ | $n_{12}$ | ... | $n_{1j}$ | ... | $n_{1l}$ | $n_{1.}$ |
| $x_2$ | $n_{21}$ | $n_{22}$ | ... | $n_{2j}$ | ... | $n_{2l}$ | $n_{2.}$ |
| . | . | . | | . | | . | . |
| . | . | . | .... | . | .... | . | . |
| . | . | . | | . | | . | . |
| $x_i$ | $n_{i1}$ | $n_{i2}$ | ... | $n_{ij}$ | ... | $n_{il}$ | $n_{i.}$ |
| . | . | . | | . | | . | . |
| . | . | . | .... | . | ..... | . | . |
| . | . | . | | . | | . | . |
| $x_k$ | $n_{k1}$ | $n_{k2}$ | ... | $n_{kj}$ | ... | $n_{kl}$ | $n_{k.}$ |
| TOTAL | $n_{.1}$ | $n_{.2}$ | ... | $n_{.j}$ | ... | $n_{.l}$ | $n_{..} = N$ |

TAB. 1.1 – Tableau de contingence.

► L'effectif $n_{ij}$ désigne le nombre de fois où la modalité $x_i$ de la variable X et la modalité $y_j$ de la variableY ont été observées simultanément.

► L'effectif $n_{i.}$ appelé effectif marginal de X, représente le nombre total d'observations de la modalité $x_i$ de X, quelle que soit la modalité de Y.

$$n_{i.} = \sum_{j=1}^{l} n_{ij}$$

► De même, L'effectif $n_{.j}$ appelé effectif marginal de Y, est le nombre total d'observations de la modalité $y_j$ de Y, quelle que soit la modalité de X.

$$n_{.j} = \sum_{i=1}^{k} n_{ij}$$

On a évidemment :

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = \sum_{j=1}^{l} n_{i.} = \sum_{i=1}^{k} n_{.j} = N$$

La distribution conjointe peut aussi être définie par les fréquences :

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

**Exemple 1.2.1** *Soit la série statistique bidimensionnelle du couple (X, Y) suivante :*

| X/Y | -2 | 0 | 2 | 3 | $n_{i.}$ |
|---|---|---|---|---|---|
| 2 | -2 | 4 | 0 | 6 | 13 |
| 3 | 4 | 3 | 3 | 2 | 12 |
| 4 | 2 | 3 | 3 | 2 | 10 |
| $n_{.j}$ | 9 | 10 | 6 | 10 | 35 |

## 1.3 Représentation graphique

### 1.3.1 Nuage de points

Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives.

Si les observations de deux variables statistiques X et Y sont connues individuellement, on commence par les visualiser en les représentant sous la forme d'un nuage de points : dans un repère cartésien, chaque observation $(x_i,\ y_i)$ est figurée par le point $M_i$ de coordonnées $(x_i,\ y_i)$, et la forme du nuage donne une information sur le type d'une éventuelle liaison.

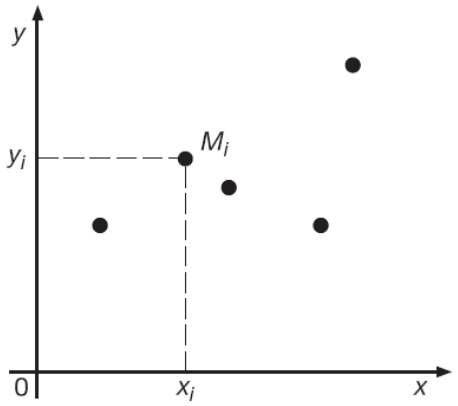


FIG. 1.1 – Nuage de points.

## 1.4 Distributions marginales

La distribution marginale est déterminée en isolant les première et dernière colonnes du tableau de contingence. La première colonne contient les modalités $x_i$ et la dernière, les effectifs correspondants. C'est-à-dire sur la marge du tableau de contingence, on peut extraire les données seulement par rapport à X et seulement par rapport à Y.

Les k couples $(x_i,\ n_{i\cdot})$ forment la distribution marginale de la variable X.

