fiches de Statistique descriptive

Rappels de cours et exercices corrigés

Farid Makhlouf

ellipses

Fiche 1

Données et variables

DÉFINITIONS

- Les **données** sont des informations et des faits sous forme de chiffres, textes ou d'images collectées. Ces données doivent être organisées en information afin de devenir interprétables dans le cadre d'une prise de décision dans un environnement risqué. Les données collectées pour répondre à une problématique prédéterminée sont appelées « ensemble de données ».
- Les **individus statistiques** sont des éléments sur lesquels les données sont collectées. Ils sont également appelés les unités statistiques ou éléments.
- Les **observations** sont des informations fournies par chaque individu ou unité statistique.
- Une **variable statistique** représente les caractéristiques d'un élément (unité statistique). Elle peut présenter des caractères qualitatifs ou quantitatifs. Un caractère étudié peut prendre plusieurs modalités. Les modalités doivent être incompatibles et exhaustives. Généralement, une variable est notée par une lettre majuscule et les valeurs qu'elle peut prendre par des lettres minuscules.
- Les **variables qualitatives** sont des variables mesurées par des labels ou des noms. Elles présentent des caractéristiques non quantifiables. Les opérations arithmétiques ne sont pas applicables aux variables qualitatives.
- Les **variables quantitatives** sont des variables qui présentent des caractéristiques chiffrables et calculables. Les opérations arithmétiques sont applicables aux variables qualitatives.
- Les **variables discrètes** sont des variables qui prennent des valeurs finies.
- Les **variables continues** sont des variables qui peuvent, en théorie, prendre une infinité de valeurs. En pratique, les variables statistiques ne peuvent pas être continues. Les mesures utilisées engendrent des discontinuités dans les résultats.
- Les **modalités** sont les différents caractères ou valeurs que peut prendre une variable. Exemple : situation matrimoniale avec 4 modalités : {marié (e) ; divorcé (e) ; célibataire ; veuf (ve)}.
- Un **échantillon** est un sous-ensemble ou une partie d'une population destinée à la représenter.

- Un **sondage** est une étude réalisée auprès d'un échantillon.
- Un **recensement** est une étude menée auprès de l'ensemble des éléments de la population.

I. Données

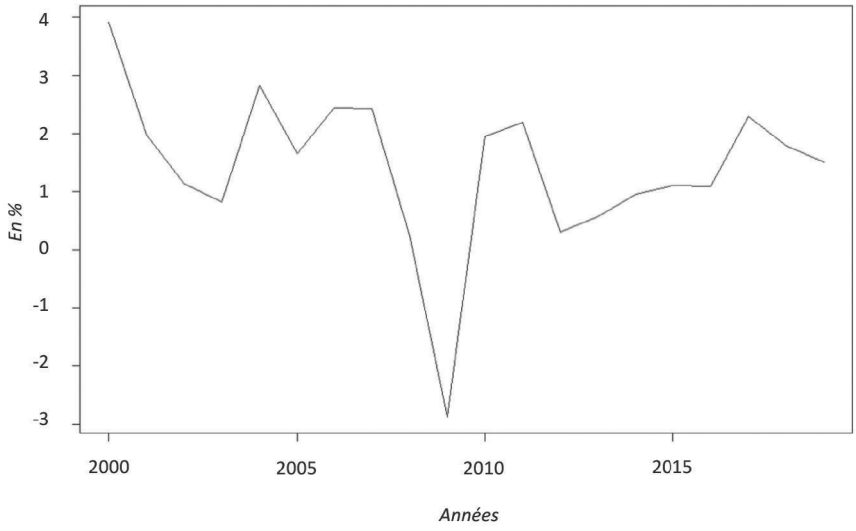
En statistiques descriptives, on distingue, généralement, trois types de données : données chronologiques, données en coupe transversale et données de panel.

A. Données dites chronologiques (séries temporelles)

Les données chronologiques sont des données observées et collectées sur plusieurs périodes pour une seule unité statistique. La variable liée aux données chronologiques est notée « X_t » avec « $t = \{1,2,3,\ldots,T\}$ ». t est un indice relatif à la période d'observation, habituellement, elle représente un moment important du phénomène étudié, elle peut être annuelle, semestrielle, trimestrielle, mensuelle, hebdomadaire, journalière, etc.

