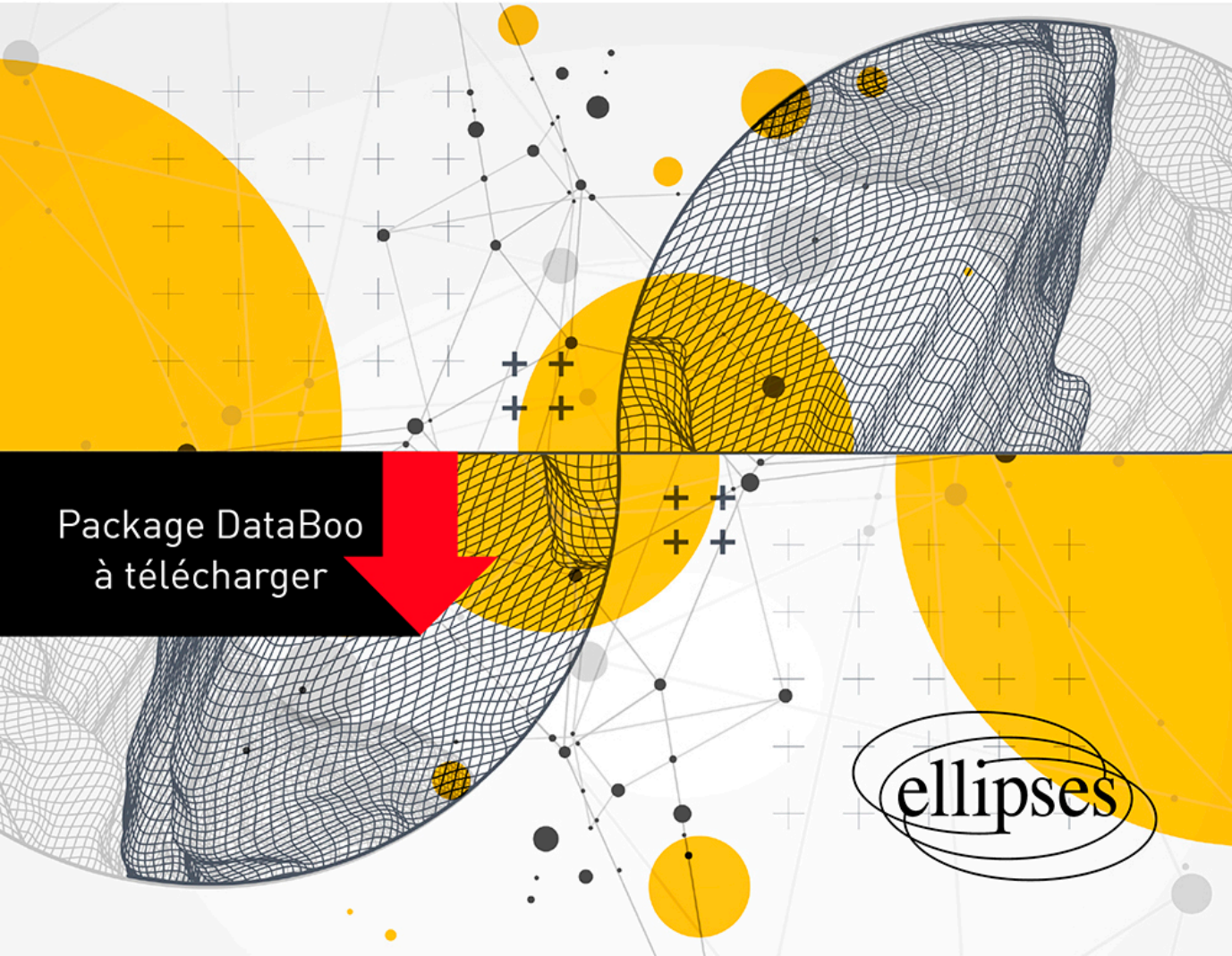
Marie-Noëlle Jubénot
Daniel Eudes

ANALYSE DES DONNÉES SOUS R

pour LES SCIENCES HUMAINES

Théories et exemples commentés



ellipses

PARTIE 1

Analyse descriptive et notions de base

Avant toute utilisation de méthodes d'analyse de données multivariées, il est toujours nécessaire de commencer une étude par des analyses statistiques descriptives univariée et bivariée. Il y a deux justifications évidentes à cette approche :

