Espaces

Dans cette partie, on présente les différents types de données et les espaces associés permettant de faire des analyses.

1 Types de variables

Avant de commencer la modélisation, il est important de bien comprendre le type des variables que l'on a. En effet, la nature des variables va influer sur l'espace dans lequel vivent nos données, et donc la distance que l'on va pouvoir utilisée pour comparer celles-ci, et par conséquent l'algorithme que l'on va implémenter. Les types de variables les plus courantes avec lesquelles on va travailler sont au nombre de quatre : les variables numériques, les variables ordinales, les variables nominales symétriques et les variables nominales asymétriques. On considère le type de variable pour la **plus petite unité statistique** de notre jeu de données.

Variable numérique

Une variable numérique est une variable dont la valeur numérique mesure quelque chose de quantifiable et dont la différence entre les valeurs reflète la différence entre les objets.
Exemples : revenu en dollars, masse, âge, …

Variable ordinale

Une variable ordinale est une variable qui ne donne pas une quantification précise d'un phénomène, mais dont les modalités peuvent être naturellement ordonnées.
Exemples : revenu faible, moyen ou élevé, ou niveau d'accord entre "tout-à-fait en désaccord", "en désaccord", "pas d'avis", "d'accord", "tout-à-fait d'accord", …

Nominale symétrique

Une variable nominale symétrique est une variable qualitative dont toutes les modalités sont aussi informatives l'une que l'autre.

Exemples : Sexe, sections d'un cours, ...

Nominale asymétrique

Une variable nominale asymétrique est une variable qualitative dont les modalités ne contiennent pas toutes le même niveau d'information, habituellement lorsque l'une des modalités est très fréquente, un peu la modalité par defaut, mais que les autres ne le sont pas.
Exemples : variable indiquant si un individu est daltonien ou non (deux individus daltoniens ont quelque chose en commun, mais deux individus non-daltoniens n'ont pas forcément quelque chose en commun), variable indiquant si une transaction est frauduleuse (généralement, elle ne va pas l'être), ...

Bien que ces types de variables soient les plus communs, on peut trouver beaucoup d'autres types de variables. Par exemple, on peut s'intéresser à de la comparaison de courbes, de mots, d'images, de graphs, etc. Dans ce cas, il est important de bien comprendre la notion d'**unité statistique**. Une unité statistique est l'unité dans laquel les mesures sont faites. Par exemple, si on s'intéresse à de l'analyse d'images, on pourrait se dire que nos images sont constitués de pixels, qui sont des variables numériques, et donc analyser nos images comme des variables numériques. Cependant, on peut faire le **choix** de considérer nos images comme notre unité statistique et donc de faire l'analyse au niveau image (et non variables numériques).

2 Espaces associés

Une fois que nos données ont été collectés, la première étape d'une analyse statistique consiste à choisir un espace mathématique dans lequel travailler. Cette espace, que l'on appelle parfois **espace d'observation** et que l'on note $\mathcal{X}$, dépend du type de données observées. Il constitue le cadre formel dans lequel nos variables prennent leurs valeurs, et il guide les choix méthodologiques qui suivront.

Cas d'une variable numérique

Lorsque l'on observe un variable numérique (e.g. la température d'un pays), l'espace naturel dans lequel travailler est l'ensemble des réels, $\mathcal{X} = \mathbb{R}$. Dans certains cas, on peut restreindre cet espace à un intervalle spécifique. Par exemple, si on s'intéresse à la taille d'une personne, on peut prendre $\mathcal{X} = [0, +\infty)$ car la variable considérée ne peut pas être négative.

Cas d'une variable nominale (ou qualitative, ou catégorielle)

Pour une variable nominal, l'espace est un ensemble fini de modalité, l'ensemble des modalités prises par la variable. Par exemple, si on étudie les résultats d'un lancer de dés, la variable peut prendre les valeurs 1 à 6, et l'espace associé sera donc $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

Lorsque les données sont plus conplexes, il faut choisir des espaces plus adaptés. Pour de l'analyse de courbes ou de signaux, on peut travailler dans un espace de fonctions. Par exemple, on peut considérer l'espace des fonctions continues sur un intervalle fermé $[a, b]$, noté $\mathcal{X} = \mathcal{C}([a, b])$. Pour de l'analyse de texte (vu comme une séquence de caractères), l'espace de travail peut être un alphabet. Par exemple, on peut considérer $\mathcal{X} = \{\mathrm{A}, \mathrm{B}, \dots, \mathrm{Z}\}$.

Souvent, on observe plusieurs variables en même temps, e.g. la taille, le poids et le sexe d'un individus. Dans ce cas, l'espace d'observation sera le produit cartésien (aussi appelé ensemble produit) des espaces associés à chaque variable :

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \mathcal{X}_p,$$

où p est le nombre de variables. Dans le cas où on observe p variables numérique, on notera plus simplement $\mathcal{X} = \mathbb{R}^p$