Le taux de croissance annuel du PIB de la France pour la période 2000-2019 noté G_t : $t = \{2010,2001,\ldots,2019\}$ représente des données chronologiques. Graphiquement, elles sont présentées par des diagrammes en lignes. La source de variation du taux de croissance est due à la variation des périodes (années). La figure 1.1 donne l'évolution du taux de croissance de la France pour la période 2000-2019.

Figure 1.1 : Taux de croissance (prix courants) données chronologiques

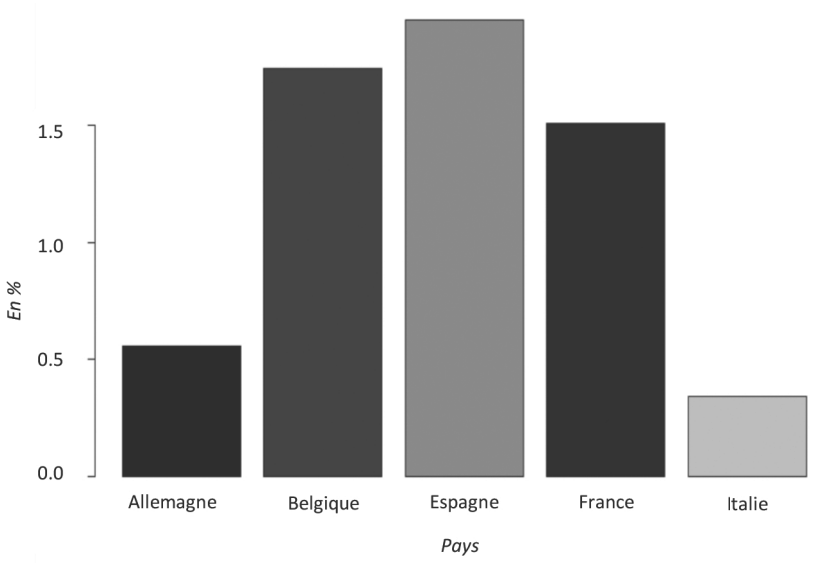


Source : Banque mondiale (WDI, 2021)

B. Données en coupe transversale

Les données en coupe transversale (instantanée) sont des données observées et recueillies à un instant donné pour plusieurs individus. Les données en coupe transversale sont notées « X_i » avec $i = \{1,2,3,\ldots N\}$ relatif au nombre d'individus. N est le nombre total d'individus pris en considération, et X est la variable étudiée. Les données en coupe transversale sont généralement utilisées dans études comparatives. Par exemple, la croissance économique de chaque pays de la zone euro en 2020 représente des données en coupe transversale. Si on considère que G_i est la variable liée à la croissance économique, G_3 représente le taux de croissance économique du pays numéro 3. Dans ce cas, la seule source de variation de la croissance économique est due à la différence des taux de croissance entre les individus (pays). Les données en coupe transversale ne présentent pas une variation temporelle du fait qu'elles sont observées dans une seule période. Régulièrement, les données en coupe transversale sont, graphiquement, présentées par des diagrammes en barres ou circulaires. La figure 1.2 montre le taux de croissance de 5 pays de l'Union européenne en 2019.

Figure 1.2 : Taux de croissance en 2019 (prix courants) données en coupe transversale