- Cela permet, d'une part, d'avoir une bonne connaissance des données. Il n'est pas possible d'interpréter correctement les résultats issus de l'analyse des données si nous ne connaissons pas par avance certaines de leurs particularités, par exemple leur niveau de variabilité, leur échelle… Il est parfois possible d'identifier préalablement des individus ou des variables que l'on pourra soupçonner d'être aberrants. Les points aberrants, atypiques**,** se définissent comme des observations peu fréquentes qui ne suivent pas les caractéristiques du reste des données. Les points aberrants doivent faire l'objet d'une analyse différenciée, car dans le cas contraire ils risquent de masquer les propriétés observables sur les autres données ;

- D'autre part, les méthodes d'analyse des données, bien qu'un peu plus sophistiquée, s'appuient sur les statistiques descriptives de base. Il n'est pas possible de comprendre leur signification et par conséquent de les interpréter si on ne comprend pas toute la portée cognitive d'indicateurs simples tels que : une moyenne, une variabilité, ou une corrélation entre variables.

Dans cette première partie, nous reviendrons sur les définitions des principaux indicateurs de statistiques descriptives, tels que les indicateurs du centre d'un nuage de points, de la variance et de l'inertie, de la distance entre individus, etc.

CHAPITRE 1. Analyse descriptive sur les variables quantitatives

Nous commencerons nos rappels d'analyse descriptive univariée par les indicateurs du centre qui nous permettront d'introduire par la suite les indicateurs de la variabilité et de l'inter-variabilité entre variables.

I.1.1. Les principaux indicateurs du centre d'une variable

Les indicateurs de la moyenne et du centre d'une variable constituent des indicateurs de base pour résumer ses propriétés. Ces indicateurs ont pour fonction principale :

- de caractériser la tendance centrale d'une variable ;
- de déterminer sa valeur la plus probable.

Pythagore, fameux philosophe et mathématiciens grec du VI[e] siècle av. J.-C., a fourni à partir d'analyses géométriques, quatre définitions de la moyenne : la moyenne arithmétique, la moyenne géométrique, la moyenne harmonique, et la moyenne quadratique.

Bien que les méthodes d'analyse des données utilisent essentiellement la moyenne arithmétique, nous choisissons ci-dessous de présenter l'ensemble de ces indicateurs, afin de pouvoir comparer leurs propriétés respectives.

La moyenne arithmétique (ou moyenne empirique) et centre de gravité

Nous allons aborder d'emblée la notion de matrices de données pour présenter les différents indicateurs. Dans une matrice de données, conformément aux conventions habituelles, les k colonnes correspondent à ses k variables, et les N lignes à ces N observations, ou individus statistiques, comme le montre le schéma ci-après :

$$Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{i1} & z_{i2} & \cdots & \boxed{z_{ij}} & \cdots & z_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Nj} & \cdots & z_{Nk} \end{bmatrix} \text{OU} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1j} & \cdots & z_{1k} \\ z_{21} & z_{22} & \cdots & z_{2j} & \cdots & z_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{i1} & z_{i2} & \cdots & \boxed{z_{ij}} & \cdots & z_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Nj} & \cdots & z_{Nk} \end{pmatrix}$$

La matrice de données brute Z comprend ainsi $N \times k$ éléments. Cette matrice peut être écrite indifféremment avec des crochets ou des parenthèses. Pour rechercher

facilement un élément, il suffit de donner son positionnement, autrement dit ses numéros de ligne et de colonne : l'élément (aussi nommé coefficient) z_{ij} désigne l'élément qui se situe au croisement de la ligne *i* et de la colonne *j*. Ce qui signifie que l'individu *i* possède la caractéristique (ou modalité) z_{ij} de la variable *j*.

Par convention, le premier indice de l'élément noté *i* correspond au numéro de ligne, le second noté *j* au numéro de colonne.

Lorsqu'un seul indice est indiqué, il s'agit de la désignation de l'ensemble des éléments de la colonne, et d'une variable donnée. Ainsi, z_j désigne la $j^{\text{ème}}$ colonne de la matrice de données Z. Cette variable comprend tous les éléments les z_{ij}, avec $i = 1, \ldots N$. L'indice *i* indique le numéro d'observation.