Les l couples $(y_j,\ n_{\cdot j})$ forment la distribution marginale de la variable Y.

Les distributions marginales peuvent aussi être données sous forme de fréquences :

$$f_{i\cdot} = \frac{n_{i\cdot}}{N} \qquad et \qquad f_{\cdot j} = \frac{n_{\cdot j}}{N}$$

De plus, on a :

$$\sum_{j=1}^{l} f_{i\cdot} = \sum_{i=1}^{k} f_{\cdot j} = 1.$$

Ces deux distributions pouvent se présenter sous forme de tableaux statistiques :

**Distribution marginale de X :**

| X | *Effectif marginal* $n_{i.}$ | *Fréquence relative marginale* $f_{i.}$ |
|---|---|---|
| $x_1$ | $n_{1.}$ | $f_{1.} = n_{1.}/N$ |
| $x_2$ | $n_{2.}$ | $f_{2.} = n_{2.}/N$ |
| . | . | . |
| . | . | . |
| . | . | . |
| $x_k$ | $n_{k.}$ | $f_{k.} = n_{k.}/N$ |
| TOTAL | N | 1 |

**Distribution marginale de Y :**

| Y | *Effectif marginal* $n_{.j}$ | *Fréquence relative marginale* $f_{.j}$ |
|---|---|---|
| $y_1$ | $n_{.1}$ | $f_{.1} = n_{.1}/N$ |
| $y_2$ | $n_{.2}$ | $f_{.2} = n_{.2}/N$ |
| . | . | . |
| . | . | . |
| . | . | . |
| $y_l$ | $n_{.l}$ | $f_{.l} = n_{.l}/N$ |
| TOTAL | N | 1 |

## 1.5 Description numérique

Disposant d'une distribution conjointe, on peut déduire les distributions marginales qui permettent d'étudier séparément chaque variable en représentant graphiquement sa distribution et s'il s'agit d'une variable quantitative, en calculant ses caractéristiques de tendance centrale et de dispersion.

### 1.5.1 Caractéristique des séries marginales

► Les moyennes marginales des variables X et Y sont :

$$\overline{x}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i\cdot}\ x_i\ = \sum_{i=1}^{k} f_{i\cdot}\ x_i,$$

et

$$\overline{y}_M = \frac{1}{N}\sum_{j=1}^{l} n_{\cdot j}\ y_j\ = \sum_{j=1}^{l} f_{\cdot j}\ y_j\ .$$

► Les variances marginales des variables X et Y sont données par :

$$var_M(x) = \overline{x^2} - (\overline{x}_M)^2 = \frac{1}{N}\sum_{i=1}^{k} n_{i\cdot}\ x_i^2 - (\overline{x}_M)^2\,,$$

et

$$var_M(y) = \overline{y^2} - (\overline{y}_M)^2 = \frac{1}{N}\sum_{j=1}^{l} n_{\cdot j}\ y_j\ - (\overline{y}_M)^2\,.$$

► Les écarts-type marginaux de X et Y sont donnés par :

$$\sigma_X = \sqrt{var_M(x)} \qquad et \qquad \sigma_Y = \sqrt{var_M(Y)}.$$

**Exemple 1.5.1** *En reprenant l'exemple 2.1.1 et on détermine la moyenne marginale de X et de Y comme suit :*

$$\overline{x}_M = \frac{1}{N}\sum_{i=1}^{k} n_{i\cdot}\ x_i\ = \frac{1}{35}\left(2\times 13 + 3\times 12 + 4\times 10\right) = \frac{102}{35} = 2.914.$$

et

$$\overline{y}_M = \frac{1}{N}\sum_{j=1}^{l} n_{\cdot j}\ y_j\ = \frac{1}{35}\left(-2\times 9 + 0\times 10 + 2\times 6 + 3\times 10\right) = \frac{24}{35} = 0.686.$$