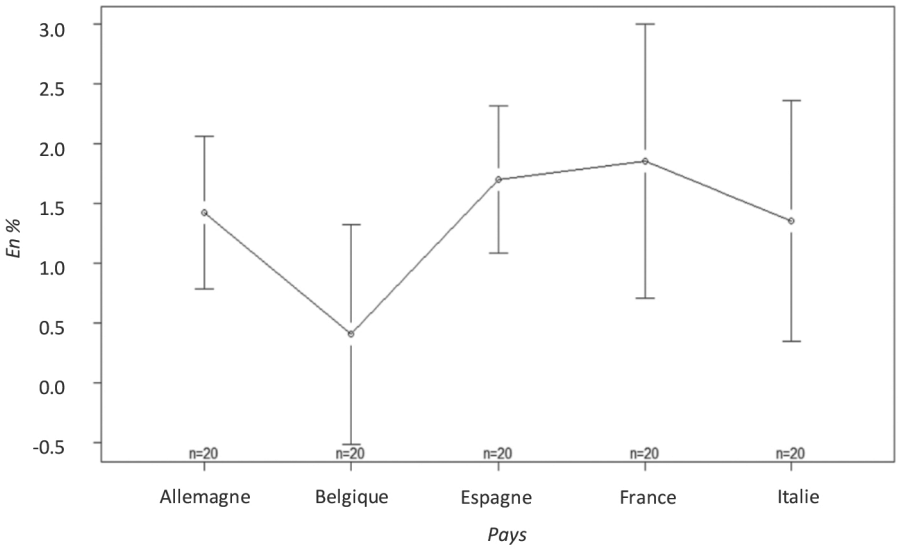
Source : Banque mondiale (WDI, 2021)

C. Données de panel (longitudinale)

Les données de panel sont des données qui possèdent deux dimensions : une dimension temporelle et une dimension individuelle. En d'autres termes, les données de panel sont observées et collectées sur une succession de périodes et sur plusieurs individus. La variable liée aux données de panel notée « X_{it} », « "i" » étant l'indice relatif à l'individu et« "t" » l'indice relatif à la période d'observation. Par exemple, les valeurs relatives aux taux de croissance des pays de la zone euro observées sur la période allant de 2010 à 2019 représentent des données de panel. $G_{1,2015}$: représente le taux de croissance du pays affecté de l'indice (1) en 2015. La variation du taux de croissance est due à la variation dans le temps et la variation entre les individus (pays). Dans les données de panel, couramment, deux types d'effets sont distingués : effet temporel et effet individuel. Graphiquement, les données de panel sont présentées par :

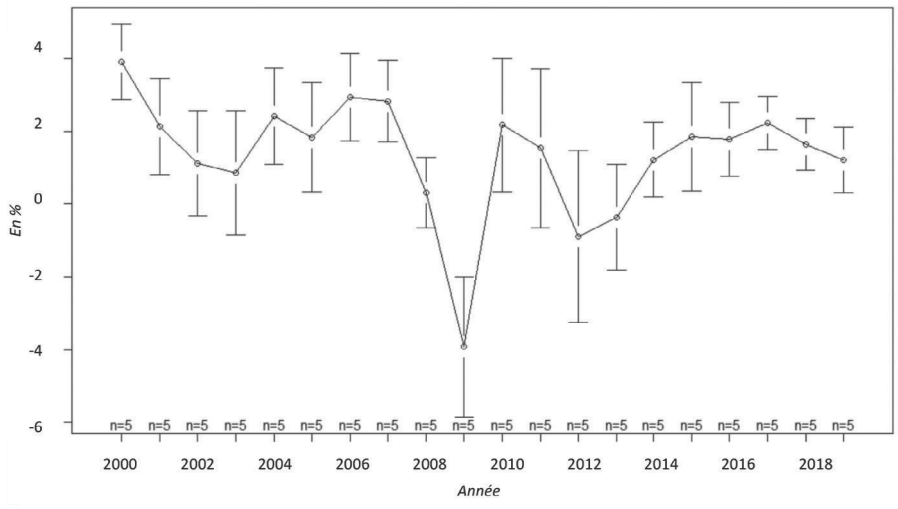
- Évolution des moyennes par pays (figure 1.3) ;
- Évolution des moyennes par période (figure 1.4) ;
- Boîte à dispersions (figure 1.5).