Voici un exemple concret d'une matrice (entre parenthèses) de données, documentée par le nom des lignes et des colonnes. Pour cette matrice N = 10, et k = 3 (chiffre de 2019) :

	Valeur du PIB	Superficie	Nbre d'habitants
1-Italie	1791	301 234	60 359
2-Espagne	1245	511 015	46935
3-France	2426	551 695	67 028
4-Allemagne	3449	357 578	83 019
5-Roumanie	223	237 500	19 402
6-Hongrie	146	93 030	9 773
7-Slovénie	48	20 273	2 080
8-Belgique	476	30 528	11 468
9-Pays-Bas	810	41 526	17 282
10-Portugal	214	92 042	10 277

Dans cette matrice, tous les chiffres sont en milliers. L'élément, $z_{7,2}$, situé à l'intersection de la septième ligne et de la deuxième colonne, contient le niveau de la population de la Slovénie, soit 2,080 millions d'habitants

Une analyse descriptive univariée des variables du tableau consiste à étudier les propriétés de chaque variable prise séparément.

Lorsque toutes les observations sont supposées avoir le même poids, situation la plus commune en analyse non supervisée, la moyenne arithmétique de la variable z_j, notée $\bar{z}_j$, a pour expression :

$$\bar{z}_j = \frac{\text{Somme de tous les éléments de la colonne } j}{\text{effectif total de la colonne } j}$$

$$= \frac{\sum_{i=1}^{N} z_{ij}}{N}$$

Plus généralement, avec des poids p_i différents ($i = 1, \dots, N$), pour les individus statistiques, selon un système de pondération découlant de leur niveau d'importance (en termes d'effectif, d'influence, de poids économiques et/ou financiers…), il est souhaitable de normer ces pondérations de sorte que $\sum_{i=1}^{N} p_i = 1$. Par exemple si on pondère les observations en fonction du nombre d'individus dans chacune des entités, alors :

$$p_i = \frac{\text{Effectif de } i}{\text{Effectif total}} \qquad \text{par conséquent} \sum_{i=1}^{N} p_i = 1$$

La formule de la moyenne arithmétique pondérée prend donc la forme suivante :

$$\bar{z}_j = \sum_{i=1}^{N} p_i z_{ij}$$

Par définition, la moyenne arithmétique correspond à la valeur que devraient prendre tous les individus d'une variable afin que le total de la variable soit égal à la moyenne multipliée par le nombre d'individu :

$$\sum_{i=1}^{N} p_i z_{ij} = N * \bar{z}_j$$

Le concept de **centre de gravité** (ou centroïde) d'un nuage de points découle directement de celui de moyenne arithmétique, lorsque l'on considère un ensemble de variables. En effet, si on calcule la moyenne de chaque variable d'une matrice Z quelconque :

$$Z = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,k} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,k} \\ \vdots & \vdots & & \vdots \\ z_{k,1} & z_{k,2} & \cdots & z_{k,k} \end{pmatrix}$$

$$\text{Moyennes} \quad \bar{z}_1 \quad \bar{z}_2 \quad \cdots \quad \bar{z}_k$$

Le concept de centre de gravité correspond au vecteur dont les coordonnées sont l'ensemble de ces moyennes :

$$G = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_k)$$

Le centre de gravité étend le concept de moyenne arithmétique dans le cadre d'une analyse multidimensionnelle.

Dans un nuage (représentation de points dans l'espace) à deux dimensions, le centre de gravité peut être facilement représenté sur un graphique. Nous pouvons représenter les individus à partir de leurs coordonnées sur les variables.

Prenons l'exemple d'un jeu de données comprenant trois individus caractérisés par deux variables et calculons le centre de gravité.

Sur le graphique ci-dessous, chaque point correspond à un individu. La représentation de l'ensemble des individus forme ce qu'on appelle un **nuage de points des individus**. Le centre de gravité (G) du nuage, ou point moyen, prend le rôle d'un individu moyen, i.e. caractérisant l'ensemble du nuage.