Les variances marginales des variables X et Y sont :

$$var_M(x) = \frac{1}{N}\sum_{i=1}^{k} n_{i\cdot}\ x_i^2 - (\overline{x}_M)^2 = \frac{1}{35}\left(4\times 13 + 9\times 12 + 16\times 10\right) - (2.914)^2 = 0.650$$

et

$$var_M(y) = \frac{1}{N}\sum_{j=1}^{l} n_{\cdot j}\ y_j\ - (\overline{y}_M)^2 = \frac{1}{35}\left(150\right) - (0.686)^2 = 3.815.$$

### 1.5.2 Distributions conditionnelles

La distribution de la variable Y, la variable X étant égale à $x_i$, est appelée distribution conditionnelle de Y pour $X = x_i$ :

| Y/X=$x_i$ | $y_1$ ... $y_j$ ... $y_l$ | Total |
|---|---|---|
| Effectif | $n_{i1}$ ... $n_{ij}$ ... $n_{il}$ | $n_{i.}$ |

Cette distribution des $n_{i.}$ observations, satisfaisant à la condition $X = x_i$, est présentée sous la forme de fréquences conditionnelles :

$$f_{j/i} = \frac{n_{ij}}{n_{i.}} \qquad \text{avec :} \qquad \sum_{j=1}^{l} f_{j/i} = 1.$$

| Y/X=$x_i$ | $y_1$ ... $y_j$ ... $y_l$ | Total |
|---|---|---|
| Fréquence | $n_{1/i}$ ... $n_{j/i}$ ... $n_{l/i}$ | 1 |

La fréquence $f_{j/i}$ parfois notées et $f_j^i$ se lit " f indice j si i ", c'est-à-dire fréquence de $y_j$ si $X = x_i$. Il y a k distributions conditionnelles de Y pour $(i = 1, \ldots, k)$.

Lorsque la variable Y est quantitative, on peut calculer pour chaque valeur $x_i$ sa moyenne conditionnelle $\overline{y}_i$ et sa variance conditionnelle $var_i$ :

$$\overline{y}_i = \sum_{j=1}^{l} f_{j/i}\ y_j \qquad et \qquad var_i = \sum_{j=1}^{l} f_{j/i}\ \ (y_j - \overline{y}_i)^2 .$$

Les k modalités de X induisant une partition des observations en k sous groupes, la moyenne peut s'exprimer comme somme pondérée des k moyennes $\overline{y}_i$ (chapitre 1) :

$$\overline{y} = \sum_{i=1}^{k} f_{i\cdot}\ \overline{y}_i$$

Symétriquement, on a $l$ distributions conditionnelles de X et on définit les fréquences conditionnelles f indice i si j :

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} \qquad avec : \qquad \sum_{i=1}^{k} f_{i/j} = 1.$$

| X/Y=$y_j$ | $x_1$ ... $x_j$ ... $x_l$ | Total |
|---|---|---|
| Fréquence | $n_{1/j}$ ... $n_{i/j}$ ... $n_{k/j}$ | 1 |

Lorsque la variable X est quantitative, on peut calculer pour chaque valeur $y_j$ sa moyenne conditionnelle $\overline{x}_j$ et sa variance conditionnelle $\sigma_j^2$ :

$$\overline{x}_j = \sum_{i=1}^{k} f_{i/j}\ x_i \qquad et \qquad \sigma_j^2 = \sum_{i=1}^{k} f_{i/j}\ \left(x_i - \overline{x}_j\right)^2 .$$

**Exemple 1.5.2** *En reprenant l'exemple 2.1.1 alors pour déterminer la moyenne conditionnelle de X quand Y=2, il suffit d'observer le comportement de X relatif à la colonne Y=2.*

| **X** | **y=2** |
|---|---|
| 2 | 0 |
| 3 | 3 |
| 4 | 3 |
| $\mathbf{n}_{.j}$ | **6** |

$$\overline{x}_{y=2} = \frac{0 \times 2 + 3 \times 3 + 3 \times 4}{6} = 3.5$$

Pour déterminer la moyenne conditionnelle de Y quand X=3, il suffit d'observer le comportement de Y relatif à la colonne X=3 :

| **y** | **x=3** |
|---|---|
| -2 | 4 |
| 0 | 3 |
| 2 | 3 |
| 3 | 2 |
| $\mathbf{n}_{i.}$ | **12** |

$$\overline{y}_{x=3} = \frac{-2 \times 4 + 0 \times 3 + 2 \times 3 + 3 \times 2}{12} = \frac{1}{3} = 0.33.$$