Figure 1.3 : Taux de croissance (prix courants) données de panel (moyenne par pays, 20 périodes)



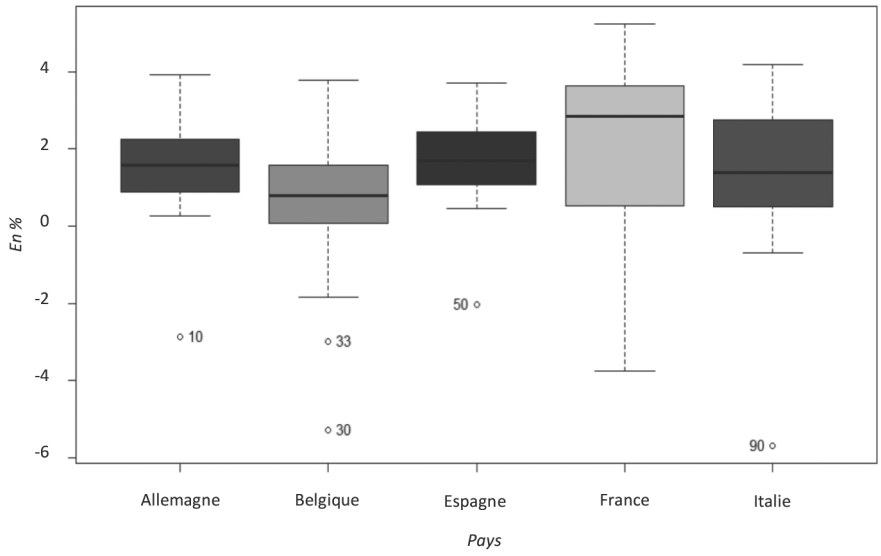
Source : Banque mondiale (WDI, 2021)

Figure 1.4 : Taux de croissance (prix courants) données de panel (moyenne par période, 5 pays)



Source : Banque mondiale (WDI, 2021)

Figure 1.5 : Taux de croissance (prix courants) données de panel (boite de dispersion)



Source : Banque mondiale (WDI, 2021)

II. Variables

Les variables peuvent être classifiées en deux principales catégories : variables qualitatives et variables quantitatives. Les variables qualitatives sont toujours des variables discrètes et les variables quantitatives peuvent être discrètes ou continues. Les opérations arithmétiques ne sont pas applicables aux variables qualitatives. Par opposition, les variables quantitatives ont des valeurs numériques et présentent un caractère dénombrable. En statistiques, il y a plus d'outils d'analyses pour des variables quantitatives que des variables qualitatives. Il est très important de comprendre et de distinguer la différence entre les deux types de variables pour bien utiliser des outils statistiques adéquats pour proposer des traitements et des analyses statistiques pertinents. Par exemple, nous ne pouvons pas calculer la variance pour une variable qualitative.

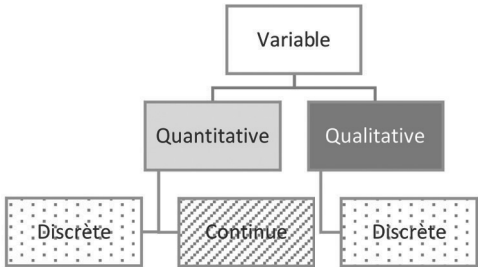
Exemples :

- Supposons que X est une variable relative au « **nombre de personnes satisfaites suite à une prestation de service donnée** ». X est une variable discrète car elle ne peut pas avoir de valeur décimale. En d'autres termes, un résultat de 3,5 de personnes qui sont satisfaites n'est pas logique puisqu'il n'existe pas une moitié d'une personne. Les résultats possibles sont des

nombres entiers naturels comme 3, 4, 5, etc. Par ailleurs la moyenne peut être un nombre décimal, du fait que la moyenne est un indicateur qui ne reflète pas un individu donné.

– Le taux de fécondité en France en 2019 est de 1,92 enfant par femme selon la Banque mondiale. La variable « **le nombre d'enfants par femme** » est une variable quantitative discrète. Elle est quantitative étant donné que nous pouvons quantifier le nombre d'enfants pour chaque femme. Discrète parce que le nombre d'enfants est un entier naturel. Le schéma 1.1 ci-dessous résume les différents types de variables.