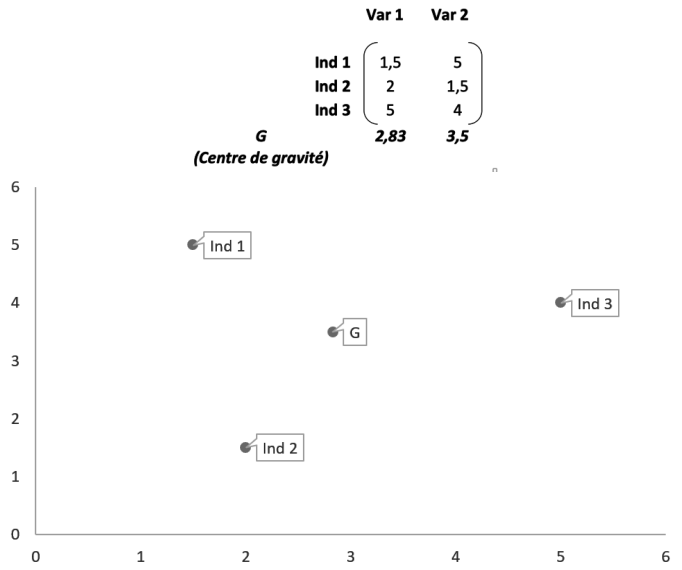


Figure 3. Représentation des individus en fonction des variables

Plusieurs interrogations se posent sur la position du nuage autour du centre de gravité, telles que :

- Quel est le niveau de la dispersion autour de ce point moyen ?
- Quelles sont les directions principales du nuage de point (celles qui expliquent la forme du nuage de points) ?

Comme nous le verrons plus tard, les méthodes d'analyse de données tentent justement de répondre à ces questions.

La moyenne géométrique

Cette moyenne géométrique correspond à la racine $N^{\text{ème}}$ de la multiplication des N éléments de la variable z_j soit :

$$\sqrt[N]{\prod_{i=1}^{N} z_{ij}}$$

Contrairement à la moyenne arithmétique, on multiplie tous les éléments au lieu de les additionner et le calcul de la racine $N^{\text{ème}}$ remplace la division par N.
Il est cependant impossible de déterminer la moyenne géométrique d'une variable contenant au moins un élément égal à zéro ou négatif. Ce qui limite donc l'application de cette moyenne à quelques domaines particuliers. La moyenne géométrique, moins sensible que la moyenne arithmétique aux valeurs les plus élevées, donne par ailleurs, dans le cas d'une distribution à « longue traîne », une meilleure estimation de la tendance centrale des données. C'est-à-dire lorsque la distribution des fréquences de l'ensemble des valeurs prises par la variable reste élevée pour les valeurs extrêmes, contrairement à la distribution de la loi normale où les traînes (ou « queues ») décroissent rapidement, et ceci de façon exponentielle.
Lorsque les observations sont affectées de poids différents, on utilise la moyenne géométrique pondérée :

$$\sqrt[\sum_{i=1}^{N} p_i]{\prod_{i=1}^{N} z_{ij}^{p_i}}$$

Moyenne harmonique

La moyenne harmonique, utilisée essentiellement pour déterminer un rapport moyen dans des domaines où ils existent des liens de proportionnalité inverse, fournit toujours la valeur la plus faible parmi les autres définitions de la moyenne de Pythagore. Elle s'écrit de la façon suivante :

$$\frac{N}{\sum_{i=1}^{N} \frac{1}{z_{ij}}}$$

Lorsque les observations ont des poids différents, on utilise la moyenne harmonique pondérée suivante :

$$\frac{\sum_{i=1}^{N} p_i}{\sum_{i=1}^{N} \frac{p_i}{z_{ij}}}$$

La moyenne quadratique

La moyenne quadratique se définit par la racine carrée de la moyenne arithmétique des carrés des éléments de la variable, soit :

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N} z_{ij}^2}$$

Comme il n'y a pas de compensation entre les chiffres de signes opposés, cette formule, s'utilise principalement pour estimer l'écart moyen (écart-type) entre

les éléments et le centre du nuage (défini par la moyenne arithmétique), ou pour calculer des distances entre des points quelconques.

Il existe encore d'autres indicateurs du centre d'un nuage, notamment les indicateurs de la médiane et du mode.