**Remarque 1.5.1** *On a la relation suivante entre la moyenne $\overline{x}$ et les l moyennes conditionnelles $\overline{x}_j$ :*

$$\overline{x} = \sum_{j=1}^{l} f_{\cdot j}\ \overline{x}_j$$

## 1.6 Covariance entre deux variables statistiques

**Définition 1.6.1** *La covariance est égale à la moyenne des écarts des couples* $(x_i,\ y_i)$ *de* $X$ *et* $Y$ *par rapport au point* $(\overline{x},\ \overline{y})$

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{N}\sum_{i=1}^{N}x_i y_i - \overline{x.y}$$

**Définition 1.6.2** *Dans le cas de données groupées dans un tableau de contingence (covariance pondérée) est donnée par :*

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}(x_i - \overline{x})(y_j - \overline{y}) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l}n_{ij}\ x_i y_i - \overline{x.y}$$

La covariance indique le sens de la relation entre les variables X et Y.

Ainsi, On peut distinguer les cas suivants :

- Si $cov(X,Y) > 0$, alors on peut dire que la relation entre les deux variables est positive.

  Dans ce cas, ces deux variables varient dans le même sens.

- Si $cov(X,Y) < 0$, alors on peut dire que la relation entre les deux variables est négative.

  Dans ce cas, ces deux variables varient en sens inverse.

- Si $cov(X,Y) = 0$, alors on peut dire qu'il n'y a pas de relation entre les deux variables.

  Dans ce cas, les variations de l'une n'entraînent pas la variation de l'autre.

**Propriétés de la covariance :**

1. $cov(X,Y) = cov(Y,X)$.
2. $cov(X,X) = var(x)$.
3. $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$.
4. $\forall a,b,x_0,y_0 \in \mathbb{R}:\ cov(aX+x_0,\ bY+y_0) = ab\ cov(X,Y) \Rightarrow var(aX+bY+c) = a^2 var(X) + b^2 var(Y) + 2ab\ cov(X,Y)$.
5. $|cov(X,Y)| \leq \sqrt{var(X).var(Y)}$.

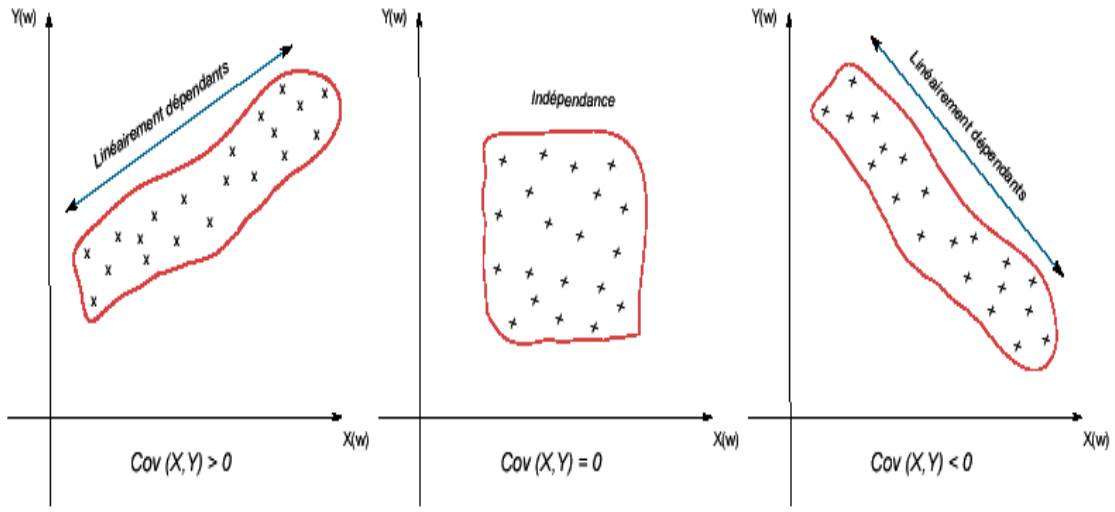


FIG. 1.2 – Covariance et la variabilité.

