Espaces

Avant de pouvoir modéliser ou analyser des données, il est fondamental de bien comprendre la nature des variables que l'on manipule. En effet, le **type de variables** détermine :

- l'espace mathématique dans lequel elle vit ;
- les mesures de distance qu'on peut utiliser pour la comparer à d'autres ;
- et les modèles pertinents à utiliser.

Dans cette section, nous présentons les types de variables les plus courants, ainsi que les espaces associés.

1 Unité statistique

Une **unité statistique** est l'élément de base sur lequel une observation est effectué. Moralement, c'est le "porteur" de l'information qui est utilisé pour déterminer le niveau d'agrégation de l'analyse. L'unité statistique est un **choix** du modélisateur.

Exemples

- Dans le cas d'une enquête sur les revenus, on peut choisir l'individu comme unité.
- Dans le cas d'une étude sur les classes d'un lycée, on peut choisir la classe comme unité.
- Dans le cas d'une base de données d'imagerie médicale, on choisir l'image comme unité.

Parfois, une même base de données peut être analysée à plusieurs niveaux.

Une image est constituée de pixels, chacun pouvant être décrit par des variables numériques (e.g. valeurs RVB, opacité, …). On peut choisir d'analyser chaque pixel, et donc prendre le pixel pour unité, ou bien analyser chaque image comme un tout, et donc prendre l'image comme unité.

2 Types de variables

On distingue généralement quatre types de variables, que l'on identifie au niveau de la plus petite unité statistique du jeu de données.

Variable numérique (ou quantitative)

Une variable numérique (ou quantitative) est une variable dont les valeurs sont des nombres représentant une quantité mesurable.
Exemples : revenu en dollars, masse, âge, …

Variable ordinale

Une variable ordinale est une variable qualitative (ou catégorielle) dont les modalités peuvent être ordonnées naturellement, sans que l'écart entre les modalités soit quantifiable.
Exemples : niveau de revenu (faible, moyen ou élevé), niveau de satisfaction ("tout-à-fait en désaccord", "en désaccord", "pas d'avis", "d'accord", "tout-à-fait d'accord"), …

Variable nominale symétrique

Une variable nominale symétrique est une variable qualitative (ou catégorielle) dont toutes les modalités sont aussi informatives l'une que l'autre.
Exemples : nationalité, filière de formation, …

Variable nominale asymétrique

Une variable nominale asymétrique est une variable qualitative (ou catégorielle) dont l'une des modalités a un statut particulier, souvent plus fréquente ou considérée comme la valeur "par défaut". Ainsi, avoir deux observations avec la valeur "par défaut" de cette variable nominale asymétrique ne nous apprend pas grand chose sur celles-ci ; alors que on peut retirer beaucoup plus d'information de deux observations qui n'ont pas la valeur "par défaut".
Exemples : présence ou absence d'un symptôme, transaction frauduleuse ou non, …

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 textes, d'images, de réseaux, etc. Dans ces situations, le choix de la représentation dépend du niveau auquel on souhaite se placer, et donc de l'unité statistique.

3 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 $\mathcal{X} = \mathbb{R}^p$