La médiane

La médiane se définit par la valeur centrale d'une série statistique en termes d'effectifs. La « médiane » est la valeur qui coupe une population en deux parties égales : la moitié des observations ont une valeur de la variable inférieure ou égale à la médiane, et la moitié une valeur supérieure ou égale.
En pratique, après avoir trié une variable par ordre croissant (ou décroissant), la médiane est la valeur telle que l'on ait autant d'éléments avec une valeur supérieure ou égale, que d'éléments avec une valeur inférieure ou égale. L'indicateur de la médiane a la propriété de mieux s'approcher du centre des distributions asymétriques, des répartitions des valeurs inégalitaires entre individus, que la moyenne arithmétique.
Plus généralement, la médiane fait partie de l'ensemble des quantiles qui divisent les données en parts égales.

Le mode

Contrairement aux autres indicateurs rappelés plus haut, l'indicateur du mode s'applique aussi bien aux variables quantitatives qu'aux variables qualitatives, puisqu'il porte sur les effectifs de réalisation de chacune des modalités. Le mode, correspond à la valeur des modalités ayant l'effectif le plus grand. **Voici, par exemple une petite série de nombre. Le mode de cette série ci-dessous, soit la valeur la plus fréquente, est 2 :**

$$(0, 0, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 5)$$

Dans les méthodes d'analyse de données, les indicateurs du centre les plus usités sont la moyenne arithmétique avec son corollaire le centre de gravité, et la moyenne quadratique pour le calcul de certaines distances, en ce qui concerne les variables quantitatives. Le critère du mode convient particulièrement aux variables qualitatives.

Après avoir, rappeler les définitions des principaux indicateurs du centre d'un nuage, nous allons aborder un autre point essentiel de l'analyse descriptive les critères de variance, de covariance et de corrélation, dans le cas d'analyse bivariées.

I.1.2. Variables quantitatives : les concepts de variance, covariance et de corrélation

I.1.2.1. Les concepts de variances et d'écart-type

Une fois avoir choisi un indicateur du centre, le plus souvent la moyenne arithmétique, il est intéressant de connaître le niveau de dispersion des valeurs d'une variable z_j autour de sa moyenne $\bar{z}_j$. La question est de savoir si les éléments d'une variable sont ou non très éloignés de sa moyenne. Et dans le cas d'une analyse avec plusieurs variables, si le nuage de points est plus ou moins dispersé autour de son centre de gravité. Chaque point, chaque élément est à une distance plus ou moins proche du centre. L'écart d'un point par rapport au centre a pour formule avec j constant :

$$z_{ij} - \bar{z}_j \qquad \text{pour tout } i = 1, \dots, N$$

Pour déterminer l'écart moyen entre chaque valeur et la moyenne, il ne convient pas de calculer la moyenne de ces écarts $\frac{1}{N}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}_j\right)$, car, comme la moyenne est au centre du nuage de points, certains écarts sont positifs et d'autres négatifs. La somme des écarts de la moyenne arithmétique s'annule par construction.

Il faut en conséquence remettre tous les écarts sous le même signe afin d'éviter leur compensation. En théorie, deux méthodes apparaissent possibles : la moyenne des valeurs absolues des écarts, ou la moyenne des écarts au carré. La deuxième solution plus pratique en termes de calcul arithmétique correspond à la définition de la variance. Ainsi, la variance de la variable j ($Var(z_j)$), que nous appellerons également $\sigma_{z_j}^2$ prend la forme suivante :

$$\boxed{\boldsymbol{\sigma_{z_j}^2 = Var(z_j) = \frac{1}{N}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}_j\right)^2}} \qquad \text{avec} \quad 0 \leq Var(z_j) \leq \infty$$

$$= \frac{1}{N}\left(\sum_{i=1}^{N} {z_{ij}}^2 + \sum_{i=1}^{N} {\bar{z}_j}^2 - 2\bar{z}_j \sum_{i=1}^{N} z_{ij}\right) = \frac{1}{N}\left(\sum_{i=1}^{N} {z_{ij}}^2 + N\,{\bar{z}_j}^2 - 2N{\bar{z}_j}^2\right)$$

$$= \boxed{\frac{1}{N}\sum_{i=1}^{N} {z_{ij}}^2 - {\bar{z}_j}^2 = \boldsymbol{Var(z_j)}}$$

Cette moyenne des écarts au carré s'interprète de la façon suivante : plus la valeur de $Var(z_j)$ est grande, et plus le nuage de points est dispersé autour de la moyenne. Une valeur de $Var(z_j)$ nulle implique que tous les individus ont la même valeur, soit la valeur moyenne de la variable.