## 1.7 Coefficient de corrélation linéaire

Nous allons calculer le coefficient de corrélation entre deux séries de même longueur. On suppose qu'on a les tableaux de valeurs suivants : $X(x_1, ..., x_N)$ et $Y(y_1, ..., y_N)$ pour chacune des deux séries.

**Définition 1.7.1** *On appelle coefficient de corrélation linéaire ou coefficient de Bravais-Pearson entre deux variables statistiques X et Y, le rapport de leur covariance par le produit de leurs écarts-types :*

$$r(X,Y) = \frac{cov(X,Y)}{\sigma_X \times \sigma_Y}.$$

**Remarque 1.7.1** *La liaison entre deux variables numériques peut être étudiée grâce au coefficient de corrélation. Néanmoins, il faut bien garder présent a l'esprit que le coefficient de corrélation de Bravais-Pearson ne mesure que des relations linéaires, et sa valeur n'est en rien le reflet de l'existence d'un lien de causalité entre les deux variables.*

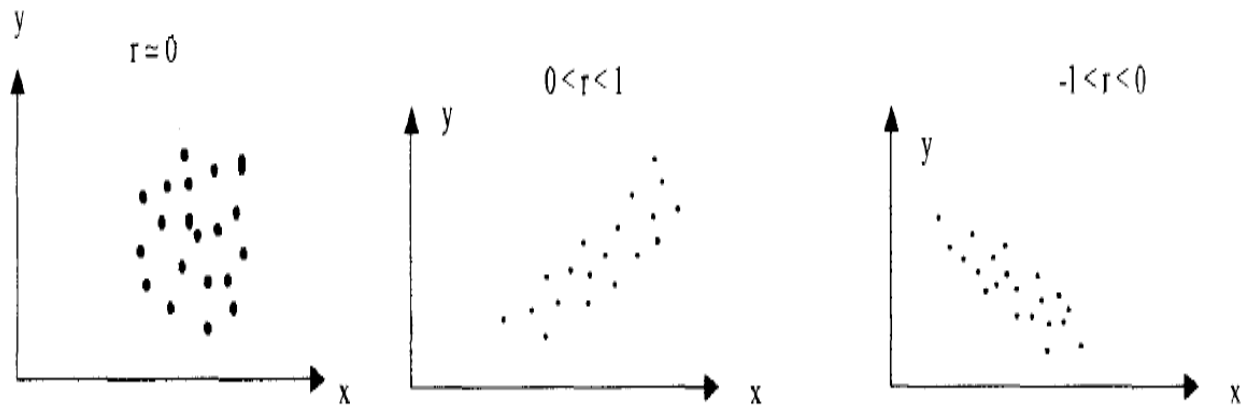


FIG. 1.3 – Exemples de nuages de points et coefficients de corrélation.

**Propriétés du coefficient de corrélation linéaire :**

1. Le coefficient de corrélation est toujours compris entre -1 et +1.
2. Si $r = +1$ alors les points se trouvent tous sur une même droite croissante, la corrélation linéaire positive parfaite.
3. Si $r = -1$ alors les points se trouvent tous sur une même droite décroissante, la corrélation linéaire négative parfaite.
4. Si $r = 0$ alors il n'y a pas une relation linéaire entre les variables X et Y.
5. On a pour tout $a,\ b,\ x_0,\ y_0 \in \mathbb{R}$ :

$$r(aX + x_0, bY + y_0) = \frac{cov(aX + x_0, bY + y_0)}{s_{aX+x_0} \cdot s_{bY+y_0}} = \frac{abcov(X,Y)}{|ab|\, s_X . s_Y}$$

$$= \begin{cases} +r(X,Y) & \text{si a et b de même signe} \\ -r(X,Y) & \text{si a et b de même opposé} \end{cases}$$