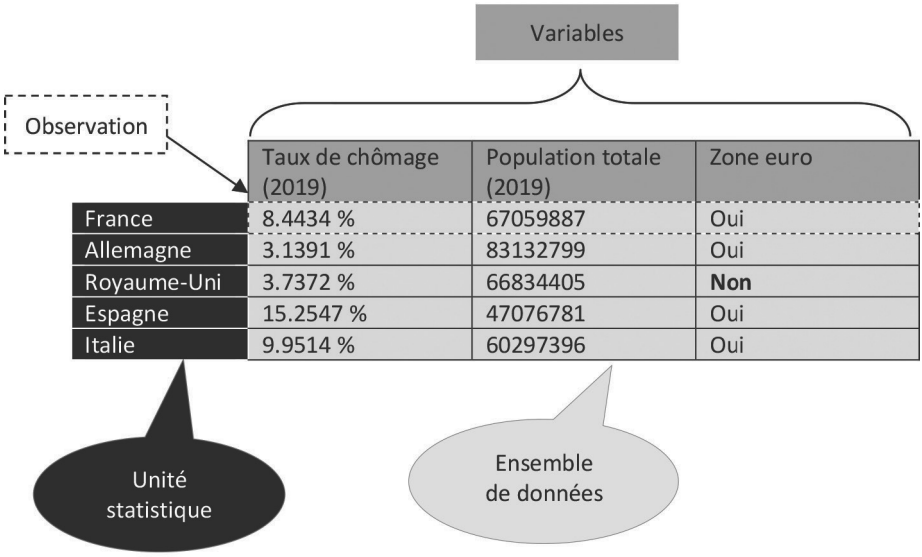
Schéma 1.1 : Variable qualitative et variable quantitative



À RETENIR

Les données fournies par les unités statistiques sont généralement organisées sous forme d'un tableau statistique. Les variables peuvent être présentées en colonnes et les individus statistiques (éléments) en ligne ou *vice versa*. L'intersection entre la colonne et la ligne nous donne une information précise de l'individu concernant la variable donnée. Par exemple, dans le tableau 1.1, le taux de chômage en France en 2019 est représenté par l'intersection de la première colonne et la première ligne. Le tableau 1.1 montre un exemple concernant les variables, les individus ainsi que les observations. En effet, il compare le taux de chômage, la population totale et il présente également l'appartenance à la zone euro. Il y a cinq pays dans le tableau 1.1 à savoir : la France, le Royaume-Uni, l'Allemagne, l'Espagne et l'Italie. Les cinq pays représentent des individus statistiques (éléments ou unité statistiques). Il y a également une variable quantitative continue (taux de chômage), une variable quantitative discrète (population totale) et une variable qualitative discrète avec deux modalités (Oui : le pays fait partie de la zone euro ; Non : le pays ne fait pas partie de la zone euro). Nous avons également cinq observations. Une lecture verticale du tableau 1.1 nous donne l'ensemble des réponses pour une seule variable et une lecture horizontale nous donne des valeurs de toutes les variables pour un seul individu (élément).

Tableau 1.1 : Individus, variables, observations et ensemble de données



Source : Banque mondiale (WDI, 2021)

POUR EN SAVOIR PLUS

- Bernard Py (1994), *Exercices corrigés de statistique descriptive*, Economica, EAN / ISBN-13 : 9782717826821.
- Denis J., Sweeney, D.J., Williams A., W., « *Essentials of Statistics for Business and Economics* », 6th edition by David R. Anderson, 2010, Cengage.
- Farid Makhlouf (2017), *Regression Analysis*. *In* : Farazmand A. (eds), *Global Encyclopedia of Public Administration*, Public Policy, and Governance. Springer, Cham. https://doi.org/10.1007/978-3-319-31816-5_476-1.
- Farid Makhlouf (2018), « Chapitre 6. Les approches quantitatives », *dans* : Frédéric Dosquet éd., *Études de marché.* Paris, Dunod, « Management Sup », 2018, p. 201-235. DOI : 10.3917/dunod.dosqu.2018.01.0201. URL : https ://www.cairn.info/etudes-de-marche--9782100781362-page-201.htm
- Hassène Siby (2017), « Introduction à la statistique et aux probabilités », Loze-Dion éditeur, ISBN : 978-2-924601-06-8.